International Journal of Computer-Assisted Language Learning and Teaching
Volume 6 • Issue 3 • July-September 2016

Constructing a Collocation Learning System from the Wikipedia Corpus

Shaoqun Wu, Computer Science Department, University of Waikato, Hamilton, New Zealand

Liang Li, University of Waikato, Hamilton, New Zealand

Ian H. Witten, University of Waikato, Hamilton, New Zealand

Alex Yu, Centre for Business, Information Technology and Enterprise (CBITE), Waikato Institute of Technology, Hamilton, New Zealand

ABSTRACT

The importance of collocations for success in language learning is widely recognized. Concordancers, originally designed for linguists, are among the most popular tools for students to obtain, organize, and study collocations derived from corpora. This paper describes the design and development of a collocation learning system that is built from Wikipedia text and provides language learners with an easy-to-use interface for looking up collocations of any word that occurs in Wikipedia. The use of this corpus exposes learners to contemporary, content-related text, and enables them to search for semantically related words for a given topic. The system organizes collocations by syntactic pattern, sorts them by frequency, and links them to their original context. The paper includes a practical user guide to illustrate how to use the system as a language aid to facilitate academic writing.

KEYWORDS

Academic Writing, Collocation Learning and Teaching, Concordancers, Corpus-based Language Learning, Wikipedia

INTRODUCTION

Collocations are of great importance for second language learners: they play a key role in producing language accurately and fluently. In recent years, corpus-based collocation learning has aroused considerable interest from teachers and researchers (e.g. Boulton, 2010, 2012; Chambers & O'Sullivan, 2004; Chang, 2014; Chen, 2011; Daskalovska, 2015; Yeh, Li, & Liou, 2007; Yoon, 2008). Concordancers, originally designed for linguists, are popular tools for students to explore corpora, particularly with a view to examining collocations. Support for learner use of corpora and concordancing is premised on the fact that exposure to a word and its associated lexical and grammatical patterns in different contexts allows learners to develop a greater sense of its form, meaning and use.

This paper describes the design and development of a collocation learning system, FlaxCLS. FlaxCLS is one of the key elements of the FLAX system (http://flax.nzdl.org), a self-access language learning system documented in Wu (2010), Wu, Franken, and Witten (2009, 2010), and Wu, Witten, and Franken (2010). FlaxCLS has two components: a collocation database built from three million Wikipedia articles comprising three billion words, and a simple interface for looking up collocations. The use of this text base allows learners to inspect typical language use in contemporary, content-related text. Wikipedia articles represent modern English in almost every area of art, life, and science, and includes emerging topics whose vocabulary is not covered by standard corpora such as the British National Corpus.

DOI: 10.4018/IJCALLT.2016070102

The term collocation has different definitions in the literature. We take a syntax-oriented approach in this paper that emphasises the grammatical structure of collocation (Firth, 1957; Nation, 2013; Nattinger & DeCarrico, 1992; Nesselhauf, 2004; Sinclair, 1991) and identifies collocations by syntactic structures (e.g. verb + noun, adjective + noun, noun + verb). FlaxCLS first downloads Wikipedia text, parses it, extracts useful syntactic-based word combinations (e.g., verb+noun, noun+noun, adjective+noun), organizes them by syntactic pattern, sorts them by frequency, and links them to their context sentences. Once this comprehensive collocation database is established, an easy-to-use and learner friendly interface is provided through which learners can seek collocations that include any given word and word type (verb, noun, adjective and adverb), or search for combinations of multiple words (e.g., play an extremely important role).

Furthermore, the concept structure of Wikipedia is used to retrieve semantically related words on a given topic, so that learners can seek topic-related key words and their collocations. For example, searching for animal testing yields related words like toxicity, drug, ethical, welfare, treatment, pain, and their collocations, such as toxicity tests, effect of the drug, ethical principles, animal welfare, potential treatment and pain relief.

The paper is organized as follows. First we examine the use of the Web corpus in collocation learning and rationalize the choice of Wikipedia articles as the primary source from which to build a collocation database. Next we consult the literature to see how concordancers are used to facilitate the inspection of collocations. We discuss limitations reported by researchers and teachers, and suggestions that have been made for learner friendly interfaces. We then describe the design principles underlying our system, including how collocations are extracted, organized and presented in a simple manner. Following that we briefly walk through how to use the online interface to explore collocations. Finally, we review a student guide that has been created to demonstrate its use in preparing essays, choosing appropriate words, using hedging and boosting devices, improving formality, and increasing text variation during writing.

USING THE WEB CORPUS

The web, a vast, contemporary, freely available corpus, has the potential to offer language learners authentic, representative language resources (e.g. Boulton, Jul 2012; Hundt, Nesselhauf, & Biewer, 2007; Kilgarriff & Grefenstette, 2003). Various concordance tools have been developed for Web search, including WebCorp (Renouf, Kehoe, & Banerjee, 2007), KwiCFinder (Fletcher, 2007), and WebBootCat (Baroni, Kilgarriff, Pomikálek, & Rychlý, 2006). De Schryver (2002) distinguishes the use of the Web through direct consultation via search engines—for example, Shei (2008)—from its use as a source of text for corpus building—for example, Wu, Franken, and Witten (2009). Shei (2008) used Google hits to identify recurrent formulaic sequences. He visualized frequencies of Google hits for up to seven consecutive words to indicate phraseology between words. If the frequency line remains at roughly the same level, the newly added word is closely related to its predecessors; if it drops substantially, the new word is no longer part of the formulaic sequence. Shei suggests that language learners can use this indication to guide their choice of collocations.

