Identifying Opaque Items on the Academic Vocabulary List

Christopher Osment
Independent researcher
Corresponding author's email: osmentchristopher@gmail.com
Received October 30, 2024; revised December 28, 2024; accepted December 29, 2024

Abstract
Numerous word lists exist, such as the Academic Word List (AWL) and the Academic Vocabulary List (AVL). However, while many of the words on these lists are understood relatively easily with the help of online dictionaries or translations, some words have multiple senses and grammatical aspects that are likely difficult for learners to readily understand. These items are termed "opaque", as their meanings cannot be clearly determined. In this paper, I identify such problematic vocabulary items to provide a more focused list since the AVL includes over 3,000 items. The meanings of the target items were examined first with six online dictionaries: Cambridge, Collins, Longman, MacMillan, Merriam-Webster, and Oxford Dictionaries. The academic section of the Corpus of Contemporary American English (COCA) was used to select 100 random entries for a target item with entries up to the year 2022. The parameters applied were a mix of qualitative and quantitative, with relevant information (collectively called item affinities) such as collocations, lexical bundles, related words, senses, and colligations included. The first 600 AVL Core Academic words were sifted for opacity, resulting in a list of 103 items. These were high frequency words with approximately 75% possessing 1 or 2 syllables and 25% having 3 or 4 syllables. It is suggested that filtering for opacity can render a vocabulary list more manageable for teachers and learners.

Keywords: opacity, word list, affinities, AVL, COCA

Word lists are a useful resource; however, there can be too many words to cover within the limited time of many university courses—vocabulary being one of the domains of language teaching competing for class time and instructor attention. Moreover, simply assigning lists to students to learn may be problematic, as it assumes that the students will be able to readily define words with the aid of a dictionary (online or paper-based) or a translation app. Assignment of vocabulary lists does not allow for the possibility of differing senses of words in different genres or texts. Additionally, work over the past decades has established other information that a person needs to enable understanding and use of a word effectively, such as collocations, colligations, genre sense, and connotations (Biber et al., 1999; Gibbs, 2007; Hoey, 2005; Nation, 2013; Schmitt, 2000; Sinclair, 1996).

Aside from the difficulty in identifying the presented meaning of a word, there is also the daunting task of sifting through lengthy word lists and determining which items should be taught. However, it is often the case that an ”average” instructor would not know how to reduce a lengthy vocabulary list to a manageable and relevant number of items, nor would they have the time to identify key information such as collocations, genre use, or lexical bundles.

With these issues in mind, this paper covers the first 600 words of the Academic Vocabulary List (AVL) - Core Academic List (Gardner & Davies, 2014), which consists of just over 3,000 words. Later work will include words 601–1,000 and then words 1,001–2,000. In exploring the see words, I use the term “opacity’” to identify vocabulary items that are difficult for a leaner to easily understand by using a dictionary or translation app. The term and idea of opacity does not originate with this work but is based on work by Watson Todd (2017) and Hsu (2014). Watson Todd examined an engineering word list for opaque items, while Hsu examined opaque formulaic sequences. Both works were useful for this present work; however, the goal of this paper is to extend the process to the AVL.

Word lists for academic purposes, such as the AVL, typically evaluate for genre, range and dispersion, or whether the identified item

appears on other lists (Gardner & Davies, 2014). The creation of these lists lead to a variety of statistics that are necessary, but unlikely to mean much to a front-line instructor or learner. Additionally, there are other considerations when dealing with student praxis, and some typical questions that arise are as follows:

1. How does an instructor select items in a meaningful and practical manner?
2. What does a student encounter when looking up a word?
3. How does a learner know which of the senses for polysemous words to choose?
4. How can researchers render their findings into manageable understandable forms?

While this work cannot fully address these four questions because of length restrictions, it can start us on the long path to providing some usable answers. Thus, how learners try to find definitions and what issues they encounter when doing so is key; however, this also works recognizes that there is no one answer to how a learner seeks a definition.

Regarding Question 1, as useful as word lists are, they present far too many items for an instructor to adequately cover in the limited time of many courses. For instance, in a class conducted by this researcher, a South Korean student related how his language teacher had provided a list of 3,000 words and told the students to memorize them for the end of term test. The student reported that he was able to memorize most of the words, but he had virtually no understanding of how to actually use the words for productive purposes. This is anecdotal in nature, yet many practitioners may recognize such an occurrence and recall similar events.

Regarding Question 2, as Watson Todd (2017) noted for Thai students, they employed Google translate or Longdo; in contrast, while working in China, Chinese students make use of Baidu, WeChat or another platform indigenous to China. That is, learners in different locations are likely to choose what is at hand; a learner may encounter

very different definitions, particularly for polysemous words, thus some form of focused instruction would be needed.

As for Question 3, the idea that a learner can guess the meaning from contextual clues is helpful at times, but as Nation (2009, 2013) has noted, it would require significant knowledge of the other words in the text and an understanding of the topic being read.

Question 4 remains problematic. There is no easy solution, and identifying opaque items is only one step in a much longer process. Successful acquisition and employment of learned words is difficult to define, yet the work of Sinclair (1996) and others (Davies, 2020; Jiang & Hyland, 2017; Schmitt & Carter, 2004) on collocations, lexical bundles, genre, etc., suggests that a definition alone will not enable a learner to adequately produce accurate discourse.

Literature Review

While several approaches could be used to discuss the topic of vocabulary lists, a historical approach is employed in this work to show the evolution of word lists. Such an approach also highlights how further evolution is needed beyond listing to reduce the challenges that practitioners and learners encounter with word lists, hoping to render such lists more effective and manageable. The utility of word lists is not the focus of this work, nor is this researcher questioning the utility of word lists.

