Using Wikipedia for Language Learning

Shaoqun WU
Faculty of Computing and Mathematical Sciences
The University of Waikato
shaoqun@waikato.ac.nz

Ian H. Witten
Faculty of Computing and Mathematical Sciences
The University of Waikato
ihw@waikato.ac.nz

ABSTRACT

Differentiating between words like *look, see* and *watch, injury* and *wound*, or *broad* and *wide* presents great challenges to language learners because it is the collocates of these words that reveal their different shades of meaning, rather than their dictionary definitions. This paper describes a system called FlaxCLS that overcomes the restrictions and limitations of the existing tools used for collocation learning. FlaxCLS automatically extracts useful syntactic-based word from three millions Wikipedia article and provides a simple interface through which learners seek collocations of any words, or search for combinations of multiple words. The system also retrieves semantically related words and collocations of the query term by consulting Wikipedia. FlaxCLS has been used as language support for many Masters and PhD students in a New Zealand university. Anecdotal evidence suggests that the interface it provides is easy to use and students have found it helpful in improving their written English.

Keywords: Collocations, Concordancers, Computer-assisted Language Learning, Data-driven Learning, Wikipedia

1. INTRODUCTION

Collocations, recurrent word combinations, have been widely recognized as an essential aspect of vocabulary knowledge and an important approach to support language production (Firth, 1957; Lewis, 2008; Nation, 2001; Nattinger & DeCarrico, 1992; Sinclair, 1991). It is widely recognized that collocations have particular importance for language learners. However, collocation knowledge is difficult to acquire even for advanced learners (Bishop, 2004; Nesselhauf, 2003) simply because there is so much of it.

Printed dictionaries and concordancers can help, but physical size restricts the number of collocations they provide. Computer-based concordancers are widely used by linguists for corpus analysis, but their interfaces are ill-suited to language learners. For example, formulating a command to retrieve verb collocates of a given word requires specialized knowledge—and varies from one concordance to another. As a palliative, some researchers advocate screening the concordancer output before presenting it to students (Varley, 2009), or providing a simplified but less powerful retrieval interface (Chen, 2011).

This paper describes a system (FlaxCLS) that extracts useful syntactic-based word combinations (e.g., verb + noun, noun + noun, adjective + noun) from Wikipedia articles (3 trillion words) and constructs a collocation database from it. The system provides a simple interface through which learners seek collocations that include any given word and word type (verb, noun, adjective and adverb), or search for combinations of multiple words (e.g., *take* and *role*). The system also retrieves semantically related words and collocations of the query term by consulting Wikipedia. For example, searching for organ donation yields related words like *donor*, *recipient*, *legal*, *ethical*, *transplantation*, and corresponding collocations like *potential donors, suitable recipient, legal death, ethical issues, organ transplantation*.

2. BACKGROUND

The importance of collocations in successful language learning has long been recognized (Palmer, 1933). Nation (2001) argues that language knowledge is collocation knowledge, because storing chunks of language in long-term memory forms the very basis of learning, knowledge and use. He supports Ellis's (2001) contention that language learning and use can be accounted for merely by associations between sequences of words, without any need to refer to grammatical rules.

Several researchers point out that many errors can be attributed to a lack of correct and appropriate use of collocations (for example, Arabski, 1979; Bahns and Eldaw, 1993; Marton, 1977). Hill (2000) further emphasizes the importance of collocation knowledge when developing accuracy of expression. Learners often use long, labored, clumsy sentences in speech and writing because they are unable to express complex ideas lexically. In many cases, the unnatural sentences or phases they produce can be succinctly replaced by collocations. Lewis (2000) suggests that teachers should encourage their students to build up so-called "islands of reliability"—formulaic chunks that often occur in fluent speech and academic writing. These help learners convey the central meaning of what they wish to say, particularly if it is complex.

Unfortunately, learning collocations is not as straightforward as one might assume. Collocation is a notoriously challenging aspect of English productive use, even for advanced learners (Bishop, 2004; Nesselhauf, 2003). Studies show that educated native English speakers know about 20,000 word families (Goulden et al., 1990). However, the size of their mental lexicon—stored as prefabricated multi-word chunks—is far larger than was first thought (Lewis, 1997). High frequency words make up about 80% of the words in running text, and the first 2000 words cover almost 90% of what we say and write (Nation, 2001). It is not so much the words themselves as the hundreds of millions of expressions, idioms, and collocations that make up the language of everyday use. The single most formidable task a learner faces is mastering a sufficiently large lexicon to achieve native-like fluency.