In contrast, instead of relying on live Web search to generate collocation and concordance data, Wu et al. (2009) work with an off-line corpus (generated from a trillion words) in the form of snippets of up to 5 consecutive words, generated and supplied by Google in 2006. Like the Web itself, the snippets are messy: they contain many non-word character strings, website names, grammatical errors, slang, and unsuitable material (racist, pornographic, etc.). The corpus is cleaned in order to render it suitable for language learning, but unfortunately it is impossible to eliminate grammatical errors

and inappropriate text. The final product is a concordance-like tool called Web Phrases that allows users to seek words that precede or follow any particular word. The system provides an option to group the results by syntactic pattern (e.g. preposition, verb, noun, adjective). It also supports wild-card search: searching for "is * responsible", "is * * responsible" and "is * * * responsible" yields phrases such as is solely responsible, is to be responsible, is not liable and responsible. A notable limitation of this tool is that at most 5-word phrases are returned: other resources are needed for studying phrases in context.

As a corpus, the Web has unique features shared by no other. It is potentially useful for language study because it contains a wealth of language examples that are contextualized and authentic. However, it has intrinsic limitations. Its content is completely uncontrolled and heterogeneous: it has been described as a "dirty corpus" (Kilgariff & Grefenstette, 2003). When using the live web, search results are inconsistent and unstable due to the continual addition of new text, not to mention changes in search engine operation. Biber and Kurjian (2007) remark that "linguistic patterns observed on the Web can vary radically — and seemingly randomly — from one search to the next." When teachers set exercises involving direct Web search they cannot predict what their students will see.

We decided to use Wikipedia text as the primary data source for our collocation database because of its sheer size and contemporary nature. Our work, however, is not restricted to this particular corpus: the system we have developed can be automatically applied (and has been applied) to other corpora. One might contend that the open source nature of Wikipedia – anyone can edit a Wikipedia article – makes it unsuitable for language learning, because grammatical errors and non-standard English can easily be introduced. However, Wikipedia text has several advantages over Web text for collocation learning. First, it provides stable content, which is fixed when it is downloaded from the website and built into a database. Second, despite constant editing by different users it contains far fewer grammatical errors and language misuse than the Web. Third, it is continually evolving. Although the content is fixed upon downloading, updating is a completely automatic process. Fourth, inlinks and outlinks to articles, together with Wikipedia's hierarchical category structure, can be exploited to provide topic-related words and collocations (see "Exploring Related Words" and "Linking to Wikipedia" below), a feature that no other online or offline concordance tool offers. An option for even more controlled language is to use the Simple English Wikipedia, which uses simple English words and grammar, but since because only contains a small collection of articles (about 116,000) we use the full Wikipedia.

CONCORDANCERS IN COLLOCATION LEARNING

Recent years have witnessed an upsurge of research in "data-driven language learning," that is, the notion of learners as language researchers (Johns (1991). Electronic corpora and corpus-based tools have created new potential for learners to explore multiword units, such as collocations (e.g. Boulton, 2010, 2012; Chambers & O'Sullivan, 2004; Chan & Liou, 2005; Chang, 2014; Chen, 2011; Daskalovska, 2015; O'Sullivan & Chambers, 2006; Yeh, Li, & Liou, 2007; Yoon, 2008; Yoon & Hirvela, 2004). Responses are universally positive: corpus use not only facilitates learning and writing, but also arouses learners' awareness of collocations and increase their confidence in language use.

However, despite their proven effectiveness, language learners rarely gain hands-on experience with corpora in mainstream education (Leńko-Szymańska & Boulton, 2015, p. 3). What keeps corpora and corpus-based tools out of mainstream classroom practice? To answer this question, we examine the affordances of these tools and review learner feedback.

Most corpus-based tools allow one to search for two-or-three-word collocations, and language activities also target short collocation learning. For example, Chan and Liou (2005) studied the use of TOTALrecall, a web-based bilingual concordance, for learning verb + noun collocations, which account for the most common word errors among Chinese EFL students. Thirty-six students were required to complete a series of collocation activities that focus on understanding the subtle meaning

of certain verbs that lack direct Chinese equivalents: synonyms (e.g. construct, build, and establish), hypernyms (e.g. create and compose) and troponyms (e.g. break and damage); de-lexicalized verbs (e.g. make, take, do); and non-congruent V-N collocations (e.g. brew tea, *pao cha* in Chinese). Yoon (2008) conducted case studies that introduced six L2 writers in an EAP writing course to the Collins COBUILD Corpus. The most frequently sought items were prepositions and verbs, and common collocation searches included verb+noun (e.g. solve the problem), adj+noun (e.g. high frequency), adv+adj (e.g. quite lower), adv+verb (e.g. greatly affect), and verb+adj (e.g. feel difficult) patterns. Daskalovska (2015) introduced the BYU-BNC concordance to a group of first year undergraduates in Macedonia and compared their performance to that of a control group to assess the effectiveness of data-driven learning for adv+verb collocations (e.g. entirely agree). Ackermann and Chen (2013) developed an Academic Collocation List (ACL) with 2500 frequent and pedagogically relevant entries; however, their list is limited to short collocations.

Corpus analysis tools, whether web-based (e.g., the Collins COBUILD Corpus, WebCorp, WebCollocate, BYU-BNC, COCA) or stand-alone (e.g., WordSmith Tools, AntConC), were originally developed for linguistic researchers with somewhat different interfaces, search functions and presentation of results. As a result, difficulties have been reported by language learners, who are ill-versed in both target language and metalinguistic knowledge, unfamiliar with complex interfaces and search functions, and feel overwhelmed by voluminous search results. Learners have to master query syntax (e.g., part of speech tags) and the idea of wild cards (Boulton, 2012; Chang, 2014; Chen, 2011), spend time analysing copious concordance lines (Boulton, 2010, 2012; Chambers & O'Sullivan, 2004; Chang, 2014; Daskalovska, 2015; Geluso & Yamaguchi, 2014; O'Sullivan & Chambers, 2006; Yeh, Li, & Liou, 2007; Yoon & Hirvela, 2004), locate collocates in concordances (Chan & Liou, 2005), and interpret the meanings of concordances, mostly in the form of keyword-in-context (KWIC) fragments and incomplete sentences (Geluso & Yamaguchi, 2014; Yoon & Hirvela, 2004). The differing interfaces and functions of corpus analysis further increase the challenge, and learners generally need to learn a new system in order to access a different corpus (Chang, 2014).