In the mid-20th century, when Michael West (1953) published his seminal work, the General Service List (GSL), the vocabulary list as it is today, based on corpus-driven work, took shape. Earlier work by Palmer (1933) illustrated the collocational aspects of language, and how such collocations were of importance in the production of natural sounding language. The concept of collocation and the pattern-based nature of language was furthered in work by Firth (1957) and Halliday (1966); additionally, these works drew attention to word and phrase frequency of words that appeared in texts; however, the limits of computer and software technology hampered the elucidation of insights at the time.

The technology to examine the frequencies of words and the patterns of language arrived in the 1960s. Subsequently, the one-million-word Brown Corpus appeared in 1967 (Kucera & Francis, 1967). In the next decade, amalgamation of the Lancaster-Oslo/Bergen Corpus and the University of Lancaster University resulted in the British English Lancaster-Oslo/Bergen Corpus, and during the 1990s collaboration between three publishers (Oxford, Longman, and Chambers) and two universities (Lancaster and Oxford), led to the British National Corpus of over 100 million words. More or less at the same time, the Collins Birmingham University International Language Database (COBUILD) presented large quantities of accessible and analyzable corpus data. Outside of the English language, numerous other corpora exist as well; for example, the Montreal French Project (Sankoff & Sankoff, 1971) and, more recently, the Quranic Arabic Corpus, which was compiled by Dukes at the University of Leeds (2011).

With the availability of personal computers and relatively ease to use software, corpus linguistics in the 1980s and 1990s enabled researchers to calibrate and clarify the earlier observations of West, Palmer, Firth and Halliday. This led to concepts such as Willis' *Lexical Syllabus* (1990), Sinclair's discussion of lexical grammar in *Trust the Text* (2004), Hoey's *Lexical Priming* (2005), and the *Longman Grammar of Spoken and Written English* (Biber et al., 1999). The data from corpora and subsequent ruminations on said data have led to multiple impacts on the field of language learning and teaching.

Then, Coxhead's (2000) much discussed and oft cited Academic Word List (AWL) entered the foray, with its 570 word families and ten sub-lists. The list was based on the examination of a wide variety of academic texts, and aimed to provide usable academic vocabulary for learners and instructors. Within a decade, work based on the Corpus of Contemporary American English (COCA) presented the Academic Vocabulary List that was derived from a more balanced set of genres than the AWL of Coxhead (Davies & Gardner, 2014). An updated version of the GSL appeared in 2013 to accommodate changes

in language use over the proceeding sixty years. It was based on an +270 million subset of the Cambridge English Corpus, in the form of the New General Service List (NGSL), which updated and expanded the original GSL (Browne, 2014).

Although the above-mentioned work on vocabulary has been of great benefit to learners and practitioners, they are not without flaws that render the lists difficult to apply in practice. First, most vocabulary lists present far too many vocabulary items for application within the limited scope of most courses. Additionally, the broad nature of such word lists makes them a little impractical for genre specific applications as noted by Hyland and Tse (2007), who argued that "the different practices and discourses of disciplinary communities undermine the usefulness of such lists and recommend that teachers help students develop a more restricted, discipline-based lexical repertoire" (p. 235). As the formulaic aspect of language has become more apparent, researchers have also noted the need for learners to understand the functions of lexcial bundles as discourse organizers, and how text is shaped by these bundles (Granger & Larsson, 2021; Jiang & Hyland, 2017; Tahara, 2020). Hyland and Tse (2007) also argued that many language forms can have different meanings and functions that depend on the contexts of the language used, thus it is possible to claim that vocabulary behaves differently across contexts and genres. Such a claim leads to a need for vocabulary lists to include more relevant data, such as colligations, collocations, and the differing senses of an item in different environments. Notably, in his GSL, West (1953) did provide such data, which, considering the lack of computer aid, was a note-worthy achievement.

Along with the limits of vocabulary lists regarding formulaic language, discourse functions, and the like, some work has focused on "whether all words … should be the focus of productive activities in EAP classes. Learners' needs for academic writing are clearly not the same as for academic reading" (Paquot, 2007, p. 127); that is, the needs of learners engaging with different domains will diverge from each

other to some degree, thus a one size fits all vocabulary list, no matter what its size or genre coverage may be, will lack specificity.

At this stage, it should be noted that considerable research has been conducted on the coverage of established lists. Work by Nation (1999) and Gilner and Morales (2008), for example, has shown that the GSL does not provide as much coverage as the BNC2000; however, Browne (2014) concluded that the GSL did provide greater coverage than Browne's (2014) NGSL in some genres. While such works are important, there still exists a significant gap in praxis for learners and practitioners regarding learning domains. Specifically, the coverage of vocabulary lists is arguably not the one and only concern of a classroom teacher conducting a first-year undergraduate writing course in Thailand or China, for example, nor a civil engineering student struggling to make sense of how to link ideas in a paragraph. Moreover, many classroom practitioners have limited training in using or engaging with word list data, and given limited time and other constraints, it should not be expected that a front-line teacher is likely to spend much time deciphering the application of a vocabulary list; researchers and materials designers have this responsibility.

