LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

EMPIRICAL STUDY

The Nuclear Word Family List: A List of the Most Frequent Family Members, Including Base and Affixed Words

Tom Cobb [a] and Batia Laufer [b]

[a]Université du Québec à Montréal [b]University of Haifa

Abstract: This article introduces the NFL7 (Nuclear Family List 7), a list of the 2,887 most frequent "nuclear" word families, that is, families that include just the most frequent family members and exclude those that constitute less than 7% of family occurrences. The NFL7 was developed by using a dedicated computer program, the Nuclear List Builder (freely available to users). To construct the list, we used that tool to reduce the complete BNC/COCA lists of the 3,000 most frequent word families from 19,062 to 7,293 word types and from 9,132 to 5,610 lemmas. Despite this reduction, the NFL7 compares favorably with other lists in terms of text coverage, and it includes a small number of the most frequent derivational affixes. We argue that the nuclearization of the list makes it suitable for nonadvanced learners, for teaching and testing both receptive and productive knowledge, and for instruction in basic morphology.

Keywords vocabulary; pedagogy; word family; word lemma; word frequency; coverage

Introduction

Learners of a first language (L1) acquire most vocabulary through input from spoken and written language and not through language instruction (Nagy, Herman, & Anderson, 1985; Sternberg, 1987). In contrast, second language (L2) learners, particularly in the classroom learning context, do not receive the kind of massive listening and reading exposure to a second language that is necessary for "picking up" a sufficient number of words. According to

To the many users of lextutor.ca who asked for a usable word list.

Correspondence concerning this article should be addressed to Tom Cobb, Dept. de Didactique des langues, Faculté des sciences de l'éducation, Local N-4205, 1205, rue Saint-Denis, Montréal (Québec) H2X3R9, Canada. E-mail: cobb.tom@uqam.ca

The handling editor for this article was Scott Crossley.

DOI: 10.1111/lang.12452

Nation (2014), words at the fifth 1,000-word frequency band can be learned in a year if learners who know 4,000 word families (base words with inflected and principle derived forms) manage to read a further 1 million words, which is a rather ambitious target. Even more concerning perhaps is Cobb's (2007) corpus analysis and subsequent computation, which showed that most words beyond the 2,000 most frequent words will not be learned in a year or two even if we assume the largest plausible amounts of reading.

Because of the limited learning opportunities available to them, most non-native speakers operate with a limited vocabulary by comparison to native speakers. It is therefore important to carefully select the vocabulary they will be exposed to in their classrooms by adhering to the cost–benefit principle, which states that learners should get the best return for their learning effort (Laufer & Nation, 2012). This means that they should be learning vocabulary that they will encounter inside and outside the classroom, that they will be able to use often, and that will leverage further independent learning: that is, high-frequency vocabulary. The compilation of a word list is an attempt to be clear about what this vocabulary is.

This realization has given rise to several word frequency lists that have been used in curricula, materials design, and testing. The most influential of these has perhaps been the 2,000-word family General Service List of English Words (West, 1953), which includes frequency information for families as a whole, based on precomputational hand counts mixed with some measure of intuition, and also for the major senses and meanings of the words that it contains and their principle derivations. It was and remains especially influential in the development of schemes for graded readers (Wan-a-rom, 2008), well into the era of more sophisticated lists extracted by computer programs from large electronic corpora. A more recent influential frequency list is the BNC/COCA word family list, which includes 25 thousand-family lists based on frequency and range data. The British National Corpus (BNC) is held by Oxford Computing and is described at http://corpora.lancs.ac.uk/BNCweb/ and can be accessed from http://bncweb.lancs.ac.uk/bncwebSignup/user/login.php; the Corpus of Contemporary American English (COCA) is by Davies (2008), with updated versions available at https://www.english-corpora.org/coca/; the BNC/COCA lists (first 10,000 headwords) are available at Paul Nation's website at https://www.wgtn.ac.nz/lals/resources/paul-nations-resources, or the complete set at the first author's website at https://www.lextutor.ca/vp/comp/. There are also lists of specialized vocabulary, such as Coxhead's (2000) Academic Word List and Dang, Coxhead, and Webb's (2017) Academic Spoken Word List, as well as lists derived from a range of subject domains and

interfacing with more general lists in various ways (Nation & Hwang, 1995), although none of these is as well-known as the general lists. (Research papers introducing a list typically include a copy of the list itself in an appendix, and the list is subsequently indicated by a reference to the paper.)

In this article, we introduce another nonspecialized word list that is smaller and, we believe, more useful than the other lists; it can be used with all learners below upper intermediate level, including beginners. The list includes three sublists, comprising the first, second, and third most frequent 1,000-word families (as defined below). It differs from other comprehensive word family lists in including only the most frequent or "nuclear" family members, whether inflections or derivations, namely those constituting at least 7% of the occurrences of the family as a whole. We thus name the list the Nuclear Family List 7 (NFL7).

Background Literature

Ways of Organizing Word Lists

The majority of word lists are organized by word families, that is, they use word families as the unit of counting. The concept of a word family is most clearly defined by Bauer and Nation (1993), who divide words into six cumulative levels: a base word (Level 1), plus inflected forms of the base word (Level 2), and four groups of derived forms (Levels 3−6). A word family is typically and often implicitly defined as a Level 6 family, and consists of a base word (e.g., *avoid*), its inflections (*avoids*, *avoided*, *avoiding*), its derived words (*avoidance*, *avoidable*, *unavoidable*), and inflections of the derived words (*avoidances*). The base word is normally the infinitive form for a verb, the singular form for a noun, or of course simply the unchanging form for a preposition, conjunction, and so forth.

Another unit of counting is the lemma, which consists of just a base word with its inflections. The lemma is thus a subset of the family, or it can be called a Level 2 family. Lemmas can be further divided into true lemmas, which treat different parts of speech for the same word form as separate lemmas (e.g., *run* as both noun and verb: *a run* and *to run*), and form-based lemmas or "flemmas," which count them as one lemma. (Flemmas are typically preferred over true lemmas as the unit of counting because true lemmas require specialized software to identify them as different parts of speech and to be counted as such by coverage software.) The example of the aforementioned *avoid* family includes four distinct true lemmas: (a) *avoid*, *avoids*, *avoided*, *avoiding*; (b) *avoidance*, *avoidances*; (c) *avoidable*; and (d) *unavoidable*. Each part of speech is assigned to a different true lemma (e.g., the lemmas headed by *avoid* and *avoidance*), and two words of the same part of

speech with a different or additional affix are assigned to different lemmas (e.g., *avoidable*, *unavoidable*). If the word form *avoid* could also function as a noun (as well as a verb), then that would constitute another true lemma. True lemmas are used as the unit of counting in most dictionary entries, as well as in some recent word lists; examples include Davies and Gardner's (2010) COCA list; Brezina and Gablasova's (2015) new-GSL or New General Service List (available at http://corpora.lancs.ac.uk/vocab/browse.php); and Gardner and Davies's (2013) Academic Vocabulary List. Flemmas are used in other recent lists, such as Dang and Webb's (2016a) Essential Word List and Browne, Culligan, and Phillips's (2013) NGSL (New General Service List, available at https://www.newgeneralservicelist.org). To avoid ambiguity when referring to the two identically named yet slightly different lists (Browne et al.'s NGSL includes 3,000 base words and Brezina and Gablasova's NGSL only 2,500 base words), we will refer to these lists as the NGSL3000 and NGSL2500, respectively.[1]

Good and Better Word Lists

What makes a word list useful for learners? The most important quality in terms of the cost–benefit principle is the text coverage that the list provides, that is, the percentage of words in any given text that can also be found in the list. Dang and Webb (2016b) compared the coverage of four high-frequency non-specialist word lists—West's (1953) GSL, Nation's (2012) BNC/COCA2000, Brezina and Gablasova's (2013) NGSL2500, and Nation's (2006a) BNC2000. The first three of these lists have been mentioned above, with the difference that the BNC/COCA version used in Dang and Webb's (2016b) study was a subset of the full 25 thousand families list. The fourth list (BNC2000; Nation, 2006a) is an equal size subset of Nation's earlier BNC-based list of 14 thousand families, which was subsequently merged with the COCA to create the BNC/COCA. The coverage of these lists was compared in 18 similarly general corpora that represented a wide range of spoken and written language and were distinct from any of those from which any of the lists had originally been derived. The lists, apart from the NGSL2500 (which is based on 2,500 lemmas), otherwise comprised the 2,000 most frequent word families. The authors concluded that the BNC/COCA2000 and the NGSL2500 provided the greatest lexical coverage and therefore might be considered the most useful lists for L2 learners. Nation (2016), however, found that the NGSL2500 had a higher coverage in academic texts, possibly because it is slightly larger, but that the BNC/COCA2000, although smaller, fared better in schoolbooks, novels, TV/movies, and both UK and US spoken language.

Another criterion of the usefulness of a list for L2 learners is inclusion of items that are relevant to the learners' experience and needs. For example, BNC/COCA's first 1,000 most frequent words include *lordship* under the headword *lord*, but do not include *absent*; yet it is questionable whether *lordship* is more relevant than *absent* in a classroom context. Dang, Webb, and Coxhead (2020) compared the BNC/COCA2000 and the NGSL2500 by examining teacher perceptions of the usefulness of words in the lists. They separated out all the words that occurred in one of the lists but did not occur in the other and asked teachers to rate the usefulness of each word for their students' needs. The results showed that the BNC/COCA2000 consistently made up a significantly larger proportion of words perceived as useful by teachers than the NGSL2500.

