International Journal of Lexicography, 2021, Vol. 34, No. 2, 206–234
doi: 10.1093/ijl/ecaa022
Advance Access Publication Date: 28 November 2020
Article

Article

SLIPPING THROUGH THE CRACKS IN E-LEXICOGRAPHY

Ana Frankenberg-Garcia*
University of Surrey (a.frankenberg-garcia@surrey.ac.uk)

Geraint Paul Rees
University of Surrey (geraintrees@gmail.com)

Robert Lew
Adam Mickiewicz University (rlew@amu.edu.pl)

*Corresponding Author: Ana Frankenberg-Garcia, University of Surrey (a.frankenberg-garcia@surrey.ac.uk)
Slipping Through the Cracks in e-Lexicography

Abstract

Despite the remarkable advances made in recent years to facilitate the lexicographer's work of interpreting and synthesizing the complexity of language uncovered by corpora, an uncritical use of cutting-edge corpus tools and resources can instill a false sense of assurance. In this paper, authentic examples pertaining to wordlist use, collocation research and example selection that arose when compiling a real-world lexical database are discussed through the lens of problems that can easily slip through the cracks in e-lexicography. In doing so, we emphasize the importance of solid training and sound lexicographic judgment when using corpora, corpus tools and corpus-derived resources, and provide an opportunity to reflect on how e-lexicography can be further refined in the future.

Key words: e-lexicography, wordlists, collocation, corpus examples, academic writing, English, writing assistant

1. Introduction

The last decades have seen truly remarkable advances in lexicography, thanks to the 'corpus revolution' which underpins present-day e-lexicography (Rundell and Stock 1992, Hanks 2012). Corpora and corpus tools allow for much better descriptions of how words are used than those that were possible in the times of pre-corpus dictionaries. Thanks to monitor corpora, it has become simpler to update headword lists systematically, by

detecting new words and new senses, and acknowledging words and senses that have fallen into disuse so as to label them accordingly. Thanks to word-frequency statistics from corpora, lexicographers can better assess what might be considered core vocabulary to be included in dictionaries for language learners, and what rarely used words would be unhelpful to employ in controlled defining vocabularies. Concordances have in turn been key to assisting lexicographers in teasing out more and better information about word meaning, grammar, phraseology and situations of use.

Although corpora with millions of words may have initially represented an overload of language data for dictionary editors to process, the natural language processing tools at the lexicographer's disposal today have evolved a great deal since the publication of COBUILD, the original corpus-based dictionary (Sinclair 1987a). With lemmatization, part-of-speech tagging, statistical measures of word association, and automated summaries of how words are used in context, it has become far more manageable for lexicographers to inspect huge amounts of language data so as to enhance definitions, synthesise syntactic patterning, give more comprehensive coverage to phraseology and lexical collocation, and label words according to specific situations of language use. Even the highly skilled work of selecting suitable illustrative examples to present to the end user has been facilitated by e-lexicography tools. These developments have become all the more important since 21st century corpora tend to be much larger, generating substantially more language data for lexicographers to process than COBUILD's original 7-million-word corpus, which was a breakthrough back in the eighties (Sinclair 1987b).

Yet, as so eloquently reasoned by Rundell (2002:138), 'good-old fashioned lexicography' is still necessary. When Michael Rundell wrote his paper on 'Human Judgement and the Limits of Automation' in lexicography nearly twenty years ago, he argued that instead of entailing more straightforward editorial work, access to large corpora of texts required lexicographers to become more skilled in interpreting the complexity uncovered by corpora and in synthesising information that is relevant to dictionary users. He illustrated this with his usual flair for highly perceptive insights, such as drawing attention to the passivity of the prepositional phrase *in front of* when examining concordances like *slumping in front of the TV* when compared with the more active role of the preposition *at* in concordances like *sitting at the computer* (p.144). Another excellent example of nuances of meaning discoverable through skilled corpus-based lexicography is noting how the adjective *old-fashioned* can signify not just outdatedness in a negative way, as in *very old-fashioned, dyed-in-the-wool ideas*, but also in a neutral way, as in *an old-fashioned barber chair*, and even in a positive way, as in *discreet, old-fashioned comfort* (p. 151).

It is not just the inherent complexities of language and the sheer volume of data that make skilled human editors essential, however. In this paper, we argue that 'the need for skilled human editors with a good grounding in relevant linguistic disciplines *and* highly developed intuitions about language' (Rundell 2002:139) is even greater, as it is only too easy to be misled by a false sense of assurance provided by our dependence on empirical language data and cutting-edge lexicography software.

To illustrate this, we present here a series of issues encountered during the development of a real-world lexicographic project: the lexical database behind ColloCaid. ColloCaid is a text editor that provides real-time collocation suggestions to support academic English writers (Frankenberg-Garcia et al. 2019a). While it would not have been possible to compile the lexical database for ColloCaid without e-lexicography applications, this reflective

commentary on a practical project draws attention to the importance of solid training when using corpora, corpus tools and corpus-derived resources, and provides an opportunity to reflect on how e-lexicography can be further refined in the future.

In the remaining parts of this paper, we begin by providing a brief introduction to ColloCaid and its general guiding principles. This is followed by a critical appraisal of our frequency-based approach to collocation coverage and of our use of corpora and sophisticated lexical analysis software to research collocations and select illustrative examples of collocations in context. The paper ends with a discussion of the ever-present need for 'good old-fashioned lexicography', and of how an analysis of issues that require manual curation can contribute not only to the education and training of dictionary editors, but also to furthering the development of e-lexicography.

2. An overview of ColloCaid

In line with recent approaches to 'bringing lexicographic information to writers instead of waiting for them to get the information they need from dictionaries' (Frankenberg-Garcia 2020:30), ColloCaid is part of a growing number of writing assistants under development (Strobl et al. 2019). Unlike more generic writing aids, ColloCaid focuses on providing users with academic English collocation suggestions (Frankenberg-Garcia et al. 2019a). Through the ColloCaid editor, writers can consult words that typically combine with other words in written academic English, like *offer a solution*, *successfully applied*, *highly effective*, and so on. Taking a data-driven-learning approach (Johns 1991), where collocations are shown rather than explained, ColloCaid aims to help writers less familiar with general academic English collocations – such as undergraduate students and researchers with less experience of written academic English – discover suitable collocations for themselves, so as to expand their academic collocation repertoire, reassess possible misconceptions about collocation saliency, and even offset collocation avoidance (Frankenberg-Garcia 2018, Tavares-Pinto et al. forthcoming). Alternatively, ColloCaid can work simply as an aide-memoire for academic writers who cannot remember the best word to use in a given syntagmatic slot, or as a revision tool to refine the vocabulary of initial drafts.

Another characteristic of ColloCaid is its emphasis on human-computer interaction in the way collocation suggestions are integrated into the writing environment (Frankenberg-Garcia et al. 2019b), and its multiple-views approach to how collocations are visualised, so that users can decide how collocation suggestions are presented to them (Roberts et al. 2020). Interested readers are more than welcome to trial a proof-of-concept prototype of our writing assistant.[1] The focus of this paper is however on issues encountered during the compilation of the lexical database behind ColloCaid, whose guiding principles are summarized below.

At a very early stage in the development of ColloCaid, a decision was made not to start the lexical database from scratch, but rather to build on existing resources. Substantial work had already been undertaken in the creation of excellent lexicographic material such as corpora and wordlists for academic English, and it would be ill-advised to ignore it.

Another key premise of our work was to focus on collocation nodes (or bases) as a starting point for collocation look-ups. When writers using the ColloCaid text editor type in a word like *system*, they will be shown that collocation suggestions are available for it and, if followed up, they will be given options like *design/establish/create a system*, and many

more. The idea is to help writers find answers to questions like 'What verb can I use with [the node] *system*?' or 'Is there a better way of saying *make a system*?'

When it came to deciding which collocation suggestions to offer, in the original development of the tool we chose to present collocations that figure in multiple disciplines so as to help a wider range of users. However, in future it should be possible to customize ColloCaid so as to provide discipline-specific coverage,

Finally, apart from collocation suggestions, we wanted to provide examples of collocations being used in authentic academic English texts. ColloCaid examples serve not only to attest usage in written academic English, but also and most importantly in a writing assistant, to enhance data-driven learning by illustrating how a collocation suggestion can be effectively integrated into a sentence. The last element of the ColloCaid lexical database therefore consists of carefully curated examples that were selected to help writers gain confidence in using collocations that they may otherwise not dare to use.

At the time of writing this paper, the ColloCaid lexical database is practically complete, with 572 academic English collocation nodes, which serve as prompts to 32,655 academic English collocation suggestions. Additionally, for the top 9,685 collocations, we have prepared a total of 29,055 contextualized examples of collocations in use (Figure 1). While it would not have been possible to achieve such coverage without the support of cutting-edge e-lexicography tools and resources, human curation was essential at all stages of development of the ColloCaid lexical database.

In the next three sections of this paper, we provide further methodological details about the compilation of the database and discuss e-lexicography issues regarding wordlists, collocation research and example selection.

3. Wordlist issues

This section explains how we employed wordlists to select the academic English nodes used as access points to the collocation suggestions offered in ColloCaid, and discusses issues we had to address in the selection.