Regarding the making of vocabulary lists more attuned to teaching, some work has already been done by Watson-Todd (2017) and Hsu (2014). Watson-Todd applied the idea of opacity to the Engineering English Corpus (Osment & Graham, 2013) to identify a more readily teachable list. Watson-Todd's work was based on the idea that "words chosen for an explicit classroom focus should be words that students are likely to have problems dealing with autonomously, and that these are polysemous words where the meaning required is not the usual meaning, in other words, opaque words" (p. 31). This current work is based upon Watson-Todd's approach and view of opacity, but in a modified form. As for Hsu's work (2014), formulaic sequences were investigated to determine if they were "non-transparent" (p. 146). Results showed that "475 non-compositional expressions of 2–5 words … ultimately chosen and formed the Opaque Formulaic Sequences (OFS) list" (p. 148); the list "encompasses 264 two-word,

152 three-word, 57 four-word and 2 five-word phrases commonly used in college textbooks" (p. 148).

Theoretical Framework

Creating a list of opaque words is subjective to an extent; however, work by Hsu (2014) and Watson Todd (2017) provides guidance. These two works examined polysemous words and formulaic phrases; that is, words or phrases that have many meanings or senses. Items that have multiple meanings may have a commonly encountered meaning; a facile example is the word "table", which in everyday life refers to the item of furniture, but in academic texts refers to some form of data containing graphic. For a learner to identify the specific meaning of a word would require the capacity to understand the discourse context well, understand the sentence or sentences surrounding the word, and decipher complex dictionary entries to find the correct sense of the word (Nation, 2013, 2015). Also, research by Nesi and Haill (2002) found that learners often focus on an incorrect sub-entry for a headword when they use a dictionary. Additionally, learners were shown to frequently select the first sub-entry of a word and stop (Boonmoh, 2003). Again, this is not unsurprising given the confusion a learner might feel when presented with the dense information of many dictionary entries.

Method

This work uses a hybrid approach; it is a mix of quantitative and qualitative. Although quantitative data provides much useful data, simply transferring it to a list is insufficient, as quantitative data does not indicate what a student or instructor encounters when looking up a word; this relates to Questions 1–3 (see Introduction). The qualitative aspect recognizes that data is interpreted and applied based on belief, experience, and knowledge, and thus enables analysis based on factors like opacity, allowing for selection that relates to Questions 1–4 (see Introduction). In other words, by incorporating a qualitative aspect, it is possible to use existing corpus data to include collocations,

colligations, and other relevant information, along with a quantitative assignation of opacity that examines multiple dictionaries and COCA for differing senses and uses. However, sometimes, a subjective decision was applied to grey areas when the data did not present a clear choice; additionally, selection of teachable points was subjective, yet based on the existing knowledge and experience of this researcher. The details of the process are outlined in the following sub-sections.

Selection Criteria

A two-step process was employed. Step one consisted of identifying the opaque items based on the criteria listed below in a–h. Step two involved more detailed work and consisted of the inclusion of other relevant information for each sense of the opaque items where possible. Much of this information was found within the COCA website, but some aspects were added, such as the CEFR ranking, since CEFR vocabulary appears in many textbooks such as Cambridge's *Touchstone* series, but the CEFR senses are typically the commonly encountered ones, while the AVL is for academic senses. Additionally, senses were added for opaque items 1–100 to illustrate the complexity of what a learner may encounter when attempting to define a word. Senses where also added to potentially aid teachers in appropriate selection for the sense occurring in their materials, texts, etc. Due to space limitations, the full list of senses is not presented in this work; instead a sampling of ten opaque items is presented in Appendix G.

A. Meanings of the words were looked up first with the dictionaries listed in C. This step helped to determine what senses of the words were since a cursory examination of the Corpus of Contemporary American English (COCA) first, resulted in different senses being missed. Moreover, it allowed for the recognition of which parts of speech a word belonged to, as it may not always be obvious without examination of the sentence context.

B. COCA was used, and the Academic Selection was chosen. Then, 100 random entries were generated. The original AVL was compiled in 2013, but entries up to the present (2022) have been

examined. The additional nine years is thought unlikely to significantly affect academic usage for the words on the AVL Core Academic List. Words that would have greater usage today such as “virus” or “woke” do not appear on the AVL Core Academic List.

C. Six online dictionaries were referenced: Cambridge, Collins, Longman, MacMillan, Merriam-Webster, and Oxford dictionaries, as these dictionaries appeared in most Google searches for a word when using search terms such as *determine meaning*, *determine definition*, *what does determine mean*. Moreover, each of the dictionaries represents a major publisher of academic materials, suggesting that a learner is likely to use a dictionary in a language class that is published by one of these publishers.

If four to six of the dictionaries had the same sense listed and part of speech listed first as the COCA entries, then it was not listed as opaque. If three or more dictionaries listed a different sense or part of speech first, the item was considered problematic for a learner and opaque, since a learner that cross-checked meanings with different dictionaries would encounter differing senses, which was judged to be potentially confusing for the learner.

Translation tools were not used since their output may vary according to where a student is located and a student’s primary language; in other words, translation tools where deemed impractical, as they would provide a list too complex to be of practical application, which is a primary goal of this paper.

D. Multiple senses were considered for overlap in meaning. If overlap occurred, they were classed as not opaque. If no overlap occurred, then they were classed as opaque. Two examples are shown below to illustrate the process, starting with the word “internal”, which has several overlapping senses and the common aspect of being “inside something/someone”.

i. existing or happening within a country, not between different countries

ii. existing or happening within an organization or institution

iii. existing or happening within something such as a process or system
iv. existing or happening inside an object or building
v. existing or happening inside your body
vi. existing or happening inside your mind

On the other hand, a word such as "directly" has multiple senses with much less overlap, as shown below.

i. involving no one else
ii. in a direct line
iii. exactly
iv. clearly and honestly
v. immediately
vi. soon

E. If more than 15% of the COCA entries were different parts of speech or senses, and thus there is a greater chance that a student may find the word difficult to define from a dictionary alone, the word was considered opaque.

