**Core Vocabulary Lists**

Authors: Dana Gablasova & Vaclav Brezina

Lancaster University

*Abstract*: This chapter discusses the notion of core vocabulary and core vocabulary lists in general language use. Core vocabulary lists are widely used in fields such as language learning and teaching, language testing, psycholinguistics, first/second language acquisition, lexicography and cognitive science. The chapter focuses on key aspects of core vocabulary list creation and interpretation, in particular, on i) the construct represented in core vocabulary lists, ii) the operationalisation of the construct, and iii) the purpose and applications of these wordlists. Examples of notable core vocabulary lists are provided, demonstrating a variety of approaches to core vocabulary list creation and applications in relation to different user needs.

*Keywords*: core vocabulary, vocabulary lists, corpus-based wordlists, word frequency, dispersion, frequency dictionaries

**Introduction**

The notion of core vocabulary in general language use has been an inherent part of language analysis for over a century. It reflects the interest in what is considered the central, or the most important, set of lexical items in a language. Identifying these lexical items can be of considerable theoretical and practical importance in different language-related fields, such as language learning and teaching, language testing, psycholinguistics, first/second language acquisition, lexicography and cognitive science. Lists of core vocabulary items, therefore, represent a valuable resource in pedagogy and research. However, in practice, the definitions of ‘central’ or ‘core’ lexical elements can vary considerably and thus result in a different inventory of words that are included in core vocabulary lists, often reflecting varying theoretical, methodological and practical approaches to the analysis of vocabulary. Specifically, there are three key components that affect the word selection in a core vocabulary list: i) the construct represented by the list, ii) the operationalisation of the construct, that is, how the target words are identified and ordered, and iii) the purpose of the vocabulary list. These are discussed in detail below.

**Defining the Construct of Core Vocabulary**

Different perspectives on what constitutes core vocabulary in general language use have guided the development of core vocabulary lists. This diversity present in the definitions of ‘core vocabulary’ reflects the multifaceted structure of vocabulary and its use in context as well as different purposes of these lists (see ‘Purpose and applications of core vocabulary lists’). Largely, the following three considerations have been employed, sometimes in concert, to define ‘core vocabulary’ (Lee, 2001): i) word frequency and pattern of use, ii) the (perceived) usefulness of words, and iii) cognitive properties associated with lexical items.

Definitions of core vocabulary commonly draw on the information about word frequency in language use, with highly frequent words often considered to be the most important for communication (Vilkaitė-Lozdienė & Schmitt, 2019). This is because frequent words are encountered and produced more often, with a relatively small set of highly frequent vocabulary accounting for a large proportion of these words in any given text. For example, the New General Service List (Brezina & Gablasova, 2015) identified 2,494 lexical items such as *the, you, very,* *think* and *time* that account for over eighty per cent of all words encountered by users in English across different linguistic settings (e.g. speech/writing, formal/informal, general/technical texts). To date, core vocabulary lists have been commonly based on words that are most frequent in the language as a whole such as illustrated by the *Frequency Dictionary of Spanish: Core vocabulary for learners* ([Davies](https://www.routledge.com/search?author=Erwin%20Tschirner) & [Davies](https://www.routledge.com/search?author=Jupp%20M%C3%B6hring), 2018). This list provides 5,000 most commonly used words in contemporary Spanish based on two corpora (a 20-million-word corpus and a 2-billion-word corpus), representing spoken and written Spanish from across different genres and regional varieties. However, core vocabulary can also be described in relation to a more narrowly defined subset of general language, such as the medium of communication, as demonstrated by research on basic spoken vocabulary (McCarthy & Carter, 2003).

Another major approach to the definition of core vocabulary has employed the consideration of the usefulness of words. Frequency of occurrence has typically played a key role in the perception of word usefulness, due to the importance of highly frequent words for communicative purposes discussed above. In addition to frequency, researchers have drawn on further criteria to determine usefulness of words for a particular purpose or to a group of users. One such criterion has been related to the ability to communicate ideas deemed necessary for a specific group of language users (Bell, 2012). The words for inclusion in such lists are defined, often in addition to word frequency, in terms of semantic criteria such as the neutrality of meaning. This excludes words with affective meaning e.g. swearwords or diminutives. Other criteria also include the focus on superordinate terms (such as *to walk*) considered to be more general, and thus more useful for communication than their hyponyms (e.g. *to stroll, to stride, to march*). Usefulness as one of the selection criteria was used in West’s (1953) General Service List, one of the most influential lists of general vocabulary, which combined corpus evidence about word frequency with additional criteria such as stylistic and emotional neutrality, necessity of use and the ease of learning. For instance, in West’s vocabulary list, the verb *to preserve* (food)was added to the list despite its relatively low frequency as it was argued to represent a necessary idea. Another example of a list of core vocabulary built with a specific aim and language users in mind is the Essential Word List (Dang & Webb, 2016). In addition to corpus-based information, the list uses ‘practicability’ as the criterion to determine the number of items to be included in the list, with practicability being decided based on the estimate of “the amount of vocabulary that can be acquired by L2 learners within a language program” (Dang & Webb, 2016, p. 159). Other vocabulary lists in this category include defining vocabulary lists constructed for lexicographic purposes in learner’s dictionaries such *Longman Dictionary of Contemporary English* (2000 words), *Oxford Advanced Learner’s Dictionary* (3000) and *Cambridge International Dictionary of English* (2000 words). Words in these lists were designed specifically for writing dictionary definitions and explaining vocabulary items in simple terms. Stein’s (2000) Common Core Vocabulary combines all three defining vocabulary lists mentioned above into one comprehensive list.

