Refining the use of the Web (and Web search) as a language teaching and learning resource

Shaoqun Wu[a], Margaret Franken[b*] and Ian H. Witten[a]

[a]Computer Science Department, University of Waikato, New Zealand; [b]School of Education, University of Waikato, New Zealand

[*]Corresponding author. Email: franken@waikato.ac.nz

Refining the use of the Web (and Web search) as a language teaching and learning resource

The Web is a potentially useful corpus for language study because it provides examples of language that are contextualized and authentic, and is large and easily searchable. However, Web contents are heterogeneous in the extreme, uncontrolled and hence 'dirty,' and exhibit features different from the written and spoken texts in other linguistic corpora. This article explores the use of the Web and Web search as a resource for language teaching and learning. We describe how a particular derived corpus containing a trillion word tokens in the form of n-grams has been filtered by word lists and syntactic constraints and used to create three digital library collections, linked with other corpora and the live Web, that exploit the affordances of Web text and mitigate some of its constraints.

Keywords: CALL, web collocations, web phrases, web pronoun phrases, Google N-grams

Introduction

In recent years, large corpora are beginning to be exploited in language teaching and learning (Yoon, 2008, p. 31). In fact as Chambers (2005, Abstract section, ¶ 1) states, "The potential of corpora as a resource in language learning and teaching has been evident to researchers and teachers since the late 1960s." At the present time, there are many free and easily accessible on-line resources for both teachers and students. For example, the Compleat Lexical Tutor from Université du Québec à Montréal is a comprehensive site that promises readers "data driven language learning on the Web" (http://www.lextutor.ca/). It allows students to exercise their vocabulary knowledge, access word frequency data, seek the meanings of words, and test their ability to detect grammatical errors. It allows teachers to generate cloze exercises, build hypertext resources, and construct quizzes to test students' knowledge of vocabulary in context. Several dictionaries offer free on-line (but often limited) access for seeking word meanings, collocations, and concordance entries. For example, the Collins website (http://www.collins.co.uk/corpus/CorpusSearch.aspx) includes a collocation function that makes use of several corpora including the Brown Corpus and British National Corpus, and offers concordance search that accesses the Collins WordbanksOnline English corpus. Meyer (2003) provides an inventory of useful corpora.

Because of its massive volume of natural text, researchers, teachers and learners are turning their attention to the Web. However, although it is clearly a potentially useful, and easily searchable, source of frequently occurring, authentic, and contextualized language samples, some writers have questioned whether or not it can be regarded as a legitimate corpus. It certainly fails to meet the rather specific criteria proposed by McEnery and Wilson (1996, p. 21), namely sampling, representativeness, finite size, machine readable form and a standard reference. However more inclusive and pragmatic definitions have been proposed. In his book on English corpus linguistics, Meyer (2002, p. xi) considers a corpus to be "a collection of texts or parts of texts upon which some general linguistic analysis can be conducted." For Kilgariff and Grefenstette (2003, p. 334), a corpus is any collection of texts that is "considered as an object of language or literary study."

This article explores the use of the Web and Web search as a resource for language teaching and learning, and describes ways in which both can be refined to serve that purpose better. As a corpus, the Web has unique features shared by no other corpora. We begin by identifying them and exploring how they both afford and constrain language study. We go on to describe how a particular derived corpus has been used to exploit the affordances of Web text and mitigate some of its constraints. The corpus in question contains a trillion word tokens in the form of n-grams (made available by Google). We filtered it and linked it with other corpora in order to enlarge the limited linguistic context that n-grams provide. Because of its vast size and all-encompassing generality we focused on particular language learning issues and created three separate sub-collections with tailored searching and browsing facilities. With the first, learners can explore word sequences associated with personal pronouns: ones starting with the word *I* appear to be particularly productive. With the second, learners explore collocations represented by syntactic patterns, drawing on a vastly greater database of examples than other systems. With the third, learners check word sequences against general usage on the Web.

The Web as corpus

The most striking, and perhaps the most compelling, feature of the Web for language teachers, and developers of language teaching resources, is its size. However, this brings its own problems. Web contents are heterogeneous in the extreme, uncontrolled and hence "dirty", and exhibit features different from the written and spoken texts in other linguistic corpora.

Size

The Web far outstrips any existing corpus and grows on a daily basis. Kilgariff and Grefenstette show this in their comparison of frequencies of a set of English phrases. For example, the phrase *perfect balance* occurs in the British National Corpus 38 times, as compared with 355,538 in Spring 2003 using AltaVista as the search engine (Kilgariff & Grefenstette, 2003, p. 337) and 3,800,000 today (Summer 2008, using Google).

The continual addition of new text has drawbacks, however, for it makes individual search results inconsistent and unstable. Indeed, Biber and Kurjian (2007, p. 112) observe that "linguistic patterns observed on the Web can vary radically - and seemingly randomly - from one search to the next". Therefore, when teachers set certain kinds of exercises involving direct Web search they cannot rely on predicting what they will retrieve or knowing exactly what their students will see. This is a serious disadvantage.

Representativeness

Most corpora are based on particular domains, genres, or collections of certain types of documents from which recurrent phrases and grammatical patterns can easily be retrieved (Stubbs & Barth, 2003). However, this certainly cannot be said about the Web taken as a whole. More than a decade ago, Kessler, Nunberg and Shütze (1997) characterized it as "a large and heterogeneous search domain". Since then it has grown many-fold in both size and diversity.

Biber and Kurjian (2007) recognize that "identifying 'register' or genre is an especially important consideration for linguistic research based on the Web" (p. 110), but acknowledge the difficulty of doing so. Search engines and other portals impose various taxonomic structures on Web items and resources. As Meyer (2002) notes, Yahoo categorizes documents and websites into fields such as *Arts and Humanities* and *Science Education*, each having further subcategories - both in terms of content itself, and of information sources such as journals or magazine articles. Similarly Robb (2003) explores limiting searches to within particular 'educated' domains using site names ending in "edu", "ac.uk", "edu.au" and "jp". However, these categories are still broad and not particularly useful for language study.

Biber and Kurjian (2007) used the two categories *Home* and *Science*, with their respective subcategories, to explore linguistic differences amongst Web-based texts. They conclude that there is wide variation within each category and subcategory, and substantial overlap in the occurrence of a large number of linguistic features. In other words, the categories imposed by search engines reflect little or no consistency between the genres of the documents that fall under them.

To what extent does the hypertext found on the Web resemble or differ from those in traditional hardcopy form? Meyer (2002, p. 63) asks the question in this way: "Are electronic texts essentially the same as traditionally published written texts?" Apart from on-line journals, newspapers, and advertising material, most of the text on the Web has not been subjected to any editorial process - for example, documents posted on personal home pages or constructed on blogs. This is in clear distinction to traditional commercially published text, for which the economics of publishing dictate quality control mechanisms that affect and to some extent normalize the writing style.

