Creating and using *A frequency dictionary of Contemporary American English: word sketches, collocates, and thematic lists*

Mark Davies and Dee Gardner

Brigham Young University

Abstract

We have recently published A frequency dictionary of Contemporary American English: word sketches, collocates, and thematic lists *(Davies & Gardner 2010). The dictionary is based on the 400 million word Corpus of Contemporary American English (COCA), 1990-2009. It contains the top 5,000 lemmas of American English, along with the top 20-30 collocates for each word, which help to paint a useful "word sketch" for each word. In addition, there are more than thirty frequency-ranked thematic lists – family terms, clothing, new words in American English, American / British contrasts, top phrasal verbs, and so on. In this paper, we discuss the creation of the dictionary and how it can be used to facilitate the teaching and learning of English vocabulary.*

1. The value of a frequency dictionary of English

"I don't know that word." "What does that word mean?" "How is that word used?" These are some of the most common pleas for help by language learners—and justifiably so.

Not knowing enough words, or the right words, is often the root cause of miscommunication, the inability to read and write well, and a host of related problems. This fundamental need is compounded by the fact that there are simply so many words to know in any language, but especially in English, which may contain well over two million distinct words (Crystal 1995)—and growing fast. Thirty years ago, who would have thought that we would be "surfing" in our own homes, or that "chips" would be good things to have inside our equipment, or that we would be excited to "*google* this" or that we would "get *freaked out* by that". Without belaboring the obvious, it is little wonder that learners, teachers, researchers, materials developers, and many others are interested in establishing some sense of priority and direction to what could easily become vocabulary chaos.

To address this issue, we have recently published *A frequency dictionary of Contemporary American English: word sketches, collocates, and thematic lists* (Davies & Gardner 2010). We wanted to know which of the vast number of English words to start with, and we also wanted to know which other words these words "hang out with" — their neighbors (or collocates) — which provides crucial information about the meaning and use of these words. Perhaps even more importantly, we wanted to know this for our current day, not for some English of the past, when punch cards were used to program computers, and when surfing was only done at the beach. In short, our goal in creating the dictionary was to create a tool that would benefit those who are trying to learn English, as well as for those who desire to assist them.

2. Overview of the dictionary content

The frequency dictionary is designed to meet the needs of a wide range of language students and teachers, as well as those who are interested in the computational processing of English. The main index contains the five thousand most common words in American English, starting with such basic words as *the* and *of*, and quickly progressing through to more intermediate and advanced words. Because the dictionary is based on the actual frequency of words in a large 400 million word corpus (collection of texts) of many different types of English texts (spoken, fiction, magazines, newspaper, and academic), the user can feel comfortable that these are words that one is very likely to subsequently encounter in the "real world."

In addition to providing a listing of the most frequent five thousand words, the entries provide other information that should be of great use to the language learner. Each entry also shows the main collocates for each word, grouped by part of speech and in order of frequency. These collocates provide important and useful insight into the meaning and usage of the word, following the idea that "you shall know a word by the company it keeps" (Firth 1957: 11). The entries also show where each of the collocates occur with regards to the head word (before, after, or both), which indicate whether they are subject, object, and so on. Finally, the entries indicate whether the words are more common in one genre of English (e.g., spoken or academic) than in the others.

Aside from the main frequency listing, there are also indexes that sort the entries by alphabetical order and part of speech. The alphabetical index can be of great value to students who, for example, want to look up a word from a short story or newspaper article, and see how common the word is in general. The part of speech indexes could be of benefit to students who want to focus selectively on verbs, nouns, or some other part of speech. Finally, there are a number of thematically-related lists (clothing, foods, emotions, etc.) as well as comparisons of vocabulary across genres and over time, all of which should enhance the learning experience. The expectation, then, is that the frequency dictionary will significantly maximize the efforts of a wide range of students and teachers who are involved in the acquisition and teaching of English vocabulary.

3. Comparison to other frequency dictionaries of English

Historically, most frequency dictionaries (also referred to as "word books" and "word lists") have been created to meet educational needs, with many designed specifically to meet the needs of foreign- and second-language learners of English. Prominent among these are *The teachers word book of 30,000 words* (Thorndike & Lorge 1944)—based on 4.5 million words from general English texts, magazines, and juvenile books, *The general service list of English words* (West 1953)—a list of the 2,000 highest frequency words (with semantic distinctions and counts) based on visual inspections by semanticists of five million words from various sources (encyclopedias, magazines, textbooks, novels, etc.), the Brown Corpus list (Francis & Kučera 1982)—based on one million words of written American English, and its British English counterpart—the LOB Corpuslist (Johansson & Hofland 1989).