Finally, there are lists that base the notion of coreness in language on the psycholinguistic properties involved in the learning, processing or production of lexical items.  Using this approach, core lexical items may be defined as words acquired, understood or used by speakers of a certain age. Vocabulary lists with the information about the estimated age when the words tend to be acquired by first or second language speakers – such as the list of age-of-acquisition ratings for 30,000 English words (Kuperman, Stadthagen-Gonzalez & Brysbaert, 2012) - play an important role in psycholinguistic and educational research. From the psycholinguistic perspective, core lexical items may also be associated with linguistic features such as concreteness, imaginability, familiarity and meaningfulness. Words with these properties (e.g. more imaginable or concrete words) tend to be considered more central to language knowledge as they tend to be acquired earlier and to be processed faster by speakers.

**Operationalising Core Vocabulary in Wordlists**

When developing a wordlist and operationalising the construct of core vocabulary, there are two key considerations that the designers need to address: i) a suitable source of evidence on which to base the selection of the words, and ii) appropriate procedures for identifying and ranking relevant words.. The information about these two components is crucial for evaluating how well the vocabulary lists are able to reflect the construct (see ‘Defining the construct of core vocabulary’).

Currently, corpus data is the primary source of evidence used in the creation of wordlists. In some cases, this is further supplemented by human judgement to reflect considerations related to usefulness for a specific group of users. When using corpora, one of the main issues to consider is the representativeness of the corpus and how its composition translates into the construct of core vocabulary. Here, we need to consider the proportion of language representing different modes of communication, genres and registers as well as different varieties of a language.

In corpus-based wordlists, frequency and distribution information have been used for identification, selection and ranking of words. In some cases, these were further complemented by additional sets of criteria. Frequency has been a key criterion in corpus-based lists of core vocabulary, reflecting the relationship between frequency, centrality and usefulness discussed above. However, absolute frequency is strongly related to the composition of the corpus and thus may not alone be sufficient to identify words that could be considered as general or core in language as a whole. For example, if a corpus contains a higher sample of texts from a certain domain (e.g. engineering or medicine), certain mode (e.g. writing) or a certain register (e.g. formal production), this may influence which words would be included in the list and how representative the list would be of other modes, domains and registers. As a result, the majority of core vocabulary lists also add information about distribution of the words, ensuring that the words in the list are widely used across different text types. Distribution of words is operationalised as dispersion, which refers to how evenly words are distributed across a corpus and its parts; dispersion can be measured statistically. For example, words such as the definite article *the* will be found in most English texts and thus would be considered more evenly dispersed than specialised words (e.g. *actuator, engine*), which may be very frequent in their domain of use but appear less often in other domains. Different measures of dispersion have been used in the creation of core vocabulary lists (e.g. Juilland’s D and DP). One of the basic measures of dispersion is *range* measured by the number of different texts or genres that a word appears in. For example, in addition to an overall high frequency, authors of a vocabulary list may set a specific threshold of a minimum number of texts or genres (e.g. 100) that the word has to occur in to be considered a ‘core’ item.

Dispersion can also be combined with frequency in the form of a measure called *average reduced frequency* (ARF). This measure calculates the distribution of words in a corpus based on the distance between individual occurrences of the word. If words are unequally distributed in a small number of clusters, the resulting frequency (ARF) of the item is reduced accordingly. The use of this measure can be seen in the New General Service List (Brezina & Gablasova, 2015), where it was used to identify and rank words that are not only highly frequent but also likely to appear in different domains, modes, and genres.

Another key methodological consideration in the creation of core vocabulary lists is the unit of counting that is used in selecting and ranking the words to be included in the list. Three major units have been employed in the compilation of core vocabulary lists: word families, lemmas and flemmas. Representing a broad approach, word families group together word forms that are related through their inflectional and derivational morphology. For instance, the following nine word forms are subsumed under a single word family: *walk, walkable, walked, walker, walkers, walkie, walkies, walking* and *walks*. Word families have traditionally been used in earlier lists of core vocabulary such as the British National Corpus/Corpus of Contemporary American English (BNC/COCA) 2000 list (Nation, 2020 [2012]) but they also appear in more recent lists such as The Nuclear Word Family List (Cobb & Laufer, 2021), which consists of a list of the 2,887 most frequent “nuclear” word families, i.e., families with just the most frequent family members. Lemma is the primary lexical unit employed in corpus-based vocabulary lists today. As a more precise and specific unit than a word family, a lemma subsumes word forms related through inflectional morphology, which belong to the same word class. For instance, the verb to *walk*, subsumes morphological variants *walks*, *walked* and *walking.* However, it is distinct from another lemma - the noun *walk*, which subsumes *walks* used as a plural noun. Lemmas are used as units e.g. in the Routledge frequency dictionary series (e.g. the *Frequency Dictionary of Spanish* or the *Frequency Dictionary of British English*). Finally, flemmas are typically used in contexts, where more sophisticated identification of lemmas is not available. A flemma subsumes all word forms based on inflectional morphology regardless of their word class membership. For example, a single flemma would subsume *walk, walks, walked* and *walking* regardless of whether these forms represent a nominal or verbal use. An example of a word list that was based on flemmas is the Essential Word List (Dang & Webb, 2016).