According to Biber and Kurjian's (2007) study, identifiable Web-based text types include: personal, involved, stance-focused narration, persuasive/argumentative discourse, addressee-focused discourse, and abstract/technical discourse. Two of these types (personal, involved, stance-focused narration; and addressee-focused discourse) appear particular to the Web. Some features that characterize the former are: first person pronouns; mental verbs such as *think*; certainty adverbials such as certainly, definitely, surely and undoubtedly; *that*-clauses; the pronoun *it*; and past tense. Some that characterize the latter are: second person pronouns, progressive verbs, desire verb + *to*-clause (Biber & Kurjian, 2007, p.116).

The complexity and variety of Web text means that searches will produce results that are anomalous with those obtained by searching corpora based on written material, which are necessarily focused and selected - and even with those based on spoken material.

Cleanliness

The Web contains a huge number of language errors such as grammatical and spelling mistakes, not to mention the use of unusual and less acceptable collocations. Kilgariff and Grefenstette (2003, p. 342) describe it as a "dirty corpus". This represents a rather serious constraint on its use for language learners, because a fundamental requirement for such texts is that they represent exemplary models of language. One response to this constraint is limiting searches to "impeccable sources" (Robb, Possible approaches

section, ¶ 1, 2003). Robb describes how to use Project Gutenberg, a huge collection of "e-texts of material that is out of copyright, particularly works of literature and texts of historical value" (Robb, Procedures section, ¶ 1, 2003).

Using the Web corpus

The Web has often been used in linguistics research to corroborate intuitions about the frequency of individual words, collocations, phrasal verbs, and idioms. As many researchers have noticed, it is a particularly valuable source of information about collocations. For instance, Guo and Zhang (2007, p. 748) suggest that it provides a convenient platform for investigating and verifying the "frequency, context and source of a combination."

Perhaps more significant than the Web *per se* is the use of search engines as a resource for language teaching and learning - as indicated by the recently coined neologism GALL, for Google Assisted Language Learning (Chinnery, 2008; Shei, 2008). Google appears to have become the search engine of choice for this purpose—partly because it can do far more than just search. As Chinnery explains, Google "has the capacity to do much more than simply facilitate basic Boolean searches" (2008, p. 3), and surveys the range of specific search tools, from a *define* command that finds definitions of words to more complex operations such as issuing a search in one language to find pages in another language, and having the results automatically translated back to the original language (Google Language Tools).

Shei (2008) used Google search results to identify the occurrence counts of consecutively truncated subsequences. This lets users study particular words and phrases to check the extent to which the text they have written represents common usage. He devised a visual tool that represents the frequency of sequential word combinations, and their subsequences. For instance, in the sequence *have been found to be infected with*, the subsequence *have been* is very much more frequent than *have been found*. Of course, frequencies inevitably become smaller as more words are included in the analysis; the point here is that the two-word sequence is *much* more frequent than the three-word one. Shei suggests that learners may use this to guide their choice of collocations. For instance, *have been found to be infected with* is a much more common collocational string than *have been found to be polluted with*.

Using search services

The ordinary facility of Web search, that search engines provide for free, can underpin valuable and imaginative services. Guo and Zhang (2007) demonstrate how search capacity can be enhanced to generate collocation and concordance data from the snapshot lines returned in search results. They combine advanced options like phrase and wild-card search into a simple interface that retrieves concordance entries live from the Web and presents them to users.

Unfortunately, this approach has a disadvantage for large-scale use: search engine companies do not support the use of their services through secondary interfaces. The reason is presumably because they wish, quite reasonably, to protect themselves from people who piggyback on their search engine to offer services that may enhance or compete with their own.

There are other practical disadvantages of using commercial search services in this way. A minor one is that the frequency counts that search engines return for words and phrases are only approximate, though they are probably a good enough indication for language learning purposes. Far more serious is the fact that although arrangements can sometimes be made with search engine companies for limited experimental usage for research purposes, these are restricted to a certain number of queries per day, which would be insufficient to support concordance-style services on a satisfactory, scalable basis.

Using Web n-grams

Instead of relying on live Web searches to generate collocation and concordance data, we work with an off-line corpus generated and supplied by Google (2006). This contains short sequences of consecutive words, called "n-grams," along with their frequencies. Unigrams comprise one word, bi-grams two, tri-grams three, and so on. The corpus contains all of these up to and including five-grams. Using this resource is an innovation that mitigates some of the constraints associated with the Web as corpus. It also provides a sound basis for operating scalable services that use Web text as a resource for language teaching and learning.

The corpus is a vast set of word n-grams in the English language, along with their frequencies. The text was collected in January 2006 from publicly accessible Web pages. The n-grams range from single words (that is, unigrams) to units of five words (5-grams). The corpus was generated from approximately one trillion word tokens of text on publicly accessible Web pages—a staggeringly large body of natural English. N-grams that occur less than 40 times were discarded (by Google, before publishing the corpus). Even so, the material comprises approximately 24 GB of compressed text files.

Table 1 summarizes its size. The number of n-grams increases as n grows beyond 1, peaks at n=4, and then begins to decay. In the files that Google supplies, each n-gram occupies one line, as in:

word_1 <space> word_2 <space>… word_n <tab> *count*

where *count* is the number of occurrences of this n-gram.

Table 1. Number of units in the n-gram corpus

Tokens	1,024,908,267,229	10^{12}
Sentences	95,119,665,584	0.95×10^{9}
Unigrams	13,588,391	0.014×10^{9}
Bi-grams	314,843,401	0.3×10^{9}
Trigrams	977,069,902	1.0×10^{9}
Four-grams	1,313,818,354	1.3×10^{9}
Five-grams	1,176,470,663	1.2×10^{9}

While the number of units in the Google n-gram collection is also vast, the units themselves are of a size that can be exploited by teachers and learners seeking to integrate "corpus consultation" (Chambers, 2003).

Cleaning the data

We found it necessary to clean up this corpus in order to make it suitable for language learning. This process had the useful side benefit of reducing its massive size to more manageable proportions.

Like the Web itself, the n-grams are messy. They include many non-word character strings, website names and grammatical errors. Unfortunately, it is virtually impossible to eliminate grammatical errors. Deficiencies in natural language processing technology makes analysis difficult and somewhat unreliable, but - more importantly - the fact that no context is available beyond the neighboring few words makes accurate parsing impossible in principle.

Nevertheless, great improvements can be made by cleaning up the text. We used the British National Corpus wordlist to remove non-words and website names. Discarding word sequences if they include words not in this list reduces the volume of text by 30%. It yields a much tidier corpus, but has the unfortunate effect of removing sequences containing neologisms (often ones coined since the British National Corpus was constructed), notably, for example, the word *google*.

We built three digital library collections from this dataset, and undertook further selection and cleaning for each one. We describe this later when introducing the collections.

