WORD LISTS IN DATA-DRIVEN LEARNING (4886 wds)

Tom Cobb
Didactique des langues
Université du Québec à Montréal
Canada
cobb.tom.3@gmail.com

Abstract
While concordances and word lists are the two usable forms of corpus data, word lists are largely absent from both practitioner awareness and ongoing research in data-driven learning (DDL). This is mainly because word lists work behind the scenes in the design of DDL instruction rather than up front as concordancers do in the hands of learners. Both are instances of DDL in the sense of language patterning exposed with computer software. Examples of these points are elaborated and a proposal made to integrate word list research within DDL.

Keywords
data-driven learning, language learning, corpus, concordance, frequency list, profiler

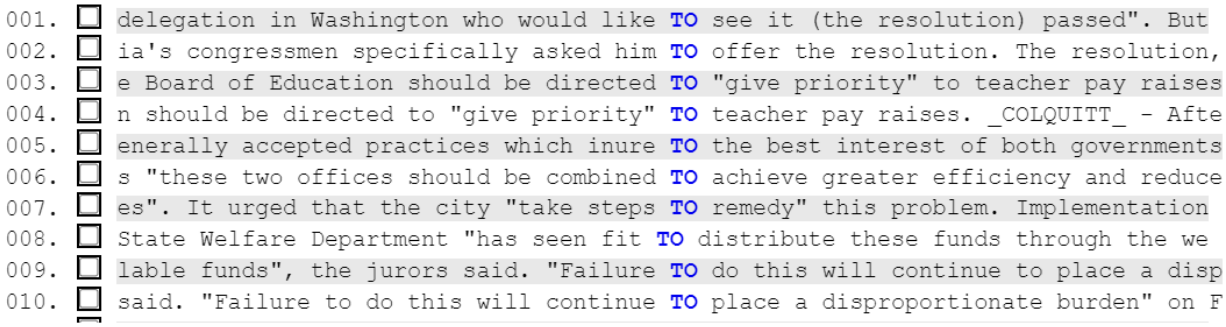
Introduction

If data-driven learning (DDL) is broadly "the use of a corpus in language learning" (Boulton & Cobb, 2017), then, since corpora of any size are unapproachable as totalities, DDL refers to the use in language learning of the two possible break-outs of a corpus, word lists and concordances. Word lists are normally frequency lists, count-ups of every word-type in a corpus (Table 1 shows the first 30 types of the Brown Corpus, 1969, sorted by frequency), or else concordances, a virtual list of every word token in a corpus, with a few words of context to either side (Figure 1 shows the first ten of 26,327 instances of "to" in the same corpus).

Table 1: Start of the Brown Corpus frequency list

1. the 69992	7. that 10783	13. with 7290	19. at 5382	25. but 4381
2. of 36472	8. is 10102	14. as 7252	20. by 5348	26. from 4370
3. and 28931	9. was 9815	15. his 6999	21. this 5146	27. or 4227
4. to 26237	10. he 9766	16. on 6766	22. had 5130	28. have 3942
5. a 23541	11. for 9499	17. be 6387	23. not 4618	29. an 3748
6. in 21419	12. it 9071	18. i 5620	24. are 4394	30. they 3648

Figure 1: Start of the concordance for "to" in Brown Corpus



From Lextutor.ca/conc/eng/

Note that a frequency list is a complete, existing artefact, composed of every word type in a corpus along with the number of its occurrences, while a concordance is typically a virtual list of every token, some part of which is created dynamically on user request. (Concordancers exist that produce one context line for every word token, e.g., Zimmerman, 1988, but they are normally called "indexers.") Despite this difference, it is useful to see concordances and lists as aspects of the same phenomenon, i.e. concordances as expanded word lists with a few context words added to each side. The original purpose of the context words was to make part of speech (POS) parsing of the keyword possible, thereby creating a more accurate list (e.g, with "run" noun and "run" verb counted separately); a more recent purpose is to improve the comprehensibility of corpus data for language teachers and learners. Seeing lists and concordances as similar corresponds to how concordances are programmed in software, i.e., the same as frequency lists but with extra characters.