On the basis of a large-scale international survey, Tribble (2015) reported that user-friendliness and no-cost access are major factors that hinder the application of corpus tools. What learners need is a tool designed specifically for language pedagogy, with a user-friendly interface that requires minimal typing and clicking and straightforward search functions that require little linguistic and metalinguistic knowledge. It should present search results in a way that that automatically classifies retrieved collocations (including multiword ones) with collocates in the correct position, and make complete example sentences readily available.

BUILDING A COLLOCATION DATABASE FROM WIKIPEDIA

We developed the collocation database from 3 million articles downloaded from the Wikipedia website. Collocations are organized according to automatically assigned syntactic patterns. The principle of syntactic organization is widely supported by the literature, and is also adopted by the *Oxford Collocation Dictionary for Students of English* (McIntosh, Francis, & Poole, 2009), *BBI Combinatory Dictionary of English* (Benson, Benson, & Ilson, 1997), and the *LTP Dictionary of Selected Collocations* (Hill & Lewis, 1997).

The key issues raised during the design were these:

1. What are the most useful collocation patterns for learners?
2. How should collocations be presented?
3. Can learners be encouraged to expand their collocation knowledge?

Table 1. Collocation Patterns

Pattern	Example	from
verb + noun(s) verb + noun + noun verb + adjective + noun(s) verb + preposition + noun(s)	*cause problems* *tackle the root cause of* *take a full responsibility for* *result in an increase in*	BBI
gerund verb + noun	*the underlying concept*	NEW
noun + noun	*tax increase*	BBI
noun + *of* + noun	*concept of power*	OCD
adjective(s) + noun(s) adjective + noun + noun adjective + adjective + noun(s) adjective + and/but + adjective + noun(s)	*abstract concept* *a solar energy system* *intensive qualitative research* *economic and social development*	BBI
noun + *to* + verb	*ability to influence*	NEW
noun + preposition + noun	*difference in opinion*	NEW
adjective + *to* + verb	*crucial to understand*	NEW
adjective + preposition + verb	*positive in their attitude*	NEW
adverb + adjective	*seriously addicted*	BBI
verb + pronoun + adjective	*make it easy*	NEW
verb + *to* + verb	*cease to amaze*	OCD
adverb + verb	*beautifully written*	NEW
verb + adverb	*rely heavily on*	OCD

Collocation Patterns

Table 1 shows the 14 collocation types we adopted, with examples of each. Collocations contain from two to five contiguous words (five is rare). The types include some from the work of Benson, Benson, and Ilson (1997) (marked BBI in the Table); some from *Oxford Collocation Dictionary Students of English* (marked OCD); and some that we added ourselves (marked NEW) — for example, the pattern gerund verb + noun (e.g. hotly debated issue, driving issue), particularly useful in academic writing, is omitted from most dictionaries. We extended some types to include more constituents of potential use to learners. For example, the noun part of a verb + noun collocation can be a complex noun phrase involving one or more nouns coupled with modifiers or prepositions: examples are take full advantage of, play an extremely important role. Collocations containing common adverbs like more, much, very, quite are omitted from the patterns involving adverb because these qualifiers can accompany most adjectives and verbs.

Presenting Collocations

The collocations should be presented so as to manage the massive volume of data without overwhelming students. Query terms are often associated with multiple collocation types: their syntactic part of speech may be ambiguous, and some collocations have many variations (e.g. the word advantage in take advantage of can be qualified by full, unfair, undue, greater advantage).

A further issue is how to organize collocations containing different inflected verb forms (e.g. taking, takes, took for the verb take). For example, take advantage of, taking advantage of, took advantage of are the three most frequent verb + noun collocations for advantage, followed by have/has/had the advantage of. Of these, the system shows take advantage of and have advantage of,

suppressing the others to move other useful collocations like gain an advantage, saw the advantage of, and offer the advantage of further up the results list.

To address these issues, we adopted a hierarchical organizational structure. Collocations are first grouped by the syntactic role of the query term (e.g., used as noun or verb). Then they are organized by syntactic pattern (e.g., all verb + noun collocations are displayed together). For collocations that contain inflected verb forms or extensions (e.g. take full advantage of is an extension of take advantage), only the most frequent one is displayed; when it is clicked, the others appear in a pop-up window. This is done by extracting two key words from the collocation, transforming them into their base form, and using this for grouping. The result is that take/taking/took advantage of and take/taking/took full/unfair/undue advantage of are all grouped under take advantage. Users see only take advantage of in the results page, because it is the most frequent, but clicking it lists all the others.

We adopted the principle of ordering collocations by frequency. This is achieved in three ways: the most frequent syntactic type of the query word, the most frequent collocation pattern, and the most frequent collocation. For example, collocations of the query benefit are first grouped under noun and verb forms; the former are displayed first because they are more frequent than the latter. Within the noun group, adjective + benefit, noun + benefit, benefit + of + noun, verb + benefit . . . are presented in descending order of frequency, and within each pattern the most frequent collocation is listed first. The same applies to the verb group.

Expanding Learners' Collocation Knowledge

We have investigated ways to encourage students to expand their collocation knowledge on topics related to their area of study.

Whenever a query is made, a selection of related words, which can be clicked in order to explore their collocations, is displayed. To do this, the Wikipedia Miner tool (Milne & Witten, 2012) is invoked to determine the Wikipedia article that corresponds most closely to the query. Wikipedia Miner uses Wikipedia's internal hyperlinks to determine the semantic similarity of any pair of articles (Milne & Witten, 2012). Of course, a single query term might match more than one article — for example, the word kiwi may refer to a bird, a fruit, a person from New Zealand, or the New Zealand national rugby league team, all of which have distinct Wikipedia entries — and in this case the most popular interpretation is chosen, popular in terms of its number of mentions in the Wikipedia itself. If such an article exists, key terms and their collocations that appear in it are returned as suggestions to the user. First the article is parsed, and its nouns, adjectives, verbs, and adverbs are designated as content words. For each such word, a score is calculated that reflects how central the word is to the article, based on the number of times it appears in it (which increases the score) and the number of times it appears in the collection as a whole (which decreases it). This metric, which is commonly used in information retrieval (called TF-IDF, and described by, for example, Witten, McNab, Jones, Apperley, Bainbridge, & Cunningham, 1999), is used to rank words related to the query, so that they can be displayed in descending order of relatedness. Collocations involving any of these related words can be obtained simply by clicking. If no Wikipedia article matches the query term, say advantage, the related words are not displayed.