Yet another criterion of list usefulness is what Brysbaert, Keuleers, and Mandera (2020) refer to as *word prevalence*, which refers to the degree of learners' familiarity with words. Using crowd-sourcing technologies, the authors conducted a large study testing learners from different countries on 61,000 lemmas. Test takers were asked to say whether or not they were familiar with a sample of target words. Based on 17 million responses, a new list of 20 levels of 1,000 word families was constructed. The authors argue that the list is useful for pedagogical purposes because the levels represent the order in which English vocabulary seems to be acquired by L2 learners around the world. This list, however, does not yet appear to be presented in a pedagogically accessible format or to have received validation in a pedagogical context. Degree of learner familiarity was also investigated by Dang et al. (2020), who found that L2 learners were more familiar with the words in BNC/COCA2000 than in the NGSL2500.

The best generic English word lists available for use at the present time are the BNC/COCA2000 and the NGSL2500, as convincingly argued by Dang and Webb (2016b), principally on account of the coverage they offer and the quality of the corpora they are derived from, and, we would add, of the service they apparently offer teachers and learners as shown by the number of download requests for them, relative to other lists available, on the Lextutor website (https://www.lextutor.ca/), where links are provided.

Family Versus Lemma

An important pedagogical decision in making or choosing a word list is the unit of counting. Are learners likely to benefit more from lemma-based lists or from family-based lists? Dang and Webb (2016a), Dang (2020), and Nation (2016) suggest that learners of different proficiencies may benefit from different units

of organization in a list. Beginners with almost no morphological knowledge would probably benefit from lists of lemmas that introduce words but spare the learners information overload.

From a language teaching perspective, both types of lists, based on lemmas and on families, have strengths and weaknesses. Bauer and Nation (1993), Brysbaert et al. (2020), and Laufer (2020a) argue in favor of family-based word lists. When the word family is an organizing principle of word lists, learners other than beginners can add related words to their vocabulary without much learning effort, as well as acquiring rules (such as derivational systems) that will help them to understand additional new words in the future. The learning burden of adding derived words to the learner's vocabulary (e.g., adding *avoidable* to *avoid*) is presumably less than that of adding new words that do not belong to the same word family (e.g., knowing *abolish* and adding *avoid*), because the basic forms and the basic meanings in the first instance are related to one another in a family. Furthermore, many affixes are regular in the sense that they add the same lexical or grammatical meaning to the base word each time they are used. Once learners understand how some affixes change the meaning of the base word, they may understand a large number of new words without the need to be taught or to discover what they mean. For example, if they understand the meanings of the affixes in *unreadable* (*un-*, *-able*) and *happiness* (*-ness*), they will probably transfer this understanding without much difficulty to *unbearable* and *emptiness*. Hence, teaching word families accords with the cost–benefit principle better than teaching only lemmas because learners get a better return for their learning effort. Another advantage of family lists is that they are more compact than lemma lists, where similar words are encountered in only slightly different forms at subsequent levels.

A downside to word families, particularly for the most frequent base words, is that they may include derived forms such as *lordship* whose usefulness is questionable for learners at beginning and even intermediate learning stages. Indeed, families are not constructed primarily as learning targets, arguably, but as a means of calculating total family coverages by computer programs (as discussed further below). For example, the *accept* family in the first 1,000 families of the BNC/COCA word list includes the following distinct lemmas: *acceptable*, *acceptably*, *acceptability*, *unacceptability*, *acceptance*, *accepting* (adjective), *accepted* (adjective), and *acceptor*. Learners acquiring basic vocabulary would probably benefit more from working with entirely new words than from about half of the derived forms in that example, many of which are quite infrequent (although the family as a whole is frequent, with frequency and hence family position in the list based on the summed frequency of all

its members). Laufer and Cobb (2019) showed that many derived (Level 3−6) family members occurred infrequently over a range of text types.

In this respect, lemma-based lists are more useful because they are organized according to the summed frequency of the members of individual lemmas only. Thus the position of *accept* is based on the summed frequencies of *accept*, *accepts*, *accepting*, and *accepted*, whereas *acceptance* will be a different lemma at another level, based on the summed frequencies of *acceptance* and *acceptances*. The NGSL2500, which provides the 2,494 most frequent (true) lemmas, includes the headwords *accept* and *acceptable* but none of the other derived forms (*acceptance*, *acceptability*, and so on); these do not appear in the list at all, although they may function behind the scenes in the computer program (at http://corpora.lancs.ac.uk/vocab/analyse.php) that calculates the coverage of the list in texts. In other words, the list's exact contents are unknowable unless one wished to submit a large number of texts into the program and assemble the results.

Some derived words, however, even if they do not share the frequency of the lemma headword, are not of low frequency and furthermore can be learned without much effort, as explained earlier. For example, *access* and *announce* appear in the NGSL2500, but without any related derived forms; yet, *accessible* and *announcement* are both frequent (all four appear on the list of the 5,000 most frequent lemmas in COCA's 450-million-word corpus), are not loaded with much new lexical information, and provide an opportunity to learn the meaning of the highly generative suffixes *-ible* and *-ment*. It is pedagogically questionable to separate frequent, useful and easily learned related words into different frequency lists, taught at different learning stages: for example, to separate *announcement* from *announce*, *accessible* from *access*, or even *walk* (verb) *from walk* (noun). Nation (2016) and Bauer and Nation (1993) use similar reasoning to support the pedagogical value of word families for English language instruction.

The Present Study

In this study, we try to reconcile the two counting principles, family-based and lemma-based, capitalizing on the advantages of both. On the one hand, we seek to include a limited number of frequent derived words (*accessible* and *announcement*), but exclude many of the infrequent derived family members that appear on complete family-based frequency lists (such as *lordship*) to arrive at a significantly reduced "nuclear" list. We do this so that learners are not overloaded with learning the meanings of infrequent members of frequent families and can focus instead on a wide range of other useful words. On the other hand,

we group rather than separate selected family members, including derivations, using corpus analysis techniques described in the following section. In doing so, we follow what we believe to be a sound pedagogical principle. We aim to expand learners' useful vocabulary and basic morphological knowledge.

There are two other suggestions to reduce family lists that we are familiar with: Nation's (2016) "Level 3 partial" and Greene and Coxhead's (2015) Academic Vocabulary for Middle School Students lists. Nation's list includes base words and their inflected forms but also four particular sets of frequent derived forms—those with affixes *un-*, *-ly*, *-er*, and *-th*—from Level 3 of the Bauer and Nation (1993) scheme. The first three of these are common in the 3,000 most frequent base words and *-th* is used with ordinal numbers. Therefore, these affixes should not require much teaching or learning effort. This list was found to provide good coverage of a 14-million-word corpus of graded readers and other introductory materials compiled by Nation to supplement the high-frequency end of the BNC/COCA (unpublished, but available for concordance runs under the name "BNC/COCA (1+2k) 14m" at https://www.lextutor.ca/conc/eng). Coverage was just 1% less than that provided by the complete lists with derived words at all six affix levels. Greene and Coxhead's lists limit family members on the basis of range and frequency in a corpus of school textbooks, but they are a set of specialized lists for native speakers at upper elementary school. Our proposal is for a single comprehensive list intended for learners of English as a second or foreign language for general purposes at beginner and intermediate levels of knowledge. The family reduction principle, determining the inclusion or exclusion of affixed words, will be based solely on frequency of occurrence in a corpus.

The list we are proposing is thus not a new list constructed on the basis of corpora and according to all the requirements of list construction (Nation, 2016), but rather a reduced version of an established list. We reduced the most complete family list currently available, Nation's (2016) 25 BNC/COCA lists of 1,000 word families. These lists are close to 100% complete, although, as Nation (2016) has stated, the families continue to expand as remote members come to his attention. The lists are based on all the forms of the majority of word families in the language, as found in two very large general corpora, the BNC and the COCA. Such completeness has its uses, for example, when profiling texts or corpora for lexical frequency using computer programs such as Range (https://www.wgtn.ac.nz/lals/resources/range), AntWordProfiler (https://www.laurenceanthony.net/software/antwordprofiler/), and VocabProfile (https://lextutor.ca/vp/comp/). A highly infrequent variant of a common word like *marry* (26,773 instances in

the current 1-billion-word update of COCA), such as *unmarriageable* (27 instances), is still worth classifying in a profile under *marry*, should it appear in a text, despite the fact that it is probably not worth drawing to learners' attention at the cost of a more useful word. As discussed earlier, the BNC/COCA lists have been shown to provide the best text coverage and are considered the most useful by experienced teachers. However, because the aim of our list is to include only the derived forms that learners are likely to need when encountering or using the most frequent word families (the "nucleus" of the family, as it were), it is expected that our list will involve a substantial reduction in the number of derived words in the most frequent families of the BNC/COCA lists. We believe our study is a timely response to Dang's (2020) call to develop word lists that better meet the needs of particular groups of learners.

We base our list on the 3,000 most frequent word families rather than some other number for two reasons. First, the complete version of these families typically covers 88−95% of vocabulary in written and spoken language (Nation, 2006b) and hence provides learners with the most basic vocabulary necessary for comprehension and as a basis for further independent learning. Second, given that the families with the greatest proportion of derived forms appear in the 3,000 most frequent word families (Laufer & Cobb, 2019; Nation, 2016) (i.e., less frequent words have fewer derivations), we believe that our list will include the majority of the most frequently used English affixes, although this is to be determined. If it is the case, then exposure to and instruction in these affixes should lead learners to recognize the meanings of affixes in more advanced vocabulary as they encounter them. Because basic vocabulary is necessary in both comprehension and production, we expect our reduced list of the most frequent family members to meet both needs.