F. At times, the part of speech listed in the AVL Core Academic List did not correspond to COCA entries; for instance, the AVL Core Academic List shows a verb, but the noun appears much more frequently in the COCA entries. In this case, an additional 100 random entries were examined to determine if the result was the same. If so, then the word may be opaque.

G. If two parts of speech appeared in the AVL Core Academic List, the 100 random entries were checked to see if the ranking in the AVL list matched the frequency of appearance in the COCA entries (i.e., if the verb form appeared first in the list, was the verb form more frequent in the random sample COCA entries?).

H. Different forms for verb lemmas were not examined for every case although future work should do so. This is because inflections, and tense and gerund/participle forms may have different senses and could prove problematic for learners. For example, the lemma "increase" (V) was compared with "increased/increasing" (ADJ) as they appear in the AVL Core Academic List, but the lemma "select" is not.

Note that inter-raters were not used since one of the core ideas of this work is to address what learners or front-line practitioners encounter when looking up a word. Whether another researcher considered the word opaque or transparent was not the issue. As for intra-rating, the approach to determine opacity included safeguards to check reliability; for example, when in doubt an additional 100 entries were generated in COCA, or for the 15% criterion in E more than one count was typically done with three counts being done if there were any issues.

Item Affinities

Although word lists are an excellent tool, they do not provide additional information that is often needed to fully understand a word's meaning and usage. This additional information is referred to as *affinities* in this work to encompass the following aspects under one term: collocations, colligation, semantic preferences, related words, and lexical bundles for the opaque items. Ideally, the above information for different senses would be included, but at present, this is too time intensive. The COCA data provided the basis for affinities as this information appears for each of the items in the AVL. Additionally, COCA data for topics indicates the semantic preference of target words. The topics (semantic preference) are words that occur in a text frequently with the target word but are not necessarily nearby as collocates would be. Davies and Gardner (2020) suggest that the topics provide a better sense of what words and ideas are related to the target word. This could provide useful information for teaching.

Additionally, wherever possible the Common European Reference for Languages (CEFR) listing will be given (e.g., A1, A2, B1, B2, or C1) as such information may be useful to instructors for ordering the target items. Notes have been written for the opaque words from the first 1–100 words in the AVL (see Appendix G) when something interesting or potentially relevant to teaching was observed; this is one of the subjective aspects of this work. These notes follow the list of opaque words.

Results and Discussion

The items that presented opaque characteristics in the first 1–600 items in the AVL were high frequency words, with approximately 75% possessing 1 or 2 syllables and approximately 25% having 3 or 4 syllables. It is unknown if the syllable count is of consequence, but it is suggested that high frequency words possess a small number of syllables to accommodate the limitations of working memory. In total, for the first 600 AVL Core Academic words there were 103 opaque items. The number of opaque items decreased as less frequent items were analyzed. A preliminary probe into the 2901–3015 section (not presented in this work) of the AVL revealed only six opaque items, while the 1–100 section revealed 25 opaque items (Appendix A). The numbers for the other sections examined in this work are shown in Table 1 below.

Table 1

A Preliminary Analysis of AVL Sections for Opaque Items

Section	# of opaque items	Appendix
1–100	25	A
101–200	21	B
201–300	15	C
301–400	19	D
401–500	13	E
501–600	10	F

This result is unsurprising, as a decline in opaque items was expected to occur further into the AVL, as the items appear to have more specific applications and meanings. Since the data sample is only for the first 600 items in the AVL, any attempt to fit the data to a clean line has been avoided, recalling that Leung et al. (2004) asserted the need for researchers to develop a conceptual framework acknowledging how messy data can be rather than trying to obscure this issue.

Determining a precise number of senses can be problematic, as some senses appear very similar; however, an approximation is possible. The opaque words typically had a wide range of senses with

some words such as "image" and "base" possessing sixteen and seventeen senses, respectively, but more when it came to their dictionary definitions. See the full information for "image" in Table 2 for an illustration of the potential challenges for a learner.

Table 2

Information for the Opaque word "image"

Word: image; **AVL Ranking:** 59; **CEFR:** A2

Semantic Preferences:	Used to indicate where the reader should look
Lexical Bundles:	images in, images from, in the image
Collocations (noun):	body, photo, digital, create, above, below
Colligations:	links clauses with noun phrases/clauses or a subject verb with a noun phrase
Related words:	imaging, imagery, imagination, imaging, self-image
Senses:	**Noun:** A visual representation of something: such as a likeness of an object produced on a photographic material; a picture produced on an electronic display (such as a television or computer screen); the optical counterpart of an object produced by an optical device (such as a lens or mirror) or an electronic device; a mental picture or impression of something; a popular conception (as of a person, institution, or nation) projected especially through the mass media; an exact likeness; a person strikingly like another person; a tangible or visible representation; a vivid or graphic representation or description; a reproduction or imitation of the form of a person or thing, especially an imitation in solid form. **Verb:** To create a representation of something; to represent symbolically; to call up a mental picture of something; to describe or portray in language especially in a vivid manner; to make appear; to make a disk image of something.

The above information highlights the potential difficulty a learner can encounter when determining the meaning of a novel vocabulary item or encountering the same vocabulary item in different genres where a differing sense may be used.

Another issue was the ranking within the CEFR, as the CEFR does not indicate senses of words, thus a word such as "table", which is an A1-level word likely refers to the common piece of furniture; this indicates that the CEFR based word lists may be of limited use in an academic English course. It is probable that in CEFR vocabulary lists for other languages a similar issue arises. Such issues further demonstrate the limitations of employing vocabulary lists without consideration of the items on a list. It should be stressed that the employment of word lists is not questioned by this work, rather it is the employment of unfiltered word lists and the omission of critical lexical features such as an item's affinities.