Further information is displayed that allows learners to explore the topic represented by the query. Having identified the closest corresponding Wikipedia article, a definition — typically the first sentence or two of the article — is extracted from it and presented to the user. Furthermore, semantically related Wikipedia articles are listed, with mouse-over definitions.

A useful way of allowing users to explore collocations related to a particular domain would be to build a domain-specific database (e.g. about "nuclear weapons," for an essay assignment) from relevant Wikipedia articles. This is simple to do with our software, and although we have not yet done so for Wikipedia collections we have experimented with separate collocation databases built from the Social Sciences, Arts and Humanities, Physical Sciences and Life Sciences partitions of

the British Academic Written English collections (Nesi & Gardner, 2012). Incidentally, a domain-specific database also reduces the problem of semantic ambiguity.

Building Procedures

We download artciles (in XML format) from the Wikipedia website and use the Wikipedia Miner tool (Milne & Witten, 2012) to retrieve individual articles which are then feed into the collocation identification process. The identification process involves involves six steps:

1. Split the text into sentences
2. Assign part of speech tags to all words
3. Match tagged word sequences against a set of syntactic patterns
4. Remove collocations containing high-frequent words e.g. very
5. Associate sample text with the collocations that have been identified
6. Build search indexes.

In steps 1 and 2, an off-the-shelf natural language processing tool (OpenNLP) is used to split the text into sentences and assign syntactic tags to the words. Then the tagged words are compared against each collocation pattern in Table 1. In step 4, collocations containing high-frequent words such as only, quite, very are discarded. Whenever a collocation is identified, its sentence are extracted and associated with it in step 5, to help students study collocations in context rather than as isolated items. Finally, collocations are grouped by pattern, and search indexes are created for all their constituent words.

EXAMPLE OF USE

The system is called FlaxCLS. To illustrate it in action, we first search for collocations of a particular word; then explore various expansions of a collocation; examine it in context; and finally retrieve related words.

Searching for Collocations

To look up collocations the user types in a word of interest, whereupon FlaxCLS retrieves collocations and displays them along with other information about the word.

Figure 1. Family words, synonyms and collocations associated with the word research

Figure 1 shows the result for the word research (only the first part of the page is shown). Inflected and derived forms (family words) of the query term appear first, along with its synonyms; these are identified using the standard WordNet resource. Here the derived forms are researched, researcher, researchers, researches and researching. The verb synonyms include search, explore and investigate; noun synonyms include investigation, investigating, inquiry and enquiry. Clicking any of these derived forms or synonyms invokes a new search using it as the query term.

Collocations are grouped by the syntactic role of the query term. In this case, research can be used as both noun and verb. There are eight patterns related to the noun form and seven to the verb form; they are shown in descending order of frequency. The excerpt in Figure 1 displays the three most popular noun patterns: research + noun, adjective + research, and noun + of + research. For each one, 50 collocation samples are retrieved, along with their frequency, and displayed ten at a time in decreasing frequency order. Here, the most frequent collocations of the above three types are research project, scientific research and area of research respectively. The more link at the bottom right shows the next ten.

Selecting any of these collocations brings up its extensions in a superimposed panel: Figure 2 shows the result of clicking scientific research. These extended collocations all contain the words scientific and research, not necessarily adjacent (although they all happen to be in this example), and have the form adjective + noun. Many extended collocations include more than one noun or adjective, such as basic scientific research, scientific research organizations, and independent nonprofit scientific research institute. Their frequencies are shown on the right. Note that there is a 5-word limit on the collocations stored in the database.

If a collocation cannot be extended into another collocation in the database, then clicking it retrieves contextual sentences from the original text. For example, Figure 3 shows the result of selecting basic scientific research (the third item in the list of Figure 2).

Searching for Multiple Words

Typing more than one word retrieves collocations containing them all, irrespective of order and intervening words. For pedagogical reasons, this differs from the usual search-engine implementation of phrase search, where the words are constrained to appear consecutively, in order. Multi-word searching is a good way to expand collocation knowledge by studying how combinations of terms are used. Many students have difficulty with the correct use of articles and prepositions: Should an article appear between take and advantage? — What prepositions can follow make sense? Searching for take advantage yields the expansions shown in Figure 4, indicating that no article intervenes between these two words, but that qualifiers such as full, maximum, unfair, greater, little are possible. It also shows that the expressions are following by the preposition of.