In the rest of the article, we first describe the tool that was developed for constructing nuclear lists, the Nuclear List Builder. We then describe the decisions involved in the construction of our particular list, the NFL7, and discuss its effectiveness.

Method

The Nuclear List Builder

The Nuclear List Builder is a special online tool we offer to users who wish to make their own nuclear lists, that is, to reduce any one of the first ten original BNC/COCA lists, or any combination of these lists, to any size users need. The tool is freely available at https://www.lextutor.ca/freq/nuclear/.

The construction of a nuclear list follows several stages, each of which is described in detail below: selection of BNC/COCA frequency lists (from 1 to

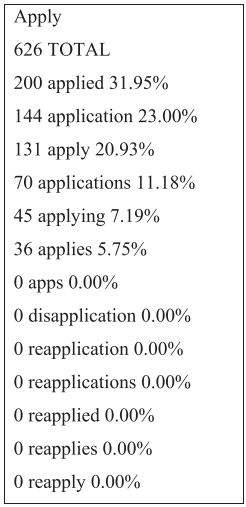

Figure 1 Results obtained from the Nuclear List Builder for the *apply* family when the British National Corpus/Corpus of Contemporary American English (BNC/COCA) second 1,000 list is crossed with the Brown corpus.

10); selection of a "cross-corpus" to cross the selected lists with; and making several decisions regarding list reduction size, mainly in terms of inclusion of family members.

The website of the Nuclear List Builder offers several cross-corpora and the option to enter the user's own corpus (of up to 850,000 words). The cross-corpus, although sizable, would typically be smaller than the BNC or the COCA and thus more representative of a learner's needs (a small corpus is unlikely to contain many very infrequent derived words, such as *unmarriageable*) or of the lexis of a domain. The user of the program chooses a BNC/COCA list from a menu and one of the cross-corpora from another menu, and then runs the program. The Nuclear List Builder "crosses" the list with the selected corpus. The original BNC/COCA list then appears with each word tagged for its frequency in the chosen cross-corpus, typically with a large number of words tagged with zeros, along with the percentage of each member within its family. For example, Figure 1 shows the results for the *apply* family when crossed with the 2-million-word combined Brown and BNC Written Sampler corpora: The base word *apply* comprises 20.93% of the *apply* family; the most frequent family member is *applied*, comprising 31.95% of the family; and seven family members do not appear in the collection at all.[2]

Typically, many individual words that are present in a complete BNC/COCA list are not found at all in a wide range of cross-corpora, or else are found very infrequently. Absences are typically of derived forms, although, in some of the specialized corpora, even entire families from the BNC/COCA list can be missing. For example, the first family in the second 1,000 list, *accent*, is entirely absent from the Engineering subcorpus of the British Academic Written English corpus (BAWE; Nesi, Gardner, Thompson, & Wickens, 2009).

After running the program as described above, the user inspects the frequencies and percentages of the list words in the chosen cross-corpus, and selects a percentage cutoff (percentage of family members to be excluded from the list). For example, in the example illustrated in Figure 1 for the *apply* word family, *applies* will be eliminated if the selected cutoff is 7%, and *applying* will be eliminated as well if the cutoff is 10%. The user then runs the program again, eliminating the words below the cutoff.

An additional feature of the tool is the option not to eliminate base words regardless of frequency. For example, in the *excite* word family, the base word *excite* does not appear on a nuclear list at 7% or at 5% cutoff because it is less frequent than its inflected and derived words *excited*, *excitement*, and *exciting*. If the users want to include the base word anyway, or even the full lemma (all inflected forms), possibly to present a more transparent list to learners, they select the option to do so.

Finally, users can also choose to exclude some words that may be overrepresented in a smaller corpus, where words used in a handful of texts can gain undue prominence. Therefore, the Nuclear List Builder offers a "frequency-in-COCA filter" option that excludes words occurring below a criterion number of occurrences (500, 750, or 1,000) in the 400-million-word version of COCA, as selected by the user. For example, the word *lordship* occurs in the first 1,000 BNC/COCA list even at 7% cutoff. However, when a "below 500" filter is applied, the word is no longer included.

Figure 2 shows the various stages (marked by numbers) of creating a nuclear family list (NFL). In the example, the original BNC/COCA list chosen to be reduced is the first 1,000 list, labeled as "1k" (Stage 1), the cross-corpus is BASE, the British Corpus of Academic Spoken English, comprising 1.9 million words (Stage 2), the selected cutoff point is 10% (Stages 3 and 4), the frequency in-COCA filter is 750 (Stage 4), and infrequent base words are not included (Stage 4). Stage 5 is running the program, and Stages 6 and 7 show the output. The output of the example shows that the number of words in the reduced list is 1,883, whereas in the original list it was 6,862. The entire

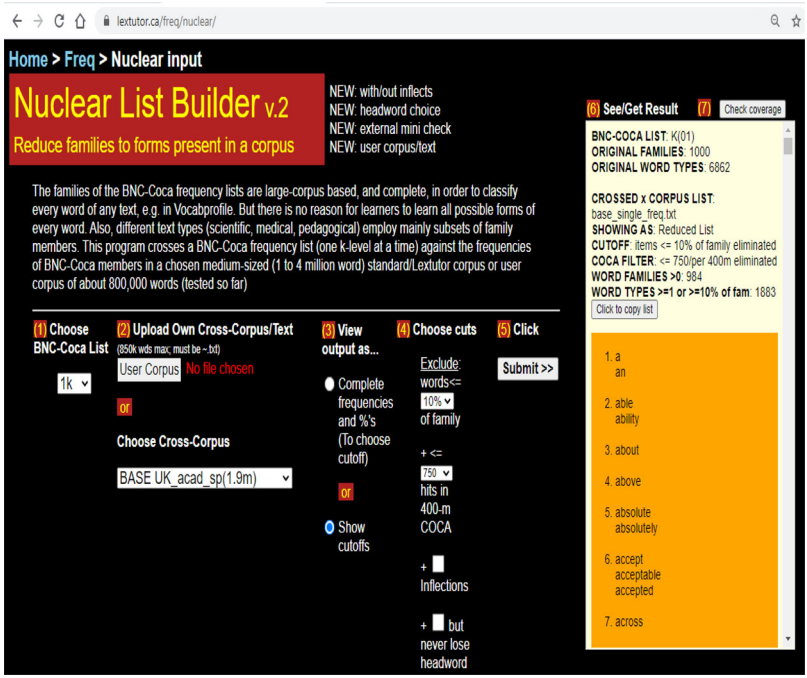

Figure 2 Nuclear List Builder screen, showing input and output. [Color figure can be viewed at wileyonlinelibrary.com]

operation can be repeated with different cutoffs until the user is satisfied with the list produced.

To sum up, the Nuclear List Builder tool can be used to construct reduced BNC/COCA lists. It provides the user with 10 original complete BNC/COCA lists, a choice of 14 cross-corpora or the option to enter a personal cross-corpus, an option to reduce the lists to different sizes (the cutoff option), an option to reduce oddities (the frequency-in-COCA filter), and an option to remove or retain headwords and inflected forms if they do not meet the cut-off choice.

Developing the NFL7

In this section, we describe the various stages in the development of the NFL7 using the Nuclear List Builder: cross-corpus analysis; testing the different cut-offs against text coverage loss in a range of test corpora; the 7% cutoff selection; and further reduction with the frequency-in-COCA filter.

Cross-Corpus Analysis

The cross-corpus assembled for developing our NFL reflects an attempt to meet several desiderata: It should be general; it should balance speech, scripted speech, and several types of writing; it should balance North American and British varieties of English; it should be large enough to be representative, yet small enough to include mainly medium- and high-frequency items; and, most important, it should represent the type and level of language that teachers and even learners themselves regard as useful for learning purposes. We know of no single corpus that meets all these criteria. We drew the components of our cross-corpus from the 10 most consulted single-file corpora on the Lextutor website over the period 2014−2019, with millions of searches recorded for each (Cobb, 2019). On the other hand, what the cross-corpus does not have to be is large or complete like those used to create the BNC/COCA in the first place; that list has already been created and, in constructing the NFL, we are pruning the BNC/COCA to meet a specific pedagogical need (and meet Dang's, 2020, call for more tailored word lists). Nor does a cross-corpus have to be particularly up to date, inasmuch as our target is high-frequency single words that are fairly constant and similarly ordered over the decades and across corpora; new or nonce items will probably not appear in the first 3,000 families, or if they do will rarely if ever amount to 7% of their families.

Believing no single corpus meets our requirements, we assembled three corpora, from among those most often consulted, that we believed would do so. The three component corpora of our cross-corpus comprise just under 4 million words: the two BNC (1985) 1-million-word samplers (https://ucrel.lancs.ac.uk/bnc2sampler/sampler.htm) of spoken and written British English, and Davies's 1.7-million-word COCA "Now" sampler corpus of US English (2019; drawn randomly from the relevant divisions of the main COCA corpus 2010–2016, as described at https://www.corpusdata.org/formats.asp and obtainable at http://corpus.byu.edu/nowtext-samples/text.zip), which is divided into the three roughly equal parts of press and academic writing, fiction and magazine writing, and US speech. (It is worth noting that our procedure thus parallels on a smaller scale the initial combining of the BNC and COCA word lists into the BNC/COCA.) An example of the Nuclear List Builder's handling of our cross-corpus for two different word families is shown in Figure 3. The total refers to the number of occurrences of the family as a whole in the cross-corpus; the percentages refer to the occurrences of each family member as a proportion of the occurrences of all family members. The lists come out sorted by number of occurrences of each family member (shown on the left).

accuse	adapt
168 TOTAL	156 TOTAL
111 accused 66.07%	45 adapted 28.85%
24 accusations 14.29%	34 adaptation 21.79%
17 accusing 10.12%	33 adapt 21.15%
8 accuse 4.76%	11 adapting 7.05%
5 accuses 2.98%	10 adaptive 6.41%
3 accusation 1.79%	9 adaptable 5.77%
0 accuser 0.00%	5 adaptors 3.21%
0 accusers 0.00%	3 adaptations 1.92%
0 accusingly 0.00%	2 adaptability 1.28%
	2 adapter 1.28%
	2 adaptor 1.28%
	0 adaptabilities 0.00%
	0 adapters 0.00%
	0 adaption 0.00%
	0 adapts 0.00%
	0 maladaptive 0.00%
	0 unadapted 0.00%

Figure 3 Results obtained from the Nuclear List Builder for two word families, *accuse* and *adapt*, when the British National Corpus/Corpus of Contemporary American English (BNC/COCA) second 1,000 list is crossed with our cross-corpus.