The low profile of lists

But if there is a connection between concordances and word lists, it is not obvious to many trainee English as a Second Language (ESL) teachers I have known. In successive rounds of (unpublished) classroom research, I have asked classes, after the DDL part of their course in Computer-Assisted Language Learning (CALL), to brainstorm in twos, then fours, then as a class, the top 10 keywords of DDL. These were typically "corpus/corpora," "specialist/generalist," "discovery/inductive/problem-based learning," "authentic/simplified," "hands-on/printout approaches," and "collocate/collocation." Word lists never came up, though their role and uses had been mentioned in sessions. I occasionally distributed my own set of DDL keyword which included "word lists," and asked groups to rank-order them, and lists were never far from the bottom.

The picture is not so different in the work of DDL researchers. Word lists are hardly unknown to researchers, since the main DDL toolkits available (Anthony, ANTCONC, https://laurenceanthony.net/software/antconc/; Cobb, LEXTUTOR, https://lextutor.ca, Cobb; Davies, COCA, https://English-corpora.org; Kilgariff, SKETCHENGINE, https://sketchengine.eu; and Smith, WORDSMITH, https://lexically.net/wordsmith) all include both corpus based lists and tools for building lists from texts and corpora. But do these lists get research attention? They should, inasmuch as word lists function behind the scenes or in the "back-ends" of many types of DDL software and hardware and are loaded with variables that involve judgments and can affect learning. One is whether word lists require parsing and/or grouping into some type or larger unit, and if so which type, to be most useful to learners in, say, corpus based dictionaries or some of the software to be discussed below. Nonetheless, in a collection of 790 research abstracts produced between 1988 and 2021, assembled by Boulton and Vyatkina (2021) and updated thereafter, there are just 21 occurrences of "word list" (including "list," "vocabulary list" and "frequency list") compared to 354 of "concordance" (with "concordances" and "concordancing"). It is as if the concordancer was the only tool in the DDL tool shed.

The equation of DDL and concordance on the part of teachers and researchers is probably due to the fact that concordances (being readable pieces of sentences linked to their original texts) are immediately comprehensible to learners in ways that lists are not, and to the fact that concordance output contains information that a list does not and cannot, notably about grammar and collocation. But lists contain important information that concordances do not, and can do work in DDL that concordances can not.

Two corpus outputs for two jobs

To explore the different roles of lists and concordances in DDL, it may help to expand the definition of DDL slightly. To "the use of a corpus in language learning…," we should add an answer to the implicit question the definition poses, "to do what?" DDL is the proposition that language, when amassed and transformed into text data, can be squeezed, sliced, and diced by computer software to (1) reveal its underlying patterns - to linguists, certainly, but also with some injection of pedagogy, to language learners; but also to (2) raise the comprehensibility and learnability of particular instances of language through features like parsing, exhaustiveness, simplification, the increased motivation of knowing a feature is frequent, the comprehension of a new word by seeing several examples of it grouped together, and others. With some overlap, concordances work up front to make language patterning available for deliberate learning, while lists work behind the scenes to make language comprehensible for implicit learning. These different loci of operation may explain the neglect of lists in the DDL conversation.

Patterns revealed in concordances

There are cases where lists and concordances can both reveal the same pattern but concordances can do it far better. The most obvious involves the display of frequency, the number of times a word or phrase appears in a text, corpus, or lifetime, which is largely opaque to language users and even teachers (McCrostie, 2007). But a frequency list reveals the frequency only of single words and adjacent-word phrases. Disjunctive collocations (with keyword and collocate separated by intervening words), which are typically the last thing learned by unassisted L2 learners (Boers *et al*, 2014), can be revealed only by a lifetime of experience in a language, or by a concordancer. Even linguists discovered the extent of collocation in language only with the help of a concordancer, and this is a clear case where the "learner as linguist" (Seliger, 1983) can follow in their footsteps (Cobb, 2018).

Similarly, grammatical patterning can be found only in concordances, not in lists. All the individual words of "He badly plays the piano*" may feature in a particular frequency list, but they will never appear in this sequence in any corpus. Indeed, the ultimate revelation of concordances that can never be provided by word lists is what it does *not* contain. Concordances are an answer to the generative linguists' famous "negative evidence" (e.g., Marcus, 1993), the lack of which in natural input argues for an innatist acquisition theory. By this they mean that no one points to a sentence like the example just above and tells you, "This is ungrammatical." But a skilled DDL coder can set up a concordance that makes the absence of such a structure clear to learners. Gaskell and Cobb (2004) show ways of responding to learners' writing errors with negative-evidence concordances.