**Purpose and Applications of Core Vocabulary Lists**

Core vocabulary lists are constructed with a range of purposes and applications in mind, as apparent from the different constructs of ‘coreness’ discussed above. While some core vocabulary lists are explicitly constructed with a specific purpose such as the Essential Word List (Dang & Webb, 2016) for beginner learners, others were developed with a primary purpose to provide descriptive information about a segment of the vocabulary, allowing for a broad range of applications. Core vocabulary lists have been, for example, employed in research on language processing and production, such as studies on reading comprehension that estimate the number of lexical items in texts likely to be known to the readers. Core vocabulary lists have also been used to distinguish general and more domain-specific vocabulary in research on specialised language use in fields such as literary studies and language for specific purposes (LSP). The lists (databases) of words ranked according to their estimated age of acquisition serve commonly as a source of stimuli selection for psycholinguistic studies or in educational settings to support early language development (e.g. Laubscher & Light, 2020). Finally, a major application of core vocabulary lists has been in language learning and teaching (Dang, 2019), with this purpose reflected either in the construct of the list (e.g. the General Service List; West, 1953) or in additional resources available for the list as illustrated by the *Frequency Dictionary of British English: Core Vocabulary and Exercises for Learners* (Brezina & Gablasova, 2023). Such lists of core items can guide the selection of the vocabulary appropriate for L2 learners at different levels of proficiency and the creation of teaching materials such as exercises, textbooks and graded readers.

Given the large number of areas that have benefitted from core vocabulary lists so far, research on identification and classification of central lexical items represents a major direction in the study of language, with significant real-life applications.

**References**

Bell, H. (2012). Core vocabulary. In Chappelle, C. (Ed.) *The Encyclopedia of Applied Linguistics*. Wiley Blackwell.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, *36*(1), 1-22.

Brezina, V., & Gablasova, D. (2023). *A Frequency Dictionary of British English: Core Vocabulary and Exercises for Learners*. Routledge.

Cobb, T., & Laufer, B. (2021). The Nuclear Word Family List: A list of the most frequent family members, including base and affixed words. *Language Learning*, *71*(3), 834-871.

Dang, T. N. Y., & Webb, S. (2016). Making an Essential Word List for beginners. In: Nation, ISP. (Ed.) *Making and Using Word Lists for Language Learning and Testing* *(pp. 153-167)*. John Benjamins.

Dang, T. N. Y. (2019). Corpus-based word lists in second language vocabulary research, learning, and teaching. In *The Routledge handbook of vocabulary studies* (pp. 288-303). Routledge.

Davies, M., & Davies, K. H. (2018). *A Frequency Dictionary of Spanish: Core vocabulary for learners*. Routledge.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior research methods*, *44*, 978-990.

Laubscher, E., & Light, J. (2020). Core vocabulary lists for young children and considerations for early language development: A narrative review. *Augmentative and Alternative Communication*, *36*(1), 43-53.

Lee, D. Y. (2001). Defining core vocabulary and tracking its distribution across spoken and written genres: Evidence of a gradience of variation from the British National Corpus. *Journal of English Linguistics*, *29*(3), 250-278.

McCarthy, M., & Carter, R. (2003). What constitutes a basic spoken vocabulary? *SELL*, *1,* 233-249.

Nation, I.S.P. (2020 [2012]). The BNC/COCA word family lists. [www.wgtn.ac.nz/data/assets/pdf\_file/0005/1857641/about-bnc-coca-vocabulary-list.pdf](http://www.wgtn.ac.nz/data/assets/pdf_file/0005/1857641/about-bnc-coca-vocabulary-list.pdf) [accessed on 19 March 2024].

Vilkaitė-Lozdienė, L., & Schmitt, N. (2019). Frequency as a guide for vocabulary usefulness: High-, mid-, and low-frequency words. In Webb, S. (Ed.) *The Routledge Handbook of Vocabulary Studies* (pp. 81-96). Routledge.

Stein, G. (2002). *Developing Your English Vocabulary*. Stauffenburg.

West, M. (1953). A general service list of English words. Longman.

**Suggested Reading**

Brezina, V., & Gablasova, D. (2017). How to produce vocabulary lists? Issues of definition, selection and pedagogical aims. A response to Gabriele Stein. *Applied Linguistics*, *38*(5), 764-767.

Dang, T. N. Y., Webb, S., & Coxhead, A. (2022). Evaluating lists of high-frequency words: Teachers’ and learners’ perspectives. *Language Teaching Research*, *26*(4), 617-641.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied linguistics*, *28*(2), 241-265.