Regarding the CEFR level of the AVL word list, the 100 most common academic words were anticipated to occur within CEFR levels A1–B1, since they are high frequency items. Indeed, upon examination, most of the words did occur within A1 (9), A2 (11), and B1 (4) levels; one item, "approach", fell into the B2 level. Overall, for the 103 opaque items, the percentages for occurrence in the CEFR are as follows, with a small percentage of items not found (NF) in the CEFR lists: A1 = 15%, A2 = 24%, B1 = 19%, B2 = 34%, C1 = 4%, and NF = 4%.

There were items such as "approach" that occurred in CEFR B2. This disparity is probably due to "approach" not being a word that one encounters in everyday discourse, but it frequently occurs in academic genres. Other items within the top 600 of the New AVL that did not appear in CEFR levels A1–C1 were the following: *given*, *developing*, *increased*, *adopt*, *establish*

Since the CEFR vocabulary list is non-academic, the occurrence of items at the lower CEFR levels and those commonly occurring on the New Academic Vocabulary List, is likely due to the multiple senses of the academic items. In other words, the CEFR listing likely relates to common daily objects and meanings that differ from the sense specific meanings in each academic genre. For instance, as previously noted, the item "table" is a CEFR A1 word, which refers to furniture, while

the item "table" in academic use often refers to some form of graphic with columns and rows, representing data.

Although a variety of senses are presented in the data for the opaque items from the first 1–100 AVL words, a classroom instructor should not be expected to give all these senses for a given vocabulary item to students; the senses are a resource of which practitioners can avail themselves. For example, the number of senses found for vocabulary such as "subject" are numerous; the presentation of these senses in the data is not intended to suggest that each one of the senses should be taught in the classroom. The intention is to highlight the complexity of defining a word for a student or instructor whose English language skills may be insufficient to do so. To present all of the listed senses would defeat a primary purpose of this work, which is to provide a condensed, more teachable academic vocabulary list. While perusing the dictionary definitions, this researcher found the lists somewhat daunting and requiring further definition. One must consider that if the act of looking up a word is a tricky task for an experienced practitioner, then the question arises: how would one expect a learner to do so?

Conclusion

Frequency in itself is insufficient for teachers and learners to organize vocabulary lists around. Current vocabulary lists are valuable tools, but incomplete. Opacity is one tool that can render a vocabulary list more manageable. Additionally, it is suggested that more research should be conducted on the productivity of an item. Item productivity would be a collation of a word or phrase with its affinities and frequency and genre breadth.

Additionally, there is a subjective aspect to opacity determination, which needs to be made more consistent through further refinement. Collocations are well investigated, yet this is only one of the affinities that a word or phrase may have. In addition, the various senses of a word and its genre associations indicate that collocations, colligations, lexical bundles, and semantic prosody (neutral words can be perceived with positive or negative associations) can vary significantly. The

affinities of words and phrases require further investigation to render the complexities of their intersectionalities understandable for learners and teachable for instructors. This was not a primary focus of this work, but it became apparent that an item's affinities (collocations, colligations etc.) require better collation with items if they are to be taught effectively and rendered accessible to learners.

Another implication of my argument is that teaching all an item's affinities would likely be confusing to learners and impractical for instructors. Therefore, I would recommend a layered approach, with the simplest aspects presented first, such as presenting the most common meanings and collocations for high frequency items and then, in subsequent levels, other more detailed item affinities related to colligation, semantic prosody, genre, and senses could be presented. As a result, an item would be encountered multiple times at different levels, which is likely to improve not only item acquisition, but item application as well.

References

Baugh, S., Harley, A., & Jellis, S. (1996). The role of corpora in compiling the Cambridge International Dictionary of English. *International Journal of Corpus Linguistics*, *1*(1), 39–59. https://doi.org/10.1075/ijcl.1.1.04bau

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* Pearson Education.

Boonmoh, A. (2003). *Problems with using electronic dictionaries to translate Thai written essays into English* [Unpublished master's thesis]. King Mongkut's University of Technology Thonburi, Bangkok.

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction*, *3*(2), 1–10. https://doi.org/10.7820/vli.v03.2.browne

Cambridge Dictionary. (2022). Retrieved August 23, 2022, from https://dictionary.cambridge.org/

CEFR Vocabulary Word List. (2024). Retrieved from https://www.esl-lounge.com/student/reference/c1-cefr-vocabulary-word-list.php

Collins Online Dictionary. (2019). Retrieved December 28, 2019, from https://www.collinsdictionary.com/

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Dang, T. N., Webb, S., & Coxhead, A. (2022). Evaluating lists of high frequency words: Teachers' and learners perspectives. *Language Teaching Research,* 617–641. https://doi.org/10.1075/itl.167.2.02dan

Davies, M. (2020). *Corpus of contemporary American English*. Retrieved April 25, 2020, from https://www.english-corpora.org/coca/

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text, 20*(1), 29–62. https://doi.org/10.1515/text.1.2000.20.1.29

Firth, J. (1957). *Papers in Linguistics 1934–1951*. Oxford University Press.

Friginal, E. (2018). *Corpus linguistics for English teachers: Online resources, and classroom activities* (1st ed.). Routledge.

