Survey of Corpus-Based Vocabulary Lists for TESOL Classes[1]

Alison M. Youngblood[2], Western Kentucky University, Bowling Green, Kentucky, USA

Keith S. Folse[3], University of Central Florida, Orlando, Florida, USA

Abstract

Language learners need a large vocabulary in order to make strides in the target language. A strong vocabulary curriculum must carefully select vocabulary items for focus during limited class time, so one way that researchers have tried to help guide vocabulary instruction is in the generation of corpus-based vocabulary lists. While the *Academic Word List* (Coxhead, 2000) is by far the most well-known vocabulary list, there is a wide array of corpus-based vocabulary lists available to teachers and material writers. This article summarizes almost 100 years of research with an overview of 31 corpus-based vocabulary lists. The lists are grouped into four categories: general, academic, disciplinary, and formulaic. In addition, the authors explain key information about the list development process and content in order to help TESOL professionals become more confident consumers of vocabulary list research.

Resumen

Los estudiantes de idiomas requieren un amplio vocabulario para avanzar en un idioma nuevo. Un currículo de vocabulario debe seleccionar cuidadosamente y enfocarse en los elementos de vocabulario durante el tiempo limitado de clase, por lo que un método que investigadores han usado para guiar la enseñanza del vocabulario se encuentra en la generación de listas de vocabulario basado en corpus. Mientras que la *Academic Word List* (Coxhead, 2000) es la lista de vocabulario más conocida, hay una amplia gama de listas de vocabulario basado en corpus disponibles para profesores y escritores. Este artículo resume casi 100 años de investigación presentando un resumen de 31 listas de vocabulario basado en corpus. Las listas se agrupan en cuatro categorías: general, académica, disciplinaria y formulaica. Además, los autores explican la información clave sobre el proceso de desarrollo de la lista y el contenido con el fin de ayudar a los profesionales de TESOL a convertirse en consumidores más seguros de la investigación de listas de vocabulario.

A young ESL student in California struggling to make sense of a 6th grade honors science textbook to stay in his or her favorite class, a college student from Saudi Arabia studying to pass the Test of English as a Foreign Language (TOEFL) or the International English Language Testing System (IELTS) in order to begin an engineering degree in the U.S., and an employee of a large corporation in Mexico City taking night classes to earn a promotion are all learning English in different environments and for different reasons, yet they each share a common learning experience in that academically and professionally oriented learners of a second or foreign language face pressure to show gains in proficiency even after relatively short exposure to the target language (Cobb, 1999; Folse, 2010). Vocabulary knowledge is one way that learners can improve second language proficiency. Research in second language (L2) acquisition has shown that vocabulary knowledge is key in dramatically improving L2 proficiency in reading (Hsueh-Chao & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010), writing (Engber, 1995; Guo et al., 2013; Laufer & Nation, 1995), listening (Stæhr, 2009; van Zeeland & Schmitt, 2013), and speaking (Hilton, 2008; Koizumi & In'nami, 2013), yet studies vary in their estimates of

[1] This is a refereed article.
[2] Alison.Youngblood@wku.edu
[3] Keith.Folse@ucf.edu

an adequate L2 vocabulary size with lower figures of 4,000 to 5,000 word families (Laufer & Ravenhorst-Kalovski, 2010) and higher figures around 8,000 to 10,000 (Hsueh-Chao & Nation, 2000; Nation, 2006). Regardless, a detailed vocabulary learning plan is critical in any L2 curriculum, especially in the beginning levels; however, actual vocabulary instruction in the classroom faces many challenges, as many teachers are often in disagreement over which teacher's responsibility it is to develop this skill (Folse, 2010). Cobb (1999) succinctly sums up the reality of vocabulary learning as: "*Students typically need to know words measured in thousands, not hundreds, but receive language instruction measured in months, not years*" (p. 345).

After decades of receiving little to no attention in TESOL, vocabulary is now at the forefront of our field, as evidenced by the growing number of vocabulary articles in major journals (e.g., *TESOL Quarterly, Applied Linguistics, Modern Language Journal,* etc.), as well as whole issues dedicated specifically to vocabulary research. Teachers, learners, and curriculum developers would like researchers to identify not only how many words are needed, but also which words might be more important and therefore worth very limited instructional time. In sum, teachers would like research-based, empirically substantiated lists of the most important vocabulary words that our learners need to know.

For TESOL, perhaps one of the most promising lexical developments in the past two decades has been the increased use of corpus linguistics to produce truly useful vocabulary lists because Nation (2001) cautions that not all vocabulary words result in equal gains in proficiency. For example, consider the words *fundamental, essential,* and *vital* as equally occurring words in a reading class at an Intensive English Program (IEP). At first glance, based on intuition, none stand out as more important to integrate into a class vocabulary list. In fact, intuition might suggest it does not matter which of these synonyms is explicitly taught. For students at an IEP, however, Coxhead's (2000) corpus-based research on academic vocabulary found that *fundamental* would give the most benefit in decoding skills, as it is highly frequent in a variety of disciplines and consequently included in her Academic Word List. In other words, corpus linguistics allows us to produce vocabulary lists that better match our learners' actual English needs.