The first step in this process was to determine how many collocation nodes to tackle. Given the time and resources available, the coverage of around 500 nodes represented an achievable ambition. Although a set of 500 nodes does not give the impression of much, as

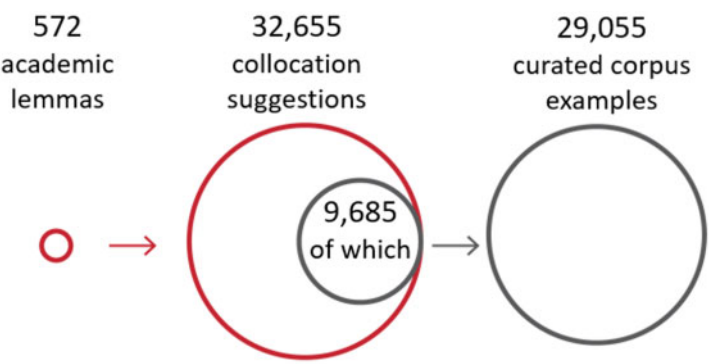

Figure 1. ColloCaid lexical database (v.06 June 2020).

seen in the previous section, the 572 nodes in ColloCaid v.06 give rise to over 32K academic collocation suggestions. This is nearly 30K more collocations than those available in the Academic Collocations List (Ackermann and Chen 2013).

Of course, quantity does not necessarily mean quality. It was important to ensure that the 500 nodes initially selected were maximally useful to the target users of ColloCaid. An immediately obvious criterion was to eliminate grammatical words like prepositions, articles and conjunctions from our list of nodes, as it is highly unlikely that writers would ask themselves questions like 'What words combine with *of* or *the* or *but*?'. For a similar reason, adverbs were excluded, since to arrive at collocations with adverbs, writers are more likely to think first of the verbs or adjectives that go with them. For example, to arrive at *increase dramatically*, writers who are unable to recall this collocation as a chunk are more likely to use the verb *increase* as a starting point for what they intend to say than to start their thoughts from the adverb *dramatically*. This is because in this case the verb conveys the central propositional meaning of the collocation. Likewise, to arrive at *highly specific*, the adjective *specific* would normally come first. We focussed thus on selecting noun, verb and adjective nodes, whose meaning is more central to collocations than that of grammatical words and adverbs.

Having defined which types of nodes to cover, precedence was given to high-frequency academic nouns, verbs, and adjectives. Because of the Zipfian distribution of natural language vocabulary (Nation and Waring 1997), words that rank higher in frequency offer more coverage of the language. Therefore, high-frequency academic words provide a better coverage of academic vocabulary.

With this priority in mind, the nouns, verbs, and adjectives from three well-established general academic English wordlists were taken as a platform from which to compile our initial node list: the Academic Keyword List (Paquot 2010), the Academic Vocabulary List (Gardner and Davies 2014), and the Academic Collocations List (Ackermann and Chen 2013).

The Academic Keyword List (AKL) was compiled by extracting keywords from a focus corpus combining published academic texts and student academic writing from various sources, using a reference corpus of fiction for contrastive purposes. The original AKL contains a total of 930 lemmas and phrases such as *according to*, *given that*, and so on. Of these, the 353 nouns, 233 verbs, and 180 adjectives listed (766 lemmas in all) were considered for inclusion in ColloCaid.

The Academic Vocabulary List (AVL) consists of 3,000 lemmas that occur across a range of academic disciplines in the 120-million-word academic sub-corpus of the Corpus of Contemporary American English (COCA) (Davies 2008). Lemmas whose frequency was 50% or more greater in the non-academic portion of COCA were excluded so as to filter out general language. Because AVL contains six times more lemmas than the 500 we aimed to cover, we decided to use a subset of 172 AVL nouns, 129 verbs, and 86 adjectives (387 lemmas in all) which Durrant (2016) found frequently and widely in the British Academic Written English (BAWE) corpus of UK university student writing (Heuboeck, Holmes & Nesi, n.d.). This was important, because there would be no point in offering collocations suggestions for nodes that only experts use. This selection is henceforth referred to as AVL-BAWE.

The Academic Collocations List (ACL) is a little different because, as the name suggests, it consists of collocations rather than individual lemmas. Its 2,469 cross-disciplinary

collocations were extracted from the 25-million-word written component of the Pearson International Corpus of Academic English (PICAE) (Ackermann et al. 2011) and then vetted by a panel of EAP experts. To incorporate this list into our node selection, we used the 704 headwords in the ACL appendix of the Longman Collocations Dictionary (Mayor 2013). Of these, 95 are verbs, 83 are adjectives and 526 (74.7%) are nouns, which is not surprising, as noun nodes normally generate more collocations than verbs and adjectives.

Together, the above three lists feature 1,139 different noun, verb and adjective lemmas, which was more than double our target coverage of 500 nodes. What we did next, therefore, was to cross-reference the three lists so as to focus on lemmas present in more than one list (Figure 2). In this way, we aimed to build on their collective strengths. The extraction method of the AKL emphasizes academic keyness, i.e., words that are particularly salient in academic English. The AVL-BAWE subset focuses on academic keywords which are prominent in novice academic writing. The ACL, in turn, prioritizes nodes that generate strong academic collocations.

The resulting 514 lemmas at the intersection of at least two of the three lists was conveniently very close to our target. These are itemized in Frankenberg-Garcia et al. (2019a)[2], and were selected as the basis for the compilation of the ColloCaid lexical database. What we want to draw attention to here, however, is how surprisingly little overlap there is between the three lists.

Considering that all three sources aim to represent general academic English vocabulary, Figure 2 shows that they are far more heterogeneous than might be expected. Only 187 of a total of 1139 different lemmas considered (16.4 %) are common to all three lists. This is a sobering confirmation that academic wordlists can vary quite significantly, depending on the corpora and extraction methods used.

AVL-BAWE was the list that overlapped the most with the other two, with 91.5% of its lemmas attested in at least one of the others. This was not surprising, as AVL-BAWE is not

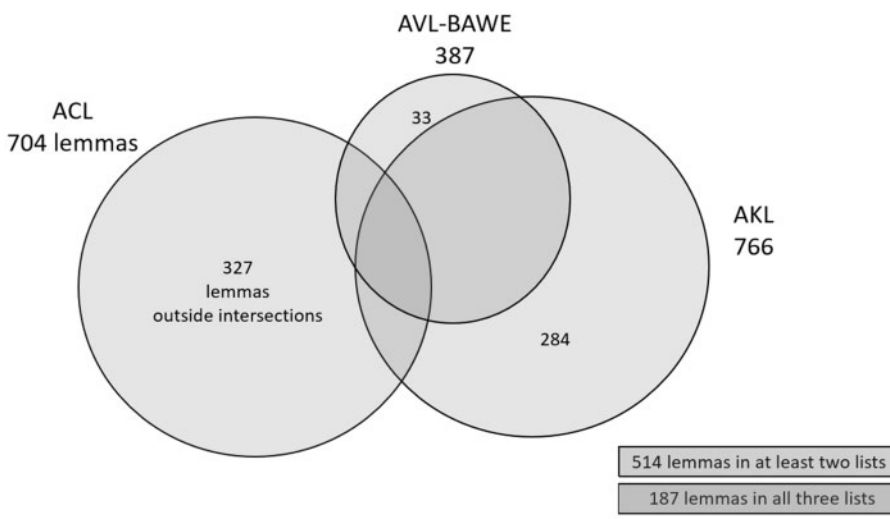

Figure 2. Noun, verb, and adjective lemmas overlapping in the Academic Collocations List (ACL) headwords, the BAWE subset of the Academic Vocabulary List (AVL-BAWE) and the Academic Keyword list (AKL) (not to scale).

only the smallest of the three, but also prioritizes lemmas used by novice academic writers.[3] Next came AKL, with 62.9% of its lemmas common to one or both other lists. The most distinct wordlist was ACL, with only 53.6% of its lemmas matching those of other lists. This again was unsurprising, as unlike the two other lists, its extraction was based on collocations rather than individual lemmas, and it did not filter out lemmas widely used in general language.

A closer examination of the lexical items left out of list intersections can provide further insights into how this affected the initial ColloCaid node selection. Appendix 1 details which nouns, verbs, and adjectives were present in only one list. While it is not our aim here to comment on the reasons why individual lemmas were part of one list but not of others, it is useful to reflect upon the relative strengths and weaknesses of our wordlist intersection approach.

By focusing on what at least two of the lists had in common, we were successfully able to filter out:

a. Academic lemmas that are not strong node candidates because they are not collocationally productive, particularly non-gradable adjectives, which do not collocate with adverbs. For example, AKL lists *absolute*, *main*, *male*, *other, parallel, so-called*, and so on. Likewise, AVL-BAWE, includes *above*, *economic, given, numerous, varying*, and so on.[4]
b. Subject-specific academic lemmas from ACL that are part of compound terms like *intercourse* (sexual), *pollution* (environmental), *species* (human), *organ* (internal) *election* (presidential), and so on, which we had deliberately intended to leave out.
c. The noun *university* from AVL-BAWE, which occurs mainly in the context of copyright notices and reference lists (e.g. *Oxford University Press*), and is not relevant to the type of collocation suggestions we had planned to provide.

However, by focussing on wordlist intersections, we left out many lemmas that generate strong general academic English collocations, such as:

d. In ACL only: *way* (e.g. alternative, appropriate, meaningful...); *know* (e.g. commonly, generally, widely...); *impossible* (e.g. virtually, practically, logically...), and so on.
e. In AKL only: *fact* (e.g. ignore, reflect, highlight...); *fail* (e.g. ultimately, consistently, largely...); *favourable* (e.g. tentatively, generally, moderately, highly...), and so on.
f. In AVL-BAWE only: *table* (e.g. summarized, listed, shown...); *observe* (e.g. directly, frequently, empirically...), *modified* (e.g. slightly, substantially, suitably...), and so on.

Although it would not have been possible to accommodate more than around 500 nodes generating strong general academic English collocations within the scope of this project, examining the collocational behaviour of lemmas such as those exemplified in (d) to (f) can be useful should there be further means and resources available to expand the already considerably large database of collocation suggestions provided by ColloCaid.

In addition to node candidates left out from our initial selection because they figured in just one of the lists inspected, we noted that a few lemmas which seemed to be very central to written academic English—like the nouns *paper* (e.g. explore, report, examine...), *step* (e.g. take, involve, follow, outline...), and *field* (e.g. dominate, advance, revolutionize...)—were conspicuous by their absence from all three academic wordlists used in ColloCaid.

