Supporting collocation learning with a digital library

Shaoqun Wu[1][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*

[1] Corresponding author. Email shaoqun@cs.waikato.ac.nz

Supporting collocation learning with a digital library

Extensive knowledge of collocations is a key factor that distinguishes learners from fluent native speakers. Such knowledge is difficult to acquire simply because there is so much of it. This paper describes a system that exploits the facilities offered by digital libraries to provide a rich collocation-learning environment. The design is based on three processes that have been identified as leading to lexical acquisition: noticing, retrieval and generation. Collocations are automatically identified in input documents using natural language processing techniques and used to enhance the presentation of the documents and also as the basis of exercises, produced under teacher control, that amplify students' collocation knowledge. The system uses a corpus of 1.3 B short phrases drawn from the Web from which 29 M collocations have been automatically identified. It also connects to examples garnered from the live Web and the British National Corpus.

Keywords: CALL, collocation learning, collocation activities, automatic answer generation, cherry-picking

INTRODUCTION

Why do language learners find it difficult to differentiate between words like *look*, *see* and *watch*, or *broad* and *wide*? Why do students who know many individual words still struggle to express complex ideas simply and precisely? Why are so many frustrated that they make little visible progress? How is it that native speakers communicate so much more effectively? The answers rest on the collocational knowledge of language learners. It is the collocates of *look*, *see* and *watch* or *broad* and *wide* that reveal their different shades of meaning, rather than their dictionary definitions (Conzett, 2000). Complex ideas are hard to express unless one can use simple vocabulary in a range of collocations (Lewis, 1993). Hill (1999) points out that students with good ideas often lose marks because they don't know the four or five most important collocations of a key word that is central to what they are writing about. Wray (2002) emphasizes that collocations are particularly important for learners striving for a high degree of competence in a second language, because they enhance not only accuracy but also fluency. Nesselhauf (2003, p.223) reiterates, "Collocations are of particular importance for learners striving for a high degree of competence in the second language but they are also of importance for learners with less ambitious aspirations, as they not only enhance accuracy but also fluency".

Although the rise of computer assisted language learning has brought a new dimension and dynamic into language learning, little research has been done on computer assisted collocation acquisition. Vocabulary learning in a computer environment often makes use of exercises that isolate target vocabulary items and remove them from their original context, and thus pay scant attention to the need for learners to learn and manipulate the form and contexts of words. Concordance data allow learners to analyze collocations, but as Peachey (2005) states "concordancers are primarily linguistic research tools. Almost all have been designed with the sophisticated researcher in mind" (Some possible problems section, ¶2). This results in learners often being overwhelmed by the vast number of collocations returned when searching for common words. It also means that

teachers may find it hard to identify sets of useful collocations for their students from large collections of text (Gabrielatos, 2005; Peachey, 2005; Stevens, 2001).

This paper describes a system that exploits the facilities offered by digital libraries to provide a pedagogically enriched collocation-learning environment. Its design is based on three processes that have been identified as leading to lexical acquisition: noticing, retrieval and generation (Nation, 2000). Teachers build collections of material they have prepared for their students; the system extracts important collocations automatically and presents them in a way that draws the attention of students and gives them an opportunity to systematically acquire core collocations for a particular subject. The system links to external sources that help students expand their collocational knowledge by examining them in exemplary text and in live Web samples. We have developed four activities whose exercises are automatically generated from collocations and their accompanying text. Teachers can create exercises specifically tailored for their students using an interface that allows them to choose appropriate material and apply quality control to the automatically-selected exercise content.

COLLOCATIONS

The importance of collocations for successful language learning was recognized over seventy years ago (Palmer, 1933). Hornby (1974) and Brown (1974) contend that oral listening comprehension and reading speed can both be improved by increasing collocational knowledge. Marton (1977) and Arabski (1979) show that collocation errors constitute a high percentage of errors committed by L2 learners. Bahns and Elaws (1993) point out that collocations present a major problem in the production of correct English, even for advanced ESL students. Hill (2000) lists nine reasons why collocations are important in terms of the lexical nature of a language, the sheer number of collocations that native speakers hold, the role of memory, and the way we think and express ideas. As Nation (2000) summarizes:

- language knowledge is collocational knowledge;
- collocational knowledge is important for developing both fluency and accuracy;
- knowing a word involves knowing its set of its collocates.