Figure 2. Collocations similar to scientific research

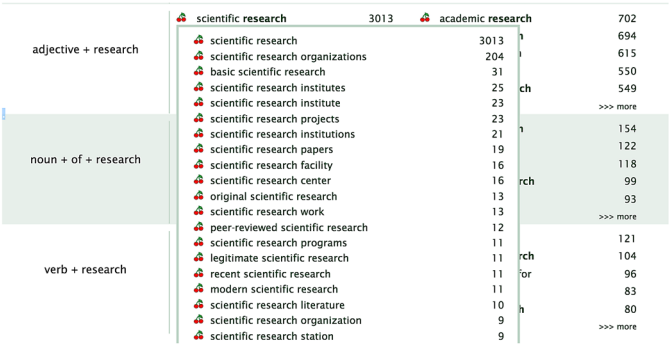


Figure 3. Text samples of basic scientific research

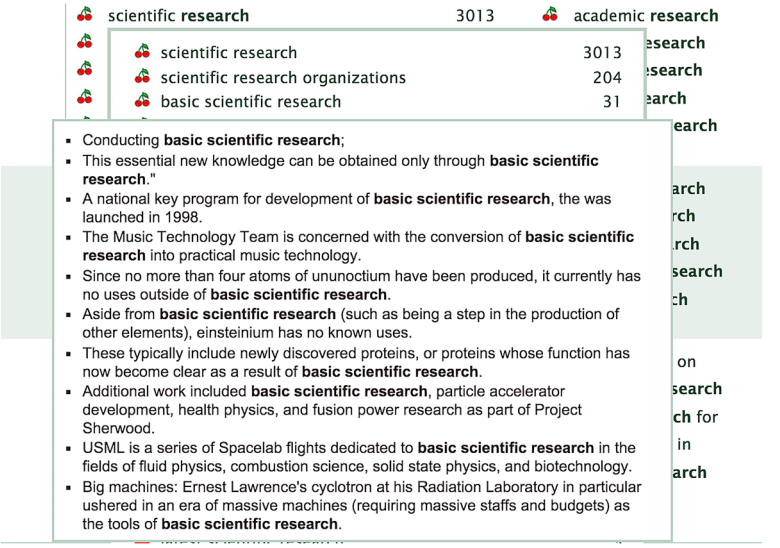


Figure 4. Collocations containing the words take and advantage

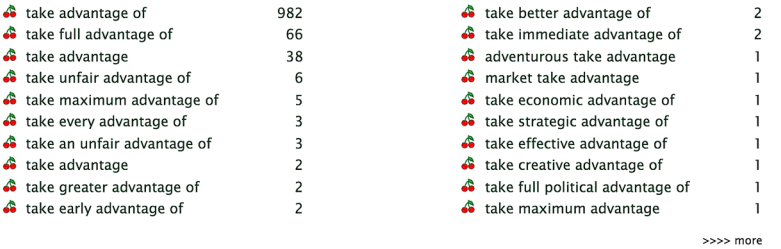
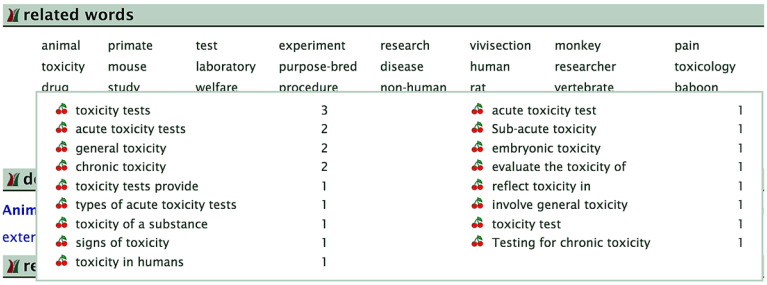


Figure 5. Collocations containing the words social and research

As another example, Figure 5 shows the result of searching for social research. It includes the three- or four-word adjective + noun collocations seen in the previous section, and many other patterns: research in social science (noun + preposition + noun), tradition of social research (noun + of + noun) and predominant in social research (adjective + noun). On this interface collocations of different syntactic patterns are displayed together, sorted by frequency. The more link at the bottom reveals further collocations containing these two words.

Figure 6. Words related to the topic animal testing and collocations associated with toxicity



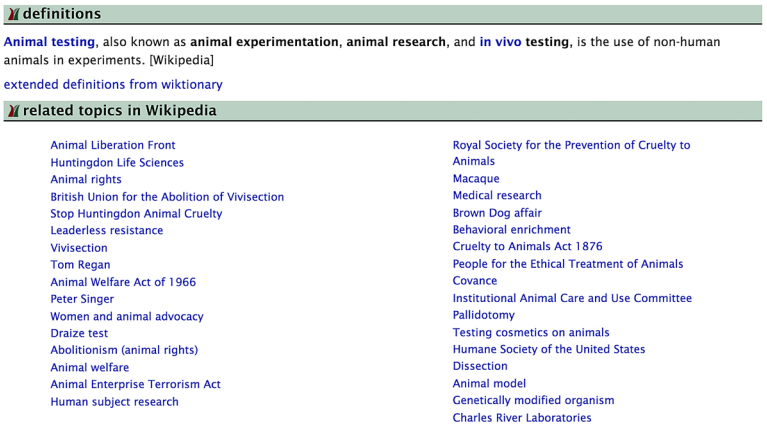
EXPLORING RELATED WORDS

Further down the query response page, of which Figure 1 shows the beginning words that are related to the query term appear. Figure 6 shows this for a different query: animal testing. Forty related words appear (partially obscured in Figure 6): animal, primates, test, experiments, research, vivisection …. The collocations associated with any of these can be viewed in a superimposed window. In Figure 6, toxicity has been selected and some of its collocates are shown: they involve adjectives acute, general, chronic, embryonic; verbs reflect, involve, evaluate; and noun phrases toxicity tests, sign of toxicity, toxicity of a substance. As usual, more related words are available via the more link. The words are sorted by TF-IDF metric mentioned earlier. Words in later panels are more general: for example, the last words related to animal testing are population, line, end, series, form, play, have, be.

Linking to Wikipedia

The panel beneath the related words displays Wikipedia's definition of the query term, animal testing in Figure 7. Following that are related Wikipedia articles: Animal Liberation Front, Huntingdon Life Sciences, Animal rights, and so on; each hyperlinked to the corresponding Wikipedia article. Up to 50 topics are displayed, sorted by their conceptual relatedness to the query. Mousing over a topic gives its definition; clicking it takes users to the Wikipedia article on that topic.

Figure 7. Animal testing: its definition and related topics in Wikipedia



GUIDE FOR STUDENT USERS

We have developed a guide for student users, based on actual writing assignments. By analyzing typical errors that students make and relating them to the possibilities the system affords, we have created five kinds of exercises. Here we will give brief descriptions of these exercises, using an essay entitled "Alcohol advertising: Should it be banned?" Students are also encouraged to develop their own search strategies beyond these five suggested applications.