To make the situation more challenging, all lexical items, expressions, and collocations are arbitrary: they are conventionalized language that simply has been used for years. Very few are consciously learnt by native speakers. EFL (English as Foreign Language) students, are not constantly exposed to the language, as native speakers are. As a primary language source, they rely heavily on course books

This quality assured paper appeared at the 6th annual conference of Computing and Information Technology Research and Education New Zealand (CITRENZ2015) and the 28th Annual Conference of the National Advisory Committee on Computing Qualifications, Queenstown, New Zealand, October 6-9, 2015. Michael Verhaart, Amit Sarkar, Rosemarie Tomlinson and Emre Erturk (Eds).

from which many features of natural language have been removed (Lewis, 1997). Another difficulty that teachers and learners face is that there are few resources for checking which collocations are correct. Many non-native teachers still use out-of-date dictionaries rather than modern ones with many thousands of corpus-based examples. Few course texts address collocations explicitly, and most teachers are forced to rely on intuition (Conzett, 2000). Therefore we should not only teach collocations but also need to teach students how to use various learning resources and strategies to find collocations for themselves (Woolard, 2000).

3. COLLOCATION RESOURCES

Dictionaries are the traditional language learning resource for finding word definitions and common usage. But dictionaries are changing. With the wide recognition of the importance of collocation learning, modern dictionaries designed for language learners give increasing attention to collocations. For example, the Oxford Advanced Learners' Dictionary (6th edition, 2000) contains about 10,000 collocations. The BBI Combinatory Dictionary of English (Benson and Benson, 1986) focuses on "essential grammatical and lexical recurrent word combinations"; its revised version (1997) contains 18,000 entries and 90,000 collocations. It claims to convey information that cannot be found in other dictionaries for second language learners, such as which verbs are used with which nouns. The LTP Dictionary of Selected Collocations (Hill and Lewis, 1997), which aims to help intermediate and advanced learners make more effective use of the words they already know, groups collocations into noun, adjective and adverb sections. It identifies the five most important collocation types as adjective + noun, verb + noun, *noun + verb*, adverb + adjective and verb + adverb, and selects a headword as the entry point for each one.

Many recent collocation dictionaries are compiled from large corpora. The Oxford Collocation Dictionary for Students of English (2009) is based on the 100 million words in the British National Corpus, and covers over 150,000 collocations for 9,000 headwords. It includes a full range of collocations like

- fairly weak collocations: see a film and an enjoyable holiday
- medium-strong collocations: see a doctor and direct equivalent
- the strongest and most restricted collocations: see reason and burning ambition.

Collins Cobuild's English Collocations, published on CD-ROM, is derived from the 200 million words in Collins' Bank of English, and provides 140,000 collocations and 2,600,000 examples. It defines collocations as frequent word combinations, including idioms, phrasal verbs, compounds, fixed phrases and grammatical patterns. To find them, the user clicks one of a list of 10,000 English words to bring up the twenty most frequent collocates that occur before or after it. Clicking a collocate shows twenty randomly selected examples; and each example can be expanded to show more contexts. The restriction to just twenty collocates, which are often diluted by common words such as any, own, and new, is disappointing—particularly considering the huge volume of underlying text.

4. CONCORDANCERS

Concordancers, traditional linguistic tools, have become popular in helping language learners study collocations. A concordancer is "a piece of software, either installed on a computer or accessed through a website, which can be used to search, access and analyse language from a corpus" (Peachey, 2005).

One accessible and user-friendly concordancer, shown in Figure 1 and available on the Web, is the *Compleat Lexical Tutor* from Université du Québec à Montréal (Cobb, n.d.). Using this tool, students can enter a word and explore what words are most likely to occur before or after it. They specify a keyword to search for, and select one of a number of different corpora to search in. They can also associate another word with the keyword, specifying a position—left, right or any. The search results are chunks of text (constrained by line width) that contain the keyword and, if specified, the associated word. Figure 2 shows the result of searching for the word cause, which is underlined. A line width parameter determines the size of the context that is displayed (here it is 45 characters).