Linking to external resources

For language learners, n-grams have the intrinsic limitation that context is lost when they are removed from the original text. Context has long been recognized as crucial for vocabulary learning (see Nagy, 1997, for an in-depth discussion of its importance). Our remedy is to use text retrieved from two sources to reconstruct suitable contexts and present them to users on demand.

The first, and (to use Robb's (2003) notion) "impeccable", source is the British National Corpus. We split this into paragraph units and built them into a searchable collection using the Greenstone digital library software.[1] Whenever the learner asks to see examples of a particular n-gram in context, we arrange for Greenstone to search the collection for occurrences and display the relevant paragraphs.

The second source is the Web. We wrote a program that, whenever a language learner requests the context of a particular n-gram, connects to a search engine, uses the words as a phrase query and retrieves sample texts in real time. We used Yahoo as the search engine because Google, as noted above, imposes some limitations, and disables automatic queries from computer programs other than Web browsers as discussed above.

---

[1] We used Greenstone version 3.03; see http://www.greenstone.org.

Yahoo has no obvious disadvantages in terms of the quality of text snippets retrieved for a particular search.

Figures 1 and 2 show samples retrieved from the Web and the British National Corpus for the phrase *I was a little disappointed*. The contemporary nature of the snippets in Figure 1 is apparent from the fact that two of the eight report the feelings of an unsuccessful 2008 American Idol contestant. Many more examples of this phrase are available on the Web and can be obtained by clicking the *next* button at the bottom of the page. In contrast, the phrase has only ten British National Corpus hits in total, of which five are shown in Figure 2. They tend to be more coherent than the Web snippets, and are presented in a fuller context.

Both sources have limitations, and the two are somewhat complementary. The British National Corpus provides far fewer examples, the number declining rapidly for longer sequences. In many cases there are none at all—even for items that occur reasonably frequently on the Web. For example, *I was very disappointed in* occurs 12,000 times in the n-gram corpus but not at all in the British National Corpus. On the other hand the Web text, being extracted from individual Web pages rather than the aggregations in the n-gram corpus, is often unclean, incomplete and repetitive.

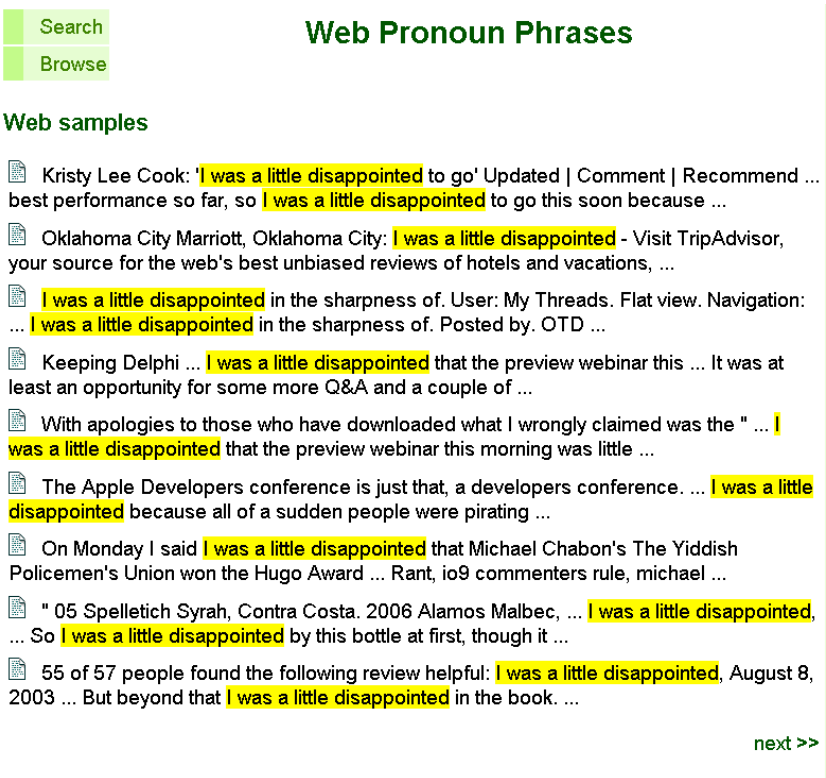


Figure 1. Samples retrieved for *I was a little disappointed* from the Web

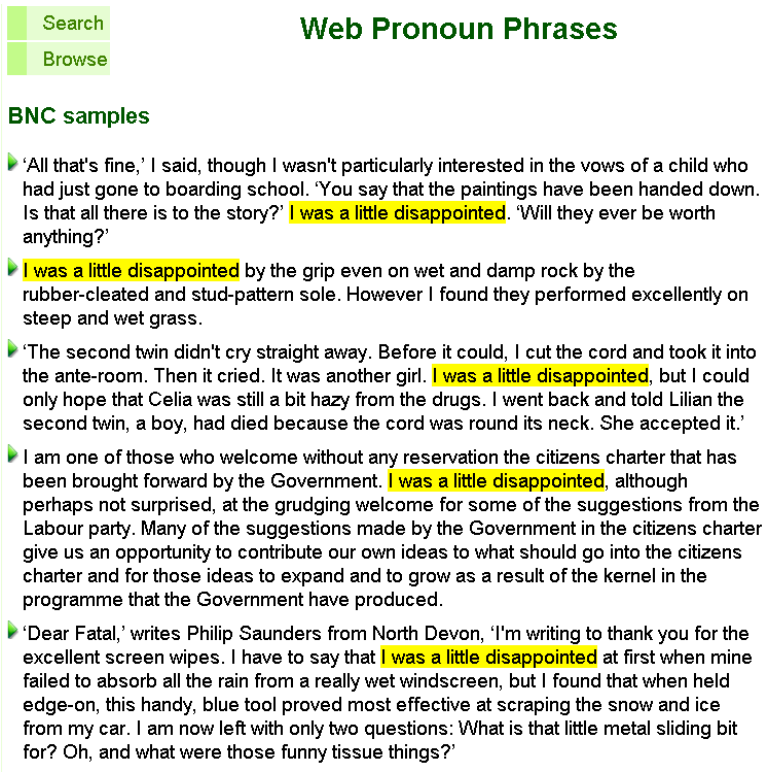


Figure 2. Samples retrieved for *I was a little disappointed* from the British National Corpus

Imposing order

The n-gram corpus was filtered as explained above, and had already been reduced by Google to eliminate items that occur less than 40 times. Nonetheless, it is a rather unstructured database and people are easily overwhelmed by the sheer number of textual examples that result from searching it.

We used the Greenstone digital library software to organize, design and build three digital library collections from different parts of the information, and serve them on the Web. These are a pronoun phrase collection, a collocation collection, and a full phrase collection.