To summarize, the sheer amount and variety of language patterning that can be revealed by concordances, shortly to be augmented by LLM (large language model) artificial intelligence, will always make the concordancer the primary learning tool of DDL . One might ask what place is left for lists.

Up-front word lists

Concordances pre-date word lists in the history of corpus research. The Bible was probably the first corpus, or the first multi-part text treated as a corpus, with just under 800 thousand words in 66 sub-corpora. The concordancing of its main concepts (not every word) was performed by hand and eye by monks, and no attempt is recorded to render the corpus as a frequency list (which would have been an unimaginable labour). But in modern digital corpus research, the word list pre-dates the concordancer. The first digital corpus, large for the time at just over 1 million words, was the Brown Corpus (Francis & Kučera, 1967), of 500 texts of 2000 words each in three broad sub-divisions, and like the Bible large

enough to be uninterpretable in itself without some sort of breaking into pieces. The Brown's first usable break-out was an unparsed frequency list of its 41,745 word types (270 printed pages at four columns/page), which though rudimentary, produced at least two notable contributions: the first large-scale confirmation of Zipf's law, that rank and frequency are inverse quantities in natural distributions like lexicons; and the selection of items for the innovative American Heritage Dictionary (1969), probably to date the most widely used corpus resource that has ever existed. It was a huge advance in lexicography to ground a dictionary in the words of current and recent usage rather than mixing historical and current items randomly or with voluminous annotation.

Concordances of the Brown came only a few years later, in the form of a very large index of every word in the corpus, written in code to conserve space, and used mainly as a means to hand-parsing the words of the list into 80 parts of speech or POS's. (Without the codes, 1 million concordance lines at 35 lines/page are almost 29 thousand pages.) The concordances were basically a broadening out of the tokenized frequency list with a few words to either side, not enough to be read continuously but enough to identify recurring immediate collocations and inform a guess at a POS. Parsing, even though problematic in the beginning and done largely by hand, eventually allowed the refinement of the frequency analysis to include, for example, separate counts for "run" (verb) and "run" (noun). Thus were the concordance lines a means to refining the list; it was probably never imagined that the concordance would largely replace the list.

Frequency lists retained a degree of prominence even in the early DDL era, roughly 1970-2000, probably because frequency software was readily available but concordance software less so, especially before it could be run over the Internet. One application adapted from Sinclair (1991) was to use a frequency list of words in a text as a measure of the text's lexical density and thus its suitability for use with language learners of various proficiency levels. Sinclair noticed that the most frequent words in a frequency list were invariably function words (articles, prepositions, pronouns, helping verbs, conjunctions) with content words appearing only well down the list, and, further, that the ranks of the first content words varied by text topic and complexity. The first content word in the general-English Brown corpus list is "said" at rank 55; in a corpus of graded stories it is "Mr" at rank 57. But in a combined corpus of physics, engineering, and math (from the BAWE; Nesi et al, 2019), "voltage" is at rank 29 with "current" and "circuit" at ranks 31 and 32. The reason for the earlier appearance of content words is that complex texts employ a greater number of them – e.g., in series lists, compound phrases, etc.

This first-content-word method of text selection was widely used in the 1980s era of English for Specific Purposes (ESP) and English for Academic Purposes (EAP), which

sought to give learners authentic but more or less comprehensible reading texts. A weakness of the method was its sensitivity to text length; a short text about Donald Trump may well make “Trump” its most frequent word. Skilled practitioners were nonetheless able to use the procedure effectively.

Lists at this time were rarely given directly to learners, as they had been in the grammar-translation and audio-visual periods. Such use of lists had come into disrepute, owing to questions about where they came from (one list circulating where I worked had “brook” and “ye” as frequent words, suggesting a Biblical origin) and to abundant research showing that simply memorizing word lists did not improve comprehension even, for the texts containing the same words (e.g., Nagy, 1988). There is only one instance of giving lists to learners in DDL work that I am aware of, and that was in one of my own studies (1999a), where learners were given explicit access to appropriate corpus-based vocabulary lists, in a computer program hyperlinked to a corpus of their own course materials. The idea was to get the best out of both lists and concordances, both “breadth and depth” of learning, in the words of a title that came out of this (1999b). The list provided the breadth, a guarantee that every word of a certain frequency would be met and considered, while the concordance provided the depth, the numerous rich contextualizations for each word that would reveal its meaning and pattern of use. It is not clear that this strategy should be called “list-based learning,” since the words on the lists were effectively suggested inputs to the concordance program, and there was no suggestion of trying to learn list words out of context.