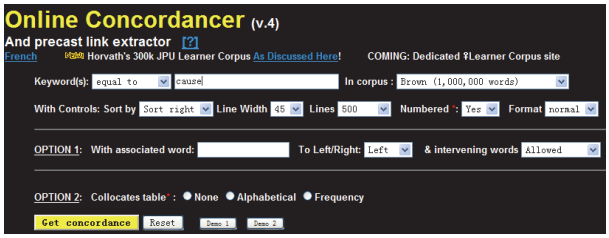

Figure 1. Online concordancer at *www.lextutor.ca*

1	n of material by hypervelocity impacts would	cause	a deviation from a linear relationship. In t
2	he vagina as the woman lies on her back) may	cause	a great deal of distress in a virgin. But du
3	vert, he devoted himself to the anti-slavery	cause.	A group of young men influenced by him enro
4	6-inch torrent of water is released, it may	cause	a lot of comment as it passes through or by
5	is no substitute for those intangibles which	cause	a man to sacrifice part of his earning poten
6	utter away the surface atoms of the dust and	cause	a slow diminution in size, with a resultant
7	rather than cuts for the texture (cuts could	cause	air pockets under the glaze creating pinhole
8	ct to stand or fall only by the merits of my	cause".	All seven recognized that independence
9	pressure and pain signals may involuntarily	cause	all the vaginal muscles to contract in an ef
10	ay fail to elicit a conditioned reaction but	cause	an increased synchrony instead of the excita
11	ve his wife the following description of the	cause	and consequence of diarrhoea: "I have bin a
12	ing action on the empirical determination of	cause	and effect provides a toughness and bravado
13	ituation in which he can perceive no visible	cause	and effect sequence, he should be alert to i
14	echanistic universe, governed by the laws of	cause	and effect, bound in chains of determinism t
15	any line of action related to the concept of	cause	and effect. He bases his approach on the bel
16	es with virtually no understanding of social	cause	and effect. Small wonder, then, that we fear
17	Pp. 64-66. Haydn C. Covington argued the	cause	and filed a brief for petitioner. H17 0550
18	e tiresome vocabulary of that lost and dying	cause,	and in the sprung syntax that is supposed t

Figure 2. Concordance Entry

More complex concordancers allow users to search using regular expressions or even discriminate between spoken and written language use. The British National Corpus website provides an example. Users use the equal (=) character to restrict the search by part-of-speech, and braces { and } to enclose a regular expression. Unlike the previous example, the result comprises a list of complete sentences, each with an associated sequence number—for example, AA9—that links to a page displaying a surrogate of the document containing the sentence, including the title, author, publisher and total word count.

Many online concordancers incorporate additional retrieval facilities. For example, WebCorp's Collocation Profile generates collocations of a given word by analyzing the first 500 web pages returned by Google's search engine for words that appear within four words of the target word, along with their frequency. The Corpus of Contemporary American English website retrieves collocations based on syntactic tags, e.g., one might specify a term along with "[v*]" to search for its verb collocates. These verbs are returned in descending order of frequency (invariably beginning with common ones such as is, was, have, are, had, before reaching verbs that are more specific to the term itself); clicking one displays concordance lines that contain both words.

Chen (2011) conducted a series of evaluations on how students and teachers retrieve collocations from three concordance systems, the Hong Kong Polytechnic Web Concordancer, the COBUILD Concordancer and BNC Sample Search tool. They identified these limitations:

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

- These tools are based on corpora that are too small for student writing needs
- They lack part-of-speech tagging and position options, e.g. seeking adjective collocates before the word improvement or nouns after enhance
- Lower and intermediate level learners find it hard to interpret concordance output
- Syntactic collocation patterns are hard to identify in search output.

He concluded that existing tools are unsuitable for retrieving collocations because they do not display whether the collocate precedes or follows the query word, nor its syntactic type (verb, noun, adjective, adverb etc.). Further improvements were suggested after he performed a further evaluation using a prototype system, WebCollocate, that overcomes some of above limitations—it was based on a large corpus (160 million words), and includes position information and part of speech tags. These included searching for phrases or multiple words, using the student's first language for queries, and clustering semantically related collocations.