Collocation learning

Collocation knowledge is difficult to acquire simply because there is so much of it. Native speakers carry hundreds of thousands—possibly millions—of lexical chunks in their heads, ready to draw upon in order to produce fluent, accurate and meaningful language (Lewis, 1997). This presents a daunting challenge to language learners.

In the classroom, collocation teaching is neglected (Farghal and Obidedate 1995)—for example, Bahns and Eldaw (1993) attribute poor collocation performance in their study to the fact that collocations are not taught explicitly. Collocation learning has been peripheral in the classroom for two principal reasons. First, grammar is the traditional focus of curriculum, especially in EFL teaching, because it is relatively easy to teach and assess. Second, identifying a set of useful collocations is a daunting task, and because of the limited resources at their disposal most teachers have to rely on intuition. This is challenging for teachers whose mother tongue is not English, but also not unproblematic for native speakers (Gabrielatos, 2005).

Without adequate guidance, learners have no means of distinguishing useful collocations from the mass of possibilities; consequently they fail to notice collocations and even to understand their existence and importance (Bishop, 2004). Collocation learning is a cumulative process that involves a great deal more than rote memorization. Students with limited study time will not learn appropriate collocations unless they are deliberately selected, prioritized, and incorporated into language material (Swan, 1996).

Resources like collocation dictionaries and concordancers therefore potentially provide a useful tool for the learning and teaching of collocations. A small number of these have been researched. Guo and Yang (2007), for instance, used the live Web as a corpus, generating collocations in the limited contexts of snapshot lines returned from search results.

However a limitation associated with these tools is that they often lack learner-friendly interfaces, or are limited by the nature of the corpus they draw on. Language activities on the Web are popular ways of helping learners practice and improve their English, but those for collocation learning are rare and inadequate.

Collocation teaching

Despite wide recognition of the importance of collocations in language learning, it is unclear how they should be taught. The general consensuses of researchers and practicing teachers includes three aspects:

1. awareness raising;
2. collocation selection;
3. learning strategies.

Many researchers believe that collocations should be learned deliberately. The first and most important step is to draw students' attention to their existence. Nation (2000) suggests that teachers encourage students to split text containing familiar items into chunks and seek patterns in them. Chunking can take place when listening to stories or during reading and writing tasks. Lewis (1997) recommends that important collocations are presented in the classroom and students trained to learn them in their entirety and break them into parts later. Gonzett (2000) advocates selecting books that include many collocations and training students to observe and note as many as possible through reading, and reinforce them in their writing.

From the tremendous number of possibilities, how should collocations be selected for students to learn? Brown (1974) speaks of "normal" and "unusual" collocations, and recommends teaching the former because they form the basis of the latter—but he does not define the distinction clearly, leaving it to the teacher's intuition. Some researchers use frequency, suggesting that when learners first encounter a word its high-frequency collocates should be presented (Channell, 1981). Others propose criteria such as *need*, *usefulness*, *productivity*, *currency*, *frequency* and *ease* (Yorio, 1980). Nation (2000) advocates *frequency* and *range*: first pay attention to frequent and immediately useful collocates; then deal with a range of related ones from different contexts.

Collocation learning is challenging, and to develop effective learning strategies learners need help. In the classroom, they consult collocation dictionaries such as LTP Dictionary of Selected Collocations (Hill and Lewis 1997), Oxford Collocation Dictionary for

Table 1. Examples of collocation activities

Purpose	Activities
Raising awareness	1. mark collocations in a text 2. insert appropriate words to reconstruct the content of a text
Learning individual collocations	1. teach common collocations when introducing new words 2. extend collocational knowledge with already-known words
Storing useful collocations	1. write down collocations in vocabulary notebooks 2. sort collocations by common key word or topic
Enhancing precision	1. rephrase, e.g. by expressing negative feelings in a variety of ways 2. uncover differences between similar words, e.g. *injury* and *wound* 3. correct collocation errors in a sentence
Improving retention	1. fill in missing parts of collocations 2. find collocation partners 3. collocation dominoes 4. odd one out 5. guess a word from some of its collocates

Students of English (2009) and record examples in their notebook while exploring text or preparing essays. Computer concordancers expose them to collocations in natural occurring contexts. Hoey (2000) suggests using concordancers to study the same collocations in different texts, and to find keywords in a text and learn how they combine with other words in context. Teachers have developed many classroom activities to help their students explore collocations, retain them in long-term memory, and expand and enrich their collocational repertoire. Lewis's (2000) book *Teaching Collocations* contains a wealth of activities contributed by researchers and practicing teachers. Table 1 shows some activities, grouped by learning objective.