Figure 3 shows that many of the family members present in the original large corpora on which the BNC/COCA lists are based are little represented or not represented at all in the cross-corpus. It further shows that even the headword may not be particularly well represented (*accuse*, with eight occurrences in 4 million words, represents only 4.76% of occurrences of its own family).

Figure 4 shows the same two families as Figure 3, but this time as nuclear families of two possible sizes. On the left, the nuclear family consists of members that comprise at least 5% of their family occurrences, preserving the headword even if it does not meet that cutoff. On the right, the members comprise at least 10% of all family occurrences and do not include the headword if it does not meet that cutoff.

As could be predicted from Figures 2 and 3, the entire nuclear lists are substantially reduced in size, in terms of the total number of words they contain. For example, with items under 5% removed, preserving base words, the second BNC/COCA list of 1,000 families with 6,371 members is reduced to 3,024 members (47.0% of the original list); with items under 10% (including the base word) removed, the original list is reduced to 2,383 members (37.4% of the original list).

5% cutoff	10% cutoff
accuse	accusations
accusations	accused
accused	accusing
accusing	
adapt	adapt
adaptable	adaptation
adaptation	adapted
adapted	
adapting	
adaptive	

Figure 4 Nuclearization of two word families at two different cutoffs.

Finding the Optimal Reduction of the NFL

In our search for the best possible cutoff point, we first created six sets of provisional lists, at family membership cutoffs as follows: 0%, 5%, 10%, 15%, 20%, and 25% of the family's total occurrences for the first three 1,000-word BNC/COCA word frequency bands. The choice of these cutoffs was based on our pilot exploration of the Nuclear List Builder's output (e.g., that shown in Figure 3), which had seemed to suggest that the optimum would be more than 5% and less than 15%. To find the optimal reduction, we weighed the reduction in the word count or list size of each candidate list against loss in text coverage over a set of test corpora. These test corpora were different from, but similar in size and character to, those comprising the cross-corpus that we used to create the provisional NFLs at different cutoff points.

The test corpora were the following:

- the LOB (Lancaster-Oslo-Bergen corpus; Johanssen, Leech, & Goodluck, 1978), containing 1 million words of written British English;
- the Brown corpus (Kucera & Francis, 1967), the US counterpart of the LOB corpus, about 1 million words of written US English of similar composition to the LOB and also used by Nation (2006b) for list validation;
- the 1.1 million words of British Academic Spoken English (BASE; Nesi, Gardner, Thompson, & Wickens, 2010) to roughly balance the written nature of the Brown and LOB corpora;
- a collection of presidential speeches compiled by a colleague of the first author (Henrichon, 2013), containing 1.1 million words, to include a component of scripted speech and also to bolster the US speech component; and
- a corpus of Oxford University Press's Bookworms series of simplified classic stories to balance the academic bias of the BASE.

Again, all of these are in the top Lextutor downloads for concordance runs and other pedagogical outputs (further discussed below) and hence are the kind of language that learners are likely to encounter because they or their teachers have chosen to encounter it.

Table 1 shows the coverage of the first three complete BNC/COCA lists, followed by reduced versions of these lists at six levels of reduction across the set of test corpora. All coverages are provided by Lextutor's Coverage Calculator at https://www.lextutor.ca/cover/ and are publicly verifiable; all coverages include elimination of proper nouns, which are typically not considered to be part of lexical knowledge in coverage studies.

The mean coverages across corpora and standard deviations are provided for the combined three lists in the rightmost column. The first dataset shows that the complete original lists comprise 19,062 individual word types and cover an average of 88.83% ($SD = 7.14$) of text lexis across the test corpora employed. Then, with the elimination of items that do not appear at all in any of the test corpora (named Nuclear/0%), the combined list size is reduced by about 30% at 13,485 words, but the mean coverage it provides is reduced by only 1.27% to 87.56 ($SD = 6.49$). Next, with the lists reduced by eliminating words comprising less than 5%, 10%, 15%, 20%, and 25% of their families' occurrences, the combined list size (rounded) is reduced to 9,000, 7,000, 6,000, 5,000, and 4,000 words, respectively, and coverage across the test corpora is reduced to mean (rounded) coverages of 86%, 83%, 78%, 76%, and 71%, respectively, each within a narrow range of standard deviations. The task is therefore to find the balance point between reduction of list size and loss of coverage.

We judged that no amount of list reduction was worth allowing coverage to go below 80% as a return for learning 3,000 families. Coverage dips to 78.43% with 15% list reduction, and we therefore focused our search for a balance point in the zone of 10%. Coverage calculations were made at each percentage point from 5% to 13%, as shown in Table 2a. Between the 6% and 12% cutoff points, the coverage loss at each subsequent cutoff point is less than 1%. Table 2a also shows that there are two points (7% and 10%) where coverage loss appears to reach a minimum and then start increasing again. It would appear that our best candidate for the NFL cutoff is either 7% or 10%.

We next explored the 6%−12% area in still more detail, calculating the cumulative list size reduction, the cumulative coverage loss, and the ratio between the two (list reduction/coverage loss). This ratio shows how many words can be reduced from the list in return for losing 1% coverage. The higher the figure at a specific cutoff point, the more efficient the reduced list is. Table 2b shows cumulative list reduction, cumulative coverage loss, and the ratio between them at

Table 1 Corpus coverage results for complete BNC/COCA lists and for Nuclear Family Lists with different cutoff points

		Corpus coverage (%)					
Lists with cutoffs	Size (types)	Brown	LOB	BASE	Graded stories	Presidents' speeches	Mean (*SD*)
BNC/COCA lists							
1k	6,859	71.47	77.49	68.68	86.71	76.98	
2k	6,344	8.57	8.36	5.56	4.97	11.53	
3k	5,859	5.12	4.77	4.25	1.54	8.13	
1–3k	19,062	85.16	90.62	78.49	93.22	96.64	88.83 (7.14)
Nuclear/0%							
1k	4,936	70.52	76.28	70.60	84.03	75.97	
2k	4,486	8.45	8.26	3.74	4.91	11.46	
3k	4,063	5.07	4.73	4.17	1.53	8.06	
1–3k	13,485	84.04	89.27	78.51	90.47	95.49	87.56 (6.49)
Nuclear/5%							
1k	2,555	69.06	74.79	69.76	82.86	74.61	
2k	3,015	8.05	7.9	5.26	4.76	10.97	
3k	3,047	4.87	4.56	1.52	1.48	7.80	
1–3k	8,617	81.98	87.25	76.54	89.10	93.38	85.65 (6.53)

(*Continued*)

Table 1 (Continued)

Lists with cutoffs	Size (types)	Corpus coverage (%)					Mean (*SD*)
		Brown	LOB	BASE	Graded stories	Presidents' speeches	
Nuclear/10%							
1k	2,040	66.93	72.45	64.87	80.70	72.28	
2k	2,350	7.54	7.45	4.98	4.53	10.43	
3k	2,408	4.62	4.34	3.81	1.41	7.50	
1–3k	6,798	79.09	84.24	73.66	86.64	90.21	82.77 (6.50)
Nuclear/15%							
1k	1,716	63.65	68.82	62.17	76.25	68.59	
2k	1,919	6.99	6.9	4.61	4.27	9.59	
3k	1,978	4.34	4.09	3.50	1.33	7.07	
1–3k	5,613	74.98	79.81	70.28	81.85	85.25	78.43 (5.88)
Nuclear/20%							
1k	1,459	61.88	66.93	60.83	74.02	65.51	
2k	1,636	6.55	6.47	4.34	4.02	9.07	
3k	1,645	4.02	3.84	3.28	1.25	6.57	
1–3k	4,740	72.45	77.24	68.45	79.29	81.15	75.72 (5.20)

(*Continued*)

Table 1 (Continued)

Lists with cutoffs	Size (types)	Corpus coverage (%)					Mean (*SD*)
		Brown	LOB	BASE	Graded stories	Presidents' speeches	
Nuclear/25%							
1k	1,293	57.79	62.35	57.07	68.73	62.03	
2k	1,402	6.08	5.96	4.04	3.74	8.47	
3k	1,391	3.74	3.58	3.09	1.18	6.16	
1−3k	4,086	67.61	71.89	64.20	73.65	76.66	70.80 (4.93)

Note. "1k," "2k," and "3k" refer to the first, second, and third 1,000 lists, respectively; "1−3k" refers to the three lists combined. "Nuclear/0%" refers to a Nuclear Family List with a 0% cutoff point (only those words are removed that did not appear in the cross-corpus at all). BASE = British Academic Spoken English; BNC = British National Corpus; COCA = Corpus of Contemporary American English; LOB = Lancaster-Oslo-Bergen corpus.