5. THE DESIGN OF A COLLOCATION LEARNING SYSTEM

We have designed and constructed a collocation learning system (FlaxCLS), which organizes collocations based on syntactic patterns. This fundamental design decision is widely supported in the literature, and is adopted by the Oxford, BBI, and LTP dictionaries noted in Section 3. For example, Hill (2000) recommends drawing the learner's attention to collocations that follow particular syntactic patterns, such as adjective + noun, noun + noun, verb + adjective + noun. He stresses the power of nouns in selecting collocations: identify key nouns in the text and then look for noun, verb and adjective collocations. Wei (1999, p. 4) supports this approach, arguing that it incorporates syntax into a predominantly semantic and lexical construct, thus encompassing a wide range of data. Many researchers (for example, Chan and Liou, 2005, Chen, 2011, Wu. S. etc., 2009) have identified student errors in particular collocation patterns (e.g. verb + noun, adjective + noun).

The key issues raised during the design of FlaxCLS were these:

- What are the most useful collocation patterns for learners?
- How can we present the most useful collocations to learners?
- How can we expand learners' collocation knowledge?

Choosing collocation patterns. Table 1 shows the collocation types that we chose. As the examples illustrate, collocations contain from two to five continuous words, five being relatively rare. The fourteen types include five from the work of Benson, Benson and Ilson (1986), marked BBI in Table 1; three from the Oxford Collocation Dictionary, marked OCD; and a further six that we added ourselves, marked NEW. For example, the pattern gerund verb + noun (e.g. *the hotly debated issue*, *the driving issue*) is very useful, particularly in academic writing, but has been omitted from most dictionaries. As Table 1 shows, we extended some of these types to include more constituents of potential use to learners. For example, the noun part of a verb + noun collocation can include 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 very common adverbs like *more, much, very, quite* are removed from the patterns involving adverb because they can accompany most adjectives and verbs.

Presenting collocations. The main challenge when presenting the most useful collocations to learners is in organizing them to manage the massive volume of data, without overwhelming students. For any given query term there are up to fourteen collocation types; many words belong to more than one type because their syntactic part of speech is 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, we chose to show *take advantage of* and *have advantage of*, suppressing the others so that other useful collocations like *gain an advantage, saw the advantage of*, *offer the advantage of* move further up the result list.

To address these issues, we adopted a hierarchical organizational structure. Collocations are first grouped by the syntactic type of the query word (e.g., used as a noun or a 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 only see take advantage of in the result page, because it is the most frequent, but can click it to see 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, with the query term *benefit*, its collocations are first grouped under its noun and verb forms. The noun collocations are displayed first because they are more frequent than the verb ones. Within the noun group, adjective + *benefit*, noun + *benefit*, *benefit* + *of* + noun, *verb* + *benefit* … are presented in frequency order, and within each collocation pattern the most frequent collocation is always listed first. The same applies to the verb group.

Expanding learners' collocation knowledge. We have investigated several ways to help students expand their collocation knowledge, especially in domain-specific areas, or on topics related to what they are studying. It is important that students can look at words and their collocations that are semantically related to the query term. In practice, it is possible to build a topic specific corpus (e.g. about "nuclear weapons," for an essay assignment) and extract collocations and key words from it. We use the publicly available Wikipedia to explore the possibilities. Given a query term, FlaxCLS consults the Wikipedia Miner tool (Milne and Witten, 2013) to determine whether there is a corresponding Wikipedia article. If so, the key words and collocations extracted from that article are returned as suggestions to the user (see Section 7.3.and 7.4).

6. BUILDING FLAXCLS FROM WIKIPEDIA

We developed FlaxCLS from 3 million Wikipedia articles, downloaded from the Wikipedia website. These articles represent modern English in many areas e.g. art, life, science, and, importantly, emerging and contemporary topics whose vocabulary is not covered by any other standard corpora such as British National Corpus. It is particularly useful for seeking topic-related key words and collocations. We use the Wikipedia Miner tool (Milne and Witten, 2013) to retrieve individual articles which are then feed into the collocation identification process. The identification process involves eight steps:

1. Splits 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. Discard "dirty" collocations
5. Calculate frequency of each collocation,
6. Sort collocations by frequency for presentation to the user,
7. Associate sample text with the collocations that have been identified, and
8. Build search indexes.