It should be noted that the potential exists to build any number of other collections or sub-collections, tailored for different teaching purposes or student groups. For instance a small sub-collection could be built for epistemic adverbs, such as *certainly* or *probably*, identified by Biber (2006) as occurring frequently in university spoken and written language. Such a collection is potentially very useful for students in EAP courses and those preparing for university study. Sub-collections that focus on a particular domain such as quantification words could support theme or function-oriented vocabulary learning. Sub-collections can also be easily built to cater for students with different levels of vocabulary size. For example, wordlist based sub-collections can be built by referencing to wordlists (such as those refined and used by Nation - the 1000, 2000 and academic word lists; and posted on the Compleat Lexical Tutor site,

(http://www.lextutor.ca/). These will eliminate low frequency items and help students to prioritise which words to attend to.

Pronoun phrases

The aim of the first digital library collection is to allow learners to study pronoun phrases in association with particular lexical items—colligational patterns. Nattinger and DeCarrico (1998, p. 178) explain that colligations are "generalizable classes of collocations, for which at least one construct is specified by category rather than as a distinct lexical item." This makes colligational patterns more free and less predictable than collocations, which (according to the same source) are "roughly predictable yet are restricted to certain specified items and thus are nameable by words." But it also makes them an important unit of analysis for language study.

We use the term "*I*-gram" for sequences that begin with the first person singular pronoun. Suppose the learner wants to write a personal statement—an *I*-gram—to express disappointment. Figure 3 shows the search results for the word *disappointed*. It shows *I*-grams that contain the word *disappointed* in inverse frequency order, grouped by tense—past, present perfect, present, future and modal. The most common sentence head is *I was a little disappointed* (47,000 occurrences), a past tense usage. The top two usages involve the hedges *a little* and *a bit*, which is useful pragmatic as well as grammatical and lexical information.

Building and using the collection

To create this collection we began by identifying n-grams that commence with the pronouns *I, he, she, you, they, we*, and *it*. We used 5-grams because these provide the largest context. Two selection steps were applied:

- select 5-grams that start with a pronoun word;
- discard grammatically incorrect sequences.

In the second step, the raw n-grams are parsed by a natural language processing tool (OpenNLP[2]) for grammar checking. For each pronoun phrase set, four wordlists were generated and sorted into inverse frequency order:

- all words regardless of type;
- main verbs;
- main adjectives;
- modal words.

The digital library collection was configured with browsing facilities that allow users to examine these wordlists. Figure 3b shows the beginning of the list of *I*-phrases. Interestingly, *think* is the most frequent word that follows *I*, and the next four most frequent verbs are *have, know, want, like*. Figure 3c gives the colligational patterns that are associated with *think* in the first person context.

[2] http://opennlp.sourceforge.net

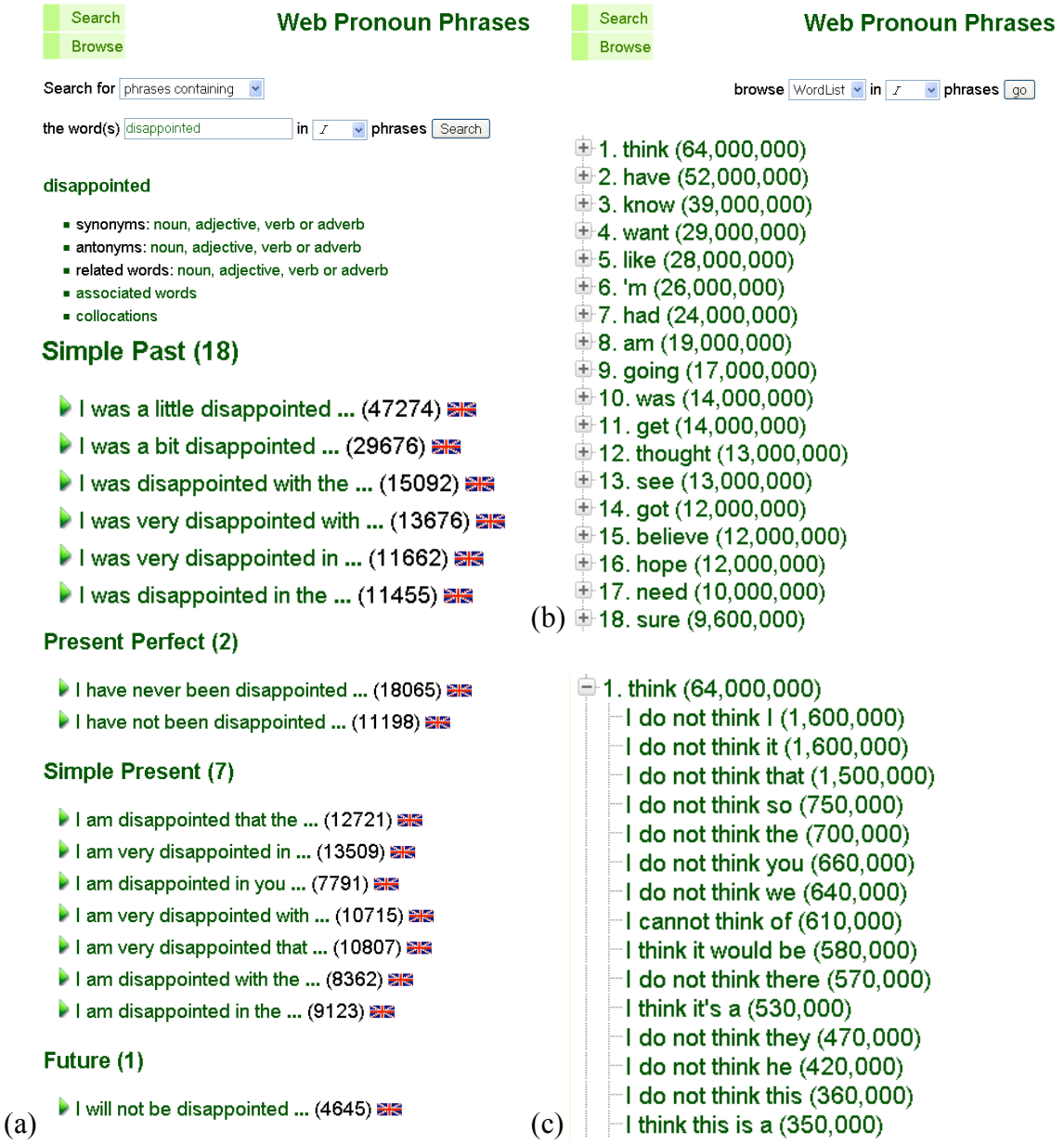


Figure 3. Searching and browsing the pronoun phrase collection

For the colligation structure corresponding to the first person singular pronoun followed by a verb (*I* + verb), *think* is retrieved as the most frequent verb. This corroborates the findings of Biber and Kurjian (2007) that frequently occurring linguistic features associated with personal involved narrative texts on the Web are the first person pronoun *I*, mental verbs such as *think,* and *that*-clauses. It also aligns with Biber *et al.*'s (1999) earlier finding that the most frequent lexical bundle in conversation consists of a subject pronoun (first person) and a verb phrase to express a personal opinion such as in the phrase *I think that*, *I think he*. However, neither study exposes the surprising fact that the pattern *I* + *think* occurs most frequently as a negative statement.