While Coxhead's Academic Word List is perhaps the most widely used vocabulary list in TESOL today, it is just one of the many vocabulary lists that interested teachers, researchers, and material writers can utilize. It is important for TESOL educators to look at the wider menu of corpus-based lists available today and consider their strengths and weaknesses. Therefore, the primary focus of this article is to provide a more extensive overview of the corpus-based vocabulary lists that are available for English learners by summarizing vocabulary list production from nearly 100 years of corpus research. A secondary objective of the article is to introduce basic corpus research terms and practices to help educators and material writers be more confident consumers of corpus research findings to best meet their classroom needs. This article does not, however, extensively discuss pedagogical uses of corpus-based vocabulary lists, as this was excellently outlined by Lessard-Clouston (2013). In addition, readers interested in a practical guide to generating word lists are encouraged to review Nation's (2016) text on making and using word lists in language learning and teaching.

Methods and Measures of Corpus-Based Vocabulary Lists

A corpus is a purposeful collection of language samples that can be analyzed to shed light on how members of a specific discourse community actually use language, which can then be shared with second language learners (Biber et al., 1999; Davies, 2010; Gardner & Davies, 2013; Nation, 2001; Nation & Webb, 2010). For example, a corpus of student compositions could tell us the vocabulary that most students are actually using, and a corpus of university lecture transcripts could be used to produce lists of vocabulary frequently used in academic lectures. The individual language samples transcribed and uploaded to a corpus should closely match the linguistic environment of the intended audience (Davies, 2010; Gardner & Davies, 2013; Nation & Webb, 2010). In addition, the appropriate language samples must be taken from a variety of authors of different text lengths to avoid bias from one particular writing style (Biber et al., 1999; Coxhead, 2000; Wray, 2002). In other words, a corpus-based list is a reflection of the language samples used to create it, and resulting word lists are only as good as the samples used to generate them.

Another important descriptor of a corpus is related to its lifespan; a corpus as a whole can be static or dynamic (Davies, 2010). In a static corpus, language samples are collected from a particular time frame, and once the samples are assembled, no new information will be added. The static corpus is a snapshot of language, and so is the list. A dynamic corpus, on the other hand, is updated yearly and can be used to monitor how a language grows and changes (Davies, 2010). The latter is perhaps of more interest to second language teachers and learners because it is a living corpus reflecting any current changes in language use. However, dynamic corpora are not as common as a great deal of resources are needed to continually update the database.

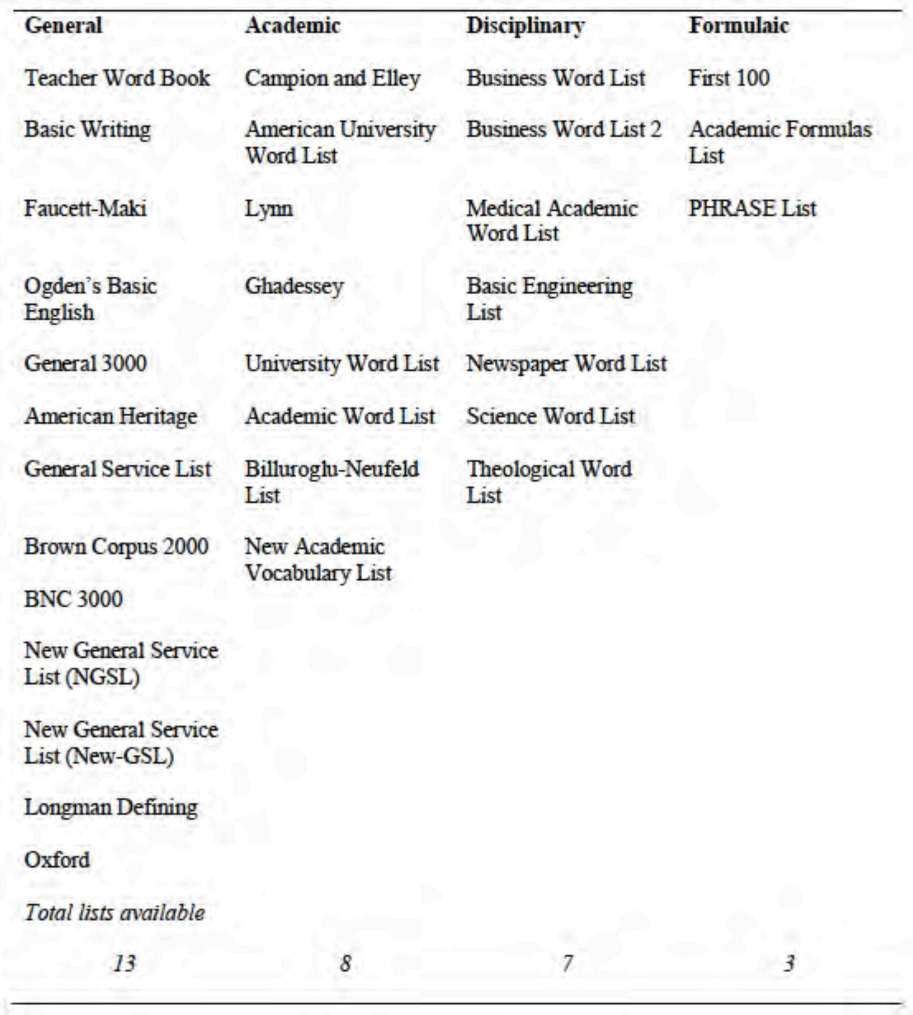
The size of the corpus is also important in order to generate a reliable list, but the ideal size largely depends on what type of vocabulary list is being created. When looking for high frequency, general vocabulary words, the corpus should contain a minimum of one to three million words (Brysbaert & New, 2009; Coxhead, 2000). Corpora with fewer than a million words can still be useful, but the results may be more appropriate for qualitative research on vocabulary-in-use or in preliminary studies (Granger, 1998). For educators, the key point to consider is how closely the context of the corpus matches their current students' needs.