Throughout this project we use the OpenNLP package for part-of-speech tagging. Released under GNU Lesser General Public license (available at opennlp.sourceforge.net), this is a collection of Java-based natural language learning tools that perform sentence detection, tokenization, part-of-speech tagging, and chunking. Step 1 and 2 consists of four sub-steps illustrated in Figure 3. The first detects sentence boundaries and splits the input into individual sentences. Then sentences are converted into tokens. The tokenizer separates punctuation: for example, *you?* becomes two distinct tokens *you* and *?*. It also detects contractions, that is, shortened forms in which a subject and an auxiliary verb, or an auxiliary verb and not, are combined into a single word, and splits them into two parts—for example, *I'm, we're, you'd, can't*. The result of these two steps on the text *"How are you? I'm fine."* is the following eight tokens:

How	are	you	?	I	'm	fine	.

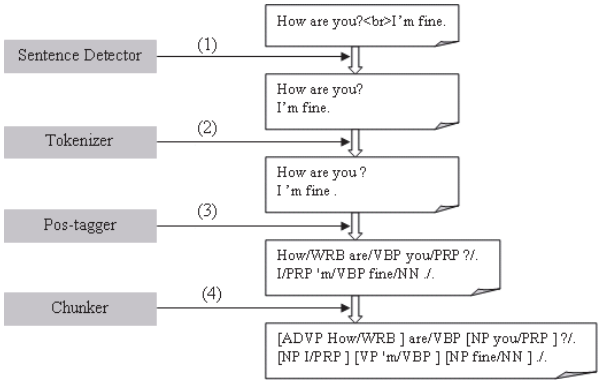

Figure 3. Parsing a document

Next the tagger performs tagging: it assigns a part-of-speech tag to each word. These tags begin with a letter that conveys the basic class and follow it with letters that qualify the class. For example, N… and V… indicate noun and verb; NN and VBP signal a singular noun and a non-third-person singular present verb. OpenNLP's tagger adopts the Penn Treebank tagset that comprises three levels: word, phrase and clause; we use only the word-level tags. Finally, the chunker assigns non-overlapping phrase and clause tags.

Then, in step 3, the tagged sentence are compared against regular expressions that specify the syntactic patterns in Table 1, and those that match are extracted as candidate collocations. For example, the pattern for verb + noun is:

*word/VB[DZP]? + (word/IN)? + (word/DT)? + (word/JJ)? + (word/NN[S]?) + (word/NN[S]?)**

A verb + noun collocation must begin with a verb (VB), which could be in base, past, or present form, followed by an optional preposition (IN), an optional article (DT), an optional adjective (JJ), a compulsory noun (NN) and optional nouns. Patterns that match any of the ten regular expressions are grouped by collocation type; ones that do not match are discarded.

Some extracted collocations are messy because they contain a haphazard mix of upper- and lower-case letters, unconventional single-character words (other than the article a or pronoun I) such as time *t*, *p* values, and *m* sections, or repeated words such as *part part*, *pain pain* and *man man*; Step 4 discards these "dirty" collocations because they are not useful for learning.

In printed dictionaries, collocations are organized by syntactic pattern and ordered in various ways. Some dictionaries show the most frequent or idiomatic ones first; others use arbitrary ordering. Given a list of collocations derived from the Wikipedia articles, our goal is to present good collocations at the top of the list and relegate poor ones to the bottom. To accomplish this, we tested the five standard statistical measures—*Frequency*, *t-test, Log-Likelihood Ratios*, *Pearson's chi-square test*, *Mutual information*— and selected the best for ranking extracted collocations. It turned out to be a particularly simple one—plain frequency of occurrence. Therefore, in step 5 and 6 collocations are sorted by frequency for presentation to the user without any manual selection.

Whenever a collocation is identified, its sentence are extracted and associated with it in step 7, 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 in step 8.

The indexes consist of index and dictionary files that are built for each collocation type and each constituent word of a collocation. A collocation type has two to four index files, each corresponding to a particular position in a collocation. For example, noun + noun has two index files, say *i0* and *i1*; where *i0* is for the first noun and *i1* for the second. The verb + noun and adjective + noun types have four index files because they are extended to include more components (see Table 1). Each word in an index file occupies one line: the word, the name of the dictionary file, and the most common collocation. A dictionary file contains all collocations of a particular word in a particular position, with their frequencies.