Evaluation

Evaluations are being conducted with several classes of General English learners in the context of personal writing tasks. Observations suggest that proficient learners can use the collection to generate text as well as revise it, whereas their more limited vocabularies may restrict learners at earlier stages to revising already written texts. However, most learners experience positive effects at the lexical, grammatical and perhaps most saliently the pragmatic level, although these are hard to measure. The text they produce appears to be more 'native-like'.

Collection 2: Collocations

We define a collocation as a sequence of words that come together more often than chance. Of course, there are many other definitions, but this statistical one is appropriate here because we identify and rank collocations based on statistical measures. There are many methods for finding collocations. Simply listing the most frequent word combinations does not work well because these tend to be overwhelmed by small structural expressions involving function words alone. The normal approach is to overcome this by applying more sophisticated statistical tests, such as the t-test, which take account of the frequency of the constituent words. However, function word combinations can be effectively filtered out by restricting collocations to certain syntactic patterns, which is desirable anyway in our application. Given such a filter, Justeson and Katz (1995) produce surprisingly accurate collocation results using frequency alone, and this simple method works reasonably well in the evaluation by Thanopoulos, Fakotakis and Kokkinakis (2002) of collocation extraction techniques. Given that our n-gram corpus is far larger than the ones used previously, and Banko and Brill's (2001) observation that "huge datasets trump sophisticated algorithms," we decided to use plain frequency as the web collocation extraction metric.

We are particularly interested in word combinations that constitute nouns, adjectives, verbs and adverbs and follow eight syntactic patterns. Table 2 gives these patterns, and samples of them.

Table 2. Collocation types, with samples

collocation type	sample
verb + noun(s) Includes verb + noun + noun verb + adjective + noun(s) verb + preposition + noun(s)	*make appointments* *cause liver damage* *take annual leave* *result in the dismissal*
verb + adverb	*apologize publicly*
noun + noun	*a clock radio*
noun + verb	*the time comes*
noun + of + noun	*a bar of chocolate*
adjective(s) + noun(s) Includes adjective + noun + noun adjective + adjective + noun(s) adjective + and/but + adjective+ noun(s).	*a little girl* *a solar water system* *a sunny beautiful day* *a funny and cute boy*
adverb + verb	*beautifully written*
adverb + adjective	*seriously addicted*

Six of the patterns were adapted from Benson and Benson (1986), and two more were added: noun + noun and adverb + verb. To make full use of n-grams, the verb + noun and adjective(s) + noun(s) patterns are extended to include more constituents that are of potential use for learners. These extensions are shown in Table 2.

Building and using the collection

The Web collocations were extracted using two- to five-grams. The identification process involved three steps:

- assign syntactic tags to the words of n-grams,
- match tagged n-grams with the syntactic patterns, and
- discard ones that occur less than 100 times.

An interface was built to allow learners to search the collocations by syntactic pattern. In practice, the interface allows users to start by specifying a word and consulting the British National Corpus's word type database for the types that match it, and then choose one to continue with. Figure 4 illustrates the process with the word *cut*.

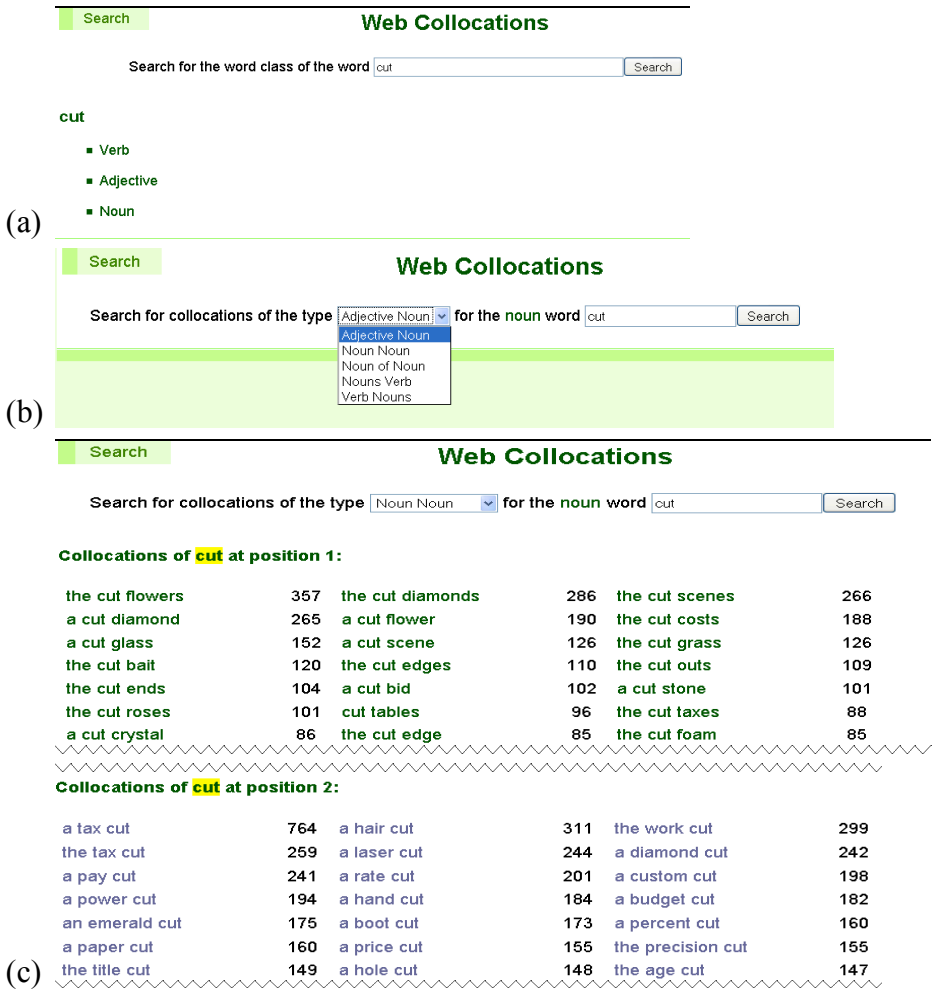


Figure 4. Searching for collocations

First the user learns that it can serve as verb, adjective and noun. Clicking the *noun* link brings up the page in Figure 4b, from which the user selects a collocation type to proceed. In this case there are five possibilities: adjective + noun(s); noun + noun; noun + *of* + noun; noun + verb; or verb + noun(s). This list varies from word to word, depending on the availability of collocations for a given word type. Figure 4c shows the noun + noun collocations of *cut* at the first and second noun-word position. The top two collocations are *the cut flowers* and *a tax cut* respectively.