For many purposes, these latter two replaced the older lists of Thorndike & Lorge. Additionally, there are several more specialized school lists, such as the *American Heritage word frequency book* (Carroll, Davies & Richman 1971) based on five million running words of written school English (grades 3 through 9), the *Academic word list* (Coxhead 2000) with 570 academic word families based on 3.5 million running words of academic texts, and the very early *A basic vocabulary of elementary school children* (Rinsland 1945), based on six million running words of children's writing samples.

A great debt is owed to the pioneering scholars who generated these and other frequency lists to facilitate English vocabulary learning, research, and description. Building on these earlier efforts, the *Frequency dictionary of American English* addresses several vocabulary needs in the field of English language education. First, and perhaps most obvious, it is based on contemporary American English, thus making it more ecologically valid in educational and research settings where American English is the target, and where many are still relying on the nearly 50-year-old Brown Corpus (Francis & Kučera 1982) for frequency information about American English vocabulary. (Note: the actual texts for the Brown Corpus were from 1961.) Second, unlike the Brown Corpus (one million words of written English only), the frequency counts in the dictionary are based on a very large and balanced corpus of both written and spoken materials (400 million words from five major genres).This increases one's confidence that the highest frequency words have indeed been determined and properly ranked, and that these words have a high degree of utility across major genres of importance to English language learners (spoken, fiction, newspapers, magazines, and academic).

Third, the inclusion of collocates (by part of speech) for each of the 5,000 high-frequency node words adds a semantic richness to the dictionary that is often lacking when only the forms of words are tallied without consideration of their potential meanings (Gardner 2007). The tightness of some of these node-collocate relationships (*big deal, bad habit, make sense, trash talk*, etc.) also highlights the phrasal nature of many English vocabulary items (Cowie 1998). Such collocational knowledge is a crucial component of what it means to know a word (Nation 2001) and has also been recognized as a characteristic difference between native and non-native language abilities (Nesselhauf 2005). Therefore, language learners and their teachers should benefit from the rich semantic and pragmatic information the collocates provide, thus taking us one step closer to Read's

(2000) call for new high-frequency word lists that are based on large electronic corpora, but which also account for the many meanings that language learners need to negotiate. Although semantic frequency is not fully realized in the dictionary, the collocates do provide some support for semantic interpretations, and will certainly aid in determining which meanings of a word form to teach and learn.

Finally, the 30 call-out boxes in the dictionary are packed with useful vocabulary information for language learners and their teachers, including words that make up many of the basic semantic sets of the language (*animals, body, clothing, colors, emotions, family, food*, etc.), words that characterize a specific genre of the language (spoken, fiction, academic, etc.), words that are new to American English, words that tend to be characteristically American or British, productive suffixes and the actual content words they are found in (nouns and adjectives), and the highest frequency phrasal verbs of American English. (Compare with Gardner & Davies 2007, which lists the highest frequency phrasal verbs of British English.) These and other call-out boxes in the dictionary can be used for self-study, teaching, assessment, materials development, and research purposes.

To our knowledge, there is only one other publicly-accessible frequency dictionary of English that is based on a large mega-corpus – *Word frequencies in written and spoken English* (Leech, Rayson & Wilson 2001). However, our dictionary is quite different in at least three major respects. First, the Longman frequency dictionary represents British, not American, English, and it bases its word-frequency information on the British National Corpus (BNC). Second, most of the texts in the BNC are at least twenty years old, while texts in the Corpus of Contemporary American English (COCA) are current through late 2008. Third, while both corpora are balanced for genre (e.g., spoken, fiction, newspaper, and academic), COCA (400 million words) is nearly four times as large as the BNC (100 million words), allowing us to have more confidence in determining the words that should "make the list" and in finding their meaningful neighbors.

In addition to the differences in focus, age, and sampling size between the two dictionaries, there are also differences in the presentation formats. The Longman dictionary is mainly composed of straight frequency lists of words and lemmas, while the dictionary is oriented specifically to language learners, supplementing the frequency listings with the unique features previously mentioned: (a) frequency-ranked collocates (co-occurring words) for each headword in the frequency dictionary—which can help learners and their teachers better understand the *meanings* and *uses* of the high frequency words; and (b) the more than 30 thematically-oriented vocabulary lists (call-out boxes) for particular semantic, grammatical, or lexical categories that would be helpful for language training purposes.