Corpus researchers use a variety of measurements to analyze their language samples and determine what vocabulary items will be included and excluded from the list. Frequency, or simply how often a word occurs, is the hallmark measurement of vocabulary lists. In addition to frequency, another counting unit called range measures the distribution of the vocabulary word throughout the different subcategories of samples in a corpus. Coxhead (2000), for example, used art, law, and science as some of her subcorpora for her study on academic vocabulary. The range criterion is just as important as the frequency count. For instance, if a researcher wants to identify vocabulary required for new university freshmen, important words need to be equally represented in textbooks across the liberal arts. Range of use and frequency counts work hand in hand.

One final point to consider is what researchers are actually counting when they calculate frequency and range. There are typically two units of vocabulary: a word family and a lemma. An example of a lemma would be the word *develop* and its grammatical variants such as *develops, developed,* and *developing*. A word family includes all the forms from the lemma and derivations like *development, undeveloped,* and *developer*. To illustrate the real importance of a word list that uses word families compared to lemmas, consider the following example taken from the Corpus of Contemporary American English (Davies, 2011). Here are two examples of the word *developing* in context: The ripple effect of international finance could turn nasty in a developing nation and Teachers will be developing students' knowledge about medical technologies*.* If generating a word list using word families, both examples would count towards the frequency of *develop*. On the other hand, if using lemmas, these sentences would count as one occurrence for the participial adjective and one for the verb. The unit of counting dramatically changes the output of the lists, and as this paper will show, more recent corpus-based lists are shifting away from word families to lemmas. Either way, a word list of 500 items actually represents an exponentially larger vocabulary learning goal for second language learners.

Vocabulary List Options for TESOL Professionals

Vocabulary lists target specific types of vocabulary items: general, academic, disciplinary, and formulaic. In this article, we will discuss thirteen general, eight academic, seven disciplinary, and three formulaic vocabulary lists available for educators and material writers.



General	Academic	Disciplinary	Formulaic
Teacher Word Book	Campion and Elley	Business Word List	First 100
Basic Writing	American University Word List	Business Word List 2	Academic Formulas List
Faucett-Maki	Lynn	Medical Academic Word List	PHRASE List
Ogden's Basic English	Ghadessey	Basic Engineering List	
General 3000	University Word List	Newspaper Word List	
American Heritage	Academic Word List	Science Word List	
General Service List	Billuroglu-Neufeld List	Theological Word List	
Brown Corpus 2000	New Academic Vocabulary List		
BNC 3000			
New General Service List (NGSL)			
New General Service List (New-GSL)			
Longman Defining			
Oxford			
Total lists available			
13	*8*	*7*	*3*

Table 1*.* Vocabulary list options in TESOL

General Vocabulary

General vocabulary includes high frequency content (*school, develop*) and function (*because, at, by*) words and make up around 80% of spoken and written language (Nation, 2001b; West, 1953). This category contains the most available word lists.

The Teacher Word Book. Thorndike created The Teacher Word Book in 1921 from a static corpus of four million words assembled by hand from the *Bible*, elementary school textbooks, hobby manuals, newspapers, and letters. The list contains 10,000 must-know vocabulary words. Thorndike's list is notable because it was the first to use range and frequency to generate a 'credit number' that would justify the ranking of the words (Fries & Traver, 1960).

A Basic Writing Vocabulary. Horn (1926) identified the 10,000 most frequent vocabulary words from a static corpus of five million words consisting of language samples from business, letters, meeting minutes, newspapers, and magazines. Horn also pioneered the concept of a range requirement by using a "*credit system*" (p. 50) that co-accounted for the frequency of occurrence and the dispersion of the word among the language samples. The Teacher Word Book and the Basic Writing Vocabulary were later combined by Faucett and Maki (1932) to create the Faucett-Maki List, which became the General Service List after further revision by West (Gilner, 2011; Schmitt 2010).

Ogden's Basic English. Ogden, a critic of Thorndike, created his own word list in 1930. It was not based on frequency or range. Instead he used a qualitative approach to eliminate what he described as the redundancy in the English language. The final list includes 850 essential words plus a sub-list of 150 additional words specifically for scientists (Fries & Traver, 1960; Bauer, n.d.). The list contains 200 names of objects that could be represented visually, 400 general names, 150 qualities, and 100 words to operationalize ideas. The list came with a set of instructions on how to combine words together to illustrate more complex ideas.

General List of 3000. Palmer (1931) generated the list using frequency and range, but he also used qualitative data from teachers to make final inclusion decisions. The list is separated into six bands of 500 words. One of the strongest innovations of Palmer's list was the grouping of common lexical derivations under a main word. Palmer was the first to use headwords for list organization and item selection, which began to shift list production towards using word families (Gilner, 2011).

The American Heritage Word Frequency Book. Created by Carrol, Davies, and Richman in 1971, this list was derived from a static corpus of five million running words of written text used in the American school system. The list is notable for two reasons. First, the nature of the corpus makes it unique as it targets general vocabulary specifically used in K-12 schools. Second, it includes range and frequency counts for words common by grade level in each subject area (as cited by Waring & Nation, 1997).