To elucidate this matter further, we inspected the frequencies of the above three nouns and of the lemmas outside list intersections exemplified in (d) to (f) in the following three academic and three general English corpora that were readily available to our team[5]:

Academic

- PICAE: written section of the Pearson International Corpora of Academic English (Ackermann et al. 2011), about 25 million tokens
- OCAE: Oxford Corpus of Academic English (Lea 2014), about 84 million tokens
- BAWE: British Academic Written English corpus of UK student university writing (Heuboeck et al. n.d.), about eight million tokens

General

- BNC – British National Corpus (Sketch Engine version tagged by CLAWS, about 112 million tokens)
- COCA – Corpora of Contemporary American English (Davies 2008), about one billion tokens
- English Web 15 – Sketch Engine web documents corpus created in 2015 (Jakubíček et al. 2013), 18.4 billion tokens

Apart from diverging in size, it must be noted that the academic texts behind the three academic corpora differ quite substantially, as do the various texts and text types that comprise the three general English corpora inspected. Our intention here was not to argue that one corpus is better than another, nor to carry out a comprehensive statistical comparison of corpus frequencies across six different corpora, but simply to raise awareness of wordlist issues affecting lexicographic work.

The results of the analysis are presented in Table 1. The figures pertaining to the corpora with the three highest frequencies of each lemma are underlined to facilitate the comparison. It can be seen that, on the one hand, the verb *know* is more frequent in all three general English corpora than in any of the academic English corpora, and that, conversely, *fact*, *favo(u)rable* and *observe* are consistently more frequent in the three academic corpora. On the other hand, however, for the lemmas *way*, *impossible*, *fail*, *table*, *modified*, *paper*, *step*, and *field*, frequency distinctions between the academic and general English corpora in Table 1 are less clear-cut. The first lesson to be learnt from the figures in Table 1 is therefore that word frequencies can vary substantially not only across different types of corpora, but also across corpora aiming to be representative of roughly similar types of language. No matter how helpful wordlists are to a lexicographic project, it is imperative not to lose sight of the fact that word frequencies can be greatly affected by corpus text selection.

Another result that stands out in Table 1 is that, in contrast to all other corpora, there are no hits for the adjective *modified* in OCAE. However, this does not mean that there are no occurrences of *modified* in OCAE. In fact, a simple search for the word returns 3608 concordances. We therefore wanted to know whether there were any adjectival uses of the word among those concordance lines. While concordances like *several items were modified or replaced* indicated that the word was being used as a verb, concordances like *a modified version of the method* reflected indeed a more adjectival use. The reason why there were no hits for the adjectival lemma is that the tagger used to annotate OCAE classifies the latter

Table 1. Selected normalized word frequencies (per million) across three academic and three general English corpora (top three underlined).

Wordlist	Lemma	Academic English			General English		
		PICAE written	OCAE	BAWE	COCA	BNC	Web 2015
ACL only	way (n)	**1034.3**	826.4	825.2	**1268.9**	**957.2**	819.2
	know (v)	591.7	570.0	457.3	**2781.0**	**1591.3**	**812.4**
	impossible (adj)	**61.4**	57.3	**75.3**	58.4	**60.9**	35.1
AKL only	fact (n)	**435.7**	**436.2**	**490.0**	402.4	372.2	239.7
	fail (v)	129.5	**145.3**	**154.0**	125.6	**141.5**	99.1
	favo(u)rable	**24.1**	**30.1**	**27.1**	10.2	13.1	11.4
AVL-BAWE only	table (n)	**305.1**	102.0	**374.9**	**240.6**	197.8	138.3
	observe (v)	**168.4**	**239.5**	**162.2**	66.7	65.6	66.5
	modified (adj)	**6.5**	0*	**22.6**	5.8	5.5	**8.1**
None of the above	paper(n)	**187.9**	151.3	**173.3**	166.2	150.7	**206.0**
	step (n)	**175.6**	**153.9**	142.6	152.5	119.7	**182.8**
	field (n)	**272.3**	**266.0**	216.5	191.8	182.1	**301.9**

cases as AVBB (past participle used attributively) instead of as adjective. A revised CQL (corpus query language) query for [word="modified" & tag="AVBB"] was therefore carried out, returning 1,325 hits, which is equivalent to a normalized frequency of 15.7 hits per million. This example serves to underscore the importance of investigating what is behind unusual corpus frequencies, which can easily slip through the cracks in e-lexicography.

A closer look at some of the other lemmas in Table 1 illustrate further the importance of sound lexicographic judgement when examining word frequency data. Take the verb *know*. The fact that it is consistently more frequent in the general English corpora than in the academic English corpora does not necessarily mean that it is less central to academic English and less relevant to a project like ColloCaid. Indeed, as exemplified in (d), the node generates strong collocations, and, as shown in Table 1, it is the second most frequent lemma in all three academic corpora examined. However, because of the ways in which AKL and AVL were compiled, excluding words that also rank very high in frequency in general language, important academic uses of the word ended up being demoted. It seems that contexts such as *I know*, *you know*, and *I don't know*, which are extremely frequent in general language, have the deleterious effect of downplaying the importance of contexts like *it is well/widely/generally/previously known* in academic English. From a pedagogical lexicography perspective, it is important to acknowledge that language users who are acquainted with the general English use of the lemma are not necessarily familiar with its academic collocations. Therefore, even though *know* is missing from two of the academic wordlists consulted, we would argue that it is a strong candidate for inclusion in a tool like ColloCaid.

The lemmas *table*, *paper*, *step*, and *field* tell a slightly different story. As shown in Table 1, in the same way as *know*, these lemmas are very frequent in both general and academic language corpora. What makes them different from the verb *know* is that while *know* is used roughly in the same sense in general and academic language, *table*, *paper*, *step*, and *field* are employed in a very specific sense in academic language, but in more

than one sense in general language. *Table*, for example, is prototypically a piece of furniture in general language, but in academic language it is more often a set of data displayed in rows and columns. As shown in Figures 3 and 4, which display the first five GDEX[6] concordances for *table* in an academic corpus (PICAE) and in a general English corpus (English Web 15), all concordances refer to the latter sense in the academic corpus, but only two concordances refer to this sense in the general language corpus (lines 1 and 4 in Figure 4). However, wordlists and keyword analyses do not normally make sense distinctions (Rees 2018). Therefore, essential academic senses of polysemous lemmas seem to be watered-down by other senses of those lemmas, and thus escape keyword frequency thresholds. From the viewpoint of pedagogical lexicography, a learner who is familiar with a noun like *table* in contexts such as *her mother set the table* may be less familiar with contexts like *the information summarized/presented/displayed in Table X*. The same applies to polysemous lemmas like *paper*, *step* and *field*, whose academic senses have slipped through the frequency thresholds of the academic wordlists consulted in ColloCaid. Thus, once again, thanks to good old-fashioned lexicographic judgment, we would argue that these lemmas are strong node candidates for academic collocation dictionaries and tools like ColloCaid.

So far, we have discussed lemmas that were left out of the initial selection of nodes for ColloCaid. Doing so drew our attention to the opposite, i.e., lemmas prominent in the academic wordlist intersections considered in ColloCaid that seemed by comparison to be much less central to general academic English than some of the lemmas left outside those intersections. For example, the nouns *application* and *code* (which are listed in both AVL-BAWE and AKL), did not seem as essential to general academic English as the nodes presented in Table 1.

Further investigation of collocations and concordances for those lemmas in academic corpora enabled us to realize that both *application* and *code* had more than one academic sense. *Application* was used in the sense of request (e.g. *a successful patent application*), software (e.g. *IT applications require significant innovative effort*), and use (e.g. *their*

1. This **table** provides relative grades for major stressful events.
2. The hidden side effects are explained in the following **table**.
3. **Table** 20.13 compares selected current access technologies.
4. **Table** 11.1 summarizes data on how faculty searched for information.
5. The fees are listed in the **table** (right).

Figure 3. First 5 GDEX concordances for *table* from an academic English corpus (PICAE).

1. A **table** explaining what these abbreviations mean is therefore essential.
2. Vendors must provide their own tent and **table**.
3. **Table** tennis is a pleasant and safe sport.
4. **Table** 5 summarizes the results from the neutral sugar analysis.
5. Her mother casually said her hello and set the **table**.

Figure 4. First 5 GDEX concordances for *table* from a general English corpus (English Web 15).

application to clinical practice). *Code*, in turn, was used to mean letters and numbers (e.g. *universal product codes*), rules (e.g. *following a moral code*), cipher (e.g. *to break wartime codes*), computer instruction (e.g. *to write a code that provides precise answers*), and labels (e.g. *we assigned codes to text passages*). This suggested that unlike an academic lemma like *table* in the sense of rows and columns, which is truly interdisciplinary, polysemous academic lemmas like *application* and *code* were multidisciplinary. Because corpus frequencies do not normally take sense distinctions into account, different discipline-specific senses like the ones above get lumped together under a single lemma and end up jumping the rank queue of academic wordlists. This is yet another reason why lexicographic judgment is key to interpreting corpus frequencies.

In the case of polysemous academic lemmas like *application* and *code*, we opted to keep the collocations behind some discipline-specific senses, but introduced a sense disambiguation step to allow users to filter out senses that were not relevant to them.

The final point we would like to discuss with regard to node selection arose not so much because of potential wordlist issues, but more because of what mattered to end users. The problem was that a few of the lemmas covered had homographs pertaining to a different word class that did not meet the selection criteria. For example, the noun *aim* lies at the intersection of the three wordlists we used, and was thus included in ColloCaid, but the verb *aim* was only listed in AKL, and was therefore not among the 514 nodes initially chosen. The same happened to several other word forms, like *aid* (v), *benefit* (n), *change* (n), *need* (n), and *objective* (n), which were included, while their homographs *aid* (n), *benefit* (v), *change* (v), *need* (v), and *objective* (adj) were not. This could give users the impression of inconsistent lexicographic coverage, as indeed happened in initial user-testing. Therefore, we decided to include homographic partner lemmas that lay outside academic wordlist intersections, provided they generated strong academic collocations.