4. The corpus

A frequency dictionary is only as good as the corpus on which it is based. The Corpus of Contemporary American English (COCA, see Davies 2009) is the largest balanced corpus of American English, and the largest balanced corpus of any language that is publicly available (http://www.americancorpus.org). In addition to being very large (400 million words; 20 million words each year 1990-2007), the corpus is also balanced evenly between spoken (unscripted conversation from 150+ radio and TV shows), fiction (e.g., books, short stories, movie scripts), 100+ popular magazines, ten newspapers, and 100+ academic journals – for a total of 150,000+ texts.

The more than 150,000 texts come from a variety of sources:

Spoken: (79 million words) Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: *All Things Considered* (NPR), *Newshour* (PBS), *Good Morning America* (ABC), *Today Show* (NBC), *60 Minutes* (CBS), *Hannity and Colmes* (Fox), *Jerry Springer*, etc).

Fiction: (76 million words) Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and movie scripts.

Popular Magazines: (81 million words) Nearly 100 different magazines, with a good mix (overall, and by year) between specific domains (news, health, home and gardening, women, financial, religion, sports, etc). A few examples are *Time, Men's Health, Good Housekeeping, Cosmopolitan, Fortune, Christian Century, Sports Illustrated*, etc.

Newspapers: (76 million words) Ten newspapers from across the US, including: *USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle*, etc. In most cases, there is a good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.

Academic Journals: (76 million words) Nearly 100 different peer-reviewed journals. These were selected to cover the entire range of the Library of Congress classification system (e.g., a certain percentage from B (philosophy, psychology, religion), D (world history), K (education), T (technology), etc.), both overall and by number of words per year.

In summary, the corpus is very well balanced at both the "macro" level (e.g., spoken, fiction, newspapers) and the "micro" level (i.e., the types of texts and the distribution of the sources) within each of these macro genres.

5. Annotating and organizing the data from the corpus

In order to create a frequency dictionary, the words in the corpus must be tagged (for part of speech) and lemmatized. Tagging of course means that a part of speech is assigned to each word — noun, verb, and so on. Lemmatization — which is crucial for the frequency dictionary – means that each word form is assigned to a particular "head word" or "lemma", such as *go, goes, going, went,* and *gone* being marked as forms of the lemma GO.

The tagging and lemmatization was done with the CLAWS tagger (Version 7), which is the same tagger that was used for the British National Corpus (http://www.natcorp.ox.ac.uk/) and for other important corpora of English as well. One of the most difficult parts of tagging, of course, is to correctly assign the part of speech for words that are potentially ambiguous. In cases like *computer, disturb, lazy,* or *fitfully*, these are unambiguously tagged as noun, verb, adjective, and adverb, respectively. But in a case like *light*, the word can be a noun (*he turned on the light*), verb (*should we light the fire?*), or adjective (*there was a light breeze*). In these cases, the tagger looks at the context in which the word occurs in each instance to determine the correct part of speech. While the CLAWS tagger is very good, it does produce errors. We have tried to correct for most of these, but there are undoubtedly still some that remain.

It of course makes sense to provide separate entries in the dictionary for words with different parts of speech, such as noun and verb. For example, the word *beat* as a noun has collocates like *hear, miss, steady, drum,* and *rhythm*. As a verb, however, it takes collocates like *heart, egg, bowl, severely*, or *Yankees*. Even in cases where the word appears as a noun and an adjective (*magic, potential, dark, veteran*), the collocates for the two parts of speech are very different, and it would probably be too confusing to conflate them into one entry. Perhaps the most problematic are function words like *since*, which appear up to three times in the dictionary. In the case of *since*, for example, it appears as preposition (*he's been here since 1942*), adverb (*several other schools have since been constructed*), and conjunction (*since they won't be here until 5pm, we'll just leave for a minute*). In these cases, we have simply followed the output of the tagger. If it says that there are multiple different parts of speech, then the word appears under each of those parts of speech in the dictionary.

6. Frequency and dispersion

After the tagging and lemmatization of the 400 million words in the corpus, our final step was to determine exactly which of these words would be included in the final list of the 5,000 most frequent words (or lemmas). One approach would be to simply use frequency counts. For example, all lemmas that occur 5,000 times or more in the corpus might be included in the dictionary. Imagine, however, a case where a particular scientific term was used repeatedly in engineering articles or in sports reporting in newspapers, but it did not appear in any works of fiction or in any of the spoken texts. Alternatively, suppose that a given word is spread throughout an entire register (spoken, fiction, newspaper, or academic), but that it is still limited almost exclusively to that register. Should the word still be included in the frequency dictionary? The argument could be made that we should look at more than just raw frequency counts in cases like this, and that we ought to look at dispersion as well, or how well the word is "spread across" all of the registers in the entire corpus.