General Service List. The GSL (West, 1953) identifies the 2,000 most useful word families in English from a static corpus of 2.5 million words. The corpus consisted of encyclopedias, textbooks, magazines, essays, novels, poetry, and science books. As the list was grounded in works by Thorndike, Palmer, and Faucett, the selection criteria included numerical requirements such as frequency and range but also more subjective measures such as the potential learning effort, necessity, register, and connotation. The resulting 2,000 words represent a combination of highly frequent items and some that are less ubiquitous, but that according to West (1953, p. x) are not easily expressed through higher frequency equivalents such as the word *preserve* to encompass *bottling, salting, freezing,* and *canning*. The list is divided into two 1,000 word bands. The first band covers an average of 75-80% of running words in a text, while the second covers an average of 4-6%.

Brown Corpus 2000. After the GSL, the hunt for the most frequent vocabulary words from other respective corpora began. The Brown Corpus 2000 was generated by Francis and Kučera (1964) to reflect the most common items from this static corpus. The Brown Corpus contains roughly one million words from 500 samples of English. The language samples come from newspaper articles, reviews, and editorials, books on religion, hobbies, and bestsellers, and other miscellaneous items like government documents. Interestingly, the researchers used the rate of publication for each category listed above during the year of assembly in order to determine what proportion of language samples should come from each (Brown Corpus, 2016; Nation, 2001).

BNC First 3000. The British National Corpus, or BNC, was once a dynamic corpus but now is static. The corpus contains 100 million running words from primarily written language samples collected between 1970s and the 1990s (Burnard, 2009; Davies, 2010). The written language samples are drawn from regional and national newspapers, specialist periodicals, journals, academic books, fiction, letters, and school and university essays. The spoken language samples, which make up about 10% of the corpus, come from transcriptions of informal conversations, formal business or government meetings, radio shows, and phone calls. The BNC First 3000 are the most common general purpose vocabulary items from this massive corpus.

Longman Defining Vocabulary and Oxford 3000. Since the advent of corpus linguistics, dictionary makers no longer rely solely on subjective measures to decide which words to include in their dictionary and how to define them. Both of these lists are used to create dictionary entries for English learners (ELs) (Oxford University Press, 2011; Waring & Nation, 1997). The Oxford English Corpus (Oxford, 2016) contains 2.5 billion words collected from web-based and some print sources. It includes language samples from literature, journals, newspapers, magazines, blogs, emails, and social media from multiple varieties of English including, but not limited to, the United Kingdom, the United States, New Zealand, Singapore, and South Africa. All of the language samples have been collected over the last 14 years, and more language is added each year, which makes it a dynamic corpus. The Longman list is used by Pearson to generate dictionary entries that reflect natural language use. The vocabulary list is based on a corpus of 330 million words from books, newspapers, and magazines (Longman, 2016).

New General Service List. In 2013, Browne, Culligan, and Phillips revisited the concept of a general service vocabulary list by looking at a larger, modern corpus and retooling the definition of a general service word. The researchers used a subsection of the Cambridge English Corpus totaling 273 million words from nine different registers: learner language, fiction, journals, magazines, non-fiction, radio, spoken, documents, and TV. The researchers identified 2,368 word families and covered 90% of the running words in the NGSL corpus compared to only 84% by the GSL. An increase in coverage is not particularly interesting as the additional 368 word families in the NGSL would logically increase the coverage rates over the GSL. What is noteworthy is that when the researchers lemmatized the NGSL, the list actually contains fewer words than the lemmatized GSL (2,818 vs. 3,623). In short, the researchers expanded the generalizability of the list while at the same time reducing the amount of individual words.

Another New General Service List. Around the same time as the NGSL was released, researchers from the United Kingdom published a New General Service List, hereafter termed New-GSL, which identified the most common lemmas in a static corpus of over 12 billion running words (Brezina & Gablasova, 2013). The samples represented both written and spoken English in a variety of registers and disciplines. The final list includes 2,494 lemmas and provided coverage for an average of 80% of running words in the sample corpus. While the coverage rate is lower than the NGSL discussed in the previous paragraph, the researchers found that 70% of the New-GSL items were equally represented across language samples of various sizes, modality, and discipline, which helps support the importance of general purpose vocabulary.

Academic Vocabulary

Academic words are those often found in textbooks, journal articles, and university lectures. They are not common to daily social interactions (*imply, perspective*). Coxhead (2000) found that academic vocabulary accounts for approximately 10% of words in scholarly settings. These lists are of special interest for academic language teachers. Many of the lists are based on corpora assembled by looking at textbooks real students were using.

Campion and Elley (1971). The Academic Vocabulary List (AVL) came from a corpus of about 300,000 words sampled from textbooks assigned in university courses. The courses were included in the program of study of the 19 most popular majors at New Zealand universities at the time based on enrollment. The first five lines of the first 200 pages of each textbook were sampled along with transcripts of a lecture and an examination from each major. Items on the list occurred at least four times in a minimum of three majors. An interesting methodological modification of the AVL was determining the "*measure of familiarity*" (Campion & Elley, 1971, p.8) by asking students to rate if each word was known or unknown. The final list includes 500 common academic vocabulary words. In a follow up analysis, Campion and Elley determined that only 31 of the items on the AVL showed up on the sample examinations from the same courses, suggesting a possible disconnect because the vocabulary students need in class and on assessments. In addition, the measure of familiarity determined that fewer than 50% of students involved in the study marked AVL items as known even though they commonly occurred in their course readings.