Preparing for Essay-Writing

The first step in essay preparation is to identify keywords for the topic; in this case, alcohol, advertising, and ban are obvious candidates. Next, collocations are sought that are germane to the topic. This can stimulate a brainstorming process during which new and inspiring ideas may be encountered. Thus it is a good idea to collect several collocations, even though some might not end up in the text. Table 2 provides four sample collocations for each keyword individually. Although exact matches may not occur, words can sometimes be substituted or added to relate a collocation to the topic. In the samples above, ban on tobacco advertising can be changed to ban on alcohol advertising, heavy advertising to heavy alcohol advertising, and legislation to ban to legislation to ban alcohol advertising.

Choosing an Appropriate Word

Many students have difficulty in finding the right words to express their ideas, because they lack collocation knowledge or are unduly influenced by their mother tongue. As a result, they tend to formulate inappropriate word combinations, or overuse general modifiers such as more, very, bad, good, etc. This is particularly noticeable in verb + noun, adverb + verb, and adjective + noun combinations, as in the following sentences, where infelicitous phrases appear in bold:

Alcohol advertising is **actively related** *to alcohol consumption, and the consumption can lead to fatalities.*
Some people argue that the alcohol product advertising should be banned and others **keep the opinion** *against it.*
While many alcohol companies are enjoying **lucrative profits***, their alcohol advertising activities are being challenged by the general public and researchers.*

Table 2. Collocations related to the topic Alcohol advertising: Should it be banned?

alcohol	advertising	Ban
alcohol consumption amount of alcohol excess alcohol addicted to alcohol	effects/power/impact of advertising heavy advertising funded by advertising	ban on tobacco advertising advertising ban legislation to ban supported the ban

Table 3. Collocations associated with the words related, opinion and profit

Related	opinion	profit
closely related highly related clearly related	express an opinion have an opinion voice an opinion	substantial profit increased profit considerable profit

In the first example, the adverb actively is used in an attempt to emphasize the strong correlation between alcohol advertising and consumption. In the second, keep is not an appropriate verb to associate with the noun opinion. The last example, lucrative profits, is a bizarre combination: lucrative is commonly used with business, market, career, etc., but not with profit.

The collocation database provides a plentiful source of plausible word combinations. It is fairly easy to locate appropriate verbs or adjectives for a particular noun, or appropriate adverbs for a particular verb. Table 3 gives some collocates of related, opinion, and profit that were retrieved using the system. In the first example sentence above, closely, highly, and clearly are all far more appropriate than actively. In the second, express, have and voice all seem to fit the context. In fact, this sentence can be further improved by including have an opposite opinion; a student can find this by examining the extensions of have an opinion. In the third example, lucrative can be replaced by substantial, increased, or considerable to express the intended idea.

Hedging and Boosting

Adding adverbs to qualify statements is a common rhetorical device, particularly in academic writing. But students often have trouble hedging or boosting statements appropriately and precisely. As a result, they overuse general adverbs (very, more, much, …) to strengthen or weaken their claim, and sometimes invalidate statements by choosing overly specific qualifiers. Consider these:

Alcohol is **very harmful** *to their physical and psychological health.*
It is a common sense that the more ads we are exposed to, the more likely we are to be seduced to drink and may drink excessively, which **inevitably leads to** *disasters while driving.*
Smart (1988) however had reviewed many other research and admitted that the link between the advertising and consumption was weak and awaiting more comprehensive research, while at the same time confirmed that alcohol drinkers **were definitely exposed to** *alcohol advertising and their consuming behaviors were in fact continuing to increase.*

The *very* in the first example is probably the most common adverb used by novice writers to add strength to a statement. Students rely on such adverbs to help voice opinions because of their restricted vocabulary knowledge. These adverbs are weak and ambiguous, and should be avoided in academic writing. In the second and third examples, the adverbs inevitably and definitely are used to express a high degree of certainty. However, they are too extreme: excessive drinking does not necessarily lead to driving disasters, and not all alcohol drinkers are influenced by liquor advertisements.

The collocation database can help writers find appropriate hedges and boosters. Table 4 shows some examples that are commonly associated with harmful, lead to and exposed to, expressing various degrees of certainty.

Table 4. Collocations associated with the terms harmful, lead to and exposed to

harmful	lead to	exposed to
potentially harmful	probably lead to	potentially exposed to
possibly harmful	easily lead to	sufficiently exposed to
apparently harmful	usually lead to	increasingly exposed to
particularly harmful	ultimately lead to	regularly exposed to
extremely harmful	inevitably lead to	constantly exposed to

Improving Formality

Formality and precision are both important features of academic writing. However, students often overuse colloquial language, and their writing comes over as informal and lacking in precision. Here are three different ways this can occur.

1. Using generic quantifiers
   a. Due to this, the consumption of alcohol product has reduced a lot.
   b. If alcohol advertising were banned then this sort of behavior would decrease.
2. Overusing general words
   a. Drinking alcohol will hurt health and make public health problems.
   b. The majority of binge drinkers do not think they are problem drinkers so they could have bad effect on their classmates.
3. Failing to employ topic-specific collocations
   a. Drinking too much alcohol can change our behaviors.
   b. Banning alcohol advertising makes people who love alcohol very much decrease.

Students can consult the system to find precise expressions that help them avoid colloquial usage. The suggestions in Table 5 relate to the example sentences above.

For sentence 1a, Table 5 suggests replacing a lot by a more expressive word: significantly, considerably, or greatly. Likewise, sort of in 1b could be replaced by undesirable, unacceptable, or deviant. For 2a, the verbs cause, raise and pose are commonly associated with the noun problem. In 2b, substituting serious, damaging or disastrous for bad adds strength. The cumbersome expressions in 3a and 3b can be replaced by topic-related collocations, heavy (or excessive, or serious) drinking instead of drinking too much alcohol, and heavy (or regular, or habitual) drinker for people who love alcohol very much.

Increasing Text Variation

A common problem in student writing is repetition, repetition, repetition. Unless deliberately used for dramatic effect, repetitive writing is boring writing. Here we illustrate how the collocation database can be used to enliven the examples below, taken from a student essay.