In doing so, we stumbled upon further wordlist glitches. As shown in Table 2, neither *aid* (n) nor *need* (v) are part of any of the wordlists consulted, yet they are in fact more frequent in three different academic corpora than their homographs *aid* (v) and *need* (n), which figure at the intersection of the three academic wordlists we used. A possible reason for this could be the widespread use of the former in general language, which, as in the case of the previous discussion about the verb *know*, could result in important academic uses of *aid* (n) and *need* (v) not reaching academic wordlist thresholds. However, the ways which these lemmas are used in academic contexts together with their frequency in academic corpora more than justifies including them in a tool like ColloCaid. Their absence is particularly conspicuous if their less frequent homographs *aid* (v) and *need* (n) are covered.

It can also be seen in Table 2 that there is very little difference in frequency between *aim* (n) and *aim* (v), yet only the former was present in all three wordlists. Importantly, all homographs that were left out of our initial node selection in Table 2 were actually more frequent than *aid* (v), which had been in all three wordlists consulted. These rather surprising findings suggest that, no matter how useful wordlists are, they may not tell the whole story. Thus, yet again, the need for sound lexicographic judgement in the use of wordlists cannot be overemphasized.

To get from the original 514 nodes that overlapped in at least two lists to our current selection of 572 nodes, we removed less helpful nodes similar to those previously highlighted in (a)-(c), and added nodes like the ones discussed, which we felt had slipped through the cracks.

Table 2. Normalized frequencies (per million) for selected homographs across three academic English corpora.

Lemma	Wordlist	PICAE Written	OCAE	BAWE	Mean
aim (n)	AKL; ACL; AVL-BAWE	79.4	84.0	151.2	104.9
aim (v)	AKL only	69.8	88.6	132.2	96.9
aid (v)	AKL; AVL-BAWE	22.4	23.6	47.0	31.0
aid (n)	None	62.0	44.9	53.7	53.5
benefit (n)	AKL; ACL; AVL-BAWE	158.1	222.3	210.3	196.9
benefit (v)	AKL only	38.7	61.4	67.8	56.0
change (n)	AKL; ACL; AVL-BAWE	617.6	703.9	608.7	643.4
change (v)	ACL only	288.1	299.9	293.5	293.8
need (n)	AKL; ACL; AVL-BAWE	329.1	326.6	309.1	321.6
need (v)	None	565.8	489.1	620.3	558.4
objective (n)	ACL; AVL-BAWE	98.2	101.3	136.3	111.9
objective (adj)	None	37.2	43.4	41.8	40.8

4. Collocation research issues

In this section, we summarize how the lexical database of collocation suggestions given in ColloCaid was compiled, drawing attention to issues in the collocation research that required us to reflect critically about e-lexicography.

As explained in Section 2, we aimed to offer general academic English collocation suggestions applicable to more than one discipline, and the focus was on researching collocations for circa 500 noun, verb, and adjective lemmas used as nodes, i.e., as the word that carries the more central propositional meaning of the collocation. We sought the collocates that writers would typically look up for each node. As exemplified in Figure 5, in addition to lexical collocations, we included prepositional collocations (sometimes referred to as syntactic patterns), which many writers tend to find troublesome (Swan and Smith 2001, Bitchener et al. 2005).[7]

The collocation suggestions for the nodes selected were researched in corpora of expert academic English consisting of published textbooks and journal articles. The main corpora used to this end were OCAE and, to a lesser extent, the written component of PICAE.[8] Note that the first language (L1) of the authors represented in these expert corpora is not necessarily English, but then L1 English is no guarantee of expertise in written academic English (Kosem 2010, Frankenberg-Garcia 2018). However, both OCAE and PICAE can be considered expert corpora insofar as the texts that make them up are published documents that have been vetted for quality and professionally edited. Another important factor that led us to choose these corpora was their availability on Sketch Engine (Kilgarriff et al. 2014), for, as shall be explained further along, its word sketch functionality was essential to our research.

Different types of corpus queries can be used to research collocations like the ones in Figure 5. The least efficient one would be to run concordance queries for each node and then inspect the results manually. A better way would be to run standard collocation queries. Figure 6 displays a snapshot of the top ten results for a collocation query showing the distribution of lemmas three tokens to the left and to the right of the noun *research* in PICAE ranked by logDice score.[9] Lexicographers wishing to look up which verbs are used

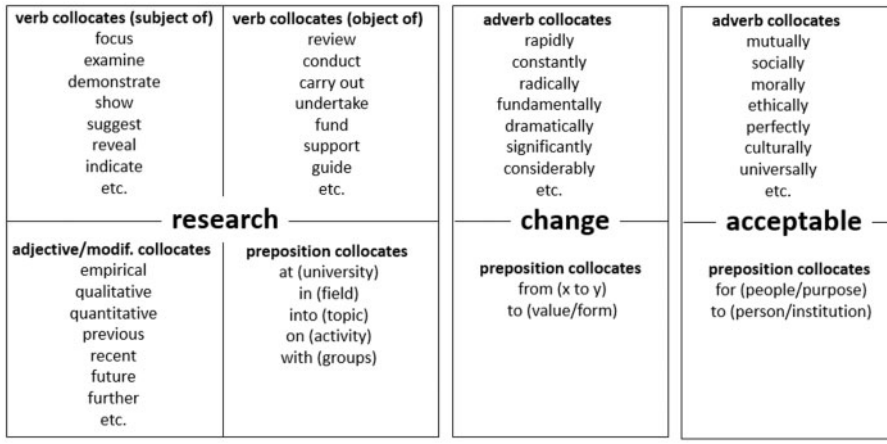

Figure 5. Examples of noun (*research*), verb (*change*), and adjective (*acceptable*) nodes and collocates.

		Lemma	Cooccurrences ?	Candidates ?	↓ LogDice
1	☐	qualitative	278	1,114	9.20
2	☐	method	360	9,481	8.95
3	☐	conduct	244	3,066	8.84
4	☐	quantitative	165	1,097	8.45
5	☐	project	203	4,983	8.43
6	☐	focus	230	7,899	8.40
7	☐	question	306	15,809	8.38
8	☐	your	359	21,131	8.37
9	☐	finding	168	2,371	8.36
10	☐	into	423	28,870	8.32

Figure 6. Top ten collocates of the noun *research* in PICAE.

with the noun *research* as object of the clause would have to scroll down that list and then consult concordances to identify which verbal lemmas listed are used with *research* as object. This would require lexicographers to manually distinguish between verbs and nouns (e.g. *research focuses* vs *the focus of the research*), as well as between verbs used with *research* as object and as subject of the clause (e.g. *the research focuses on something* vs *they focus on research*). It could take quite a while.

In ColloCaid, collocation research was greatly facilitated by Sketch Engine's word sketch functionality (Kilgarriff et al. 2004). Word sketches are of invaluable help to lexicographers, as they automatically summarize the lexical profile of a node according to grammar relation. This is illustrated in Figure 7, which shows a partial word sketch from PICAE highlighting the top ten verbs that collocate with the noun *research* as the object of the clause.[10] As can be seen, the adjective, noun, preposition, and pronoun collocates from Figure 6 have been filtered out, as have verbs like *focus*, which are more commonly used with *research* as subject, leaving only verbs used with *research* as object. Word sketches enable lexicographers to immediately visualise which collocations are relevant for a given node-collocate relation (e.g. verbs used with the noun node *research* as object).

Whereas there is no doubt that word sketches make compiling a collocation database like that of ColloCaid much more efficient, it is essential to bear in mind that their results hinge on a number of factors operating silently behind the scenes. First, as in any corpus query, the output of a word sketch depends on the quality of the corpus. If the corpus used is not representative of the language or type of language sought, the word sketch results can be skewed and misleading. Second, as in any corpus query that depends on part-of-speech annotation to differentiate between nouns, verbs and so on, the output of a word sketch is contingent on the accuracy of the tagger used to classify the words in the corpus. Third, as in any collocation query, the statistics used to rank collocates in terms of word-association strength are important. The logDice score used in word sketches is considered to be particularly suitable for lexical collocations when compared with other measures of word association such as the T-score and MI (Gablasova et al. 2017, Frankenberg-Garcia 2018). Last but not least, the output of a word sketch depends on the sketch grammar behind it, i.e., the set of CQL rules used to capture the grammatical relations between words in a sentence and generate the word sketch. For example, the CQL rule to detect which verbs are used with the noun *research* as object of the clause needs to capture not just contiguous collocations like *conduct research*, but also filter possible determiners, possessives, adjectives, adverbs, and modifiers in between. This would identify strings like *conducted their*

object_of			
conduct	106	10.14	•••
undertake	55	8.9	•••
fund	14	7.78	•••
sponsor	8	7.36	•••
publish	22	7.35	•••
summarize	7	7.08	•••
cite	11	7.02	•••
stimulate	8	6.89	•••
review	9	6.64	•••
support	29	6.62	•••

Figure 7. Partial word sketch showing top ten verbs used with *research* as object in PICAE.

research, *conducting some research*, *conduct initial research*, *conducting entirely new research*, *conducted field research*, and so on. Then the rule needs to be further defined to capture passive constructions such as *research conducted by the Institute*.