DIGITAL LIBRARIES

A digital library is usually a collection of texts (although it can also contain other resources including images, sound files, etc), and can function as a searchable corpus. It can also provide a language resource from which teachers construct activities. Digital libraries have a central, but as yet relatively unexplored, role to play in language education. They allow teachers (or students) to build collections that are relevant to their study, and can include both written and spoken text. To avoid overwhelming students, teachers can control collection size simply by importing the right amount of material into the library. Materials can come from conventional sources such as textbooks, audio and video tapes, newspapers, the Internet, and teachers themselves. They can transcend conventional library resources to include information produced by special interest groups: personal papers, collections, essays, and home pages. A particularly useful source is student assignments, suitably anonymized. Studying work by peers enhances awareness of language and helps develop critical reading skills. It also gives the class opportunities to learn from one another, and narrows the language ability gap.

Digital libraries can provide a safe learning community for learners and teachers. Learners can meet their peers, exchange learning ideas, and engage in competitive or

collaborative tasks. Teachers can share thoughts, tips and lesson plans, and organize collaborative task-based, content-based language projects. Pedagogically tuned search and browse facilities can meet the special needs of individual learners and teachers without bogging them down in fruitless tangential explorations. Wu and Witten (2006) describe eight activities, automatically generated from digital library content, that utilize search and retrieval facilities to illustrate new ways of supporting language study.

Digital libraries can provide authentic, focused material that is carefully selected and organized, exposing learners to contemporary language usage. Subject-specific collections give the opportunity to encounter texts that exhibit particular patterns of both word choice and grammar. For example, student knowledge of business language is greatly enriched by basing learning on a corpus of business reports and product reviews (Fuentes, 2003).

Wu and Witten (2007) describe an automatically created collection of business articles from Wikipedia, from which material such as keywords and business-related terms and definitions were identified by mining Wikipedia's structured format and richly linked hypertext using standard natural language processing tools. Three learning activities were implemented that draw attention to the salient vocabulary of a particular topic, increase student encounters with relevant topic-related vocabulary, and help sustain motivation and interest through collaboration with peers.

SUPPORTING COLLOCATION LEARNING

This section sketches how a digital library can support learning by automatically extracting important collocations from readings provided by teachers (or learners) and presenting them alongside the text; the details are covered in the remainder of this paper. Students read the material to gain a degree of familiarity with particular collocations, study them in different contexts, and record ones that interest them. Then they undertake various learning activities based on the same material, presented in the form of exercises. The system is designed to help learners notice important collocations, develop language sensitivity, and transfer from short- to long-term memory. The description below is structured around the three aspects of *noticing*, *retrieval*, *generation* identified by Nation (2000).

Noticing

Learning begins with noticing, which occurs when a learner pays attention to an item as part of the language rather than as part of a message. Noticing is affected by factors such as the item's salience and usefulness, its presentation, the learner's interest and motivation, their mindset—for example, focusing on individual words rather than larger chunks of language—and the learning environment. Attention can be drawn to important collocations in two ways. First, they can be highlighted typographically. Second, they can be presented in awareness-raising activities.

Examples of language activities that promote noticing are:

1. finding collocations in a text and recording them in notebooks;
2. reconstructing the content;
3. correcting common mistakes.

A *finding collocations* exercise might ask learners to select all nouns in a text, identify the verbs that are used in conjunction with them, pick out phrases they think are collocations, and sort them by significance. This activity can, of course, be applied to other syntactic types. Recording and organizing collocations in notebooks helps students consolidate what they have noticed. In a *Reconstructing the content* exercise, collocations are removed from the text and students must reinsert them to reconstruct the original text. In *correcting common mistakes*, learners correct collocation errors in text. For example, given *I was completely* ***disappointed*** *when I failed my exam*, students need to look up the collocates of the word in bold and pick one that is appropriate in the given context—for example, *bitterly*.

Retrieval

Retrieval, the process of remembering items, involves three aspects. First, learners must understand an item in the context in which it occurs, perhaps by guessing its meaning from the context, looking it up in dictionaries, or constructing their own interpretation through discussion with peers or teachers. Second, the item's meaning must be retrieved whenever it is met during reading or listening. Third, it must be used in circumstances that are semantically and pragmatically appropriate.