Evaluation

The primary obstacle to evaluating this collection is finding an authoritative database to serve as ground truth. The Collins (2008) collocation sampler seems ideal, but its output is restricted to 100 collocates regardless of their word types. The online Compleat concordancer (Cobb, 2008) is one of the best on the Web, and its use is free, but it is based on a collection of rather small corpora ranging from 80,000 to four million words. After investigation, we decided to build a baseline collocation database from the 100 million words of text in the British National Corpus. We applied the same collocation extraction algorithm as described above to this text and used all extracted collocations for the evaluation. No collocations were discarded, because most of them occur only once.

Table 3. Collocation types with statistical data from two corpora

	Web n-grams			British National Corpus		
collocation type	collocations	words	collocations / main word	collocations	words	collocations /main word
verb + noun(s)	7,000,000	23,000	300	1,700,000	20,000	85
verb + adverb	370,000	11,000	34	140,000	10,000	14
noun + noun				660,000	40,000	17
noun + verb	440,000	31,000	14	180,000	21,000	9
noun + of + noun	5,000,000	32,000	156	810,000	24,000	34
adjective(s) + noun(s)	6,200,000	30,000	207	1,900,00	36,000	53
adverb + verb	320,000	3,600	89	210,000	4,300	49
adverb + adjective	370,000	3,800	97	130,000	4,000	33

We evaluate the Web collocation collection by comparing it with the collocations extracted from the British National Corpus, both in quality and quantity. The results underscore the massive and diverse nature of the Web collection. For each corpus and collocation type, Table 3 shows the total number of collocations, the number of words, and the average number of collocations for each main word. The "main" word is identified (manually and somewhat arbitrarily) as the most important word in that collocation type. For example, the verb is chosen for verb + noun(s) and the second noun is chosen for noun + noun. The intention is to give some idea of how many collocations exist given a particular word. In practice, however, users are allowed to search for any part of a collocation, not just the main word.

As the Table shows, the collections cover a similar number of words, which is not surprising because the Web n-grams have been filtered by applying the very same wordlist. However, 2–6 times more collocations were extracted from n-grams than from the British National Corpus—despite the fact that Web collocations whose frequency fell below 100 were discarded—and a similar increase is reflected in the average number of collocations available for a particular main word.

Table 4 shows the top ten *cause* + noun(s) collocations from the Web n-gram collection alongside those from the British National Corpus; we also include results from the online Compleat concordancer for reference. The first contains 15,000 collocations, which were extracted from the n-gram collection with a frequency cut-off of 100. The second has 2200, of which 84% occur only once and 8% twice. The third has 54, most of which appear just once. Interestingly, *cause problems* is the most frequent entry in all three cases. Upon further examination, it seems that *cause* is used mostly in a negative sense and associated with problems, damage, decease, and so on.

Table 4. Top ten *cause* + noun collocations in three concordances

Web n-grams 15,000 collocations		British National Corpus 2200 collocations		Compleat concordancer 54 collocations	
samples	frequency	samples	frequency	samples	frequency
cause problems	1,800,000	*cause problems*	163	*cause problems*	5
cause actual results	1,600,000	*cause trouble*	71	*cause suffering*	4
cause damage	920,000	*cause damage*	48	*cause damage*	2
cause harm	570,000	*cause difficulties*	40	*cause offence*	2
cause injury	420,000	*cause cancer*	35	*cause death*	2
cause cancer	410,000	*cause injury*	32	*cause distress*	2
cause death	320,000	*cause death*	28	*cause a great increase*	2
cause confusion	310,000	*cause confusion*	27	*cause another war*	1
cause a denial	280,000	*cause harm*	23	*cause deactivation*	1
cause a lot	250,000	*cause offence*	22	*cause a deviation*	1

The Web n-gram collocations demonstrate great diversity in the language patterns they represent. For example, there are 268 variations of c*ause problems*, including *cause serious problems*, *cause major problems* and *cause unpredictable problems*, while the British National Corpus contains only 56 variations, half of which occur only once. Table 5 below gives four more examples. While the sheer volume of examples could present a challenge for less proficient learners, we believe that it is very valuable for more advanced learners who wish to expand the range of collocational options so as to be able to express propositions in quite a specific precise and authentic way.

Table 5: Web and British National Corpus entries for a collocation template

Collocation	Web	BNC	Examples
cause + problems	268	56	*cause serious problems, cause major problems*
cause + damage	248	54	*cause permanent damage, cause significant damage*
cause + harm	146	24	*cause irreparable harm, cause no harm*
cause + injury	90	14	*cause physical injury, cause substantial injury*
cause + death	68	14	*cause sudden death, cause premature death*

Some Web collocations are anomalous because the processing is constrained by the length of n-grams. Parsers work at the sentence level, using context to infer each word's syntactic tag—for example, whether *cut* is being used as verb or noun. No automatic parsing technique is perfect, and errors occur more frequently on n-grams because of the restricted context. This results in incorrectly identified collocations. It also accounts for partial collocations like *a beautiful skin* and *cause different side*, which should be *a beautiful skin color* and *cause different side effects* respectively.

Collection 3: Full phrases

The third digital library collection we have built contains all one- to five-grams, after filtering out non-word strings and website names. This is the largest subset of the original n-gram collection. It contains about 50,000 unique words, 14 million two-grams, 420 million three-grams, 500 million four-grams and 380 million five-grams. It allows free exploration of combinations, unconstrained by grammatical class. We build on Shei's (2008) innovative work, described earlier, which allowed users to study particular words and phrases to check whether and to what extent the text they have written represents common usage.