1. Ackoff and Emshoff (1975) confirmed that the increase of advertising activity on the alcohol brand was positively linked with the sales, hence the increasing consumption of the product. Smart (1988) however admitted that the link between the advertising and consumption was weak and awaiting more comprehensive research. Saffer (1997) focused on alcohol consumption and motor vehicle fatalities and revealed positive link between the two.

Table 5. Collocations that can be used to improve formality

1a. reduce a lot	2a. make public health problems	3a. drinking too much alcohol
significantly reduce considerably reduce greatly reduce	cause the problem raise the problem pose the problem	heavy drinking excessive drinking serious drinking
1b. sort of behavior	**2b. have bad effect on**	**3b. people who love alcohol very much**
undesirable behavior unacceptable behavior deviant behavior	have serious effect on have damaging effect on have disastrous effect on	heavy drinker regular drinker habitual drinker

2. Some people will argue that some alcohol products also have some benefits such as the use for medicine. However, everything has both sides, it is up to how people use. Even though some alcohol products have some benefits, the drawbacks of alcohol products overweight the benefits. Therefore, the alcohol product advertising should be banned.
3. In the long run, it has more advantages to ban alcoholic product advertising on the whole in terms of the healthier and sustainable development of the country, although it may have big impact on the sales of alcohol companies as frequently argued as their evidence by the opponents. For example, … It is unwise to invest even one dollar on alcohol advertisements, which have bad impact on people's health.

First, deploy synonyms to avoid overusing the same word. For example 1, the Synonyms button (Figure 1) shows that associate and relate are synonyms of the word link. Further checking the collocations of these two words and their noun forms (association and relation) yields useful phrases: associate with or association between and relate to or relation between. These are plausible alternatives for link with and link between.

Second, consider using other members of the same word family (e.g., verb, noun, adjective and adverb). The word benefit is frequently overused in student writing, particularly its noun form — as in the phrase have benefits. Searching for benefit generates the family word beneficial, and also verb usages such as benefit consumers, benefit greatly from, able to benefit from, and benefit from the use of.

Third, have + adjective + impact on occurs several times in the example essay in conjunction with weak adjectives like big, bad, small, and great. Searching for phrases by putting multiple words in the query box — in this case have impact — provides an effective way of finding alternatives, such as enormous, considerable, significant, little, adverse, and minimal. Other verbs associated with impact on include assess, examine, consider, minimize, reduce, and measure.

CONCLUSION

Collocations are one of the most challenging aspects of language learning. Native speakers rely on years of accumulation through constant exposure in authentic contexts. Corpus consultation with concordancers have been recognized as a promising way for learners to study and explore collocations at their pace and in their own time. However, learners face difficulties using existing tools, which are designed for linguists; moreover, existing corpora rarely satisfy learners' diverse needs. Effective collocation retrieval tools are required that are designed for language learners.

We have designed and built a collocation system that draws material from three million Wikipedia articles, along with an easy to use interface that is suitable for student use. Collocations are retrieved simply by typing in the word or words of interest. To minimize the volume of data the user needs to process, results are organized according to syntactic patterns, conflated by word family, and displayed in descending order of frequency. The system is linked to a publicly available knowledge database — Wikipedia — to retrieve words and collocations that are semantically related to the query terms. Finally, we have designed a guide based on an actual academic writing assignment to illustrate how students can use this resource to prepare, compose and review their text during the writing process.

Like most collocation learning resources (e.g. dictionaries and corpus-based tools), FlaxCLS is primarily designed as a self-study tool for intermediate to advanced learners who already have a sufficient repertoire of individual words but lack the knowledge of co-occurring words. Additional training is needed when students are introduced to FlaxCLS, because part-of-speech knowledge is essential to understand the wording on the interface, such as used as a noun, adjective + knowledge, and verb + preposition + knowledge. Sentence samples from Wikipedia can be difficult to understand for lower-level learners who have limited vocabulary in technical terms and proper names (e.g. endocrinologist, harpsichord and Chomsky).

FlaxCLS has been used at the University of Waikato for language support for many postgraduate students, and has received positive reviews from students and teachers. An initial study of 15 Chinese postgraduates suggests that it is easy to use and learner friendly compared to other corpus tools such as COCA, particularly for seeking noun + of + noun (e.g., focus of public attention, principle of equality), or verb + proposition + noun (e.g., speak on behave of, vary in size) collocations. Students tend to pick up longer chunks when being asked to collect useful collocations of a word (e.g. take full advantage of instead of take advantage, for the word advantage). However, to fully understand its potential to support collocation learning, comprehensive user studies are needed. We call for participation from teachers and researchers, and believe that this will lead to further refinement of the system.

REFERENCES

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) - A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247. doi:10.1016/j.jeap.2013.08.002

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: A web tool for instant corpora. In *Computational Lexicography and Lexicology* (pp. 123-131). Retrieved from http://www.euralex.org/proceedings-toc/euralex_2006/

Benson, M., Benson, E., & Ilson, R. (1997). *The BBI dictionary of English word combinations*. Amsterdam, Netherlands: John Benjamins Publishing Company. doi:10.1075/z.bbi1(2nd)

Biber, D., & Kurjian, J. (2007). Towards a taxonomy of web registers and text types: A multi-dimensional analysis. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 109–132). Amsterdam, Netherland: Editions Rodopi. doi:10.1163/9789401203791_008

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning*, *60*(3), 534–572. doi:10.1111/j.1467-9922.2010.00566.x

Boulton, A. (2012). Beyond concordancing: Multiple affordances of corpora in university language degrees. *Languages. Cultures and Virtual Communities*, *34*, 33–38.

Boulton, A. (2012, July). Wanted: Large corpus, simple software. No timewasters. *Paper presented at theTaLC10: 10th International Conference on Teaching and Language*, Warsaw, Poland.