Backend word lists

Since then lists have largely disappeared behind the scenes of other DDL software, a disappearance so total that users are often unaware it is a word list that is performing the magic in their software. A good example of this magic is the lexical frequency profiler (LFP; with just four occurrences out of 790 in the research abstracts cited above). A profiler reads in a text one word at a time, comparing every word against a set of frequency lists and recording the result, and at the end yielding a profile or summary of the frequency level of the whole text. Profilers include Anthony’s ANTPROFILER (https://laurenceanthony.net/software/antwordprofiler/); Browne’s NGSLPPROFILER (https://ngslprofiler.com); Cobb’s VOCABPROFILE (https://lextutor.ca/vp); Finlayson, et al’s MULTILINGPROFILER (https://www.multilingprofiler.net/); Nation et al’s RANGE (https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs), all of which are widely used in the language industry.

The lists running inside a profiler are normally organized by frequency but can also be organized by subject area, etymological origin, or in combinations of primary and

secondary lists (say, frequency then etymology, so that for a given frequency level it is clear which particular words are cognate). A typical way of organizing a profile is by the number of 1,000 lemma or family sets needed to achieve 95% and 98% coverage in the input text, these being key points in the development of reading comprehension (Laufer & Ravenhorst-Kalovski, 2010). If just 1,000 word families achieves 98% coverage in a text, then the text is (other things being equal) easy to read for most learners; in the Brown corpus of general English, 10 thousand families are needed to reach this coverage . Profiles are widely used in text selection, text simplification, examination control, and identification of items for pre-teaching.

Such judgments can be made more accurately if teachers know the lexical profile of their learners as well as the profiles of their texts. If they know this, their learners have probably taken a vocabulary 'size' or 'levels' test, probably one produced by Nation and colleagues (e.g., Nation & Beglar, 2007) or Meara and colleagues (e.g., Meara & Buxton, 1987) or their successors (e.g., Sasao & Webb, 2018). All such tests are based on test items dawn from a sequence of corpus-based frequency list (or phrase lists, Martinez & Schmitt, 2012). For example, Nation and Beglar's VOCABULARY SZE TEST samples 10 items from each of the first 14 thousand levels of Nation's BNC/COCA word families based word lists (described and available at https://www.victoria.ac.nz/lals/about/staff/paul-nation) and provides a size estimate by multiplying the score at each level times 1,000 and totaling the levels. The procedure is simple but effective. Since the same lists are behind both test and profiler, it is possible to tune texts to learners in a variety of useful configurations. For learners with receptive knowledge of the first 2,000 word families, a 3,000 level text provides an intensive reading activity, while 1,000 provides a fluency activity, and so on.

Another DDL tool that is list based is KEYWORD, which inputs a text and delivers its key words, the handful of words or phrases that truly characterize its "aboutness," what the text is mostly about (Scott, 2001; Bondi & Scott, 2010). If teachers know what these keywords are, they can draw learners' attention to them as an aid to comprehending the text. This is especially important to do if the keywords are also beyond the learners' current vocabulary level, as is usual, since texts' lower frequency words tend to carry their main ideas (Kučera, 1982). For this reason, it is logical for keyword and profiler to work together (e.g., at https://lextutor.ca/key/). The keywords of a text are determined by breaking the text into a frequency list and then comparing it to the frequency list of a standard or non-specialist corpus. Words that are 25 times (a commonly used but arbitrary figure) more frequent in the text than in the corpus are probably its keywords. For example, in the Oxford Bookworms graded story *Dracula,* the word "coffin" occurs 29 times in 7,613 words, or, scaled up to a corpus of 10 million words, 38,000 times. By comparison, in a general corpus of 10 million words, "coffin" occurs 73 times, giving the word a keyness in Dracula

of 520. And further, “coffin” is in the fourth 1,000 of Nation’s BNC/COCA word list, as are its fellow keywords “hammer,” “garlic,” “diary,” and “carriage”; these are all mid-frequency rather than basic items, and this is a text where the first 2,000 families provide 98% coverage. Keyword analysis is thus important even in simplified texts, but it is crucial to the establishment of domain-specific word lists. And again, the count for “keyword” in the research abstracts is just two out of 790 (though the search is error-prone in that “keyword” also appears in the unrelated phrase “keyword in context,” KWIC).