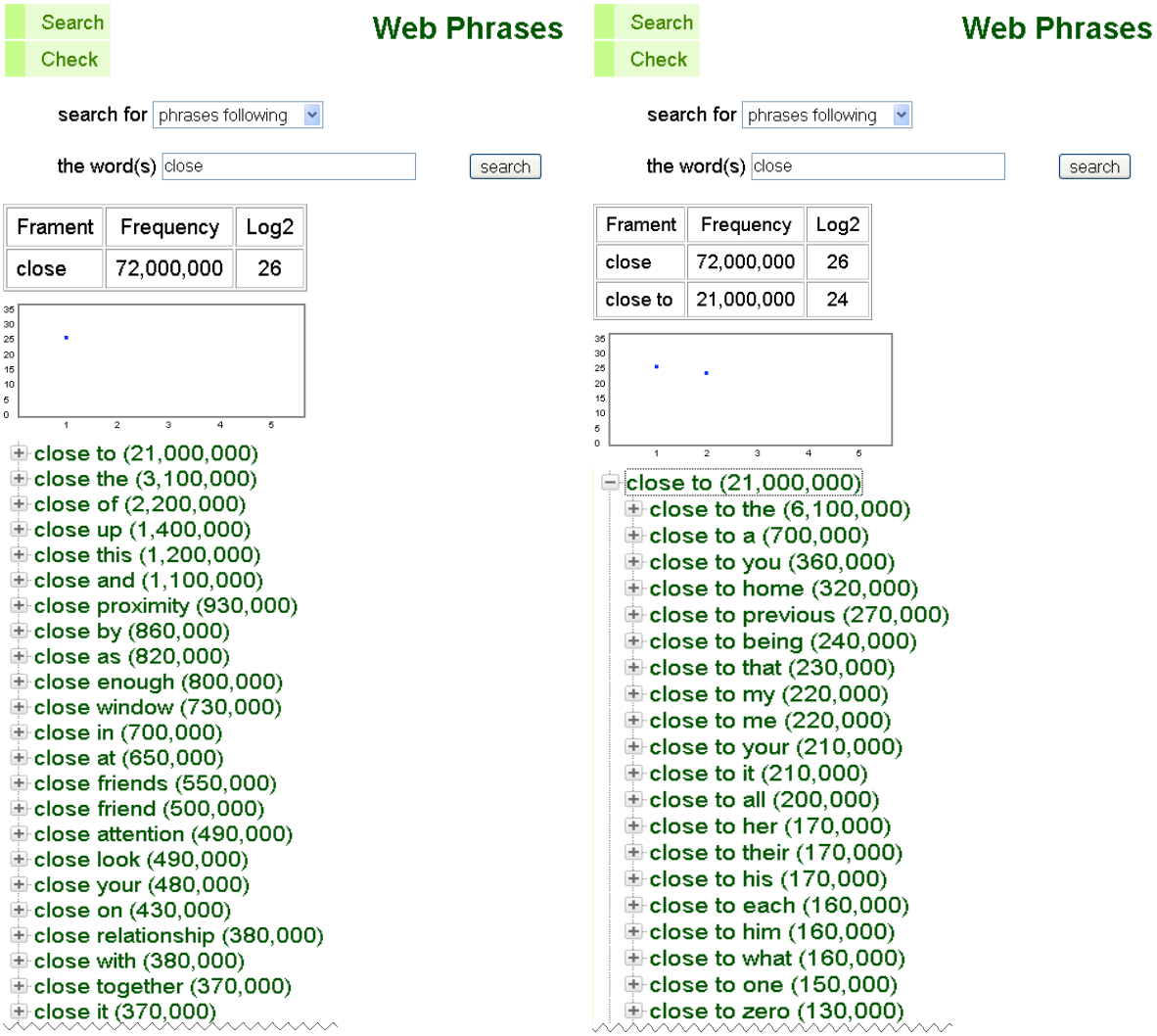
Searching word combinations

Users often want to know what words most commonly follow a particular word. Figure 5a illustrates this for the word *close*. The interface contains three parts. A statistical table gives the frequency count, and its base-2 logarithm, for the query term or phrase. Below is a graph that indicates visually how the frequency (represented by its logarithm) decays as words are added. Underneath is an expandable tree that displays associated phrases in reverse frequency order, along with the logarithm of their frequency count.

The most frequent word following *close* is *to.* Clicking *close to*, the tree expands and displays the phrases associated with this phase, as shown in Figure 5b, and the table and graph update accordingly. A phrase can be expanded up to five words, or until no further extensions are found in the collection. Samples of text that use the phrases can be retrieved from the Web, and from the British National Corpus, as described above for the pronoun collection.

Although Figure 5 illustrates a single-word search, users can specify phrases as well. Furthermore, extensions can be displayed in both forward and backward modes. For example, one could browse around successive words that *precede* a particular word or

phrase instead of ones that follow it (as in Figure 5). Interestingly, *close to* and *to close* are the most frequent combinations of the word *close* in either direction.



(a) (b)

Figure 5. Search facilities provided by the phrase collection

Checking word combinations

Learners often use search engines to check whether, and in what contexts, phrases they have written appear on the Web. To do so, they surround the words with quotation marks and perform a phrase search. The number of hits is interpreted as some indication of the "representativeness" or authenticity of the sequence. If there are no hits, no one has ever used that text before—at least on the Web. This might be good news for creative and confident writers, but for most language learners it is a negative reflection on what they have written. The Web phrase collection has the potential to come up with more constructive feedback.

Users enter text into a box on the interface and submit it to the phrase checker. The system first chunks the text into phrases (using the OpenNLP package mentioned above). For example, the sentence *I like to play tennis in that court* is divided like this:

[NP I] [VP like to play] [NP tennis] [PP in] [NP that court]

Square brackets indicate phrases, introduced by phrase level tags. This fragment contains noun phrase *I*, verb phrase *like to play*, noun phrase *tennis*, prepositional phrase *in*, and noun phrase *that court*.

Then five-word units are reconstituted from these phrases, five being the maximum length of n-grams available to the phrase checker. In this example the following three units will be constructed

I like to play tennis

like to play tennis in

tennis in that court

The procedure is to take each phrase and keep appending text until either the word limit or the end of the text is reached. Finally these units are checked against the n-grams to retrieve the frequency and an alarm is raised if it falls below a given threshold.

We illustrate the process using a small segment of student text. Figure 6a shows the result of checking *As the Internet become all-pervading*. The parsing process groups these five words into a single unit, which is displayed in square brackets. First, the frequency of the entire unit is retrieved—zero in this case. Then successively longer prefixes are constructed—*As*, *As the*, *As the Internet*, and *As the Internet become*—and their frequency count is displayed, along with its logarithm. There is nothing wrong until *become* is added, whereupon the count of 0 indicates that *As the Internet become* does not appear in the collection. The system highlights *become* and allows the user to browse alternative continuations for *As the Internet*—or in the other direction, those that precede *become*. Clicking *Verb* retrieves all verbs that follow *As the Internet*, while *all* means all words regardless of type.

Figure 6b shows the result of clicking *Verb*. Apparently the grammatical error caused by *become* can be avoided by using *becomes*, *grows*, *has*, *continues*, etc. Further useful results are revealed after clicking *As the Internet becomes*; indeed, the system's suggestions of *larger*, *mobile*, *ubiquitous* or *pervasive* are perhaps better lexical choices than *all-pervading*.