American University Word List. Praninskas (1972) needed to develop a vocabulary plan for a remedial English course for native Arabic speakers at a university. The corpus for the American University Word List (AUWL) consisted of ten textbooks; one from each course completed by first-year university students at the researcher's host institution. The researcher then sampled the tenth page of each textbook. Praninskas eliminated words occurring on the GSL or anything pre-emptively eliminated by West for inclusion on the GSL, proper nouns, abbreviations, and foreign words. Praninskas used both a frequency and range requirement. The final AUWL includes 10% of the most frequently occurring words from each book resulting in a list of 507 base words and 840 derived forms of the base words.

Lynn's Textbook Annotation List. Lynn (1974) generated this early academic word list to help students who received lectures in their native language (Mandarin) decode their textbooks written in English. Lynn assembled his corpus by borrowing 52 different textbooks from 50 business students to analyze which words the students annotated while completing assigned readings. Lynn's list includes 120 word families, which accounted for 20% of all student hand-written definitions in this corpus.

Ghadessey's Textbook Annotation List. Ghadessey (1979) assembled a corpus of almost half a million words from 20 textbooks used by non-native English speaking freshman studying biology, chemistry, or physics at a university. He recorded the words that students wrote definitions for in the textbook, the student-provided definition, and parts of speech in context. The list includes 795 of the most frequently glossed items in this sample of university textbooks. Thus, this list is based on how often students needed to look up the meaning of a word, not necessarily its frequency.

University Word List. Xu and Nation (1984) made use of the lists of Campion and Elley (1971), Ghadessy (1979), Lynn (1974), and Praninskas (1972) to compile their University Word List (UWL). The list contains a total of 836 words divided into 11 sublists based on frequency and range to make the list more manageable for teachers and students (Xu & Nation, 1984). There is no overlap between the GSL and UWL, as the list is seen as a complement to the GSL for academically-oriented language learners. According to Nation (1990), the UWL accounts for 8% of all words in a typical academic text.

Academic Word List. The Academic Word List (AWL) (Coxhead, 2000) contains 570 word families for academically-minded English learners to master, regardless of their intended field of study. Coxhead generated a static corpus of 3.5 million words taken from art, business, law, and science. These four disciplines contained 28 subfields, seven in each discipline. In order to appear on the AWL, a word family must not be represented on the GSL, occur at least 100 times in the corpus overall, and must occur in all four of disciplines, including appearing in 15 of the 28 subfields. The last two criteria control the range of the words to ensure that the results are generalizable to a wide range of students. Coxhead found that the AWL accounted for 10% of running words in her corpus and 8.5% of words in a smaller comparison corpus of academic language. When she analyzed a non-academic language sample, the AWL accounted for less than 2% of the language used, which supported the findings that the word families on the AWL are unique to academic discourse.

Billuroglu-Neufeld List. This list is difficult to classify under general or academic as it combines the GSL, the AWL, the Brown Corpus 2000, the BNC 3000, and Longman's dictionary building database (Neufeld & Billuroğlu, 2005). The purpose of the list is to represent the core vocabulary that language learners need by combining both general and academic vocabulary lists (Neufeld & Billuroğlu, 2005). The BNL contains all of the GSL and AWL items plus 183 new words. Interestingly, many items on the GSL and AWL were equally spread out on the list; they did not all make the top of the list when ranked by frequency. The researchers suggest that the BNL provided a much more "*natural*" profile of the vocabulary ELs need to know (Neufeld and Billuroğlu, 2005, p. 15). When the BNL was used to determine text coverage, the researchers found that it provided a 90% coverage rate, which surpassed the AWL and GSL combined coverage rates.

New Academic Vocabulary List. The New Academic Vocabulary List (AVL) (Gardner & Davies, 2013) is derived from 120 million words of written academic language from the Corpus of Contemporary American English, a dynamic corpus between the years of 1990 and 2015. The AVL includes 2,000 lemmas identified through several measures of frequency and range. An important innovation of this list is the notion of expected frequency. For example, the frequency of each word had to be 50% higher in the academic corpus than a comparison corpus of non-academic writing. Additionally, the list uses a measurement of how equally a word is distributed throughout the entire corpus to better detect general vocabulary and eliminate disciplinary vocabulary.

For comparison purposes, and perhaps to make the list more user-friendly, Gardner and Davies retroactively assembled the AVL into word families. To test how the AVL compared to the AWL, they randomly selected 570 AVL word families and found that it covered nearly 14% of the corpus compared to around 7% coverage from the AWL.

Disciplinary Vocabulary Lists

Disciplinary words are not common in social exchanges or academia at large, but rather in specific disciplines (*scalpel, incision*). Nation (2001b) estimates that technical words account for around 5% of words in any given exchange. There are many works that profile influential lists like the GSL and AWL in disciplinary corpora (see for example Martinez, Silvia, & Panza, 2009). This article, however, is only concerned with research resulting in a vocabulary list. Studies that compile disciplinary dictionaries are also not included.

Business Word List. Konstantakis (2007) developed this list to supplement the GSL and the AWL in order to give students the extra 5% of text coverage they needed to reach the 95% goal for reading comprehension (Laufer & Ravenhorst-Kavlowski, 2010). Konstantaskis used 33 business English textbooks to generate a static corpus of 600,000 running words. All words from the GSL and AWL were eliminated along with proper nouns, Latin words, nationalities, acronyms, and interjections. From the remaining items, Konstantakis found the word families that occurred at least ten times in five or more of the textbooks. The 560 headwords that met these criteria added 3% to the coverage rates, so it still fell short of the 95% target.