There are two effective ways to help learners remember a collocation: repetition and use. Repetition can be achieved by exercises that recycle collocations in different contexts. Readings and important collocations are presented side by side, and follow-up activities use the same material to gradually increase familiarity with its language features. Typical word usage and salient collocations can be recycled in different types of exercise to expose learners to them repeatedly. For example, sentences containing collocations of the commonly confused words *broad* and *wide* can be used in a *reconstructing the content* exercise that asks learners to fill in a blank to form a valid collocation, while the same data can be used in a *correcting common mistakes* exercise. Repetition also occurs when learners are asked to record and organize collocations that they think are useful for an essay assignment or oral presentation.

Recall of a collocation is strengthened when it is used. Activities that require students to use a particular collocation to construct sentences or conduct a conversation can be designed to consolidate and extend what has been learned.

Generation

Generation is the process of enriching and stretching the learner's knowledge of an item, and occurs when the item is met in different forms and contexts. For example, the word *heavy* has different meanings when used in *heavy rain* and *heavy smoker*; its adverbial form is *heavily*. Generation can be achieved by incorporating material from various sources into a rich contextual environment that enables learners to discover and analyze new meanings of lexical items and use them in different ways.

External material can serve to illustrate language use in different contexts, enriching the learner's lexical knowledge and promoting generative and creative use. For example, exercises can be supplemented by material collected from reference corpora such as the

British National Corpus[2] and the Web itself. These are incorporated into the system described below to provide authentic samples of language use that serve as hints for students when doing certain exercises.

BUILDING COLLOCATION-ENRICHED DIGITAL LIBRARY COLLECTIONS

This section describes how a collocation database can be established from the textual content of a digital library; the next section shows how to use it to generate questions and provide answers. We also describe two auxiliary collections that are built from a large corpus of *n*-grams.

The Greenstone digital library software lets end users build large collections of documents and metadata and serve them on the Web.[3] For demonstration purposes, this paper uses a dozen short articles of general interest, in which the only metadata available are titles.[4] The standard Greenstone system allows such a corpus to be built into a digital library collection, equipped with a full-text index and metadata browsing facilities. We have enhanced this process to automatically identify collocations in the text and organize them to support collocation searching, browsing and learning. Such collections can be created from any body of text, including text supplied by teachers, but the mechanics of building collections in Greenstone lies beyond the scope of the paper (see Witten et al., 2010). Here we focus on how collocations are identified in given documents.

Identifying collocations

We think of collocations in the same way as expressed by Benson et al. (1986, p.ix): "In any language, certain words combine with certain other words or grammatical constructions. These recurrent, semi-fixed combinations, or collocations, can be divided into two groups: grammatical collocations and lexical collocations." We focus on lexical collocations, which have structures like verb + noun, noun + verb, adjective + noun, noun + noun, adverb + adjective, adverb + verb (ibid, p. ix). 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.

We use the above six patterns and add four more from the Oxford Collocation Dictionary: noun + *of* + noun, verb + adverb, verb + adjective, and verb + *to* + verb. We allow determiners and possessive pronouns such as *the*, *a*, *any*, *some*, *his* to precede noun words, and extend four of the types to include further constituents of potential use to learners. Table 2 shows the ten collocation types, along with their extensions. As the examples illustrate, collocations contain from two to five words, five being relatively rare.

The process of identifying collocations 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. Discard sequences that do not occur in the *Web phrases* collection (see below)

[2] *http://www.natcorp.ox.ac.uk/*
[3] *http://www.greenstone.org*
[4] These articles are from the University of Waikato Pathway College's IELTS course.

Table 2. Collocation types and examples

Type	Example
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*
noun + verb *includes*: noun + verb with present tense noun + *be* + gerund noun + *be* + past participle	 *the time comes* *the time is running out* *the time is spent on*
adjective(s) + noun(s) *includes*: adjective + noun + noun adjective + adjective + noun(s) adjective + *and*/*but* + adjective + noun(s)	*a little girl* *a solar energy system* *a beautiful sunny day* *economic and social development*
noun + noun	*a clock radio*
adverb + adjective	*seriously addicted*
adverb + verb	*beautifully written*
noun + *of* + noun	*a bar of chocolate*
verb + adverb	*apologize publicly*
verb + adjective *includes*: verb + preposition + adjective verb + noun + adjective	*make available* *take up more* *take it easy*
verb + *to* + verb	*cease to amaze*

5. Associate sample text with the collocations that have been identified
6. Build search indexes and browsing structures.