Table 2a List size vs. coverage: the search for a balance point

				Loss	
Cutoff	Coverage (%) of 1−3k	*SD*	Size (words)	Size	Coverage
5%	85.65	6.53	8,617	4,868	1.910
6%	84.89	6.71	8,099	518	0.762
7%	84.39	6.67	7,709	390	0.494
8%	83.57	6.58	7,343	366	0.822
9%	83.11	6.57	7,017	326	0.462
10%	82.77	6.50	6,798	219	0.342
11%	81.99	6.71	6,480	318	0.776
12%	81.45	6.32	6,227	253	0.538
13%	79.96	6.40	6,008	219	1.498

Table 2b List size vs. coverage (cov.): the search for a balance point, taken further

Cutoff	Cov. (%) of 1−3k	Cumulative cov. loss	List size (words)	Cumulative list size reduction	List reduction per 1% cov. loss
5%	85.65		8,617		
6%	84.89	0.76	8,099	518	681
7%	84.39	1.25	7,709	908	726∗
8%	83.57	2.07	7,343	1274	614
9%	83.11	2.53	7,017	1600	634
10%	82.77	2.87	6,798	1819	633
11%	81.99	3.65	6,480	2137	584
12%	81.45	4.18	6,227	2390	571

Note. The asterisk indicates the highest list reduction/coverage loss ratio, which occurs at the 7% cutoff point.

each percentage cutoff point. It appears that the highest list reduction/coverage loss ratio is at 7% cutoff (asterisked in the table), which makes 7% the cutoff we were looking for.

An additional reason for choosing the 7% cutoff is pedagogical. Because the list is intended for nonadvanced learners, with knowledge of fewer than 3,000 words, we specifically noted the coverage of the Bookworm graded readers provided by lists at various reduction levels (Table 1). We chose 7% because it provided sufficient coverage of graded readers. Together with 2−3% proper nouns in graded readers (Nation, 2006b), the NFL at 7% cutoff provides

Table 3 Nuclear Family List 7 coverages across test corpora (NFL/7% with COCA filter)

NFL list	Size	Brown	LOB	BASE	Stories	Speeches	Mean (*SD*)
1k	2,271	68.22	73.74	65.74	81.89	73.63	
2k	2,504	7.83	7.62	5.11	4.61	10.74	
3k	2,518	4.72	4.39	3.90	1.42	7.61	
1−3k	7,293	80.77	85.75	74.75	87.92	91.98	84.23 (6.67)

Note. "1k," "2k," and "3k" refer to the first, second, and third 1,000 lists, respectively; "1−3k" refers to the three lists combined. BASE = British Academic Spoken English; COCA = Corpus of Contemporary American English; LOB = Lancaster-Oslo-Bergen corpus.

90% text coverage. There is now preliminary evidence that at 90% coverage, readers can infer the meanings of enough unfamiliar words to increase their comprehended text vocabulary to 95% and thus achieve text comprehension (Laufer, 2020b).

Having produced a preliminary NFL at 7% cutoff, comprising 7,709 word types, we scanned it to intuitively identify any oddities (which, as mentioned earlier, are always possible in a smaller corpus, where words used in a handful of texts can gain undue prominence). Having found a small number of these (e.g., *pacers* and *potters*), we applied the frequency-in-COCA filter of occurrences to eliminate any words with frequencies of less than 1,000 per 400 million. The results confirmed our intuitions, with 725 COCA occurrences of *pacers* (0.00018% of the words in the COCA corpus) and 551 of *potters* (0.00013%; verifiable at https://lextutor.ca/vp/coca/). Application of this filter led to a further reduction of 416 words to 7,293 words, providing a mean coverage of 84.23%, *SD* = 6.67 (Table 3).

In a final list nuclearization procedure, we reunified, across the three lists, roughly 100 families that had been separated in the original BNC/COCA lists, reflecting Nation and Webb's (2011) practice of including in a family only members containing the base word as a free or independent morpheme "able to stand as … independent word[s]" (pp. 136−137). For example, 1k item *apparent* by this definition is a different family from *appear* because it does not literally contain *appear*, and *appar-* is not an independent morpheme. Our rationale for modifying these separations was pedagogical: The relationships between the separated items are normally transparent. Accordingly, we placed the items into sets that we judged could be grouped around a single stem or base word without posing a learning problem. For example, in the original lists

Table 4 BNC/COCA3000 and the Nuclear Family List (NFL) in types, families, and lemmas

	Types		Families		Lemmas (flemmas)	
	BNC/COCA	NFL7	BNC/COCA	NFL7	BNC/COCA	NFL7
1k	6,859	2,310	1,000	975	3281	1839
2k	6,344	2,537	1,000	976	2996	1915
3k	5,859	2,446	1,000	936	2855	1856
Total	19,062	7,293	3,000	2,887	9,132	5,610

Note. "1k," "2k," and "3k" refer to the first, second, and third 1,000 lists, respectively. BNC = British National Corpus; COCA = Corpus of Contemporary American English.

theory and *theories* composed one family in the third thousand list, while *theoretical* and *theoretically* are another, because these two do not literally contain the base word or stem *theory*; we unified these four items into a single family, and the total number of families was reduced accordingly. The number of word types remained the same and we continued to employ the names "1k" and so on. We also decided not to include any headwords that had not achieved the criterion proportion of their families, in order to preserve a strictly frequency-based method for testing our list. Table 3 shows the coverage of the NFL7 across test corpora after the reduction process, application of the frequency-in-COCA filter set to 1,000 occurrences, and consolidation of about 100 families.

Our NFL list thus organized comprises 7,293 word types, which form 2,887 families or 5,610 lemmas (strictly, flemmas, because part of speech was not considered), as calculated by the Familizer/Lemmatizer tool (available at https:// lextutor.ca/familizer/). This can be contrasted with the original 3,000 families and 19,062 word types in the complete BNC/COCA3000. Our final NFL list can be viewed in Appendix S2 in the Supporting Information online; it can also be viewed and checked for coverage in the test corpora at https: //lextutor.ca/cover. Table 4 shows the sizes of the original BNC/COCA3000 lists and the reduced NFL7. The NFL7 consists of 38% (7,293/19,062) of the BNC/COCA word types and 61.4% (5,610/9,132) of its lemmas.

Results

Evaluating the NFL7

The evaluation of the NFL7 revolved around two research questions:

1. How does the NFL7 compare, for text coverage and list size, across a range of corpora representing text types that learners would be

likely to encounter, to the following word lists: (a) BNC/COCA3000, (b) BNC/COCA3000 Level 3 partial, (c) the NGSL3000, and (d) the NGSL2500?

2. What derivational affixes are included in the NFL7, and are they among the most frequent affixes as found by Laufer and Cobb (2019)?

The first question examines the NFL7 in light of the cost–benefit principle, asking whether it can provide text coverage that is similar to other lists, some of which are longer and more complete. The second question examines the morphological dimension of the NFL7, asking whether it includes the most useful derivational affixes as identified by Laufer and Cobb (2019). We will discuss each question separately in terms of methodology and results.

The NFL7 and Other Word Lists: List Size and Text Coverage

Part of the answer to our first research question is implicit in the cutoff selection described already, whereby the original BNC/COCA3000 of 19,062 items gave a mean coverage of 88.83% across the test corpora, whereas NFL7 was reduced to 7,293 items while still providing a mean coverage of 84.23% ($SD = 6.67$). It is traditional, however, to test proposed word lists against texts or corpora other than those used in their development. The corpora we chose for comparing the lists will be referred to as the "comparison corpora." As in the case of the cross-corpus and test corpora used in the development of NFL7, in selecting the comparison corpora, we have again made choices based on a range of dimensions (US, UK; speech, writing; literary, expository; formal, informal) and typicality in what learners are likely to encounter at different stages and types of learning. To represent what learners are likely to encounter, we have again chosen top teacher and learner corpus selections on the Lextutor website, as well as other collections and texts, as follows:

- the combined BASE and BAWE samplers (British academic spoken and written sampler corpora of 2.2 million words, described at http://www.coventry.ac.uk/bawe);
- a corpus of Wikipedia entries assembled by trainee teachers of English as a second language and intended for use by learners, divided into 10 topic areas (1,052,000 words);
- a long-running television series (*House, M.D.*): entire eight-season script of 800,000 words of US scripted speech in episodic sequence);
- the US speech section of Davies's previously used COCA "Now" sampler (387,000 words);

- two collections of texts designed especially for learners at two specified lexical levels, namely, assemblies of Nation's parallel text sets of Mid-Frequency Graded Readers at the 4k and 8k levels of 553,000 and 555,000 words, respectively (described by Nation & Anthony, 2013); and
- a single extended work of fiction, Lawrence's *Lady Chatterley's Lover*.

These corpora or collections each meet at least two out of the following five learner usability conditions, as follows:

1. They are frequently selected by learners in concordance searches on Lextutor (BASE, BAWE, Wikipedia).
2. They are recommended in a manual of English as a second language that employs a data-driven learning approach (Karpenko-Seccombe, 2021; BASE, BAWE).
3. They were designed or chosen as appropriate for learning by MA students training to be teachers of English to speakers of other languages (*House, M.D.*, Wikipedia, Mid-Frequency readers, Presidential speeches).
4. They were used in previous pedagogical coverage studies (*Lady Chatterley's Lover*, Mid-Frequency readers).
5. Their lexis comprises an atypically high proportion of items from the first two 1,000-levels of the BNC/COCA (90% or more, as termed comprehensible input for nonadvanced learners by Laufer, 2020b; COCA Speech, Mid-Frequency readers, *House, M.D.*, Presidents, *Lady Chatterley's Lover*).