Table 2 shows excerpts from index and dictionary files: *i0* is the index file for the first words of adjective + noun collocations and *c029* is the dictionary file of the adjective front.

Table 2. Example of index and dictionary file

***i0* (index file)**		
word	dictionary file	most common collocation
front	c029	*the front line*
broken	c041	*a broken link*
c029 (dictionary file)		
collocation		**frequency**
the front line		1602
the front door		1582

We also extracted keywords from each Wikipedia article using a heuristic method commonly deployed in information retrieval (TF-IDF, described by, for example, Witten et al., 1999). First, documents are parsed, and the nouns, adjectives, verbs, and adverbs are designated as content words. For each such word, a score is calculated that reflects how important the word is to the document, based on the number of times it occurs in the document (which increases the score) and the number of times it occurs in the collection as a whole (which decreases it). This is used to rank words related to a query term (see Section 7.3).

7. USING FLAXCLS TO EXPLORE COLLOCATIONS

This section illustrates the use of FlaxCLS when seeking collocations of a particular word, exploring collocation expansions, looking at the original context, and retrieving related words.

7.1 Searching for collocations

To look up collocations, the user types in the word of interest and selects a database: standard, academic, or contemporary English. In Figure 1 the word is *research*. The system retrieves and displays collocations and other information about the word.

Inflected and derived forms (family words) of the query term appear first, along with its synonyms and antonyms. Clicking one of these will re-invoke a search using it as the query term. For the query research, the family words researched, researcher, researchers, researches and researching are displayed. A standard resource (WordNet) is used to identify words that are related to or associated with a particular query term. For research, verb synonyms include search, explore and investigate; noun synonyms include investigation, investigating, inquiry and enquiry.

As Figure 4 shows, 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; these are shown in frequency order. Figure 1 displays the first three most popular patterns: research + noun, adjective + research, and noun + of + research. The interface contains two columns: syntactic pattern and corresponding collocations. For each pattern, up to fifty

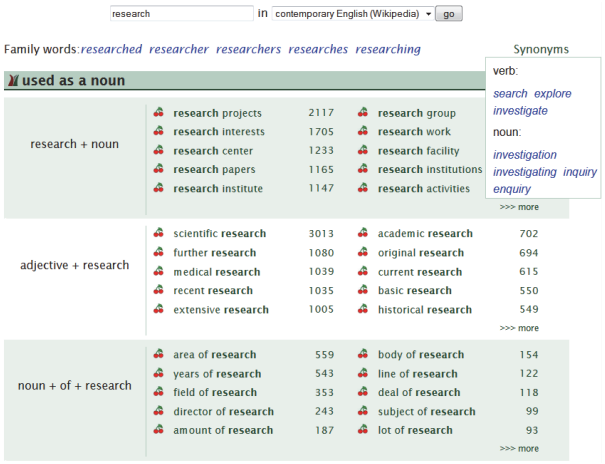

Figure 4. Family words, synonyms and collocations associated with the word *research*

collocation samples and their frequency are retrieved and displayed, ten at a time. Here, *research project*, *scientific research* and *area of research* are the most frequent collocations of the above three types. The more button at the bottom right reveals the rest.

Figure 5. Collocations similar to *scientific research*

Clicking one of these—in this case *scientific research*—brings up the superimposed window shown in Figure 5. It displays similar collocations in two columns, along with their frequency. These all contain the words *scientific* and *research*, whether adjacent or not; are of the adjective + noun type; and have up to five words. Many include more than one noun or adjective, such as *basic scientific research*, *scientific research organizations*, and *independent nonprofit scientific research institute*. Clicking a collocation retrieves samples in context from the original text: Figure 6 shows ten samples for *basic social research*.

Figure 6. Text samples of basic scientific research

7.2 Searching for phrases

Typing more than one word shows collocations containing all the query terms, irrespective of order. This is a good way to expand collocation knowledge by studying correct usage of combinations of the query terms. Many students have difficulty using correct grammar items like articles and prepositions—for example, the article between take and advantage, or prepositions following make sense. Searching for *take advantage* yields the list of expansions shown in Figure 7, indicating that no article intervenes between these two words, that associated adjectives include *full, maximum, unfair, greater, little*, and that the expressions are commonly following by the preposition *of*.