In steps 1 and 2, an off-the-shelf natural language processing tool is used to split the text into sentences and assign syntactic tags to the words.[5] Then the tagged words are compared against patterns defined for each collocation type. In step 3 teachers can specify a subset of the collocation types in Table 2 for their students before the collection is built; otherwise all ten are used. Step 4 matches the sequences that are identified in the text against the *Web phrases* collection described in the next subsection and discard ones that do not appear, because they are likely to be incorrect or infelicitous.[6] We also use the frequency recorded in this collection for ranking collocations when presenting them to students, to help them prioritize learning.

Whenever a collocation is identified, its sentence and the one before and after are extracted and associated with it in step 5. These allow students to study collocations in context rather than as isolated items, and are used in the learning exercises described below. To facilitate searching and browsing, step 6 builds indexes on the constituent

[5] We use the OpenNLP tagger, *http://opennlp.sourceforge.net*

[6] This step can be disabled when creating the collection, which might be desirable if collocations are expected to contain neologisms (such as the word *google*) that do not appear in the British National Corpus and have therefore been omitted from *Web phrases*.

words of each collocation, and creates browsing structures that group collocations by the words they contain, and by their type (Table 2).

The process of identifying collocations is not perfect: its accuracy reflects that of the underlying syntactic tagger. Taggers apply complex algorithms to resolve syntactically ambiguous words like *cut*. Despite extensive research, no algorithm yields perfect results because of the complexity of human language, and so collocations such as (for example) noun + noun are occasionally mistakenly construed as verb + noun. Taggers at the current state of the art achieve around 95% accuracy. Preliminary indications are that this does not seriously impact system performance, but further evaluation with language teachers is needed to confirm this.

Supplementary collections

Two supplementary digital library collections are used in this work; Wu, Franken and Witten (2009) give more details of their construction. The first, *Web phrases*, is built from a corpus supplied by Google,[7] which they created in January 2006 from a trillion words of publicly accessible English-language Web pages. It contains short sequences of consecutive words called "*n*-grams" ranging in size from one to five words, along with their frequencies—a good match with the two- to five-word collocations illustrated in Table 2. We intersected the items in this corpus with the vocabulary from the British National Corpus in order to remove ones that include mis-spelled words, proper names, rare and unusual terms, and other non-standard items. The resulting collection contains 145,000 unique words, 14 million two-grams, 420 million three-grams, 500 million four-grams and 380 million five-grams. Each phrase is stored, along with its frequency, in a searchable digital library collection.

The second auxiliary collection, *Web collocations*, contains fragments extracted from the *Web phrases* just discussed, organized into the ten collocation types of Table 2. The same tagger mentioned above is applied to 5-grams from the Web rather than complete sentences.[8] *Web collocations* contains a total of 29 M collocations, ranging from 90,000 examples of the smallest type—verb + adjective—to several million examples of the dominant types—verb + noun, noun + *of* + noun, adjective + noun, and noun + noun.

Apart from its sheer size, this collection has several advantages over traditional printed collocation dictionaries. First, it is fully searchable, so that users can search on any constituent of a collocation. Second, collocations are sorted by frequency to help students prioritize learning. Third, each collocation has many variants. For example, considering the verb *cause*, there are 268 variants of c*ause problems*, including *cause serious problems*, *cause major problems*, *cause unpredictable problems*, etc. Fourth, students can learn to use articles correctly by studying the collocations of a particular word. For example, we say *make a difference*, not *make difference*; *make sense*, not *make a sense*. Last but not least, students often find it difficult to decide when to use the plural or singular form. For some nouns, both forms are appropriate and depend on the context—*make a decision* and *make decisions*—but for others, one is more dominant—for example, *make a living* is 10,000 times more frequent than *make livings*.

[7] The Google n-gram collection is available on six DVDs from http://www.ldc.upenn.edu

[8] Of course, the limited context makes this a less reliable, although still useful, procedure.