When compiling the database of collocation suggestions for ColloCaid, we noticed that a few collocations that were intuitively evident to our team were not being adequately identified in word sketches. For example, even though *carry out research* seems to be a perfectly valid general academic English collocation, the verb *carry out* was not listed as one of the verbs used as object of *research* in PICAE. The problem was not the contents of the corpus nor its tagging. A concordance query for *carry out* with *research* in the context of three tokens to the left and to the right returned 160 hits, and most lines indeed exhibited the collocation *carry out research*, as exemplified in Figure 8, where only the first and eighth concordance lines of the ten shown are false positives. Additionally, *carry out* was correctly identified as a verb followed by a particle by the tagger used in PICAE.[11] It was therefore surprising that *carry out* was not displayed in the word sketch in Figure 7, where far less frequent collocations like *summarize research* and *sponsor research* are listed. The absence of *carry out* serves to show that although word sketches are extremely helpful, they will not always be perfect. In this case, the problem is that sketch grammars are not sensitive to multi-word units, and the sketch-grammar rule was confounded by the particle *out* in the phrasal verb *carry out*.

Further examples of collocations involving phrasal verbs that were conspicuous by their absence from word sketches included *set up an experiment*, *set out criteria*, and *rule out the possibility*. While in computational linguistics it is widely acknowledged that multi-word units such as these can pose problems in tokenisation and further along the natural-language processing pipeline (Michelbacher 2013), we would like to draw attention to them because they can be easily missed by the end users of e-lexicography tools and slip through the cracks in practical lexicography projects.

Another problem affecting collocation research is polysemy. In the same way as polysemy distorts rank in academic wordlists, it also skews word association ratings. When

Industry Canada, in conjunction with others, is	carrying out	. </s><s> The research program seeks to identi
location called Koobi Fora: </s> <s> research	carried out	between 1967 and 1975 very rich collection of fo
the Awash Valley (Ethiopia): </s> <s> research	carried out	between 1972-1976 in 1973, discovery of most
hominids From 1976 to 1979, MARY LEAKEY	carries out	research at site of Laetoli, in Tanzania: </s><s>
<s> project based experimental research work	carried out	by third-year students as part of their degree prc
non Schaffer. </s><s> Part of the research was	carried out	with financial support from the British Council, S
><s> A team of graduate research assistants is	carrying out	basic research for the project in libraries and els
Selection of suitable research participants was	carried out	using purposive sampling (Shaughnessy & Zech
ide their facilities to allowfaculty now in place to	carry out	their research . </s><s> Reasons For The Decli
ars merely to allow their faculty now in place to	carry out	their research effectively. </s><s> This is nearly
Space Administration) depend on universities to	carry out	research and education directly related to agen

Figure 8. Concordances for *carry out* with *research* in the context of three tokens to the left and right in PICAE.

consulting word sketches for adjectival collocates for the noun *development*, for example, the lemma *late* scores high in co-occurrence frequency and logDice. However, when examining the actual concordances behind those numbers, it becomes clear that the senses *later* (happening at some time in the future), *latest* (most recent), and *late* (happening after the correct time) are conflated, which contributes to the polysemous lemma *late* ranking higher than its individual senses would.

The next issue we noted about sketch grammars during the compilation of the ColloCaid database is that depending on the tagger and the sketch grammar used with a given corpus, the resulting word sketches may sort collocates in different ways. For example, word sketches in PICAE present adjectives and nouns used as modifiers separately, so noun+noun collocations like *quality research* are displayed in a different grammar relation from adjective+noun collocations like *qualitative research*. However, in OCAE, the two collocations are presented under the same grammar relation, i.e., modifiers and adjectives are grouped together. Both solutions are valid, but one solution may be better than the other, depending on what word sketches are being used for. In ColloCaid, as previously shown in Figure 5, we had planned to present adjective and modifier collocation suggestions together, because they are part of the same collocation gap in the minds of writers. The formal difference between adjectives and modifiers is not really relevant to a writer seeking an appropriate word to qualify a noun, which is why it makes sense for a pedagogical tool aiming to fill in this gap to conflate the two. So when looking up adjective and modifier collocation suggestions in PICAE, we had to analyse two separate grammar relations, but when using OCAE, the data we needed was neatly summarized together. Thus, in this regard, the OCAE word sketches were better suited to our project. What we would like to highlight here, though, is that it is quite easy for non-trivial differences like these to be overlooked in a practical lexicography project, which could lead to oversights and errors.

Reflecting about what was relevant to present to writers in a collocation tool led us to notice further aspects of language complexity that are less straightforward to capture through word sketches. For example, to fill in a collocational gap like *research* _____ *something*, it is possible to use not just prepositions like *on*, *into*, or *about*, but also verbs like *regarding, concerning*, *surrounding*, *involving, related to*, and *pertaining to*. However, sketch grammars do not discriminate between when such verbs behave like typical verbs and when they are used to fill preposition gaps. This is illustrated in Figure 9, with a word sketch for verbs used with *issue* as subject of the clause. As the concordances exemplified show, some of the collocates listed are indeed used mostly as verbs (e.g. *arise* and *affect*), others are mostly fulfilling a role analogous to prepositions (e.g. *surround* and *relate*), and others function in a hybrid way, such that depending on the context they are either more verbal or more prepositional (e.g. *involve*). Thus, classifying verbs like *surround* and *relate* as verbs when they are in fact used predominantly in prepositional contexts distorts the word sketch output for verbs used with *issue* as subject of the clause. Our strategy, in ColloCaid, was to offer verbal collocates such as these together with other verb collocations when they occurred in a verbal syntagmatic gap, but to shift them along to the group of prepositional collocation suggestions whenever they occurred as verbs fulfilling a preposition gap. This need for manual curation is yet another example of the importance of watching out for the possibility of e-lexicography tools not always capturing the complexity of natural languages.

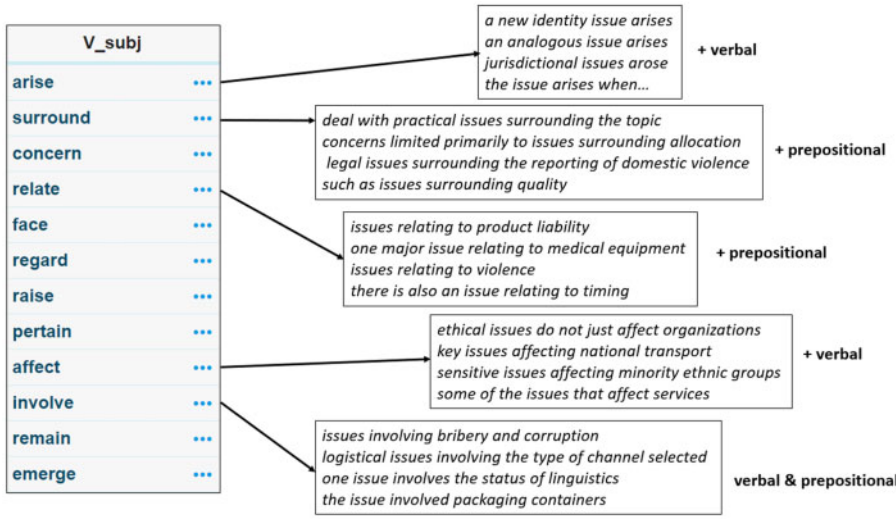

Figure 9. Partial word sketch showing top twelve verbs used with *issue* as subject in PICAE and examples of concordances generated by the grammar relation.

5. Example selection issues

This section outlines our approach to selecting examples of collocations in context so as to help writers gain confidence in using the collocation suggestions offered by ColloCaid, and draws attention to issues we had to watch out for when curating examples from corpora.

Concordances from PICAE and OCAE were used as authentic sources of expert academic English texts from which ColloCaid examples were extracted. However, the original concordances underwent editorial selection and adaptation before being incorporated in ColloCaid.

With regard to selection, care was taken to use discipline-neutral examples or examples that would be intelligible to writers working in different disciplines. Assuming writers would not employ words whose meaning they do not know, our primary concern was not to provide contextual clues to elucidate the meaning of a collocation suggestion. We did prioritize, however, examples that could help writers assess the suitability of a collocation in a given context. Wherever relevant, ColloCaid examples also illustrate typical colligation patterns, so as to nudge writers towards incorporating the collocation in similar ways. For instance, whenever noun+verb collocations occur typically in passive voice constructions, passive voice examples were preferred.

Light editorial adaptation was then carried out to shorten the concordances selected to excerpts that were easy to read and that would not occupy too much screen space in the ColloCaid text editor. Any author names cited in the concordances were also removed to ensure anonymity.

Since exposure to multiple examples can promote data-driven learning more effectively than a single example (Frankenberg-Garcia 2012, Frankenberg-Garcia 2014, Frankenberg-Garcia 2015), we wanted to enable writers to consult more than one example contextualizing the same collocation suggestion. It would not have been feasible, however, to curate

multiple examples for over 32 thousand collocation suggestions. Therefore, we focused on providing three examples per collocation for the strongest eight collocations of each node-collocate grammar relation covered in ColloCaid. As previously shown in Figure 1, this amounted to 29,055 curated examples for 9,685 different collocation suggestions.[12]

The compilation of such a large database of collocation examples would not have been possible without corpora and corpus software. Our starting point was the same word sketches seen in the previous section. Clicking on a collocate listed in a word sketch generates concordances with the specified collocation. For example, clicking on *to* in the word sketch of prepositional collocates of *attitude* in PICAE displays the 966 concordances listed under this grammar relation (Figure 10), which greatly facilitates locating suitable examples. However, while focusing on identifying short, intelligible concordance excerpts, it is easy to overlook the fact that sketch grammars are not 100% accurate, and that the concordances automatically generated from a word sketch may occasionally include misrepresented grammar relations. Of the ten concordances displayed in Figure 10, one is not quite what it seems at first sight. Careful reading reveals that the sketch grammar let slip Line 5, where the preposition *to* is not governed by the node *attitude*, but rather by the verb *convey*. Occasional examples of node-collocate relation misrepresentations also affected lexical collocations. For example, the concordance *connectedness between communities develops* was generated by the collocation *community+develop*, but the subject of the clause is *connectedness* rather than *community*. It is precisely this kind of attention to detail that is required in present-day e-lexicography, where so much is facilitated that it is easy to miss some of the intricacies of language that are part and parcel of 'good old-fashioned lexicography'.