All of the four word lists we have chosen to compare to our NFL in terms of their size and coverage are based on large corpora, were constructed during the past 10 years, and are roughly commensurate in size with NFL's 3,000 families; in addition, three of them have appeared in recent coverage research. The lists are as follows:

- Nation's BNC/COCA 3000 (lists of the 3,000 most frequent words in the original BNC/COCA 25,000);
- Nation's partial version of these same lists, set to a selection of derived words with high-frequency Level 3 affixes;
- Brezina and Gablasova's (2015) lemma-based NGSL2500; and
- Browne's (2013) flemma-based NGSL3000.

The last of these, although large-corpus-based and much used by practitioners (as shown by Lextutor's Vocabprofile statistics), has not been extensively studied for coverage.

When we compare the coverage each list provides, the basis of comparison is the total number of items in each list, that is, the total number of family or lemma members, without reference to the overall number of lemmas or families. Thus we examine the coverage of texts by 7,293 words of NFL7; 19,062 words of the BNC/COCA3000; 10,644 words of the BNC/COCA partial; 8,666 of the NGSL3000 (described at https://www.newgeneralservicelist.org/); and 5,115 of the NGSL2500 (Brezina & Gablasova, 2015).[3] The number of words in the NGSL2500, however, is an estimate because the true total number is unknown; the published list consists only of lemma headwords whereas the dedicated software needed to run it clearly contains other lemma members in addition, though not all of them. Brezina and Gablasova (2015) state that the entire list consists of 5,115 words including headwords and inflected words, but we do not know which inflected words these are. When we input several word families into the LancsLex: Lancaster Vocabulary Analysis Tool, as provided by the authors to accompany the list (available at http://corpora.lancs.ac.uk/vocab/analyse.php), we realized that not all traditional members of every lemma are recognized by the tool: For example, *abandoned* is counted as "off-list" in the input *abandon*, *abandoned*, *abandoning*, *abandons*; *admits* and *admitting* are off-list in the input *admit*, *admits*, *admitted*, *admitting*; and similarly, *continued* is off-list, whereas *continue*, *continues*, and *continuing* are in the list.

The coverages of the comparison corpora by the lists were determined by a Lextutor routine called Coverage Calculator (available at https://lextutor.ca/cover/), which allows the user to choose a list from a menu or paste in another list, then choose a corpus or text for coverage analysis, and decide whether the coverage should include or eliminate proper nouns.

To compare the four vocabulary lists and the NFL7 for size and coverage, we have developed a comparison index or ratio, which we believe to be an innovative measure, although a simple one. We divide the number of words in a list by the coverage it provides, calculating how many words on average are necessary to cover 1% of a particular corpus. For example, if a word list contains 8,000 words and achieves 80% coverage in a corpus, then the size/coverage ratio is 100; that is, an estimated average of 100 words in the list cover 1% of the corpus. The lower the index, the more efficient the list, because fewer words in the list cover 1% of the words in the corpus. This index allows us to compare lists of different sizes for both learning cost (the number of words to learn) and learning benefit (the coverage that learning these words provides). The results of the list comparison appear in Table 5, which shows the coverage and coverage/size index for each of the five word lists (identified in the column

Table 5 Coverage (cov.) and coverage/size index in each of seven corpora across five lists of different sizes

	BNC/COCA3000 (19,062 words)		BNC/COCA Partial (10,644 words)		NFL7/3000 (7,293 words)		NGSL2500 (5,115 words)		NGSL3000 (8,342 words)	
Corpus	Cov.	Index	Cov.	Index	Cov.	Index	Cov.	Index	Cov.	Index
BASE/BAWE	81.9	232.7	80.5	132.2	77.8	93.7	83.6	61.2	78.5	106.3
COCA speech	89.8	212.3	88.6	120.1	84.4	86.4	90.0	56.8	86.6	96.3
House, M.D.	89.6	212.7	88.4	120.4	83.7	87.1	91.9	55.7	86.5	96.4
M-F Readers 4k	95.3	200.0	93.7	113.6	88.1	82.7	93.2	54.9	90.8	91.9
M-F Readers 8k	94.0	202.8	92.5	115.1	86.9	83.9	92.4	55.4	89.8	92.9
Wiki Corpus	87.6	217.6	84.3	126.3	83.2	87.7	79.8	64.1	82.0	101.7
Lady Chatterley	92.2	206.7	90.6	117.5	87.4	83.4	90.9	56.3	88.3	94.5
Mean	90.0	**211.1**	88.4	**120.7**	84.5	**86.4**	88.8	**57.8**	86.1	**97.1**
SD	4.5	10.9	4.6	6.5	3.5	3.7	5.1	3.5	4.4	5.1

Note. Mean indexes are shown in bold. BASE = British Academic Spoken English; BAWE = British Academic Written English; COCA = Corpus of Contemporary American English; NFL = Nuclear Family List; NGSL = New General Service List.

headings, along with their sizes) in each of the seven corpora or text collections (identified in the leftmost column). Coverage means and size/coverage index means, with their standard deviations, are in the bottom two rows.

In terms of average size/coverage ratio, the bottom part of the table shows the NGSL2500 to be lowest (best) at 57.8, followed by the NFL7 at 86.4, the NGSL3000 at 97.1, BNC/COCA-3000 Partial at 120.7, and the BNC/COCA3000 Complete at 211.1. However, the NGSL2500 will not feature in the rest of the discussion, because, as mentioned, the size is an estimate and the full list is not available to practitioners except as embedded in the authors' software. The NFL7 is thus the smallest accessible word list with the highest coverage in texts that language learners are likely to encounter.

The NFL7 and Derived Family Members

Our second research question was as follows: What derivational affixes are included in the NFL7, and are they among the most frequent affixes as found by Laufer and Cobb (2019)?

To answer this question, we analyzed the NFL7 using Morpholex (available at https://lextutor.ca/morpho/lex/), a text analysis tool that extracts all the affixes from input lists or texts with their absolute and relative frequencies in the input. (For a detailed description of this software, see Laufer & Cobb, 2019, who used it to discover that the proportion of derived forms in texts is much smaller than often expected and hence pedagogically manageable.) The Morpholex analysis of the NFL7 showed that it contained a total of 52 derivational affixes in 1,404 derived words, but that only 22 of these occurred 15 times or more in the entire list, in at least 1% of all the derived words, comprising just under 85% of the total number of affixes. Just 12 affixes comprise more than 70% of the total.

The full list of affixes is available in Appendix S1 in the Supporting Information online. Table 6 presents the 22 most frequent affixes, and the frequency of each affix, in terms of raw score and cumulative percentage, in each of the three NFL7 sublists (first, second, and third 1,000) as well as in the entire NFL7.

Table 6 shows that some affixes appear very frequently and evenly in the three NFL7 sublists (e.g., *-ly, -er, -ment*), whereas others tend to appear in relatively less frequent third 1,000 words (e.g., *-ion, -al, -ity*). The overall (total) number of affixed words seems to increase with a decrease in word list frequency. In the first 1,000 words, only six affixes occur more than 10 times; in the second 1,000, an additional three, making nine altogether; in the third 1,000, an additional five, making 14 altogether. This shows that at each 1,000

Table 6 The most frequent derivational affixes in the Nuclear Family List 7 (NFL7), sorted by total number of occurrences

Rank	Affix	1k	2k	3k	Total	%	Cum. %
1	-ly	69	70	98	237	16.88	16.88
2	-ion	12	50	105	167	11.89	28.77
3	-al	11	30	53	94	6.70	35.47
4	-er	25	28	29	82	5.84	41.31
5	re-	3	39	37	79	5.63	46.94
6	-y	25	33	16	74	5.27	52.21
7	-ation	8	17	33	58	4.13	56.34
8	-ment	13	18	17	48	3.42	59.76
9	-ive	2	12	26	40	2.85	62.61
10	-ity	2	10	28	40	2.85	65.46
11	in-	5	6	27	38	2.71	68.16
12	-or	5	8	18	31	2.21	70.37
13	-ist(s)	5	2	20	27	1.92	72.29
14	-able	6	9	10	25	1.78	74.07
15	-ic	1	8	14	23	1.64	75.71
16	-ance	4	9	6	19	1.35	77.07
17	-ness	7	8	3	18	1.28	78.35
18	-ful	7	9	2	18	1.28	79.63
19	un-	7	8	2	17	1.21	80.84
20	pro-	2	6	9	17	1.21	82.05
21	ex-	5	5	5	15	1.07	83.12
22	-ence	1	5	9	15	1.07	84.19

Note. "1k," "2k," and "3k" refer to the first, second, and third 1,000 lists, respectively. Cum. = cumulative.

level, learners will have to learn a relatively small number of derived words that are formed with a limited number of affixes. For example, at the first 1,000 level, there are 155 words that are formed with the six most frequent affixes and 225 words formed with all of the 22 affixes shown in Table 6. Twenty-one of the affixes in the list are among the 22 most frequently used affixes found by Laufer and Cobb (2019). Their research showed that a small number of frequent affixes, together with base words and inflections, provided the necessary lexical coverage for basic comprehension, specifically that 95% of text coverage was reached with three or four derivational affixes in academic and newspaper texts, one affix (*-ly*) in novels, and none at all in graded readers.

Discussion

In the Method section we described the construction of the NFL7 by finding the optimal reduction of BNC/COCA3000, and in the Results section we evaluated the list by comparing it to other word lists in terms of list size/coverage efficiency and by exploring its derivational morphological composition in terms of its most frequent affixes. In this section we will discuss the unique properties of NFL7 as demonstrated by its construction and evaluation.