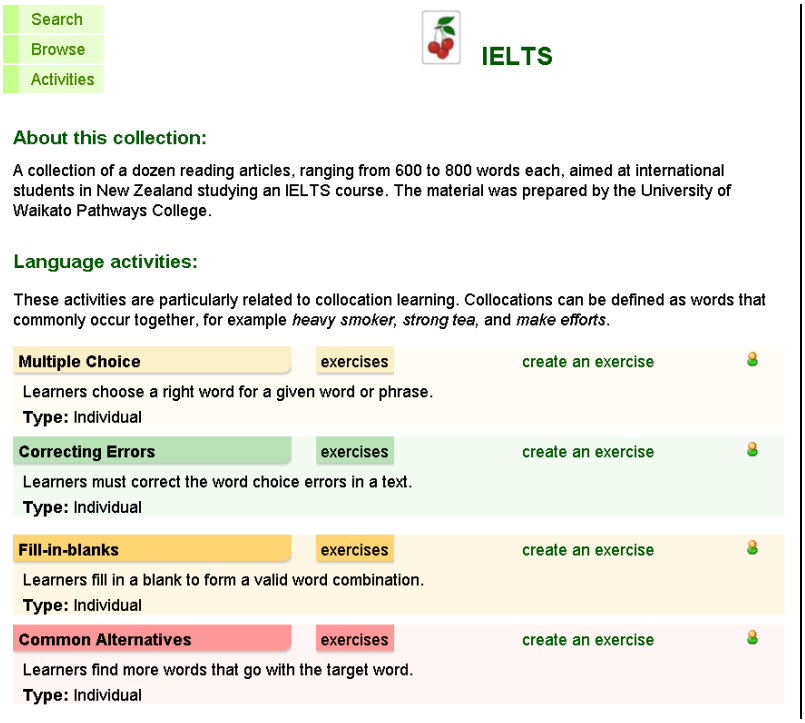

Figure 1. The example collection's "About" page

USING THE DIGITAL LIBRARY

Figure 1 shows a digital library collection built from the dozen short articles mentioned above. This "About" page displays the collection's title, description, and a list of learning activities. The *search* button allows users to seek documents and collocations containing particular words or phrases; they can also browse documents by title and language level, and browse collocations by word and collocation type. Here we focus on collocation-related facilities.

Searching and browsing collocations

Three ways are provided to access the collocations: in the context of an article; locating partners of a particular word; and browsing collocations by word and type.

As in any digital library, users can find articles by searching or browsing, and read them. Here, an alternative version is provided with collocations highlighted, to help students notice them and study their context. In the example shown in Figure 2, collocations related to stamp collecting—*collect stamps*, *new stamps*, *overseas stamps*, *stamp dealers*, *start a stamp collection*, *stamp club*, *stamp items*, *swap stamps*, *stamp competitions*—stand out from the rest of text, attracting the student's attention. The collocation *extremely high* has been clicked to reveal four small icons. The last three of these expand student knowledge by retrieving relevant material from other resources, as described in the next subsection. Following that, the function of the first small icon, "cherry-picking," is described.

From the first button at the top of Figure 1, "Search," users can seek collocations in the collection that contain a particular word. Figure 3 shows the beginning of the result for