In our dictionary, we have used Juilland's "D dispersion index" (see Juilland & Chang-Rodriguez 1964). A score of 1.00 means that the word is perfectly spread across the corpus, so that if we divided the corpus into one hundred equally-sized sections (each with four million words, in the case of our 400 million word corpus), the word would have exactly the same frequency in each section. A dispersion score of .10, on the other hand, would mean that it occurs a lot in a handful of sections, and perhaps not at all or very little in the other sections.

As a clear example of the contrast between "frequency" and "dispersion", consider the Table 1. All of the words in Table 1 have essentially the same frequency – an average of about 3,000 occurrences in the corpus. The words to the left, however, have a dispersion score of at least 0.95, which means that the word has roughly the same frequency in all of the 100 sections of the corpus that we used for the calculation. The words to the right, on the other hand, have a much lower dispersion score. Most would easily agree that the words shown at the left would be more useful in a frequency dictionary, because they represent a wide range of texts and text types in the corpus. Therefore, as we can see, frequency alone is probably not sufficient to determine whether a word should be in the dictionary.

Table 1: Differences in dispersion (PoS = Part of speech, D = Juilland's D dispersion index)

good dispersion				*poor dispersion*			
freq	*lemma*	*PoS*	*D*	*freq*	*lemma*	*PoS*	*D*
3134	CONVINCING	j	0.96	4653	HEALTHCARE	n	0.56
3107	SENSIBLE	j	0.95	4282	ELECTRON	n	0.58
3041	HONESTY	n	0.96	4181	SKIER	n	0.43
3033	UNUSUALLY	r	0.95	4113	COMPOST	n	0.31
3020	CONFUSING	j	0.97	3685	WATERCOLOR	n	0.41
3014	EXAGGERATE	v	0.96	3769	SKI	v	0.47
2950	DISTRACTION	n	0.95	2028	NEBULA	n	0.46
2922	RESENT	v	0.96	2547	PALETTE	n	0.57
2891	WRESTLE	v	0.95	2536	ANGLE	v	0.55
2876	URGENCY	n	0.96	2479	ALGORITHM	n	0.52
2873	HINT	v	0.96	2437	PASTEL	n	0.25
2842	OBSESSED	j	0.95	2388	SOCKET	n	0.60
2833	GENUINELY	r	0.96	2350	NASAL	j	0.44
2813	RESPECTED	j	0.95	2281	CACHE	n	0.43

7. The final calculation

The calculation to determine which words are included in the frequency dictionary was a fairly straightforward one. The formula was simply:

score = frequency * dispersion

For example, consider the words near 3210 in the frequency dictionary, as shown in Table 2. The word *furthermore* has a higher frequency (9,594 tokens) than the other two words, but it has lower dispersion (.86). *Orange*, on the other hand, has a lower frequency (8,881 tokens) but it has better dispersion across the corpus. *Taxpayer* (frequency of 9,140 and dispersion of .90) is in the middle of both of these. But with the formula that takes into account both frequency and dispersion, these three words end up having more or less the same score.

Table 2: Effect of frequency and dispersion on rank order

ID	*lemma*	*PoS*	*frequency*	*dispersion*	*score*
3199	ORANGE	j	8881	0.93	8270
3201	TAXPAYER	n	9140	0.90	8256
3205	FURTHERMORE	r	9594	0.86	8235

The 5,000 lemmas with the top score (frequency * dispersion) are those that appear in the frequency dictionary.

8. Collocates

A unique feature of the frequency dictionary are the listing of the top collocates (nearby words) for each of the 5,000 words in the frequency listing. These collocates provide important and useful insight into the meaning and use of the keyword. To find the collocates, we did the following. First, we decided which parts of speech to group together in order to rank the collocates and show the most frequent ones. In the case of verbs, we grouped noun collocates (subject: *the evidence supports what she said*, and object: *this supports the claim*), and all other collocates were grouped as miscellaneous (e.g., *with, directly, difficult*, and *prepare* for the verb *deal*). For nouns, we looked for adjectives (*green grass*), other nouns (*fire station*), and verbs (e.g., *desire to succeed*). For adjectives, we looked for nouns (*fast car*) and all other collocates were grouped as miscellaneous (*completely exhausted, willing to stay, black and white*). Finally, for adverbs and other parts of speech, we see collocates from all parts of speech listed together (*sharply reduce, fewer than, except for*).