The idea of family reduction is not entirely new (having been broached by Greene & Coxhead, 2015, and Nation, 2016), but our reduction principle differs from previous approaches in three ways. First, the reduction of derived family members follows an objective criterion of frequency of the derived words, as opposed to Nation's (2016) BNC/COCA Level 3 partial, where the inclusion of derived words is based on a linguistic hierarchy. Second, unlike Greene and Coxhead's (2015) Academic Vocabulary for Middle School Students, our list was designed for nonnative learners, providing them with a limited number of frequent and useful derived words in English.

Third, through nuclearization, the amount of list reduction can be adjusted by the algorithm we provide to users wishing to produce their own lists suitable for different stages of learning. For example, learners looking for a focus on a lower frequency vocabulary zone like the BNC/COCA tenth 1,000 list would see 2,982 family members reduced to 943, or 32% of the original, through nuclearization at the 7% cutoff (in the same cross-corpus as described earlier). Or the same BNC/COCA list can be reduced to just the main membership in a first set of course materials, and then to a second, more expansive set of members in subsequent materials, simply by selecting a different cutoff in the cross-corpus. Which list would be appropriate in a particular case would depend on the level of the learners and the purpose of the list: whether as a vocabulary syllabus, a control on examination scripts, the content of a set of flash cards, or many other possibilities. None of the other approaches to list building has this flexibility.

Compared with other fully accessible word lists, the NFL7 offers the best balance of size and coverage, although, as mentioned, users can easily make their own lists to suit other priorities. When compared with the complete BNC/COCA3000, the NFL7 provides an average of just 5.5% less coverage for a list 38% of the size of the original. In terms of coverage/size ratio, the NFL7 is more than doubly efficient, with a size/coverage ratio of 86.4 as opposed to 211.1. The BNC/COCA3000 lists perform less well in this regard because their goal is not size minimization, but rather the opposite, exhaustiveness. The NGSL3000 performed well, very similarly to the NFL7, but with

an additional 1,049 word types to learn and without offering the adaptability that NFL7 does. The BNC/COCA Level 3 partial did not perform as well as the other lists. The fact that the size/coverage index of the NFL7 is better than that of the BNC/COCA Level 3 partial suggests a broader principle, that individual word frequency (a bottom-up approach) is a better basis for list reduction than wholesale elimination/inclusion determined by a theoretical level (a top-down approach).

We do not claim that the coverage provided by the NFL7 is sufficient for comprehension of authentic, advanced texts. A coverage of 98% is considered optimal and 95% minimal for that purpose (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt, Jiang, & Grabe, 2011), although this has arguably been revised down recently to 90% in certain cases (Laufer, 2020b). However, none of the other lists at the 3,000 level offer the necessary coverage either. The highest coverage in the comparison corpora of 90% is provided by the full BNC/COCA list, against 84.5% by the NFL7. We believe that 5.5% loss in text coverage in exchange for about 11,800 fewer word types to learn is a good trade-off.

As for the morphological makeup of the NFL7, a small and manageable number of affixes appeared 15 or more times in all three sublists. Of these 22 affixes (Table 6), 21 are among the 22 most frequently used affixes identified by Laufer and Cobb (2019). Hence, through nuclearization, the family members that were left in the list included derived words with the most useful affixes. By comparison, in an analysis of similarly sized lists, the NGSL3000 and NGSL2500, again using the Morpholex tool, we found that the number of affixes appearing in more than 15 words was only seven in the NGSL3000 and six in the NGSL2500. In other words, these lists contain far more different affixations than the NFL7. Dang and Webb (2016a) argue that for beginning learners of English, a list based on the lemma/flemma is better than a family list, because these learners have very limited morphological knowledge, and word families contain both high- and low-frequency members. This is certainly true of complete family lists, but not of the NFL7. In our judgment, the NFL7's 22 affixes per 3,000 word families does not constitute a learning overload, but rather sets a modest yet generative target of derivational morphology.

Despite the aforementioned advantages of the NFL7, a critic could argue against constructing a pedagogical list on the basis of frequency only, disregarding features like usefulness and learner knowledge. Although we acknowledge the importance of word usefulness, as suggested by Dang et al. (2020), we believe that different words may be useful in different learning contexts.

We therefore suggest that educators who adopt our list can add or delete words according to what they regard as useful in their specific contexts or, better, put their own cross-corpus into the Nuclear List Builder to create a version that may better reflect their learners' needs. As for the learner knowledge principle, as in familiarity-based lists (Brysbaert et al., 2020), we advise caution in determining an order of teaching that is based on what learners in a particular learning context know. Sometimes words that are not known by many learners are precisely the ones they need to know and that should be attended to. For example, L1-French learners of English are familiar with many cognates of Latin origin that happen to be among the less frequent English words; and yet what these learners do not know but need to know is the more basic and frequent Anglo-Saxon vocabulary (Cobb, 2000). Similarly, some words are more difficult to learn than other words due to inherent difficulty or crosslinguistic difference (Laufer, 1990; 1997; Peters, 2020). These words may not necessarily be known well by many learners, which is why they deserve to be noticed by educators.

Pedagogical Implications

Our primary motivation for constructing the NFL7 was pedagogical. We excluded many derived words of low-frequency, high-learning burden, and questionable usefulness from high-frequency word families. On the other hand, we did not separate closely related family members into separate lemmas that often appear in different frequency bands of lemma-based lists. In keeping them together, we believe we have adhered to a sound pedagogical principle, that seeing relationships between different lemmas sharing the same or similar form is typically quite easy (Nation, 2016).

We believe that nuclearization of a comprehensive word list can contribute to the teaching and testing of productive as well as receptive word knowledge. An argument that is leveled against family-based tests is that such tests cannot assess productive knowledge. Although learners may recognize that words with the same stem are related in meaning, the knowledge of one of them does not necessarily mean that the other derived forms can be produced successfully in either speech or writing. Even supporters of word-family-based tests (Nation & Beglar, 2007; Webb, Sasao, & Balance, 2017) concede that such tests are suitable for assessing receptive knowledge only. The limitation to receptive knowledge is justified in the case of traditional, extended families. For example, the BNC/COCA's *adapt* family includes *adapt*, *adaptability*, *adaptable*, *adaptation*, *adapters*, *adapted*, *adapting*, *adaption*, *adaptive*, *adaptor*, *maladaptive*, and *unadapted*. Teachers will probably not teach some of these words,

and learners may never meet them in language input. With nuclear families, however, it is not unreasonable to teach an entire family whose words include the most frequent members only and are constructed with a limited number of affixes. The nuclear family of *adapt* is composed of *adapt*, *adaptation*, *adapted*, and *adapting*—three inflected and one derived form. The selective inclusion of the most necessary words makes the nuclear family a good candidate for the teaching and testing of productive knowledge.

We also believe that, among the available lists, the NFL7 is the best source of derivational morphology instruction and practice. As mentioned earlier, a common justification for lemma-based lists and tests is that learners do not possess morphological knowledge, even receptive knowledge, let alone productive. We contend that if this is indeed the situation, the solution is not to avoid English morphology and consequently perpetuate the problem, but to teach it. But which and how many affixes could reasonably be taught to beginners and intermediate learners? Sasao and Webb's (2017) complete list of 118 affixes, or the BNC/COCA's 81 affixes, seem to be unrealistic in terms of focused teaching. At the other end of the spectrum, the list of four derivational affixes *un-*, *-ly*, *-er*, and *-th* that feature in the BNC/COCA Level 3 partial does not appear challenging enough. As an alternative to both, teaching the limited number of frequent affixes in the NFL7 should provide the necessary morphological knowledge without imposing learning overload.

Conclusion

The last two decades have witnessed the development of many high-quality lists for general and specific language-learning purposes, all deriving from work in corpus building, computer hardware expansion, software development, and learner needs analysis. However, we believe, as a result of the nuclearization process behind the NFL7, that our list has advantages that the others lack: the inclusion of the most frequent derived family members, an optimal list size/text coverage ratio, and the potential for developing derivational morphology awareness. The NFL7 reconciles two main counting principles, family and lemma-based, capitalizing on the advantages of both. Our unit of counting words is neither the lemma/flemma nor the traditional, extended word family, but the nuclear family. An added value of the NFL approach is a special online tool for users who wish to make their own nuclear lists (NFL10 etc.). Our nuclearization algorithm can produce any reduction, with any cutoff points, in any BNC/COCA list, and in conjunction with the Coverage Calculator can quickly indicate the size/coverage parameters.

In the introduction to his chapter "Critiquing a Word List," Nation (2016, p. 119) comments, "Word lists are a bit like a black hole that seems to absorb hours and hours of work for little obvious improvement." We hope that the hours and hours of work absorbed in constructing the NFL7 will contribute to better understanding and use of word-family-based lists, materials, and tests.

Final revised version accepted 2 January 2021

Notes

1 The NGSL3000 word list is arranged by flemmas, that is, lemmas that do not distinguish parts of speech with a similar form (e.g., *jump* as a verb and a noun).

2 For example, in early trials of the software it was found that, whereas the first 1,000 BNC/COCA families include 6,866 family members in total, with headwords, inflected members, and derived members combined, even a general corpus like Brown (comprising 1 million words of 1970s US English) uses only 4,723 of these; a Graded Readers corpus (of 900,000 words) uses only 4,160; a specialist corpus like BNC-Medical (1.4 million words) only 3,459; BAWE Engineering (British Academic Written English; 440,000 words) only 3,283; and BNC Law (2.2 million) only 2,569. (All corpora used in this article are described and can be consulted at https://www.lextutor.ca/conc/eng/).