Gardner, D., & Davies, M. (2014). A New Academic Vocabulary List. *Applied Linguistics*, 35, (3), 305–327. https://doi.org/10.1093/applin/amt015

Gibbs, R. W., Jr. (2007). Idioms and Formulaic Language. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford handbook of cognitive linguistics* (pp. 697–725). Oxford University Press.

Gilner, L., & Morales, F. (2008). Corpus-based frequency profiling: Migration to a word list based on the British National Corpus. *The Buckingham Journal of Language and Linguistics*, *1*, 41–57.

Granger, S., & Larsson, T. (2021). Is core vocabulary a friend or foe of academic writing? Singleword vs multi-word uses of THING. *Journal of English for Academic Purposes, 52*, Article 100999.

Halliday, M. (1966). Lexis as a linguistic level. *Journal of Linguistics, 2*(1), 57–67.

Halliday, M. (1975). Learning how to mean. In E. H. Lenneberg & E. Lenneberg (Eds.), *Foundations of language development: A multidisciplinary approach* (pp. 239–265). Elsevier. https://doi.org/10.1016/B978-0-12-443701-2.50025-1

Hoey, M. (2005). *Lexical priming.* Routledge.

Holec, H. (1981). *Autonomy and foreign language learning*. Pergamon Press.

Hsu, J. (2010). The effects of collocation instruction on the reading comprehension and vocabulary learning of college English Majors. *The Asian EFL Journal Quarterly*, *12*(1), 47–87.

Hsu, W. (2014). The most frequent opaque formulaic sequences in English-medium college textbooks. *System*, *47*(4), 146–161. https://doi.org/10.1016/j.system.2014.10.001

Hunston, S. (2002). *Corpora in Applied Linguistics.* Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern Grammar: A corpus-driven approach to the lexical grammar of English.* John Benjamins Publishing.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quaterly*, *41*(2), 235–253. https://doi.org/10.1002/j.15457249.2007.tb00058.x

Jiang, F., & Hyland, K. (2017). Metadiscursive nouns: Interaction and cohesion in abstract moves. *English for Specific Purposes*, *46*, 1–14. https://doi.org/10.1016/j.esp.2016.11.001

Kucera, H., & Francis, W. (1967). *Computational analysis of present day American English.* Brown University Press.

Leung, C., Harris, R., & Rampton, B. (2004). Living with inelegance in qualitative research on. In B. Norton & N. Toohey (Eds.), *Critical pedagogies and language* (pp. 242–267). Cambridge University Press.

Longman Dictionary of Contemporary English. (2019). Retrieved December 28, 2019, from https://www.ldoceonline.com/

McCarthy, M., McCarten, J. & Sandiford, H. (2020). *Touchstone*. Cambridge University Press.

MacMillan Dictionary. (2019). Retrieved December 28, 2019, from https://www.macmillandictionary.com/

Meriam-Webster Dictionary. (2022). Retrieved August 21, 2022, from https://www.merriam-webster.com/

Nation, I.S.P. (2009). *Teaching ESL/EFL reading and writing*. Routledge.

Nation, I.S.P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I.S.P., & Wang, K. M. (1999). Graded readers and vocabulary. *Reading in a Foreign Language*, *12*(2), 355–380.

Nesi, H., & Haill, R. (2002). A study of dictionary use by international students at a British University. *International Journal of Lexicography*, *15*(4), 277–305. https://doi.org/10.1093/ijl/15.4.277

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Osment, C., & Graham, D. (2013). *CEEM: Corpus-driven Engineering English Materials*. n.p.

Oxford Languages. (2022). Retrieved August 20, 2022 from https://languages.oup.com/google-dictionary-en/

Palmer, H. (1933). *Second interim report on English collocations*. Kaitakusha.

Paquot, M. (2007). Towards a productively-oriented academic word list. In J. Walinski, K. Kredens, & S. Gozdz-Roszkowski (Eds.), *Practical applications in language and computers* (pp. 127–140). Peter Lang.

Schmitt, N., & Carter, R. (2004). Formulaic sequences: Acquisition, processing and use. In N. Schmitt (Ed.), *Formulaic sequences in action* (pp. 1–22). John Benjamins Publishing.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.

Sinclair, J. (1987). Grammar in the dictionary. In J. Sinclair (Ed.), *Looking up: An account of the COBUILD project* (pp. 104–115). Harper-Collins.

Sinclair, J. (1996). The search for units of meaning. *Textus, 9*(1), 75–106.

Tahara, N. (2020). Roles of metadiscursive nouns in L2 English writing. *International Journal of Languages, Literature and Linguistics*, *6*(2), 85–92.

McCarthy, M., McCarten, J., & Sandiford, H. (2020). *Touchstone*. Cambridge University Press.

Tzeng, M. Y. (1985). *Effect of chunking material as an aid to ESL*. Retrieved December 30, 2019, from https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=8903&context=rtd

Urquhart, S., & Weir, C. (1998). *Reading in a second language: Process, product and practice.* Longman.

Uzawa, K. (1996). Second language learners' processes of L1 writing, L2 writing, and translation from L1 into L2. *Journal of Second Language Writing*, *5*(3), 271–294. https://doi.org/10.1016/S1060-3743(96)90005-3

Watson-Todd, R. (2017). An opaque engineering word list: Which words should a teacher focus on? *English for Specific Purposes*, *45*, 31–39. https://doi.org/10.1016/j.esp.2016.08.003

West, M. (1953). *A general service list of English words*. Longman, Green and Co.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge University Press.

Appendices

Appendix A

Note: for the CEFR levels, some words were not found on the A1-C1 word lists and are marked as NF (not found). Additionally, the part of speech is at times different; for example, in Appendix A *present* (verb) and *present* (adj) are indicated as A1, but it is likely that the CEFR ranking is for the adjective form although it is possible it is the noun form.