Business Word List Revisited. Hsu (2011) expanded the corpus to over seven million words from 2,200 research articles representing 20 sub-specialties to generate this updated Business Word List (BWL2). The BWL2 also departs from BWL in that it does not eliminate words from the GSL and the AWL. Instead, it excludes the most frequent 3,000 words in the BNC. The BWL2 includes 426 word families that covered almost 6% of words in the BWL2 corpus, which would get students past the 95% threshold for reading a business text.

Medical Academic Word List. Wang, Liang, and Ge (2008) established the Medical Academic Word List (MAWL) from a million-word static corpus of medical research articles representing 32 sub-specialties from anesthesiology to urology. Wang et al. used Coxhead's AWL inclusion criteria modified to fit the proportions of their own corpus. The MAWL excludes all GSL items and includes word families that have a frequency of at least 30 in a sub-specialty and meet the frequency requirement in at least 16 out of the 32 sub-specialties. The 623 word families on the MAWL cover around 12% of running words in the research articles that make up the corpus. The top five word families include *cell, data, muscular, significant, clinic,* and *analyze.*

Basic Engineering List. Ward's (2009) Basic Engineering List (BEL) contains 299 of the most frequent words in a smaller corpus of textbooks used by undergraduate engineering majors at the researcher's university. Ward collected five textbooks from each of the sub disciplines (chemical, civil, electrical, industrial, and mechanical) and sampled 25 pages from each text at random. Words on the BEL occur at least five times or more in each disciplinary subsection of the 271,000 word corpus. The top five items on the BEL are *load, stress, current, motor,* and *beam* (Ward, 2009, p. 178). Ward found the BEL coverage rates ranging from 15% to 20% for engineering textbooks from the corpus.

The BEL is different from other lists discussed so far in two important ways. First, the list does not use the lemma or word family as a counting unit. They treated each word as its own entity. Ward (2009) used word type as the counting unit because he envisioned the words on the list as ultimately a list of sight words needed for students with developing language proficiency. In this case, inflected and derived forms need to be recognized in their own right without relying on context clues or morphological awareness for decoding. Secondly, the BEL does not exclude words from the AWL and GSL before identifying engineering-specific vocabulary.

Newspaper Word List. Chung (2009) created a vocabulary list to supplement the GSL when reading newspaper articles. The static corpus consisted of 868 articles from three major newspapers in the U.S., U.K., and New Zealand. From the 600,000 running words, word families were identified that occurred at least 20 times in all four newspapers and in six or more of the subsections of each newspaper (sports, international, editorial, etc.). The NWL includes 588 word families that provide almost 7% coverage of the articles in the corpus.

Science Word List. Coxhead and Hirsh (2007) used 1.5 million words sampled from science textbooks in 14 different specialty areas to aid vocabulary development for students in English for academic purposes. The final list includes 318 words divided into six sub lists. The first five items on the list include *cell, species, acid, muscle,* and *protein*.

Theological Word List. Lessard-Clouston (2010) developed this list of 100 items from recordings of academic lectures on theology. An important innovation in the corpus development was the inclusion of words from handouts used during class and those written on the board. The final list is divided into two groups based on frequency. Example words include *ecclesiology, Gnosticism, omnipotence,* and *theodicy* (Lessard-Clouston, 2013, p. 294).

Formulaic Vocabulary Lists

The fourth category, formulaic language, is a relatively new category of vocabulary lists. Wray (2002) argued that "*words do not go together, having first been apart, but, rather, belong together, and do not necessarily need separating*" (p. 212). The most general conceptualization of formulaic language is "*a sequence of words that seems to be prefabricated and is stored and retrieved as a group from memory*" (Wray, 2002, p. 9). While idiomatic expressions such as *on the fence* and discourse markers like *on the other hand* are easily identified as cohesive chunks of language, formulaic language expands this notion to include less transparent groups of words that commonly co-occur in an effort to expose learners to more natural language.

First 100. Shin (2008) created a vocabulary list of the most common collocations of the nouns, verbs, adjectives, and adverbs from the top frequency band of the spoken section of the BNC. Collocations, or words that co-occur more often than by chance, had to occur at least 30 times in the BNC's ten-million word corpus of spoken English and have the same sense. The most common collocation reported is *you know*.

The Academic Formulas List. The Academic Formulas List (AFL) includes 200 formulaic sequences common to academic speaking, 200 for academic writing, and 200 core sequences common to both registers (Simpson-Vlach & Ellis, 2010). The researchers identified three, four, and five word sequences of vocabulary words from a two corpora each totaling 2.1 million words that represented academic speaking and writing. To identify sequences for the AFL, the researchers used a combination of measurements for frequency, range, and strength of association between the words in a formula. Finally, the researchers asked an independent group of English teachers to rate the value of each sequence. All of this data was used to determine which formulas belonged on the list. Example expressions common to both speaking and writing include *a variety of, as an example,* and *different types of*.

PHRASE List. The Phrasal Expressions List, or PHRASE (Martinez & Schmitt, 2012), consists of 505 formulaic sequences extracted from the BNC. The researchers relied on frequency and a qualitative-like investigation of meaning cohesiveness to determine which repeated combinations of words would improve the receptive and productive abilities of an L2 learner. Sequences on the PHRASE list range from two to four words in length, such as *a bit, as well, in fact*. An interesting secondary finding is that almost all of the sequences on the Phrase List are made up of the 2,000 most frequent single-words of the BNC (Martinez & Schmitt, 2012).