To find the collocates for a given word, a computer program searched the entire 400 million word corpus and looked at each context in which that word occurred. In all cases, the context (or span) of words was four words to the left and four words to the right of the node word. The overall frequency of the collocates in each of those contexts was then calculated, and the collocates were examined and rated by at least four native speakers.

Obviously, common words like *the, of, to*, etc. were usually the most frequent collocates. To filter out these words, we set a Mutual Information (MI) threshold of about 2.5. The MI calculation took into account the overall frequency of each collocate, so that common words were usually eliminated from the list.

Using MI is sometimes more an art than a science. If the MI is set too low, then high frequency "noise words" show up as collocates, whereas if it is set too high, then only highly idiomatic collocates are found. As an example, the most frequent collocates of *break* as a verb – when the MI score is set high at 5.5 – are: *deadlock, logjam, monotony, and stranglehold*. These are quite idiomatic and don't really show well the core meaning of *break*. On the other hand, the most frequent collocates when the MI threshold is set very low at 1.0 are *down, into, up*, and *off*, which again do not provide a good sense of its meaning. Finally, however, when we set the MI threshold to 2.5, we find the most frequent collocates are: *heart, silence, rules, loose, leg, and barriers*, which (for native speakers, at least), probably do relate more to the core meaning and usage of *break*. But getting the MI threshold set just right for each of the 5,000 headwords was a bit daunting, to say the least.

9. The main frequency index

Chapter 2 of the dictionary contains the main index in the dictionary – a rank-ordered listing of the top five thousand words (lemma) in English, starting with the most frequent word (the definite article *the*) and progressing through to *warehouse, paradise*, and *nominate*, which are the last three words in the list. The following information is given for each entry, as in (1):

(1)

rank frequency (1, 2, 3, …), lemma, part of speech

collocates, grouped by part of speech and ordered by frequency

raw frequency, dispersion (0.00 – 1.00), (indication of register variation)

As a concrete example, let us look at the entry (2) for the verb *break*:

(2)

494 break *v*

n. law, heart, news, .rule, silence, story, .ground, .barrier, leg, bone, .piece, .neck, arm, .cycle, voice. *misc*.into, .away, .free, .apart, .loose up marriage, .fight, boyfriend, meeting., girlfriend, union, band, pass, .demonstration, .monotony down.into, .barrier, car., .cry, .door, .tear, talk., enzyme., completely, negotiation.out war., fight., fire., sweat, fighting., riot., violence., .laugh, .hive off piece, talk, .engagement, negotiation, branch, abruptly, .relation

72917 0.97

This entry shows that word number 494 in our rank order list is the verb *break*. The last line of the entry shows the raw frequency for the lemma (72,917 tokens) and the dispersion (.97 in this case). The collocates are given in the intervening lines. As can be seen, they are partially grouped by part of speech. In the case of verbs, we see the noun collocates and then other parts of speech (miscellaneous).

Note also that for some collocates, there is an indication of the placement of the collocate. When the "." is before the collocate, this means that the node word (headword) is typically found before that collocate (*break the law, break into pieces*). When the "." is after the collocate, this means that the node word is typically found after the collocate (*her voice broke, all hell broke loose*). This symbol can provide useful information, for example, on whether the collocates are subjects or objects of a given verb, or whether the node word noun acts as a subject or object of the verbal collocate. (Note, however, that with passives and relative clauses, the noun that is object of a verb will occur before the verb, which does confuse things a bit.) In order to display the "." symbol, 80% or more of the tokens of a given collocate had to occur either before or after the node word. In the case of ADJ / NOUN and NOUN / ADJ, word order is typically so consistent (*blue house*, never **house blue*) that the "." is not used to show placement.

Finally, as is seen above, in the case of some verbs that can act as phrasal verbs (*break up, turn down, cut off*, etc.), these are listed in bold (with their own collocates) at the end of the regular collocates list for verbs. Phrasal verbs are only listed when they have a frequency of at least 1,000 in the corpus, and when there are at least three collocates with a frequency of at least five occurrences each.

Let us consider one other example:

(3)

3396 hypothesis *n*

j null, following, consistent, alternative, working, general, initial, original, theoretical, competing *n* study, support., result, test, research, testing, evidence, analysis, method, set *v*.predict, suggest, reject, examine, confirm, base, develop, formulate, .state, .explain

9282 0.82 A

This entry is for *hypothesis* (word number 3396 in our list). As before, the collocates are listed in frequency order and grouped by part of speech. In this case, however, note that there is an "A" at the end of the entry. This indicates that the lemma *hypothesis* occurs at least twice as frequently in the Academic genre as it does overall in the corpus (Spoken, Fiction, Magazines, Newspapers).