There are also lists at work within concordance routines themselves. For example, when the search input space offers the options “lemma” or “family,” that means the program can access a list of all the fleshed-out lemmas or families of the language which will be incorporated into the search term through a regular expression that matches not just “see” but any of “see|sees|seeing|seen|saw.” Lists are also employed in normalizing texts prior to concordance analysis, for example reducing contractions to separate words rather than piling up separate counts of “can” and “can’t” as if they were unrelated items. On a much larger scale, Davies’ 1 billion word Corpus of Contemporary American English (COCA) is “really” just an unreadable list of numbers until it is formatted into comprehensible language following data assembly. Davies explains that this is the only way truly massive corpora can be searched or the results assembled in a reasonable amount of time (from a piece entitled ‘Architecture’ at https://www.english-corpora.org/help/architecture.pdf).

Back at the user level, one visible integration of lists and concordances is the ability of some concordancers to use resident frequency lists to work out the average frequency of every line in the concordance output and then sort it from high to low, or, normally, readable to challenging. Thus the lines at the top of the output are composed mainly of common words and have a chance of being readable despite the relative difficulty of the overall corpus. This feature is widely used in ESP courses as a way to give learners input that is authentic yet comprehensible without simplified. For instance an electrical engineering student can meet the word “current” in a sentence like “The heater was connected to a heat monitor and power supply and was switched on while a note was made of its CURRENT and voltage” (first 1,000 rating) rather than, “In 1888 Tesla demonstrated his brushless alternate-CURRENT induction motor to the Institute of Electrical Engineers (IEEE)…” (sixth 1,000). The main concordancers offering this service are SketchEngine and Lextutor, the latter employing the 1000-families scheme of the BNC/COCA thereby integrating concordancer and profiler.

A final example involves a typical DDL tutorial activity, in this case a cloze passage builder in which a teacher puts in a text and gets out a cloze passage with every n-th word removed for the learner to replace. Cloze is a comprehension focused activity in the sense

previously discussed, focusing on comprehension of the text as a whole though the task is to replace single words or adjacent phrases. There is magic that can be added to this routine and it involves recourse to word lists. For example, using a short list of prepositions, or articles, the program can choose just those items to remove for replacement. A more interesting example involves focusing the cloze builder on morphology instead of lexis, without requiring the teacher to parse or tag the text, changing all the inflectable items in the text (nouns, verbs, adjectives in the case of French) to a pulldown menu of inflection options for the learner to choose from. For example, "The boy [run, runs, ran, running] home when he heard his mother calling." To find the inflectables within the chose n-range, the program first determines programmatically that each candidate word is not a proper noun, then through list comparison that the word is not present in the lists of function words and interjections but *is* present in the list of inflected nouns or verbs and builds the gap accordingly from its resident lemma lists. One routine that can do this is N-WORD CLOZE BUILDER at https://lextutor.ca/cloze/n/. The examples of list magic could be multiplied.

So, to summarize, word lists are doing a lot of the lifting in DDL.

What would list research look like?

It is not quite true that no one is doing research on word lists and their role in language learning. There is strong ongoing research on these topics, just categorized under vocabulary research within applied linguistics, and to some extent CALL, rather than DDL. One possible reason for this is the relative recentness of DDL research; the compilation of research abstracts referenced above (790 studies) is probably pretty close to all the studies there are. Another is the lack of a dedicated DDL journal; some of the list work in applied linguistics is probably being done by DDL researchers.

Word list research within applied linguistics is rich and varied and has compared and categorized lists on questions like their text coverage (Cobb & Laufer, 2021; Dang, 2020); their usability by teachers (Dang, Webb, & Coxhead, 2020); their role in exam writing (Marsden et al, 2023); whether the lemma or the family is the more useful unit of word counting and list organization (McLean, 2017); or whether token lists or family lists provide the better predictor of writing proficiency (Crossley & Cobb, 2013). All these issues have a potential bearing on which lists should be used within DDL. It seems implicit in this research that the best lists, as progressively identified, will be used mainly in behind-the-scenes software, i.e. in tests, profilers, and cloze passage builders, rather than given up-front to learners out of context.