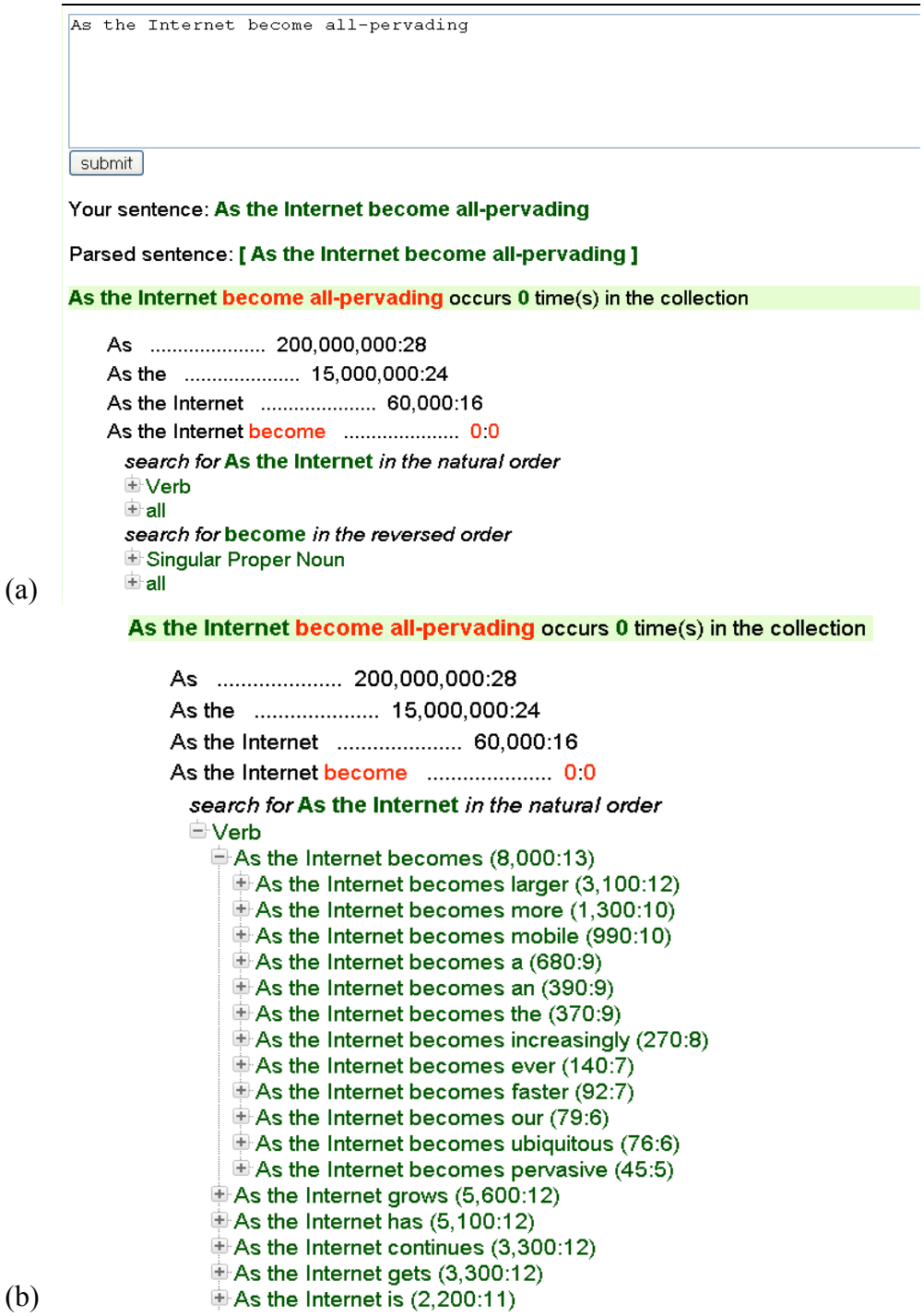


Figure 6. Checking facilities provided by the phrase collection

The approach has several limitations. First, users can search at most four words ahead (or behind), whereas the number is virtually unlimited if a search engine is used. Second, common words like *the, a, of,* and *to* are dominant constituents of phrases, which makes it hard for users to glean useful language patterns. Third, the collection is based on a historical dump of the Web, and has been further filtered: as noted earlier this falsely rejects some acceptable phrases—for example, ones containing neologisms like *google*.

Fourth, grammatical errors in Web text may confuse less advanced learners, and the situation is aggravated when they occur reasonably frequently—for example, *may not suitable* appears 602 times in the collection. Fifth, some training is required before students can use the interface productively to identify and correct errors.

Evaluation

In order to evaluate the usefulness of the two latter collections, the authors have recently trialed an intervention which involves the highlighting of collocations and phrases in second language graduate students' texts which are deemed by the teacher to be problematic and/or able to be improved significantly. The highlighted text sections are then checked by the students using the collections and text changes are made. Theoretically, these three steps align with Nation's (2001) psychological processes for vocabulary learning, namely, noticing, retrieval and generative use.

Conclusion

The size of the Web, its messiness, and the complexity and diversity of its contents, are constraints that have all, to some degree, been mitigated by the functions and interfaces described in this article. Using the Greenstone digital library software, we have managed to impose some degree of order on the raw data by building searchable collections, with particular browsing functions, from sub-collections of the n-gram corpus. Teachers and learners are exposed to examples of common usage that are stable, grammatically clean and contextualized. Links are provided to other databases to allow users to examine both exemplary text and live Web samples that are contemporary and pragmatically rich.

The specific systems we have designed allow learners to generate collocations for particular types of syntactic combination, explore colligational patterns both preceding and following a particular lexical item, and generate and review their own text with reference to contextualized samples from the Web and the British National Corpus. Further evaluation of these innovations will lead to refinements in both the data the system generates and the interfaces through which teachers and learners use it. Most importantly however, it will contribute to the important need for "research to underpin the integration of corpora and concordancing in the language-learning environment (Chalmers, 2003, Abstract section, ¶ 1).

References

Banko, M. & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France (pp. 26-33).

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Longman.

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes, 5*(2), 97-116.

Biber, D. & Kurjian, J. (2007). Towards a taxonomy of web registers and text types: A multi dimensional analysis. *Language and Computers*, *59*(1), 109–130.

Chambers, A. (2005). Integrating corpus consultation in language studies. *Language Learning & Technology, 9*(2), 111-125. Retrieved November 28, 2008, from http://llt.msu.edu/vol9num2/chambers/default.html

Chinnery, G. M. (2008). You've got some GALL: Google-assisted language learning. *Language Learning and Technology, 12*(1) 3–11.

Guo, S. & Zhang, G. (2007). Building a customised Google-based collocation collection to enhance language learning. *British Journal of Educational Technology, 38*(4), 747–750.

Justeson J., & Katz, S. (1995). Principled disambiguation: Discriminating adjective senses with modified nouns. *Computational Linguistics*, *21*(1), 1–27.

Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid (pp. 32–38).

Kilgariff, A. & Grefenstette, G. (2003). Introduction to the social issue on the web as corpus. *Computational Linguistics, 29*(3), 333–347.

Meyer, C. F. (2002). *English corpus linguistics. An introduction*. Cambridge: Cambridge University Press.

McEnery, T. & Wilson, A. (1996). *Corpus linguistics*. Cambridge: Cambridge University Press.

Nagy, W. E. (1997). On the role of context in first- and second-language vocabulary learning. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary description, acquisition and pedagogy* (pp. 64-83). Cambridge: Cambridge University Press.

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

Robb, T. (2003). Google as a quick 'n dirty corpus tool. *TESL-EJ, 7*(2). Retrieved November 28, 2008, from http://www-writing.berkeley.edu/TESl-EJ/ej26/int.html

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.

Thanopoulos A., Fakotakis, N., & Kokkinakis, G. (2002). Comparative evaluation of collocation extraction metrics. In *3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain (pp. 620–625).

Yoon, H. (2008). More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language and Learning Technology*, *12*(2), 31–48.