10. Thematic vocabulary (call-out boxes)

Placed throughout the main frequency-based index are approximately thirty call-out boxes, which serve to display in one list a number of thematically-related words. These include thematic lists of words related to the body, food, family, weather, professions, nationalities, colors, emotions, and several other semantic domains. There are also lists of words that are much more common in each of the five main genres (spoken, fiction, popular magazines, newspapers, and academic) than overall, as well as comparisons of frequent American and British vocabulary, as well as new words in the language. Finally, there are lists related to word formation issues, such as irregular past tense and irregular plurals, and common suffixes to create nouns, adjectives, and verbs. In each case, the entries are of course ordered by frequency.

Partial lists for seven of the more than thirty thematic lists are given in the Appendix.

11. Electronic version

In addition to the basic 5,000 word list with 20-30 collocates each, expanded word frequency and collocates lists that are based on the same corpus can also be found at http://www.wordfrequency.info. This site contains lists of the top 10,000 and 20,000 words with 20-30 collocates each. In addition, it also includes the top 200-300 collocates for either the top 10,000 or top 20,000 words. For the largest list, then, there are about 4,300,000 node word / collocate pairs, along with their frequency and Mutual Information score. Finally, this site also contains the full set of trigrams from the corpus along with their frequency, a total of more than 155 million unique three-word sequences.

12. Delimitations

The following are the major delimitations of the frequency dictionary:

1. Frequency is form-based (lemma), not semantically-based (homographs -- *bank, run*, heterophones – *lead* 'metal' vs. *lead* 'be in front', *contract* vs. *contract,* etc.). But our approach is an improvement over many similar frequency listings because the collocates give some indication of potential variant meanings. For example, the entry for *lead* (n) [entry 1598] and *bow* (n) [entry 4139] show that for *lead*, there are collocates for the two meanings 'metal' and 'in front' and for *bow* there are collocates for *bow* in the context of 'ship', 'arrow', 'hair', and 'violin'.
2. Except in the case of high-frequency phrasal verbs, only single-word nodes were included. When a lemma occurs almost exclusively in a given multi-word expression (*as far as, in charge of, lots of*), that multi-word expression is listed as part of the entry.
3. All collocates are single word collocates. In cases like *in terms of, by means of*, etc., each of the collocates are listed separately.
4. The most frequent form of a given collocate lemma may be an inflected form, not the head word form as listed (e.g., *long* as a collocate of *no* almost always appears as *longer* in the corpus).
5. In general, proper nouns were not included in the dictionary, either as node words or collocates. However, a few highly salient proper noun collocates were included for certain node words (e.g., *Iraq* as a collocate of *invade*; *China* as a collocate of *export*).

13. Summary

A frequency dictionary of Contemporary American English: word sketches, collocates, and thematic lists (Davies & Gardner 2010) is the first frequency list of its type for English. It is based on the 400+ million word Corpus of Contemporary American English (COCA), which is the largest, freely-available corpus of English. Because it is balanced between spoken, fiction, popular magazines, newspaper, and academic, and because it contains texts from 1990 to the present time, users can be sure that the words in the dictionary are those that they will actually encounter in the "real world".

Rather than just containing a list of the most frequent words or lemmas (as do some other frequency lists and dictionaries), this dictionary provides much more. Each entry contains the top 20-30 collocates for the head word (sorted by part of speech and ordered by frequency), which provides extremely valuable insight into meaning and usage. Entries also indicate whether the word is more common in one genre than another. Finally, there are more than thirty thematically-organized call-out boxes, which contain information not just on thematic vocabulary such as clothing and family terms, but also lists on American /British differences, "new words" in American English, genre-based differences, and also several grammatical topics, such as word formation and phrasal verbs – all organized by frequency, of course. In summary, we believe that the dictionary offers a range of features that make it highly useful for learners and teachers of English.

References

Carroll, J.B., P. Davies & B. Richman (1971), *The American Heritage word frequency book*. New York: American Heritage Publishing Co., Inc.

Cowie, A.P. (ed.) (1998), *Phraseology: theory, analysis, and applications*. Oxford: Clarendon Press.

Coxhead, A. (2000), 'A new academic word list', *TESOL quarterly*, 34(2): 213-238.

