NOVATEUR PUBLICATIONS
JournalNX- A Multidisciplinary Peer Reviewed Journal
ISSN No: 2581 - 4230
VOLUME 6, ISSUE 11, Nov. -2020

CORPUS-BASED VOCABULARY LISTS FOR SECOND LANGUAGE CLASSES

TOSHPULATOVA MEHRINISO QILICHEVNA
Teacher at the Department of English Language Teaching Methodology Termiz State University
mehrinisotoshpulatova@gmail.com, 998975537061

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 with an overview of 31 corpus-based vocabulary lists. The lists are grouped into four categories: general, academic, disciplinary, and formulaic. But it is explained only two categories in this article. 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.

INTRODUCTION:

METHODS AND MEASURES OF CORPUS-BASED VOCABULARY LISTS:

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 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 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 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 berepresented 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, andcanning. 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 fromother 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).

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.

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:

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

2) Brown Corpus. (2016, October 26). In Wikipedia, The Free Encyclopedia. Retrieved 23:09, October 26, 2016,
3) Burnard, L. (2009, January). What is the BNC? British National Corpus. Retrieved from http://www.natcorp.ox.ac.uk/corpus/index.xml
4) Coxhead, A. (2000). A new academic word list. TESOL Quarterly, 34(2), 213-238. doi: 10.2307/3587951
5) 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
6) Davies, M. (2011, March). Corpus of Contemporary American English. Retrieved from http://corpus.byu.edu/coca/
7) 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
8) Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge, UK: Cambridge University Press.
9) 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.