Discussion

Over the past century, corpus-based vocabulary list research has taken three general routes. First, new lists build on previous findings by using new, more sensitive, measurements based on advances in corpus software analysis tools and cognitive research (see for example The New Academic Vocabulary List). Secondly, as technology advances, the size of corpora have grown, thereby creating the possibility for more updated vocabulary analyses based on these new 'mega corpora' (see the New General Service List). Finally, as evidenced by the section on disciplinary vocabulary and formulaic language, corpus-based vocabulary lists have evolved to target more and more niche learning communities not only to increase students' reading skills in respective disciplines (see the Medical Academic Word List), but in the case of lists of formulaic language, to make students' writing skills more natural by mirroring vocabulary patterns found in the writings of native speakers (see the Academic Formulas List).

The most notable takeaway from this review of vocabulary lists is best seen by looking at the perceived role of these lists across the decades. When Thorndike released the Teacher Word Book in 1921, Ogden criticized the undertaking because "*[corpus research] created an army of 'word counters' applying arbitrary [measures]… to the construction of the recommended vocabularies…*" (as cited by Fries & Traver, 1960, p. 24), yet development continued and ultimately created the General Service List in 1953. Later, during the height of the Natural Approach in the 1980s, explicit language instruction, including grammar and vocabulary, was very much discouraged. Word lists were seen as unhelpful. However, lexical research in the 1990s showed that the use of word lists by teachers or students was not detrimental to lexical learning (Folse, 2004, pp. 35-45). Soon after, the Academic Word List helped frame vocabulary goals for academically-oriented students. In short, these criticisms did not deter researchers, teachers, and material writers.

Vocabulary lists are a well-established feature of second language research. A more pressing question after reviewing the menu of vocabulary lists is how to translate these lists into measurable gains in vocabulary development through curriculum and instruction. The *Vocab@Tokyo Conference* in 2016, for example, was rich with discussion about the role of word lists in language teaching. Tono (2016) stated that "*It is about time for us to consider how a wordlist can be integrated into a core language teaching syllabus in a meaningful way*" (p. 15). Schmitt (2016) cautioned that "*with the growing number of word lists now on the market, it may be time to pause and consider what it is we want from lists and what the field needs from their compilers*" (p. 17). The number of corpus-based lists does not seem to be slowing down. Quantitative and qualitative research on the application of these vocabulary lists to real curriculum does not seem to be as far along. Now more than ever, English language teachers' classroom-based research on the application of corpus-based vocabulary lists is needed to fill a gap in our understanding of real vocabulary development in the L2 classroom.

Conclusion

Even before Thorndike published his Teacher Word Book in 1921, many teachers and learners used lists of words as a tool for building vocabulary. These early lists, however, were based on one person's perceptions of which words were frequent. Today, we are able to use multi-million word corpora that more specifically match the language needs of our students. Coxhead (2000) raised the bar for practical word lists with her Academic Word List, which has become one of the most widely known options. However, there are many other word lists, including lists of individual words (e.g., Academic Vocabulary List and the New General Service List) as well as phrases (First 100 and PHRASE). We believe the taxonomy of word lists presented in this article offers an important and very practical summary of the resources available to teachers and curriculum developers. Identification of the vocabulary specifically needed by a group of learners can help teachers and curriculum writers design better, more useful materials.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow, UK: Longman.

Folse, K. S. (2004). Vocabulary myths: Applying second language research to classroom teaching. Ann Arbor, MI: University of Michigan Press.

Folse, K. (2010). Is explicit vocabulary focus the reading teacher's job? *Reading in a Foreign Language,* 22(1), 139-160. Retrieved from http://nflrc.hawaii.edu/rfl/April2010/articles/folse.pdf

Bauer, J. (n.d.). Ogden's Basic English (2012, March 24). Retrieved from http://www.basic-english.org/privacy.html

Brezina, V. & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the *New General Service List. Applied Linguistics Advance Access.* doi: 10.1093/applin/amt018

Brown Corpus. (2016, October 26). In *Wikipedia, The Free Encyclopedia*. Retrieved 23:09, October 26, 2016, from https://en.wikipedia.org/w/index.php?title=Brown_Corpus&oldid=746360244

Browne, C. (2013). The New General Service List: Celebration 60 years of vocabulary learning. *The Language Teacher, 37*(4), 13-16.

Brysbaert, M. & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990. doi: 10.3758/BRM.41.4.977

Burnard, L. (2009, January). What is the BNC? British National Corpus. Retrieved from http://www.natcorp.ox.ac.uk/corpus/index.xml

Campion, M.E., & Elley, W.B. (1971). *An academic vocabulary list*. Wellington, NZ: New Zealand Council for Education Research.

Chung, M. (2009). The newspaper word list: A specialized vocabulary for reading newspapers. *JALT Journal, 31*(2), 159-183. Retrieved from http://www.jalt-publications.org/jj/articles/263-newspaper-word-list-specialised-vocabulary-reading-newspapers

Cobb, T. (1999). Breadth and depth of lexical acquisition with hands-on concordancing. *Computer Assisted Language Learning, 12*(4)*,* 345-360.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238. doi: 10.2307/3587951

Coxhead, A., & Hirsch, D. (2007). A pilot science-specific word list for EAP. *Review Française de Linguistique Appliquée, 12*(2), 65-78.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*(4), 447-464. doi: 10.1093/llc/fqq018

Davies, M. (2011, March). Corpus of Contemporary American English. Retrieved from http://corpus.byu.edu/coca/

Engber, C.A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155. doi: 10.1016/1060-3743(95)900004-7

Faucett, L., & Maki, I. (1932). *A study of English-word values statistically determined from the latest extensive word counts*. Tokyo, Japan: Matsumura Sanshodo.