The curation of examples for ColloCaid was facilitated not only by word sketches, but also by GDEX (Kilgarriff et al. 2008), a Sketch Engine tool designed to help find good dictionary examples from concordances. GDEX automatically penalizes concordances with long sentences, rare words, more than one capital letter or non-alpha-numeric characters, and anaphoric reference (which would probably require further context to be intelligible). At the same time, GDEX prioritizes whole sentences, sentences where the target collocation is presented in the main clause, sentences exhibiting other collocates frequently occurring

Figure 10. Partial word sketch for prepositions used with *attitude* and sample concordances for *attitude to* from PICAE.

with the primary collocation, and sentences where the collocation is towards its end (to allow for more context before the collocation is presented). Thanks to these automatic algorithms, prioritized concordances are presented first, while concordances that have been penalized are pushed back towards the end of the list. This makes lexicography more efficient, as suitable examples are likely to be presented first, reducing the number of concordance lines that need to be inspected for example selection. Figure 11 shows the first ten concordance lines for *undertake* plus *activity* as the object of the clause in OCAE in a standard key-word-in-context (KWIC) display followed by the same query in a default GDEX display. As can be seen, the latter concordances are much easier to read and select examples from than the former.

One downside, however, is that the GDEX display can make colligation less evident. Figure 12 summarizes the distribution of word forms of the verb *undertake* in the collocation *undertake* + *activity*, where it can be seen that the verb is predominantly used in passive constructions. However, as shown in Figure 11, the tendency for this collocation to occur in the passive voice can be more easily spotted in standard KWIC concordances than in a default GDEX full-sentence display.[13] Therefore, when assisted by GDEX, it is important for editors not to let colligation patterns that could be relevant to end users go undetected.

Fortunately, the recent 'longest-commonest match' (LCM) algorithm (Kilgarriff et al. 2015) can help. It processes all concordances for a given collocation, and selects LCMs, i.e., matching strings of contiguous words with a minimum 25% occurrence. LCMs can

en for economies like Desta 's , where much economic	**activity**	**is undertaken**	in non-market institutions . It was by impu
iciently against failure , they are reluctant to **undertake**	**activities**		offering a chance of huge success if there is also an acc
and it is clear what has to happen first , interventions (	**activities**	**undertaken**	in the community school , such as after-sch
owed a greater predilection to expand the set of online	**activities**	**undertaken**	in a year 's time . The Pew Internet survey
back in the workplace and in the myriad of other human	**activities**	**undertaken**	in daily life . Metacognition is ' thinking abo
eaching carers on how to correctly **undertake** handling	**activities**		to maximize patient function and safety and allow for the
en ambient temperatures and humidity rise , prolonged	**activity**	**is undertaken**	(e.g. intense physical exertion) , and flu
hings about you is theoretically able to **undertake** CRM	**activities**		. However , many organizations are not able to identify t
gic fit must be held by both organizations . The types of	**activity**	**undertaken**	under key account management will vary a
ular areas , and / or by directly **undertaking** production	**activities**		. This is particularly important in open economies with tr

It must also encompass the effects of the rule on the wider environment in which the **activity** is **undertaken** .
The efficiency of the marketing **activity undertaken** is thus an important issue on which to focus control .
The types of **activity undertaken** under key account management will vary as each relationship is unique .
Companies **undertake** marketing **activities** in order to elicit some kind of response from buyers .
What difficulties are likely to be encountered in attempting to control the marketing **activities undertaken** ?
There is no requirement in this final definition for firms to **undertake** developmental **activities** .
Advise the patient not to **undertake** strenuous **activity** for 2 weeks .
Having difficulty **undertaking** a particular **activity** is inherently a subjective determination .
There are several strategic issues related to the purchasing **activities undertaken** by organizations .
Personal selling is an **activity undertaken** by an individual representing an organization .

Figure 11. Standard KWIC (top) and default GDEX (bottom) display of first ten concordances for *undertake* plus *activity* as object in OCAE.

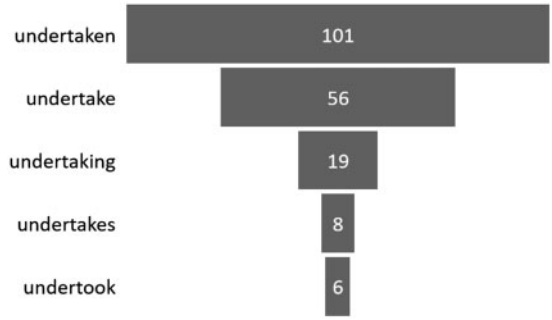

Figure 12. Distribution of word forms of *undertake* in the collocation *undertake* + *activity* in OCAE.

Figure 13. Partial word sketch for verbs used with *activity* as object in OCAE with longest-commonest match option visible.

then be visualised directly from a word sketch, without users having to inspect concordance lines, as shown in Figure 13, which displays a word sketch for verbs used with *activity* as object in OCAE. The LCM *activities undertaken* means more than 25% of the concordances generated by the collocation *undertake* + *activity* contain this form. Lexicographers can thus see that this is a pattern to look out for when selecting examples from a GDEX display if they wish examples to show how the collocation is typically used.

The verbs in the Figure 13 which do not display LCMs are verbs for which no particular contiguous strings of words stand out. The only other LCM shown is *activities carried out* under *activity* + *carry*. Interestingly, as discussed in Section 4, although sketch grammars are not sensitive to multi-word units, which would misleadingly indicate *activity* collocates with *carry* instead of with *carry out*, the LCM can mitigate the problem.

In summary, although example selection can be greatly facilitated by recent advances in e-lexicography, it is important to understand the limitations of how the concordances from

which examples can be sourced are generated and remain vigilant about slips that may occur.

6. Discussion and conclusion

This paper has argued that despite the remarkable advances of e-lexicography in recent years, an over-reliance on how language is processed by sophisticated algorithms can make us inadvertently overlook certain omissions or errors that may slip through the cracks of such systems.

We have demonstrated this in practice through a series of authentic problems we were confronted with during the compilation of the lexical database that supports ColloCaid. The issues encountered included a surprisingly large amount of variation among wordlists extracted from corpora designed to cover broadly the same type of language. Wordlists are created and shared so that they can be reused, and we are grateful that we were able to benefit from AKL, AVL-BAWE, and ACL in the compilation of the ColloCaid lexical database. However, an analysis of the distribution of noun, verb, and adjective lemmas in the three academic English vocabulary lists consulted revealed that only 16.4 % were common to all three lists. This alone is a powerful reminder of the importance of acknowledging that wordlists are heavily dependent on the corpora and extraction methods used to compile them. Because wordlists can vary so much, careful thought must be given to whether and how to reuse them in a given lexicographic project.

Inspection of how a sample of academic lemmas were distributed across six separate corpora (three academic and three general) provided further insights. We were not expecting to find so much variation in frequency across similar types of corpora. The immediate lesson to take home is that when relying on wordlists, it is important to bear in mind that word frequencies can vary not just across markedly different corpora, but also across corpora aiming to cover approximately the same type of language.

Corpus frequencies depend on not just the texts that make up a corpus, but also on corpus-processing systems. Our investigation of why there were no hits for the adjective *modified* in one of the academic English corpora consulted, for example, revealed that the lemma was simply being tagged differently in that corpus. This is precisely the kind of detail that can be easily missed in an e-lexicography project, which is why learning to question abnormal corpus frequencies should be an integral part of a lexicographer's training.

Our discussion of wordlist issues also highlights the importance of lexicographic judgment when relying on keyword analyses to determine headword selection. Although keyword analyses can be extremely useful in lexicography, as they help to detect words that are exceptionally frequent in a specific type of language (Paquot 2010), they do not normally take into account essential information about how words are used in context. As seen in Section 3, verbs like *know* and *need* —which do not feature in academic wordlists because they are also very frequent in general language—should arguably still be considered part of core academic vocabulary because they are used differently in academic contexts from the ways in which they are used in general language.

The limitations of keyword analyses with regard to contextual differences are even more evident when sense distinctions come into play. We observed that lemmas with essential academic senses like *paper*, *field* and, *step* were missing from academic wordlists, probably because they also have very frequent non-academic senses. In contrast, lemmas like *code*

and *application*, which appeared to be comparatively less central to general academic English, did feature in academic wordlists, probably because they have more than one academic sense. Examples like these suggest that, on the one hand, polysemous academic lemmas that have very frequent non-academic senses risk being overlooked because they tend to be demoted in academic keyword rank. On the other hand, polysemous lemmas with more than one academic sense risk being overrated by comparison, since their individual senses get conflated and thus promoted in keyword rank. It is therefore important to acknowledge that sense distinctions are not normally taken into account in wordlists, and that the relative importance of individual senses can be distorted as a result. Lexicographers thus need to be prepared to make informed judgments to compensate for eventual wordlist rank misrepresentations brought about by polysemy.

The compilation of the ColloCaid lexical database brought to our attention a number of collocation research issues as well. Our collocation research was greatly facilitated by Sketch Engine's word sketches, which automatically summarize the collocational profile of a lemma in a corpus. It would not have been possible for our team to compile such an extensive database of academic collocation suggestions without the assistance of word sketches. However, no matter how helpful they are, word sketches should not distort lexicographic judgment. As with all types of software, a broad understanding of how they work and of their limitations is key.

One of the first problems we noticed in ColloCaid was that word sketches are not very good at detecting collocations involving multiword units such as *carry out* + *research*. Multiword units present problems at the level of corpus tokenisation, and any e-lexicography algorithm that relies on token identification will be affected. Knowledge that the sketch grammars behind word sketches do not list multiword units as collocates allowed us to look out for them in other ways so as to mitigate glaring omissions.