Chambers, A., & O'Sullivan, Í. (2004). Corpus consultation and advanced learners' writing skills in French. *ReCALL*, *16*(1), 158–172. doi:10.1017/S0958344004001211

Chan, T.-, & Liou, H.-C. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations. *Computer Assisted Language Learning*, *18*(3), 231–250. doi:10.1080/09588220500185769

Chang, J.-Y. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL: the Journal of EUROCALL*, *26*(2), 243–259. doi:10.1017/S0958344014000056

Chen, H.-J. H. (2011). Developing and evaluating a web-based collocation retrieval tool for EFL students and teachers. *Computer Assisted Language Learning*, *24*(1), 59–76. doi:10.1080/09588221.2010.526945

Daskalovska, N. (2015). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, *28*(2), 130–144. doi:10.1080/09588221.2013.803982

De Schryver, G.-M. (2002). Web for/as corpus: A perspective for the African languages. *Nordic Journal of English Studies*, *11*(2), 266–282.

Firth, J. R. (1957). Modes of meaning. In J. R. Firth (Ed.), *Papers in linguistics 1934-1951* (pp. 190–215). London, United Kingdom: Oxford University Press.

Fletcher, W. H. (2007). Concordancing the web: promise and problems, tools and tehniques. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 25–45). Amsterdam, Netherland: Editions Rodopi. doi:10.1163/9789401203791_004

Geluso, J., & Yamaguchi, A. (2014). Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL*, *26*(2), 225–242. doi:10.1017/S0958344014000044

Hill, J., & Lewis, M. (1997). *LTP dictionary of selected collocations*. Hove, United Kingdom: Language Teaching Publications.

Hundt, M., Nesselhauf, N., & Biewer, C. (2007). Corpus linguistics and the web. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 1–5). Amsterdam, Netherland: Editions Rodopi. doi:10.1163/9789401203791_002

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. *English Language Research Journal*, *4*, 1–16.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, *29*(3), 333–347. doi:10.1162/089120103322711569

Leńko-Szymańska, A., & Boulton, A. (2015). *Multiple affordances of language corpora for data-driven learning*. Amsterdam, Netherlands: John Benjamins Publishing Company. doi:10.1075/scl.69

McIntosh, C., Francis, B., & Poole, R. (2009). *Oxford collocations dictionary for students of English*. Oxford, United Kingdom: Oxford University Press.

Milne, D. N., & Witten, I. H. (2012). An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, *194*, 222–239. doi:10.1016/j.artint.2012.06.007

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

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford, United Kingdom: Oxford University Press.

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge, United Kingdom: Cambridge University Press.

Nesselhauf, N. (2004). What are collocations? In D. J. Allerton, N. Nesselhauf, & P. Skandera (Eds.), *Phraseological units: Basic concepts and their application* (pp. 1–21). Basel, Switzerland: Schwabe.

O'Sullivan, Í., & Chambers, A. (2006). Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing*, *15*(1), 49–68. doi:10.1016/j.jslw.2006.01.002

Renouf, A., Kehoe, A., & Banerjee, J. (2007). WebCorp: An integrated ssytem for web text search. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 47–67). Amsterdam, Netherland: Editions Rodopi. doi:10.1163/9789401203791_005

Shei, C.-C. (2008). Discovering the hidden treasure on the Internet: Using Google to uncover the veil of phraseology. *Computer Assisted Language Learning*, *21*(1), 67–85. doi:10.1080/09588220701865516

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, United Kingdom: Oxford University Press.

Tribble, C. (2015). Teaching and language corpora: Perspectives from a personal journey. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 37–62). Amsterdam, Netherland: John Benjamins Publishing Company. doi:10.1075/scl.69.03tri

Witten, I. H., McNab, R. J., Jones, S., Apperley, M., Bainbridge, D., & Cunningham, S. J. (1999). Managing complexity in a distributed digital library. *Computer*, *32*(2), 74–79. doi:10.1109/2.745723

Wu, S. (2010). Supporting collocation learning (Doctoral dissertation). Hamilton, New Zealand: University of Waikato. Retrieved from http://hdl.handle.net/10289/4885

Wu, S., Franken, M., & Witten, I. H. (2009). Refining the use of the web (and web search) as a language teaching and learning resource. *Computer Assisted Language Learning*, *22*(3), 249–268. doi:10.1080/09588220902920250

Wu, S., Franken, M., & Witten, I. H. (2010). Supporting collocation learning with a digital library. *Computer Assisted Language Learning*, *23*(1), 87–110. doi:10.1080/09588220903532971

Wu, S., Witten, I. H., & Franken, M. (2010). Utilizing lexical data from a Web-based corpus to expand productive collocation knowledge. *ReCALL*, *22*(1), 83–102. doi:10.1017/S0958344009990218

Yeh, Y., Li, Y.-H., & Liou, H.-C. (2007). Online synonym materials and concordancing for EFL college writing. *Computer Assisted Language Learning*, *20*(2), 131–152. doi:10.1080/09588220701331451

Yoon, H. (2008). More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning and Technology.* Retrieved from http://llt.msu.edu.ezproxy.waikato.ac.nz/vol12num2/yoon.pdf

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, *13*(4), 257–283. doi:10.1016/j.jslw.2004.06.002

Shaoqun Wu is a lecturer in the Computer Science Department at the University of Waikato in New Zealand, and is the main developer of the FLAX language project. Her research interests include Computer Assisted Language Learning, Mobile Language Learning, Supporting Language Learning in MOOCs, Digital Libraries, Natural Language Processing and Computer Science education.

Liang Li is a PhD student in Te Hononga School of Curriculum and Pedagogy and Department of Computer Science at the University of Waikato. Her research interests lie in the area of corpus linguistics, L2 academic writing, and computer-assisted language learning.

Ian H. Witten is Professor of Computer Science at the University of Waikato in New Zealand. His research interests include language learning, information retrieval, and machine learning. He has published widely, including several books, such as Managing Gigabytes (1999), Web Dragons (2007), How to Build a Digital Library (2009), and Data Mining (2011).

Alex Yu is a Senior Lecturer at Centre for Business and Information Technology Enterprise at Waikato Institute of Technology. I am also a core developer of the FLAX project. My research interests include computer assisted language learning, MOOCs, mobile language learning, and data mining.