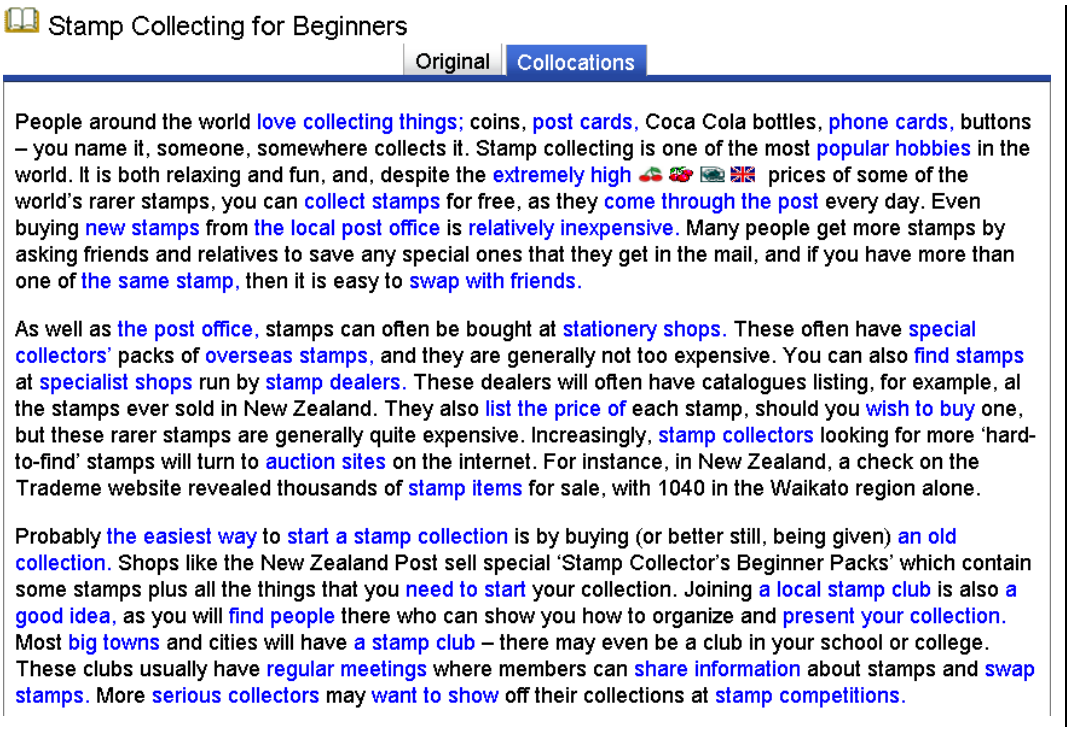

Figure 2. A document in the collection, with collocations highlighted

the word *family*, sorted by frequency on the Web. The contexts of each occurrence—here there are five instances of the first collocation, *family members*—are gathered together to acquaint learners with different usage. For *family*, the most dominant collocation types are noun + noun and noun + *of* + noun: *family members*, *family history*, *family tree*, *family relationships*, *generations of family*, *side of a family*, and *encouragement of family*.

Collocations are organized by word and type to facilitate browsing, invoked by the "Browse" button in Figure 1. When browsing by word, an alphabetic selector leads to the word in question—clicking the letter *f*, followed by the word *family*, obtains the collocations shown in Figure 3. Browsing by type retrieves all collocations of a particular type. Figure 4 shows some verb + noun examples: *take advantage of*, *take into account*, *lose weight*, *save time*, *share information*, etc.

Expanding collocational knowledge

The three small icons shown alongside each collocation in Figures 3 and 4 present additional resources associated with it. The first shows related items from the *Web collocations* collection described earlier; the others retrieve relevant text samples from the Web and the British National Corpus respectively. The last three of the four small icons following *extremely high* in Figure 2 lead to the same information.