Crystal, D. (1995), *The Cambridge encyclopedia of the English language*. New York: Cambridge University Press.

Davies, M. (2009), 'The 385+ Million Word Corpus of Contemporary American English (1990-2008+): design, architecture, and linguistic insights', *International journal of corpus linguistics*, 14: 159-90.

Davies, M. & D. Gardner (2010), *A frequency dictionary of Contemporary American English: word sketches, collocates, and thematic lists*. London: Routledge.

Firth, J.R. (1957), *Papers in linguistics 1934-1951*. London: Oxford University Press.

Francis, W.N. & H. Kučera (1982), *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin.

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

Gardner, D. & M. Davies (2007), 'Pointing out frequent phrasal verbs: a corpus-based analysis', *TESOL quarterly*, 41(2): 339-359.

Juilland, A. & E. Chang-Rodriguez (1964), *Frequency dictionary of Spanish words*. The Hague: Mouton de Gruyter.

Johansson, S. & K. Hofland (1989), *Frequency analysis of English vocabulary and grammar: based on the LOB Corpus, Volume 1: tag frequencies and word frequencies*. Oxford: Clarendon Press.

Leech, G., P. Rayson & A. Wilson (2001), *Word frequencies in written and spoken English: based on the British National Corpus*. London: Longman.

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

Nesselhauf, N. (2005), *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Read, J. (2000), *Assessing vocabulary*. Cambridge: Cambridge University Press.

Rinsland, H.D. (1945), *A basic vocabulary of elementary school children*. New York: The Macmillan Company.

Thorndike, E.L. & I. Lorge (1944), *The teacher's word book of 30,000 words*. New York: Columbia University, Teacher's College.

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

Appendix

Examples of (partial) thematic lists:

Clothing: (suit) *n* $_{35280}$, shoe $_{26007}$, (ring) *n* $_{21069}$, shirt $_{20470}$, dress *n* $_{20389}$, hat $_{17462}$, (tie) *n* $_{17060}$, coat *n* $_{15297}$, jacket $_{14938}$, boot *n* $_{14436}$, belt *n* $_{13966}$, pants $_{11412}$, glove $_{9852}$, uniform *n* $_{9809}$, skirt *n* $_{8858}$, jeans $_{8037}$, sock *n* $_{6486}$, sweater $_{5179}$, robe $_{4751}$, shorts $_{4626}$, gown $_{4387}$, scarf *n* $_{3551}$, (slip) *n* $_{3545}$, vest $_{3313}$, blouse $_{3110}$, underwear $_{3037}$, necklace $_{2914}$, diaper $_{2738}$, earring $_{2720}$, cloak $_{2624}$, bracelet $_{2394}$, bra $_{2378}$, apron $_{2205}$, sneakers $_{2181}$, stocking $_{2027}$, slipper $_{1943}$, blazer $_{1932}$, pajamas $_{1491}$, bikini $_{1418}$, sweatshirt $_{1342}$,

Family: child $_{323005}$, mother $_{163282}$, father $_{140176}$, (kid) $_{126054}$, parent $_{115339}$, wife $_{80380}$, son $_{77848}$, (baby) $_{65615}$, brother $_{60753}$, husband $_{57625}$, daughter $_{57551}$, sister $_{45904}$, mom $_{39021}$, dad $_{34364}$, uncle $_{18091}$, twin $_{15073}$, aunt $_{13184}$, grandmother $_{13045}$, daddy $_{12748}$, cousin $_{11640}$, mama $_{10838}$, grandfather $_{10663}$, ancestor $_{6336}$, sibling $_{5864}$, bride $_{5677}$, (widow) $_{5265}$, grandparent $_{5147}$, grandchildren $_{4846}$, grandma $_{4529}$, papa $_{4007}$,

Transportation: car $_{128671}$, train *n* $_{43971}$, ship $_{38312}$, plane $_{32339}$, truck $_{31140}$, boat $_{31036}$, bus $_{25488}$, van $_{22988}$, bike $_{17795}$, jet $_{13873}$, helicopter $_{9660}$, airplane $_{8339}$, automobile $_{6652}$, bicycle $_{6354}$, cab $_{6320}$, (metro) $_{5776}$, taxi $_{4635}$, subway $_{4485}$, ferry $_{4455}$, ambulance $_{4020}$, motorcycle $_{3480}$, jeep $_{3364}$, tractor $_{3100}$, carriage $_{3049}$, SUV $_{2813}$, convertible $_{2048}$, limo $_{1810}$, pickup truck $_{1716}$, space shuttle $_{1634}$, limousine $_{1583}$, minivan $_{1512}$,