Opaque Words from the AVL Core Academic List (1-100)

level	8	A2 noun	subject	60	A1 noun
use	13	A1 noun	material	62	A2
change	18	A1 noun	produce	63	A2 verb
table	19	A1	performance	68	B1
develop	27	A2	approach	71	B2 noun
suggest	28	A2	nature	78	A2
low	30	A2 adj	product	84	A1
practice	38	A1 noun	goal	86	A2
report	43	A2	note	88	A1 verb
figure	46	A2 noun	represent	89	B1
need	50	A1 noun	determine	95	B1
base	51	B1 verb	common	98	A1
image	59	A2 noun	subject	60	A1 noun

Appendix B

Opaque Words from the AVL Core Academic List (101-200)

present	102	A1 verb	associate	149	B2 verb
term	103	A2	address	153	A1 verb
movement	107	A2	benefit	156	A2
establish	114	B2	apply	159	A2
standard	121	B1	association	164	B2
argue	125	A2	status	169	B2
degree	126	A2	present	173	A1 adj
state	129	A2 verb	conduct	177	B2
act	139	A2	critical	178	B2
reflect	141	B1	principle	191	B2
recognize	142	A2			

Appendix C
Opaque Words from the AVL Core Academic List (201-300)

test	204	A1 verb	limit	253	B1 verb
mean	212	A1 noun	directly	255	B1
application	214	B1	vision	258	B2
potential	227	B2	influence	261	B1
following	230	A2	claim	272	B1
labor	238	B2 noun	perceive	298	B2
contribute	232	B2			
assume	233	B2			
view	248	A2 verb			

Appendix D
Opaque Words from the AVL Core Academic List (301-400)

increased	303	NF	above	372	A1
select	305	B2	volume	375	B2
conclude	315	B1	limited	377	B2
standard	326	B1	code	382	A2 noun
adopt	322	NF	waste	386	B1 noun
employ	328	A2	mechanism	391	B2
contact	331	B1 noun	discipline	395	B2 noun
account	335	B2 verb	construct	396	B2 verb
exchange	352	B1 noun			
objective	354	B2			
flow	367	B1 noun			

Appendix E
Opaque Words from the AVL Core Academic List (401-500)

depression	425	B2	gain	489	B2 noun
developing	429	NF	settlement	491	C1
recognition	431	B2	index	499	B2 noun
resolution	446	B2			
display	452	B2 verb			
initiative	462	B2			
regard	464	B2			
testing	474	B2 noun			
passage	479	B2			
introduction	483	A2			

Appendix F
Opaque Words from the AVL Core Academic List (501-600)

relative	506	B1
shift	508	B1
joint	519	B2
resolve	531	B2
establishment	546	C1
given	554	NF
reflection	563	C1
encounter	576	B2
utility	580	C1
function	589	B1

Appendix G
Additional Data on Opaque Words: 1-10

This data has been compiled with information from COCA (https://www.english-corpora.org/coca/). Only the information for the first ten words has been presented due to space limitations. For the full data, please contact the researcher. For some senses, information from dictionaries (online versions) was included as well. At some points, such as which senses to include and colligations, subjective decisions were made, and another researcher might render slightly different results.

Word: level; **AVL Ranking:** 8; **CEFR:** A2

Lexical Bundles:	level of interest in, levels of physical activity, level of support for, to the next level, at the local level, at the national level
Collocations:	correlate, study, statistically, finding, risk, researcher, sample, e.g., disease, blood high/higher/highest + level/levels level/levels in level/levels + for/with/at
Related words:	high-level, entry-level, low-level
Senses:	position on a scale of intensity; relative degree or position of value in a graded group; specific identifiable position in a continuum or series or especially a process; floor in building; part of computer game or other game; for checking if flat; particular height

Word: use; **AVL Ranking:** 13; **CEFR:** A1

Additional Information:	Noun (55-60% of occurrences in COCA entries) Verb (40-45% of occurrences in COCA entries)
Lexical Bundles:	Verb: used in this study
Collocations:	**Noun:** file, user, form, used, following, page **Verb:** create, click, select use of (noun) used in/used it/used for/used by (verb)
Colligations:	passive voice (verb) The verb is often employed as in the passive voice for describing how one aspect of a system or process
Related words:	user, useful, used, usage, usual, useless, reuse, unused reusable
Senses:	**Noun:** put into service; take or consume regularly/habitually, especially with some form of drug; use up, consume fully **Verb:** do something with that tool, by means of that method etc., for a particular purpose; to take an amount of something from a supply of food, gas, money etc.; to take advantage of a situation; to say or write a particular word or phrase

Word: change; **AVL Ranking:** 18; **CEFR:** A1

Lexical Bundles:	Most lexical bundles occur with the verb form, not the noun form, and are transparent
Collocations:	**Verb:** mind, subject, behavior, course, climate, attitude, clothes, page, file, option, setting, tab, user, default, color, image, heart, name
Related words:	exchange, changing, unchanged
Senses:	**Noun:** situation in which something becomes different or you make something different, especially regarding thoughts, actions and behaviors; situation in which one person or thing is replaced by another; new activity or experience that is different and enjoyable **Verb:** process by which things become different

Word: table; **AVL Ranking:** 19; **CEFR:** A1
(likely refers to the object, not the academic use)