Phrase searching is another good way to expand collocation knowledge. Figure 8 illustrates the example *scientific research* used in the previous section. Typing these two words retrieves not only the three or four word adjective + noun collocations seen in the previous section, but also many other patterns: *research in scientific field* (noun + preposition + noun), *amount of scientific research* (noun + *of* + noun). Clicking the more button at the bottom reveals more collocations containing these two words.

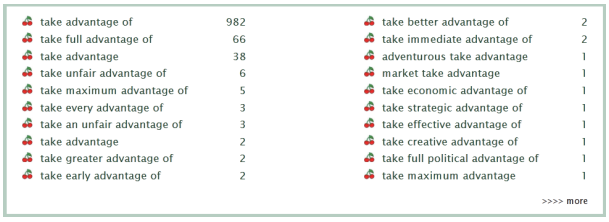

Figure 7. Collocations containing the words *take* and *advantage*

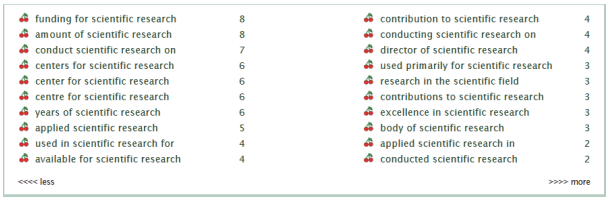

Figure 8. Collocations containing the words *scientific* and *research*

7.3 Exploring related words

Figure 9 shows the first 40 words related to the query *animal testing*, sorted by TF-IDF score, including *animal, primates, test, experiments, research, vivisection*, etc. Clicking one, say *toxicity*, reveals collocations associated with that word. The Figure gives some of its adjective collocates (*acute, general, chronic, embryonic*), verb collocates (*reflect, involve, evaluate*), and noun phrases (*toxicity tests, sign of toxicity, toxicity of a substance*). More words can be displayed by clicking the more button. Towards the end of the list the words becomes more general: for example, the last group of words related to animal testing are *population, line, end, series, form, play, have, be.*

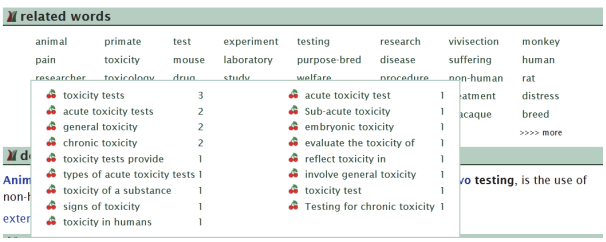

Figure 9. Words related to the topic *animal testing* and collocations associated with *toxicity*

To retrieve words related to a give word or phrase, the system first finds the best matching Wikipedia article using the Wikipedia Miner tool. 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. We use the one with the highest relatedness score, calculated using the method described by Milne and Witten (2012). Once an article is identified, the keywords and collocations of that article are retrieved, and collocations are grouped by the keywords they contain.

7.4 Linking to Wikipedia

The panel beneath the related words displays Wikipedia's definition of the query term, typically the first sentence or two of the article. Figure 10 shows the definition of animal testing; following that are the related topics in Wikipedia, Animal Liberation Front, Huntingdon Life Sciences, Animal rights, and so on, each hyperlinked to the corresponding Wikipedia article. Up to 50 topics are displayed for a query term, again sorted by the Milne and Witten (2013)'s relatedness score. Mousing over a topic gives its definition; clicking it leads uses that topic to retrieve collocations.

Figure 10. *Animal testing:* its definition and related topics in Wikipedia

8. CONCLUSION

Learning collocations is 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 language learners to study and explore collocations at their pace and in their own time. However, existing tools are designed for linguists or professionals, and learners face difficulties when using them. Effective collocation retrieval tools are required that are designed for language learners.

We have designed and built FlaxCLS, a collocation learning system that draws material from 3 trillion words Wikipedia articles. Collocations are retrieved simply by typing in the word or words of interest. To minimize the amount 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. FlaxCLS 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 for students 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.