Academic (more than twice as frequent in academic than in other genres):
[noun] student $_{163046}$, study $_{108068}$, teacher $_{73873}$, education $_{67286}$, level $_{61510}$, research $_{61043}$, community $_{56722}$, result $_{55682}$, process $_{53533}$, development $_{51798}$, use $_{51365}$
[verb] provide $_{73643}$, suggest $_{40118}$, develop $_{39775}$, require $_{37581}$, base $_{33053}$, indicate $_{31471}$, describe $_{30752}$, identify $_{29310}$, represent $_{26974}$, increase $_{26032}$, present $_{24867}$
[adjective] social $_{80721}$, political $_{63690}$, economic $_{43584}$, significant $_{37691}$, cultural $_{30243}$, environmental $_{29711}$, physical $_{27783}$, specific $_{23780}$, similar $_{23717}$, individual $_{23127}$
[adverb] however $_{70754}$, thus $_{39524}$, (for) example $_{36060}$, therefore $_{21538}$, e.g. $_{18417}$, significantly $_{15657}$, generally $_{14601}$, highly $_{12450}$, (in) addition $_{12087}$, relatively $_{12025}$

Word increasing most in frequency, 1990s to 2000s:
[noun] e-mail $_{14326}$, terrorism $_{10360}$, terrorist $_{8889}$, affiliation $_{8713}$, adolescent $_{7212}$, homeland $_{4157}$, website $_{3909}$, wireless $_{3492}$, insurgent $_{2433}$, globalization $_{2354}$,
[verb] host $_{5535}$, click $_{4094}$, e-mail $_{2139}$, download $_{1851}$, preheat $_{1364}$, bully $_{916}$, makeover $_{853}$, freak $_{649}$, partner $_{638}$, mentor $_{587}$, morph $_{349}$, vaccinate $_{286}$, restart $_{258}$
[adjective] online $_{9219}$, terrorist $_{7908}$, Afghan $_{2776}$, Taliban $_{2095}$, Shiite $_{2052}$, Pakistani $_{1967}$, same-sex $_{1397}$, sectarian $_{1138}$, upscale $_{1057}$, embryonic $_{1036}$, Islamist $_{929}$

[adverb] online $_{6034}$, famously $_{1173}$, postoperatively $_{226}$, offline $_{158}$, wirelessly $_{141}$, healthfully $_{72}$, preemptively $_{66}$, intraoperatively$_{58}$, triply $_{33}$, day-ahead $_{30}$

American vs. British vocabulary: (Amer. V): call, report, focus, guess, sign, step, figure, roll, fire, hire, file, oppose, wrap, interview, accomplish, testify, bake, track, evolve, violate, target (Brit. V): ensure, suppose, regard, voice, bind, retain, undertake, phone, allocate, knit, book, abolish, envisage, incur, fancy, commence, enclose, enquire (Amer. N): student, president, percent, kid, guy, nation, photo, arm, American, Republican, phone, movie, store, lawyer, Democrat, professor, expert, senator (Brit. N): council, minister, union, pound, scheme, shop, principle, village, provision, sector, appeal, parliament, mum, tea, lord, cabinet, pension, flat (Amer. ADJ): American, federal, tough, native, Iraqi, crazy, smart, Israeli, Mexican, congressional, elementary, online, gifted, athletic, ongoing (Brit. ADJ) British, European, English, royal, French, industrial, Scottish, lovely, working, bloody, parliamentary, alright, statutory, keen, Welsh, Tory, socialist

Phrasal verbs: go on $_{56638}$, come back $_{40285}$, come up $_{37168}$, go back $_{35719}$, pick up $_{33613}$, find out $_{28067}$, come out $_{26705}$, go out $_{26094}$, grow up $_{25268}$, point out $_{23949}$, come in $_{23563}$, turn out $_{23284}$, set up $_{21711}$, end up $_{19900}$, give up $_{18558}$, make up $_{18207}$, be about $_{17658}$, sit down $_{17375}$, look up $_{17356}$, come on $_{16116}$, get up $_{15388}$, take on $_{14697}$, go down $_{14609}$, figure out $_{14505}$, show up $_{14368}$, get back $_{13668}$, come down $_{12733}$, go up $_{12618}$, get out $_{12600}$, stand up $_{12559}$, work out $_{11512}$, be back $_{11233}$, wake up $_{10733}$, look back $_{10698}$