Awareness that the tagging and sketch grammars associated with different corpora may result in different word sketches (even for corpora in the same language) is another point we would like to emphasize. Because word sketches look rather much the same on the surface regardless of the corpus used, it is easy to miss that the collocations they provide for two different corpora may not be neatly sorted into equivalent categories. Although this may seem elementary to sketch grammar developers, it is not something obvious to word sketch end users working on a practical project. In ColloCaid, we did our collocation research based on two different corpora, and the word sketch output for the two differed in that adjective-noun and modifier-noun collocations were presented together in one corpus but separately in the other one. It was not possible to change that in our project because the corpora we used were not ours to change. However, it is worth bearing in mind that certain sketch grammars and tagging options may suit some lexicographic projects better than others. Additionally, it is useful for lexicographers to know that the programming rules behind word sketches are not set in stone, and that it may be possible to refine them in order to better suit specific lexicographic needs.

Despite the possibility of adapting sketch grammars according to pre-defined lexicographic criteria, certain aspects of natural language complexity may be quite hard to capture automatically. As seen in Section 4, the word sketches used in our project did not distinguish between verbs like *regard* in verbal contexts like *decisions regarded as...*, and the same verb in prepositional contexts like *decisions regarding something*. Although the difference is evident to trained lexicographers inspecting concordance lines, the lexical

summaries provided through word sketches could result in letting these subtleties go undetected. Thus, in spite of the undeniable advantages of word sketches, their intermediary role between lexicographers and actual texts needs to be held in check.

The need for such monitoring became even more evident when selecting examples for our lexical database. On the one hand, the use of word sketches as a starting point to retrieve concordances illustrating a given node-collocate relation greatly facilitated our work. On the other hand, this also served to lay bare word sketch issues like the *regard* example above, as well as occasional node-collocate relation misrepresentations like *convey the attitude to someone* being classed as *attitude to* rather than *convey to*.

Extra care in lexicographic curation was also required when using Sketch Engine's GDEX tool to help us select suitable examples of collocations in context. Despite the undisputable gains in efficiency brought about by GDEX, we found it harder to notice colligation patterns when inspecting lexicographer-friendly GDEX concordances than when inspecting unfiltered KWIC concordances. Fortunately, thanks to the 'longest commonest match'—yet another remarkable algorithm, which automatically highlights recurrent word forms within a given node-collocate relation—it was possible to get an advance warning about colligations to look out for when inspecting GDEX concordances.

We only had room to discuss in this paper a few selected examples of problems that can arise when using e-lexicography uncritically. We nevertheless believe the practical examples given serve to demonstrate why it is important for end users of e-lexicography tools and resources to develop a broad understanding of how they work and of their limitations. The word list, collocation, and example selection issues raised also illustrate how an overreliance on e-lexicography can inadvertently let good old-fashioned lexicographic judgment slip through the cracks. This is why although practical lexicography projects can benefit enormously from recent advances in e-lexicography, the need for specialist lexicography skills has become arguably even greater.

At the same time, many of the points raised in this paper can serve as feedback and inspiration for ways in which e-lexicography can be further refined. In future, it should be within the remit of e-lexicography to carry out comparative analyses of the frequency of a given word across several corpora like the ones conducted manually in Tables 1 and 2, to help identify possible frequency deviations occurring in a given corpus. Similarly, it would be useful to visualize a tag summary of how a given word has been tagged in different corpora so as to facilitate the detection of problems resulting from tagging differences like the adjective *modified* referred to above. We would also like to explore ways in which the 'collocationality' tool proposed by Kilgarriff (2006) could help to automatically distinguish between collocationally productive and non-productive academic words. Additionally, it would be very useful to explore ways of incorporating automatic word-sense disambiguation (Schütze 1998), so as to help lexicographers compare the frequencies of different senses of polysemous lemmas, run sense-specific word sketches, and extract sense-specific concordances.

Even though the issues raised in this paper were limited to those encountered in the course of the compilation of one specific lexical database, we believe they can help both experienced lexicographers as well as those who are new to the field look out for similar problems in other practical lexicography projects. In other words, our message to lexicographers is 'trust the text' (Sinclair 2004), but beware of the rest. Meanwhile, we hope this paper will contribute to fostering further dialogue between the developers and end-users of e-lexicography tools and resources.

Notes

1. Available at www.collocaid.uk
2. Due to an oversight, the total number of overlapping nodes published in Frankenberg-Garcia et al (2019) was 513 instead of 514, with the noun *event* missing from the analysis.
3. Note that had the complete 3000 lemmas of the original AVL been used, it would have overlapped a lot less with the other lists.
4. See Kilgarriff (2006) for a discussion of how certain words have stronger collocational tendencies and of how 'collocationality' can eventually be measured.
5. COCA is available from https://english-corpora.org (accessed between 20 and 28 May 2020). The five other corpora were accessed via Sketch Engine at https://www.sketchengine.eu/ (between 20 and 28 May 2020).
6. See Section 5 for further information on GDEX.
7. It can be seen that nouns tend to be collocationally more productive than verbs and adjectives, which is in accordance with our previous observation about the reason why the ACL, whose extraction was based on collocations, included a greater proportion of nouns than the other academic wordlists, whose extraction was based on key academic lemmas.
8. We are grateful for the permission given to use these corpora in our research.
9. LogDice is a statistical measure of word association. The higher the score, the stronger the collocation. To the left of the logDice score, 'cooccurrence' values represent the number of times the collocate appears in the span of three tokens left or right of the node, and 'candidates' are the total number of occurrences of the same lemma in the entire corpus.
10. Each verb is followed by the frequency it is used with research as object (e.g., conduct+research = 106 hits), followed by the strength of association score for this relation according to the logDice measure (e.g., conduct+research = is the strongest collocation for this grammar relation, with a logDice score of 10.14).
11. In the English Penn Treebank tagset developed by Helmut Schmid, *carry out* is tagged *V.* RP*
12. Additionally, all collocation suggestions given in ColloCaid are linked to external examples from SkELL (https://www.sketchengine.eu/skell/)
13. Even though is it possible to convert sentence to KWIC display in GDEX, which makes colligation patterns more evident, the trade-off is that KWIC makes examples selection less efficient.

Acknowledgments

This research was funded by the UK Arts and Humanities Research Council (AHRC) (AH/P003508/1).

References

A. Dictionaries

Mayor, M. 2013. *Longman Collocations Dictionary and Thesaurus*. Harlow: Pearson Education.
Sinclair, J. 1987a. *Collins COBUILD English Dictionary for Advanced Learners*. London: Collins.

B. Other resources

Ackermann, K. and Y.-H. Chen. 2013. 'Developing the Academic Collocations List (ACL) – A Corpus-Driven and Expert-Judged Approach'. *Journal of English for Academic Purposes* 12: 235–247.

Ackermann, K., J. de Jong, A. Kilgarriff and D. Tugwell. 2011. The Pearson International Corpus of Academic English (PICAE).

Bitchener, J., S. Young and D. Cameron. 2005. 'The Effect of Different Types of Corrective Feedback on ESL Student Writing'. *Journal of Second Language Writing* 14.3: 191–205.

Davies, M. 2008. Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/ (1 April, 2020).

Durrant, P. 2016. 'To What Extent Is the Academic Vocabulary List Relevant to University Student Writing?'. *English for Specific Purposes* 43: 49–61. doi:https://doi.org/10.1016/j.esp.2016.01.004.

Frankenberg-Garcia, A. 2012. 'Learners' Use of Corpus Examples'. *International Journal of Lexicography* 25: 273–296. doi:10.1093/ijl/ecs011.

Frankenberg-Garcia, A. 2014. 'The Use of Corpus Examples for Language Comprehension and Production'. *ReCALL* 26: 128–146.

Frankenberg-Garcia, A. 2015. 'Dictionaries and Encoding Examples to Support Language Production'. *International Journal of Lexicography* 28: 490–512. doi:10.1093/ijl/ecv013.

Frankenberg-Garcia, A. 2018. 'Investigating the Collocations Available to EAP Writers'. *Journal of English for Academic Purposes* 35: 93–104. doi:10.1016/j.jeap.2018.07.003.

Frankenberg-Garcia, A. 2020. 'Combining User Needs, Lexicographic Data and Digital Writing Environments'. *Language Teaching* 53.1: 29–43. doi:10.1017/S0261444818000277.

Frankenberg-Garcia, A., R. Lew, J. C. Roberts, G. Rees and N. Sharma. 2019a. 'Developing a Writing Assistant to Help EAP Writers with Collocations in Real Time'. *ReCALL* 31.01: 23–39. doi:10.1017/S0958344018000150.

Frankenberg-Garcia, A., R. Lew, G. Rees, J. Roberts, N. Sharma and P. Butcher. 2019*b*. Collocations in e-lexicography: lessons from Human Computer Interaction research. Workshop presentation. Paper presented at the Pre-conference workshop on collocations at eLex 2019, Sintra.

Gablasova, D., V. Brezina and T. McEnery. 2017. 'Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence'. *Language Learning* 67: 155–179. doi:10.1111/lang.12225.

Gardner, D. and M. Davies. 2014. 'A New Academic Vocabulary List'. *Applied Linguistics* 35: 305–327. doi:10.1093/applin/amt015.

Hanks, P. 2012. 'The Corpus Revolution in Lexicography'. *International Journal of Lexicography* 25.4: 398–436. doi:10.1093/ijl/ecs026.

Heuboeck, A., J. Holmes and H. Nesi. 'The BAWE Corpus Manual'. http://www.reading.ac.uk/internal/appling/bawe/BAWE.documentation.pdf (14 June, 2020).

Jakubíček, M., A. Kilgarriff, V. Kovvář, P. Rychlý and V. Suchomel. 2013. 'The Ten Ten Corpus Family'. In **Hardie, A. and R. Love** (eds.), *Proceedings of the 7th International Corpus Linguistics Conference*. Lancaster: Lancaster University, 125–127.