FlaxCLS has been used by the Pathways College at the University of Waikato for language support for many Masters

and PhD students. We have received positive feedback from students and teachers. Anecdotal evidence suggests that the interface it provides is easy to use and students have found it helpful in improving their written English. 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.

9. REFERENCES

Arabski, J. (1979). *Errors as indicators of the development of interlanguage*. Katowice: Uniwersytet Slaski.

Bahns, J. & Eldaw, M. (1993).“Should we teach EFL students collocations? *System*, 21(1), 101–114.

Benson, M. & Benson, E. (1986). *The BBI combinatory dictionary of English: A guide to word combinations*. Amsterdam/Philadelphia: John Benjamins.

Bishop, H. (2004). The effect of typographic salience on the look up and comprehension of unknown formulaic sequences. In N. Schmidt (Ed.), *Formulaic sequences*: Acquisition, processing, and use, 227–244. Philadelphia, PA, USA: John Benjamins Publishing Company.

Chan, T-P & Liou, H-C (2005) Effects of Web-based Concordancing Instruction on EFL Students' Learning of Verb – Noun Collocations. *Computer Assisted Language* Learning, 18:3, 231-251.

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.

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

Ellis, N. C. (2001) Memory for language.” in P. Robinson (Ed.), *Congnition and Second Language instruction*. Cambridge: Cambridge University Press.

Fuentes, C.A. (2003). The use of corpora and IT in a comparative evaluation approach to oral business English. *ReCALL*, 15(2), 189–201.

Hill, J. (2000) Revising priorities: form grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching collocation*, 70–87, LTP, England.

Conzett, J. (2000). Integrating collocation into a reading and writing course” In M. Lewis (Ed.), *Teaching Collocation*, 70–87, LTP, England.

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells? *Teaching English as a second or foreign language*, 8(4). Retrieved March 12, 2009, from http://tesl-ej.org.ezproxy.waikato.ac.nz/ej32/a1.html

Goulden, R., Nation, P. & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, 11, 341–363.

Lewis, M. (1997). *Implementing the lexical approach: putting theory into practice*. Hove: Language Teaching Publications.

Lewis, M. (2000).Learning in the lexical approach. In M. Lewis (Ed.), *Teaching Collocation*, 155–184, LTP, England.

Hill, J. & Lewis, M. Eds. (1997). *LTP Dictionary of Selected Collocations*, LTP.

O’Sullivan, I., & Chambers, A. (2006). Learners’ writing skills in French: corpus consultation and learner evaluation. *Journal of Second Language Writing*, 15(1), 49–68.*Oxford Advanced Learners’ Dictionary* (6th Edition) (2000), Oxford University Press.

Oxford Collocation Dictionary for Students of English (2nd Edition) (2009), Oxford University Press.

Palmer, H.E. (1933). *Second interim report on English Collocations*. Tokyo: Kaitakusha.

Peachey, N. (2005). Concordancers in ELT. In British Council teaching English. Retrieved October 28, 2008, from http://www.teachingenglish.org.uk/think/articles/concordancers-elt.

Marton, W. (1977). Foreign vocabulary learning as problem no. 1 of language teaching at the advanced level. *Interlanguage Studies Bulletin*, 2(1), 33–57.

Milne, D. & Witten, I.H. (2013) An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, (194), pp. 222-239, January.

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

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

Nesi, H., & Gardner, S. (2012). *Genres across the Disciplines*. Cambridge University Press.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), 223–242.

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

Sinclair, J. McH. (2004). *Trust text: language, corpus and discourse*. Routledge, London.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language*, 10(1), 61–104.

Varley, S. (2009) I'll just look that up in the concordancer: integrating corpus consultation into the language learning environment. *Computer Assisted Language Learning*, 22:2, 133-152.

Wei, Y. (1999). *Teaching collocations for productive vocabulary development*. (Report No. FL 026913). Developmental Skills Department, Borough of Manhattan Community College, City University of New York.

Witten, I.H., Moffat, A. & Bell, T.C. (1999) *Managing gigabytes: compressing and indexing documents and images* (second edition). Morgan Kaufmann, San Francisco, CA.

Woolard, G. (2000). Collocation—encouraging learner independence. In M. Lewis (Ed.) *Teaching Collocation*, 28–46, LTP, England.

Wu, S., Franken, M., & Witten H.I (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.