Lexical Bundles:	table of contents, table on page (number would follow) **Note:** most lexical bundles found related to the everyday, furniture/ restaurant sense.
Collocations (noun):	see, in, preformatted, statistically, score, summarize, variance, sample, category, variable
Related words:	tablet, timetable
Senses:	**Noun:** set of data arranged in rows and columns (academic); piece of furniture (fiction, magazines); table for people to eat at in a restaurant **Verb:** present formally for discussion or consideration at a meeting

Word: develop; **AVL Ranking:** 27; **CEFR:** A2

Lexical Bundles:	in order to develop, more likely to develop, develop an interest in
Collocations:	researcher, technology, skill, disease, development, learning, student, scientist, research, curriculum, developed, program, plan, technology, relationship
Colligations:	present passive voice- has been developed/have been developed
Related words:	development, developer, developing, developed, developmental
Senses:	make something new such as a product, a mental or artistic creation; for a skill or ability-it becomes stronger or more advanced; for a disease or illness-you start to have it; for a problem or difficult situation- it begins to happen or exist, or it gets worse; to begin to have a physical or other type of fault; to make an argument or idea clearer, by studying it more or by speaking or writing about it in more detail; to use land for the things that people need, for example by taking minerals out of it or by building on it; to make a photograph out of a photographic film, using chemicals

Word: suggest; **AVL Ranking:** 28; **CEFR:** A2

Lexical Bundles:	suggests the need for, suggest that we should, suggests that we may, it has been suggested
Collocations:	researcher, study, finding, disease, research, likely, scientist, risk, evidence, evolutionary
Colligations:	used with modals to create a hedge; exophoric reference to other peoples' ideas or comments, past tense, present perfect passive voice
Related words:	suggestion, suggested, suggestive
Senses:	make a proposal; declare a plan; refer to another person's plan, idea or action; drop a hint, i.e., state something in an indirect way; imply as a possibility; to tell someone your ideas about what they should do or what action should be taken; to make someone think that a particular thing is true

Word: low; **AVL Ranking:** 30; **CEFR:** A2

Lexical Bundles:	one of the lowest, low interest rate/s, at an all-time low
Collocations:	rate, price, upper, market, risk, level, correlate, percent, income, score, average relatively, significantly, extremely, slightly, historically, substantially, artificially reduce, associate, score, rate, level, price, cost, income
Colligations:	often is preceded by a modifying adverb
Related words:	lower, low income, low-cost, low-level, lowered, lowland **Note:** there are numerous words that employ 'low' usually in a two word hyphenated form such as 'low-cost'
Senses:	less than normal in degree, amount or intensity; bad, or below an acceptable or usual level or quality; not in a high position socially; less than desired acceptance or support; having a top that is not far above the ground; relating to a supply of something - there is not much of it remaining; lighting that is not bright; unhappy and without much hope for the future

Word: practice; **AVL Ranking:** 38; **CEFR:** A1

Lexical Bundles:	theory and practice, practice of law, practice what you preach (idiom), practice/s as well as, put into practice
Collocations (noun):	practitioner, coach, e.g., patient, systematic, learning student, player, training, teacher, classroom, learner, teaching, clinical the preposition 'in' frequently occurs before and after the word practice: in practice/ practice in practice in + period of time (century, lesson, etc.), place, field of work/study, approach in practice – indicates what really happens rather than what should happen or what people think happens; also relates to the actual application of a method, idea, plan, best, private, common, business, clinical
Genres:	often used in social work, education, health and medical writing
Related words:	malpractice, practicing
Senses:	customary way or behavior; systematic training through multiple repetitions; work of a profession (often legal, medical or teaching); knowledge of how something is customarily done; good example of how something should be done (e.g. it is good/best practice to)

Word: report; **AVL Ranking: 43**; **CEFR:** A2

Semantic Preferences:	report on (indicates a topic), report from (indicates information from another source or agency report by (indicates information another source or agency report to (indicates a person or agency that must be informed or a person officially informing)
Lexical Bundles:	contributed to this report, according to this report, according to a report, according to the report, report by/of the national
Collocations:	investigation, agency, committee, official, police, department, release, data, attorney, oversight, incident
Colligations:	used to refer to information, data, findings of another; often used as hedge as the information may not have been confirmed or verified yet
Genres:	often used in relation to news, political events or official announcements often used in business
Related words:	reporting, reporter, reportedly, reported, unreported

Word: report; **AVL Ranking: 43**; **CEFR:** A2 **(Cont.)**

Senses:	**Noun:** written document describing the findings of some group or person; short account of the news; an act of informing verbally; information that something has happened, but has not been verified yet; written evaluation of performance **Verb:** give a spoken or written account of something that one has observed, heard, done, or investigated; present oneself formally as having arrived at a particular place or as ready to do something

Word: figure; **AVL Ranking: 46**; **CEFR:** A2

Lexical Bundles:	shown in figure + number, figure of speech, in the figure, facts and figures, illustrated in figure, presented in figure
Collocations (noun):	graph, table, diagram, numbers, below, show, illustrate, skate, public
Genres:	Although used as part of a phrasal verb informally in conversation (figure out, figure on), it occurs primarily in academic work, referring to diagrams of some sort. Its greater frequency in academic work may be an artifact of COCA's input. There is not a great deal of conversational English in COCA. The spoken English is mostly interviews and more formal forms.
Related words:	configuration, fig., figurative, figurehead
Senses:	**Noun:** diagram of picture illustrating data or other material under discussion; number representing an amount, especially an official number; number from 0 to 9, written as a character rather than a word; someone who is important or famous in some way; shape of a person's body, especially a woman; a pattern or movement in skating; someone with a particular type of appearance or character, especially when they are far away or difficult to see **Verb:** be a significant and noticeable part of something; think, consider, or expect to be the case.