Johns, T. 1991. 'Should You Be Persuaded: Two Samples of Data-Driven Learning Materials'. *English Language Research Journal* 4: 1–16.

Kilgarriff, A. 2006. 'Collocationality (and how to measure it)' In *Proceedings of the 12th EURALEX International Congress*. Torino: Edizioni dell'Orso, 997–1004.

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovvář, J. Michelfeit, P. Rychlý and V. Suchomel. 2014. 'The Sketch Engine: *Ten Years On*'. *Lexicography* 1: 7–36.

Kilgarriff, A., V. Baisa, P. Rychlý and M. Jakubíček. 2015. 'Longest–Commonest Match'. In **Kosem, I., M. Jakubíček, J. Kallas and S. Krek** (eds.), *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., 397–404.

Kilgarriff, A., M. Husák, K. McAdam, M. Rundell and P. Rychlý. 2008. 'GDEX: Automatically Finding Good Dictionary Examples in a Corpus'. In **Bernal, E. and J. DeCesaris** (eds.), *Proceedings of the 13th EURALEX International Congress*. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, 425–432.

Kilgarriff, A., P. Rychlý, P. Smrž and D. Tugwell. 2004. 'The Sketch Engine'. In **Williams, G. and S. Vessier** (eds.), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud, 105–116.

Kosem, I. 2010. *Designing a model for a corpus-driven dictionary of Academic English*, Ph.D. Thesis, Aston University.

Lea, D. 2014. 'Making a Learner's Dictionary of Academic English'. In **Abel, A. C. Vettori and N. Ralli** (eds), *Proceedings of the XVI EURALEX International Congress: The User in Focus*. Bolzano: EURAC research, 181–190.

Michelbacher, L. 2013. *Multi-word tokenization for natural language processing*, Ph.D. Thesis, University of Stuttgart.

Nation, P. and R. Waring. 1997. 'Vocabulary Size, Text Coverage and Word Lists'. In **Schmitt, N. and M. McCarthy** (eds.), *Vocabulary: Description, Acquisition, and Pedagogy*. Cambridge: Cambridge University Press, 6–19.

Paquot, M. 2010. *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London: Continuum.

Rees, G. 2018. *A Phraseological Multi-discipline Approach to Vocabulary Selection for English for Academic Purposes*. Ph.D. Thesis, Universitat Pompeu Fabra.

Roberts, J. C., P. W. S. Butcher, R. Lew, G. Rees, N. Sharma and A. Frankenberg-Garcia. 2020. 'Visualising Collocation for Close Writing'. In **Kerren, A., C. Garth and G. E. Marai** (eds), *EuroVis 2020 - Short Papers*. The Eurographics Association, 181–185.

Rundell, M. 2002. 'Good Old-Fashioned Lexicography: Human Judgment and the Limits of Automation'. In **Corréard, M.-H.** (ed.), *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*. Grenoble: EURALEX, 138–155.

Rundell, M. and P. Stock. 1992. 'The Corpus Revolution'. *English Today* 8.3: 21–32. doi:10.1017/S0266078400006520.

Schütze, H. 1998. 'Automatic Word Sense Discrimination'. *Computational Linguistics* 24.1: 97–123.

Sinclair J. 1987*b*. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins ELT.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Strobl, C., E. Ailhaud, K. Benetos, A. Devitt, O. Kruse, A. Proske and C. Rapp. 2019. 'Digital Support for Academic Writing: A Review of Technologies and Pedagogies'. *Computers & Education* 131: 33–48. doi:10.1016/j.compedu.2018.12.005.

Swan, M. and B. Smith. 2001. *Learner English: A Teacher's Guide to Interference and Other Problems*. (2nd edition.). Cambridge: Cambridge University Press.

Tavares-Pinto, P., G. Rees and A. Frankenberg-Garcia. forthcoming. 'Identifying Collocation Issues in English L2 Research Article Writing'. In **Charles, M. and A. Frankenberg-Garcia** (eds), *Corpora in ESP/EAP Writing Instruction: Preparation, Exploitation, Analysis*. London: Routledge.

Appendix 1

Itemized list of lemmas that do not overlap in the Academic Collocations List (ACL) headwords, the BAWE subset of the Academic Vocabulary List (AVL-BAWE) and the Academic Keyword list (AKL)

ACL only (259)	AKL only (92)	AVL-BAWE only (9)
abuse, acceptance, accuracy, activism, adjustment, administration, affairs, agency, agenda, agreement, ancestor, answer, appearance, area, arena, arrangement, array, article, audience, authority, autonomy, average, background, body, bond, boundary, business, capitalism, care, career, chapter, circle, climate, clue, comment, competence, competition, conference, connotation, consciousness, consent, construct, cost, coverage, crime, critique, currency, customer, database, death, decade, degradation, democracy, detail, disaster, discipline, discourse, display, dispute, domain, dominance, doubt, draft, duration, economy, edition, education, effort, election, emissions, employment, encounter, energy, entity, equality, equilibrium, equivalent, essay, expectation, expenditure, expert, expertise, exploitation, expression, faith, family, feedback, field, flexibility, focus, format, formula, foundation, fraction, framework, freedom, frequency, funds, gender, generation, goal, goods, government, harm, health, help, heritage, ideology, illness, illustration, implementation, impression, incidence, income, inequality, initiative, innovation, inquiry, inspection, instability, instructions, intelligence, intensity, intent, intercourse, internet, interview, investment, investor, involvement, item, journal, judgment, language, law, leader, legislation, life, line, living, location, look, management, market, marketplace, measurement, meeting, memory, merits, message, methodology, mobility, mobilization, nation, objectivity, obligation, offence, opinion, opposition, order, organ, organism, orientation, overview, panel, paradigm, paragraph, parameters, participation, party, peace, peak, people, philosophy, picture, pilot, planning, player, pollution, portion, power, precedence, prediction, premise, presentation, press, priority, probability, proceedings, productivity, profile, prospects, prosperity, proximity, public, qualification, quantities, reaction, recommendation, record, reflection, reform, remarks, representation, resemblance, responsibility, right, roots, safety, score, scrutiny, sector, security, setting, side, speaker, species, spectrum, sphere, stability, staff, stage, state, statement, status, step, stereotype, symptom, teaching, tenet, test, text, thinking, thought, threat, time, trade, training, transaction, transformation, transport, treatment, treaty, trust, turnover, unemployment, usage, variable, village, violence, wage, war, way, weapon, welfare, wisdom, worker, writing, year	addition, adoption, adult, advice, age, analogy, assertion, balance, being, bias, birth, character, classification, commitment, committee, compromise, conception, conduct, construction, content, contradiction, convention, creation, criticism, decline, defence, destruction, determination, disadvantage, discovery, division, doctrine, effectiveness, essence, establishment, evidence, evolution, exposure, extreme, fact, female, formation, future, gain, guideline, hypothesis, idea, instance, kind, list, loss, maintenance, male, manipulation, mankind, mode, motivation, notion, observer, occurrence, operation, option, output, parallel, parent, past, person, personality, proposition, publication, reader, reasoning, relevance, representative, reproduction, resistance, resolution, respect, rise, search, selection, separation, series, similarity, spread, stimulus, team, tension, tolerance, uncertainty, validity, viewpoint	appendix, century, flow, inclusion, manner, project, region, table, university

Nouns
Verbs

ACL only (36)	AKL only (100)	AVL-BAWE only (9)
agree, ally, behave, believe, call, change, charge, communicate, correlate, disagree, discover, disperse, document, educate, embed, find, grow, ignore, intend, know, mention, motivate, oppose, populate, read, realize, resemble, root, share, structure, suit, think, tie, understand, value, work	act, advance, advocate, aim, allocate, allow, appear, assert, assist, attain, attend, attribute, avoid, be, become, benefit, can, cause, claim, clarify, classify, coincide, combine, compete, concentrate, conform, constitute, contrast, convert, damage, deal, decline, destroy, differentiate, diminish, direct, dominate, eliminate, emerge, exceed, exclude, exemplify, explain, expose, facilitate, fail, favour, finance, follow, formulate, gain, impose, induce, introduce, investigate, isolate, label, lead, may, neglect, operate, overcome, participate, perceive, pose, possess, precede, preserve, prevent, prove, pursue, quote, record, regulate, reinforce, reject, remain, render, replace, reproduce, resolve, separate, should, show, solve, specify, stimulate, strengthen, stress, study, submit, suffer, supply, sustain, tackle, term, undermine, undertake, write, yield	address, calculate, estimate, justify, modify, observe, range, situate, test

Adjectives

ACL only (32)	AKL only (90)	AVL-BAWE only (15)
aware, blurred, changing, constant, controversial, dangerous, desirable, emerging, established, exclusive, few, growing, impossible, involved, little, occurring, plausible, popular, powerful, preceding, problematic, proportional, rare, sensitive, skilled, sophisticated, straightforward, striking, unchanged, unclear, variable, visible	absolute, abstract, active, adequate, applicable, arbitrary, average, certain, complete, comprehensive, considerable, conventional, crucial, difficult, dominant, early, equivalent, experimental, extensive, extreme, far, favourable, final, formal, frequent, fundamental, great, immediate, inadequate, incomplete, indirect, inferior, inherent, interesting, large, late, leading, local, logical, main, major, male, maximum, mental, minimal, misleading, mutual, normal, original, other, parallel, partial, passive, past, permanent, physical, possible, prime, principal, productive, profound, progressive, prominent, psychological, radical, random, rapid, rational, real, realistic, related, representative, restricted, scientific, secondary, severe, sexual, single, so-called, special, strict, successive, symbolic, systematic, theoretical, unlike, unsuccessful, varied, visual, wide	above, accurate, beneficial, broad, continuous, current, detailed, economic, existing, external, given, global, increased, numerous, varying