3 As the published version of the NGSL2500 provides headwords only without their inflectional forms, a coverage comparison with other lists is imprecise and may be biased in favor of the NGSL2500. When we lemmatized the headword list by Lemmatizer (at https://www.lextutor.ca/familizer/) and included all the inflected forms of the lemmas, the size we obtained was 6,491 words. However, we have included the published version of the list size in view of its widespread use.

Open Research Badges

This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. All data and materials that the authors have used and have the right to share are available at https://lextutor.ca/freq/nuclear and http://www.iris-database.org. All proprietary materials have been precisely identified in the manuscript.

References

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*, 253–279. https://doi.org/10.1093/ijl/6.4.253

Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the *New General Service List*. *Applied Linguistics*, *36*, 1–22. https://doi.org/10.1093/applin/amt018

Brezina, V., & Gablasova, D. (2015). LancsLex: English vocabulary analysis tool [Computer program]. Retrieved from http://corpora.lancs.ac.uk/vocab

Browne, C., Culligan, B., & Phillips, J. (2013). *A New General Service List*. Retrieved from https://www.newgeneralservicelist.org

Brysbaert, M., Keuleers, E., & Mandera, P. (2020). Which words do English non-native speakers know? New supernational levels based on yes/no decision. *Second Language Research*, *37*(8), 207–231. https://doi.org/10.1177/0267658320934526

Cobb, T. (2000). One size fits all? Francophone learners and English vocabulary tests. *Canadian Modern Language Review*, *57*, 295–324. https://doi.org/10.3138/cmlr.57.2.295

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, *11*(3), 38–63. http://doi.org/10.1125/44117

Cobb, T. (2019, June). What do teachers and learners actually do with a corpus? *Symposium plenary: BAWE 10 Years On*. Coventry University, London Campus:

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

Dang, T. N. Y. (2020). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 288–303). New York, NY: Routledge. https://doi.org/10.4324/9780429291586

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The Academic spoken word list. *Language Learning*, *67*, 959–997. https://doi.org/10.1111/lang.12253

Dang, T. N. Y., & Webb, S. (2016a). Making an essential word list for beginners. In P. Nation (Ed.), *Making and using word lists for language learning and testing* (pp. 153–167). Amsterdam, The Netherlands: Benjamins. https://doi.org/10.1075/z.208.15ch15

Dang, T. N. Y., & Webb, S. (2016b). Evaluating lists of high-frequency words. *ITL – International Journal of Applied Linguistics*, *167*, 132–158. https://doi.org/10.1075/itl.167.2.02dan

Dang, T. N. Y., Webb, S., & Coxhead, A. (2020). Evaluating lists of high- frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, 1–25. https://doi.org/10.1177/1362168820911189

Davies, M. (2008). Corpus of contemporary American English *[data set with search software*]. Consulted May 13, 2019. Retrieved from https://www.english-corpora.org/coca/

Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. New York, NY: Routledge. https://doi.org/10.4324/9780203880883

Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, *35*, 305–327. https://doi.org/10.1093/applin/amt015

Greene, J. W., & Coxhead, A. (2015). *Academic vocabulary for middle school students*. Baltimore, MD: Brookes.

Henrichon, P. (2013). Discours des présidents américains. Montreal: Personal collection donated for use with Lextutor concordancer.

Hu, M., & Nation, P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, *23*, 403–430.

Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of information to accompany the Lancaster-Olso/Bergen Corpus of British English, for use with digital computers*. Department of English, University of Oslo.

Karpenko-Seccombe, T. (2021). *Academic writing with corpora: A resource book for data-driven learning*. London, UK: Taylor and Francis.

Kucera, H., & Francis, W. (1967). *A standard corpus of present-day edited American English, for use with digital computers*. Providence, RI: Brown University Press.

Laufer, B. (1990). Words you know: How they affect the words you learn. In J. Fisiak (Ed.), *Further insights into contrastive linguistics* (pp. 573–593). Amsterdam, The Netherlands: Benjamins. https://doi.org/10.1075/llsee.30.35lau

Laufer, B. (1997). What's in a word that makes it hard or easy? Intralexical factors affecting the difficulty of vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 140–155). Cambridge, UK: Cambridge University Press.

Laufer, B. (2020a). Evaluating exercises for learning vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 351–368). London, UK: Routledge. https://doi.org/10.4324/9780429291586

Laufer, B. (2020b). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly*, *54*, 1076–1085. https://doi.org/10.1002/tesq.3004

Laufer, B., & Cobb, T. (2019). How much knowledge of derived words is needed for reading? *Applied Linguistics*, *41*, 971–998. https://doi.org/10.1093/applin/amz051

Laufer, B., & Nation, P. (2012). Vocabulary. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 163–176). London, UK: Routledge.

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

Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly*, *20*, 233–253. https://doi.org/10.2307/747758

Nation, P. (2006a). *The BNC word family lists*. Retrieved from https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs

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

Nation, P. (2012). The BNC/COCA word family lists. Retrieved from https://www.wgtn.ac.nz/__data/assets/pdf_file/0005/1857641/about-bnc-coca-vocabulary-list.pdf

Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language*, *26*, 1–16.

Nation, P. (2016). *Making and using word lists for language learning and testing*. Amsterdam, The Netherlands: Benjamins. https://doi.org/10.1075/z.208

Nation, P., & Anthony, L. (2013). Mid-frequency readers. *Journal of Extensive Reading*, *1*, 5–16.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13.

Nation, P., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, *23*(1), 35–41. https://doi.org/10.1016/0346-251X(94)00050-G

Nation, P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle, Cengage Learning.

Nesi, H., Gardner, S., Thompson, P., & Wickens, P. (2009). *British Academic Written English Corpus*. Warwick, UK: Warwick University.

Nesi, H., Gardner, S., Thompson, P., & Wickens, P. (2010). *British Academic Spoken English Corpus*. Warwick, UK: Warwick University.

Peters, E. (2020). Factors affecting the learning of single-word items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 125–142). New York, NY: Routledge. https://doi.org/10.4324/9780429291586

Sasao, Y., & Webb, S. (2017). The word part levels test. *Language Teaching Research*, *21*, 12–30. https://doi.org/10.1177/1362168815586083

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(11), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Sternberg, R. J. (1987). Most vocabulary is learnt from context. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 89–105). Hillsdale, NJ: Erlbaum.

Wan-a-rom, U. (2008). Comparing the vocabulary of different graded-reading schemes. *Reading in a Foreign Language*, *20*(1), 43–69.

Webb, S., Sasao, Y., & Balance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL – International Journal of Applied Linguistics*, *168*, 33–69. https://doi.org/10.1075/itl.168.1.02web

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

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. List of Affixes in the NFL7.
Appendix S2. Sublists of the NFL7.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

A New Type of Word List to Reduce the Learning Burden

What This Research Was About and Why It Is Important

Frequency lists are an important tool in vocabulary research and instruction, especially as used in text profiling computer software. They make it possible to sequence and plan the lexical component of language learning, for example, analyzing texts in terms of lexical difficulty. But there is an issue about how these lists should be structured. Words in lists clearly have to be "grouped," otherwise the lists would be enormous with many of their items nearly identical. The main groupings are word lemmas (with just inflected, i.e., grammatical forms, like the verb *hunt, hunts, hunting*, and *hunted*) and word families (those, plus other related words of different parts of speech, called derived forms, like *hunter* and *huntress*). Each has strengths and weaknesses in a pedagogical context. Lemmas are normally recognizable as forms of the same word, but their use entails impractically large lists of similar but separated items (*hunting* and *hunter*); families produce lists with fewer units, but containing many items of widely different frequencies that learners may not recognize as related (*huntress*). To resolve this dilemma, we built a list which, while family based, contains only the most frequent family members, whether inflected or derived; that is, the nucleus of each family. We describe a way of building such a list and we compare its "text coverage" against that given by other published lists (both family and lemma based); that is, we checked how many words from each type of list are found in different texts. We describe our method of balancing list size against text coverage. We also analyze the derived forms in our list for teachable patterns. The Nuclear Frequency List (NFL) has a similar coverage to lists more than double its size; its derived forms employ a small number of morphological patterns; and it thus resolves the "grouping dilemma."

What the Researchers Did

The researchers first built a computer program that produces the NFL.

- The NFL reduces a complete family list to just the family members that are frequently used. The complete list they reduced uses the frequency information from large digital collections of words taken from real usage: The British National Corpus (BNC) and the Corpus of Contemporary American English (COCA).
- The researchers described the decisions that need to be taken when reducing a complete list to its nucleus.
- They evaluated the NFL in terms of the coverage it provides of the types of texts learners are likely to encounter.
- The researchers further analyzed the derived-word component of the NFL, in order to determine whether there exists a core of frequent affixes.

What the Researchers Found

- The 3,000 word families of the NFL contain 7,293 individual words and give an average coverage of 84.5% over a range of learner-oriented corpora, compared to the same 3,000 families in the BNC/COCA (from which it is derived) that contain 19,062 words and give a coverage of 90%.
- More than 85% of the derived forms in the NFL employ just 22 affixation patterns.

Things to Consider

- In cost–benefit terms, the loss of 5.5% coverage for about 11,000 fewer words to learn is a good trade.
- 22 affixation patterns are learnable, compared to the 100s of a complete family list.
- The NFL provides the nucleus of a lexicon that can serve both productive and receptive knowledge.

Materials, data, open access article: The Nuclear Family List-builder is available at https://www.lextutor.ca/freq/nuclear/, and the list itself is available at iris-database.org

How to cite this summary: Cobb, T., & Laufer, B. (2021). A new type of word list to reduce the learning burden. *OASIS Summary* of Cobb & Laufer (2021) in *Language Learning*. https://oasis-database.org