Francis, W. & Kucera, H. (1964). *Brown corpus manual*. Providence, RI: Department of Linguistics Brown University.

Fries, C. C., & Traver, A. A. (1960). *English Word Lists.* Ann Arbor, MI: George Wahr Publishing.

Gardner, D. & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics.* Advance access. doi: 10.1093/applin/amt015

Ghadessey, M. (1979). Frequency counts, word lists, and materials preparation: A new approach. *English Teaching Forum, 17,* 24-27.

Gilner, L (2011). A primer on the General Service List. *Reading in a Foreign Language, 23,* 65-83. Retrieved from http://nflrc.hawaii.edu/rfl/April2011/articles/gilner.pdf

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A.P. Cowie (Ed.), *Phraseology* (145-160). Oxford, UK: Clarendon Press.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18,* 218-238. doi: 10.1016/j.asw.2013.05.002

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *The Language Learning Journal, 36*(2), 153-166. doi: 10.1080/09571730802389983

Horn, E. (1926). *A basic writing vocabulary: 10,000 words most commonly used in writing*. Iowa City, IA: University of Iowa College of Education.

Hsu, W. (2011). A business word list for prospective EFL business postgraduates. *The Asian ESP Journal, 7*(4)*,* 63-99.

Hsueh-Chao, M.H. & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf

Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research, 4*(5), 900-913. doi: 10.4304/jltr.4.5.900-913

Konstantakis, N. (2007). Creating a business word list for teaching business English. ELIA: *Estudios de Linguistica Inglesa Aplicada, 7,* 79-102. Retrieved from http://institucional.us.es/revistas/elia/7/7.%20konstantakis.maq.pdf

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

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322. doi: 10.1093/applin/16.3.307

Lessard-Clouston, M. (2010). Theology lectures as lexical environments: A case study of technical vocabulary use. *Journal of English for Academic Purposes, 9*, 308-321. doi: 10.1016/j.jeap.2010.09.001

Lessard-Clouston, M. (2013). Word lists for vocabulary learning and teaching. *The CATESOL Journal, 24*(1), 287-304.

Longman Dictionaries Online U.S.A. (2016). The Longman American defining vocabulary. Retrieved from http://www.longmandictionariesusa.com/res/shared/vocab_definitions.pdf

Lynn, R. W. (1973). Preparing word lists: A suggested method. *RELC Journal, 4*(1), 25-32.

Martinez, I., Silvia, C., & Panza, C.B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes, 28*(3), 183-198. doi: http://dx.doi.org/10.1016/j.esp.2009.04.003

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299-320. doi: 10.1093/applin/ams010

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

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59-82. doi: http://dx.doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing.* Philadelphia, PA: John Benjamins Publishing.

Nation, I. S. P. & Webb, S. (2010). *Researching and analyzing vocabulary.* Boston, MA: Heinle, Cengage Learning.

Neufeld, S., & Billuroğlu, A. (2005). In search of the critical lexical mass: How 'general' is the GSL? How 'academic' is the AWL? Retrieved from www.lextutor.ca/vp/tr/BNL_Rationale.doc

Oxford University Press (2011). The Oxford 3000™. Retrieved from http://oald8.oxfordlearnersdictionaries.com/oxford3000/

Oxford University Press (2016). The Oxford English corpus. Retrieved from http://www.oxforddictionaries.com/us/words/oxford-english-corpus

Palmer, H. (1931). *Second interim report on vocabulary selection submitted to the Eighth Annual Conference of English Teachers under the auspices of the Institute for Research in English Teaching.* Tokyo, Japan: IRET.

Praninskas, J. (1972). *American university word list*. London, UK: Longman.

Schmitt, D. (2016). Beyond caveat emptor: Applying validity criteria to word lists. *Vocab@Tokyo: Current Trends in Vocabulary Studies Digital Program.* http://vli-journal.org/vocabattokyo/vocabattokyo_handbook_2016.pdf

Schmitt, N. (2010). Researching vocabulary: A vocabulary research *manual.* London, UK: Palgrave Macmillan. Retrieved from http://www.palgraveconnect.com/pc/doifinder/view/10.1057/9780230293977

Shin, D. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal, 62*(4), 339-348. doi: 10.1093/elt/ccmo91

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512. doi: 10.1093/applin/amp058

Stæhr, L. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition, 31*(4), 577-607.

Tono, Y. (2016). Toward the integration of a wordlist with a common framework of English: The case of the CEFR-J. *Vocab@Tokyo: Current Trends in Vocabulary Studies Digital Program.* Retrieved from http://vli-journal.org/vocabattokyo/vocabattokyo_handbook_2016.pdf

Van Zeeland, H. & Schmitt, N. (2012). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34*(4), 457-479. doi: 10.1093/applin/ams074

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list, *English for Specific Purposes, 27,* 442-458. doi: 10.1016/j.esp.2008.05.003

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes, 28,* 170-182. doi: 10.1016/j.esp.2009.04.001

Waring, R., & Nation, I.S.P. (1997). Vocabulary size, text coverage, and word lists. In N. Schmit and M. McCarthy (eds.), Vocabulary: Description, acquisition and pedagogy (6-19). Retrieved from http://www.victoria.ac.nz/lals/staff/Publications/paul-nation/1997-Waring-Vocab_size.pdf

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

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge, UK: Cambridge University Press.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3*(2), 215-229.