Is it a problem that official DDL confines itself to concordancing while applied linguistics/CALL gets the rest of DDL? In other words, is it a problem that DDL researchers and applied linguists appear to know each other so little? It could be a problem. DDL's main successes so far have been, as shown in meta-analyses by Boulton and Cobb (2017) and Ueno & Takeuchi (2023), mainly in vocabulary and collocation, yet largely without input from or engagement with the vocabulary research on the same topics. Conversely, many of the recent successes of vocabulary research have been dependent on corpus work, yet possibly without sufficient engagement with DDL expertise, where the possibilities of the technologies that these involve may be better known. The problem of "siloing" exists in many branches of applied linguistics, not just this one, and the relative youth and evident high energy level of DDL workers probably means this work will eventually be incorporated as a branch of CALL, and may even provide its long awaited theoretical basis.

Conclusion

Word lists are as much a part of DDL as concordances, and it is a limitation in the development of DDL not to include list research within our research agenda. Doing so will involve working with and learning from our colleagues in applied linguistics where list research is flourishing.

References

The American Heritage Dictionary of the English Language (1969). New York: Houghton-Mifflin-Harcourt.

Boers, F., Lindstromberg, S., & Eychmans, J. (2014). Some explanations for the slow acquisition of L2 collocations. Vigo International Journal of Applied Linguistics 11, 41-62.

Bondi, M. & Scott, M. (Eds.) 2010. Keyness in Texts. Amsterdam/Philadelphia: John Benjamins Publishing Company. 251 pp. ISBN 978-90272-8766-3.

Boulton, A. & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. Language Learning 65 (2), 1-46.

Boulton, A. & Vyatkina, N. 2021. Thirty years of data-driven learning: Taking stock and charting new directions. Language Learning & Technology, 25(3), 66-89. https://doi.org/10125/73450

Cobb, T. (1999a). Applying constructivism: A test for the learner-as-scientist. Educational Technology Research & Development, 47 (3), 15-33.

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

Cobb, T. (2018) From corpus to CALL: The use of technology in teaching and learning formulaic language . In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), Understanding formulaic language: A second language acquisition perspective (pp. 192-211). New York: Taylor & Francis.

Cobb, T. & Laufer, B. (2021). A nuclear word family list: The most frequent family members, base and affixed words. Language Learning 71 (3), 834-871.

Crossley, S. & Cobb, T. & McNamara, D. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. System 41 (4), 965-981.

Browne, C. (2024). NGSL Profiler [Computer Software]. Tokyo, Japan:. Available from https://www.ngslprofiler.com.

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., 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

Francis, W. & Kučera, H. (1967). Computational Analysis of Present-Day American English. Providence, RI: Brown University Press.

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? System 32 (3), 301-319.

Kučera, H. 1982. The mathematics of language, in The American Heritage Dictionary. Boston: Houghton Mifflin.

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

Marcus, G. (1993). Negative evidence in language acquisition. Cognition 46 (1), 53-85. https://doi.org/10.1016/0010-0277(93)90022-N

Marsden, E., Dudley, A., & Hawkes, R. (In press). Use of word lists in a high-stakes, low-exposure context. The Modern Language Journal.

Martinez, R & Schmitt, N. (2012). A phrasal expressions list. Applied Linguistics 33 (3), 299–320.

McCrostie, J. (2007). Investigating the accuracy of teachers word frequency intuitions. RELC Journal 38 (1), 53-66.

McLean, S. (2017). Evidence for the adoption of the flemma as an appropriate word counting unit. Applied Linguistics 39, 823–45.

Meara, P. & Buxton, B. (1987). An alternative to multiple choice will vocabulary testing. Language Testing 4 (2), 142-154

Nagy, W. (1988). Teaching vocabulary to improve reading. Urbana, Ill., NCTE.

Nation, P., & Beglar, D. (2007). A vocabulary size test. The Language Teacher 31(7), 9–13.

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

Sasao, Y., & Webb, S. (2018). The guessing from context test. ITL - International Journal of Applied Linguistics, 169(1), 115-141.

Seliger, H. (1983). The language learner as linguist: Of metaphors and realities. Applied Linguistics 4 (3), 179-191.

Sinclair, J. (1991). Corpus, concordance, collocation. London: Oxford University Press.

Scott, M. (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs. Small corpus studies and ELT, 47-67.

Ueno, S. & Takeuchi, O. (2023). Effective corpus use in second language learning: A meta-analytic approach. Applied Corpus Linguistics 3, 1-11.

Zimmerman, M. (1988). Texas indexer/browser, v. 0.27. Silver Spring, Maryland.