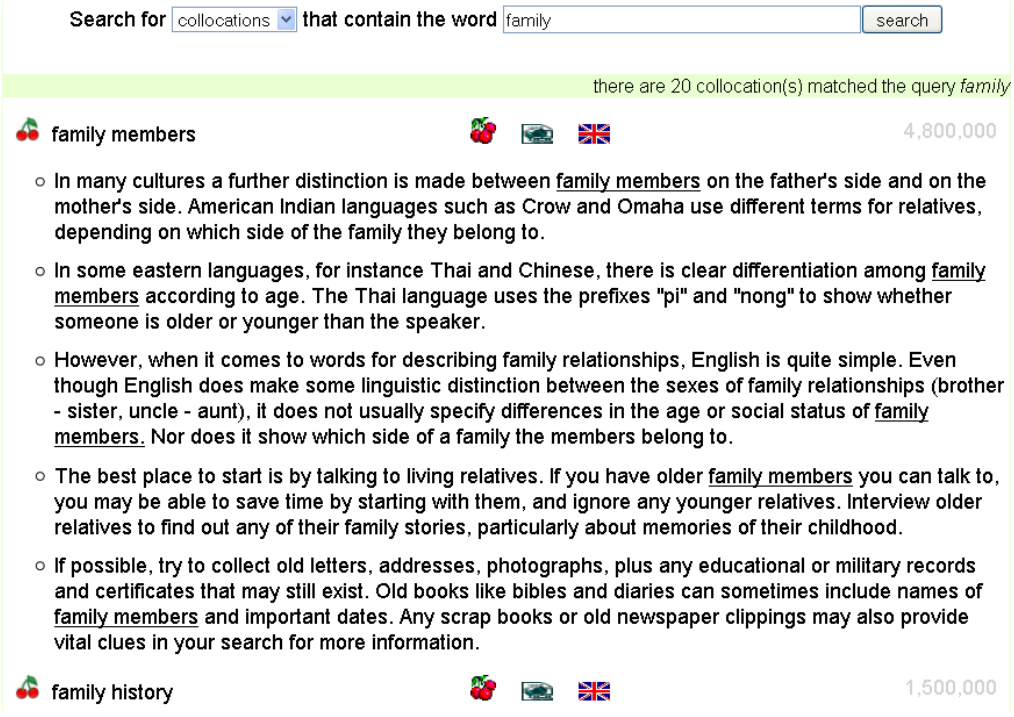

Figure 3. The collocation results of searching for the word *family*

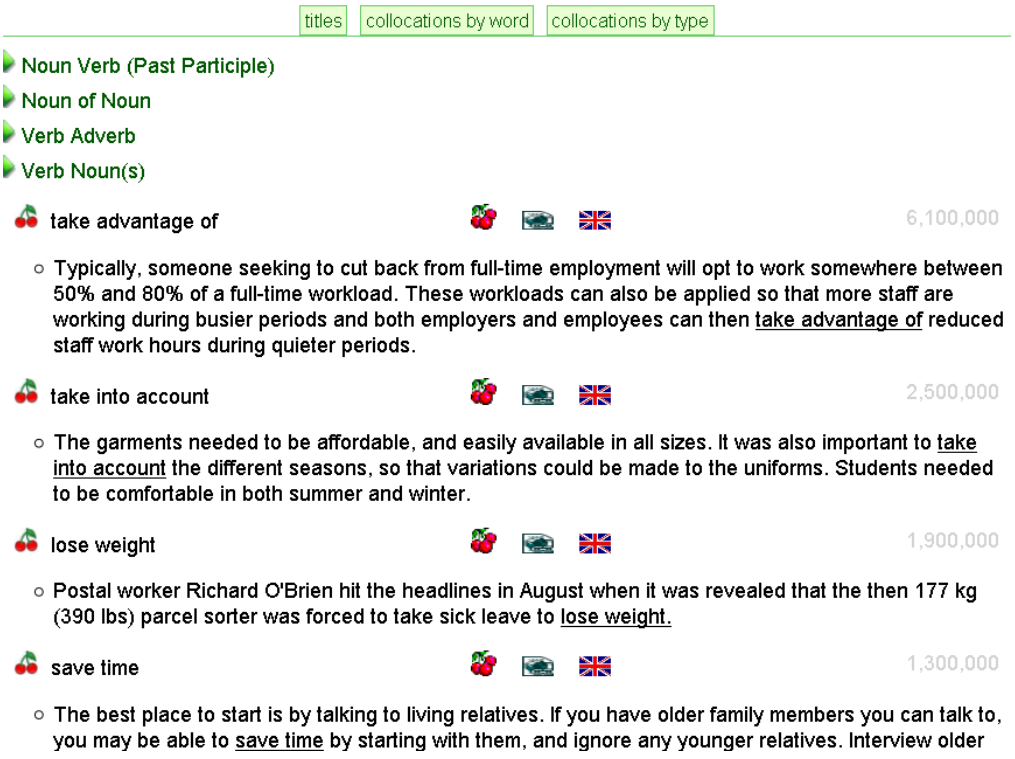

Figure 4. Browsing by collocation type

The first function, invoked by clicking the second of the four little icons in Figure 2 or the first of the three little icons in Figures 3 and 4, opens a popup showing different collocations that have the same first and last word respectively. Figure 5 gives the output for *extremely high*: the 20 most frequent related Web collocations, sorted by frequency. For the first word they include *extremely important*, *extremely difficult*, *extremely low*, *extremely useful*, and so on; for the last we see *relatively high*, *unusually high*, *fairly high*, and *consistently high*. The "more …" button at the bottom leads the user to a page on which more of these collocations can be found and studied.

The second function gives access to examples of the collocation on the Web. The system connects to a search engine, uses the collocation as a phrase query, and retrieves sample texts in real time. The third function gives access to examples from the British National Corpus. From this corpus, we extracted the written text, split it into paragraphs, and built a searchable collection, again using Greenstone. Whenever the learner clicks the British-flag icon in Figures 2, 3 and 4, Greenstone searches this collection for occurrences of the collocation and displays the relevant paragraphs.

The Web and the British National Corpus both have limitations, but they are complementary. The latter provides far fewer examples, the number declining rapidly for longer collocations. In many cases there are none at all—even for items that occur reasonably frequently on the Web. For example, the collocation *educational and informational purposes* occurs 240,000 times in the *Web collocation* collection but not at all in the British National Corpus. On the other hand, the Web text is often unclear, incomplete and repetitive—but the examples it provides are authentic and contemporary.

Cherry-picking

Bates (1989) introduced the idea of "berry-picking" to model the behavior of real users of information retrieval systems: choosing juicy documents from the briar patch. We adapt this as "cherry-picking" to describe how students can gather useful collocations while reading an article, or when searching and browsing collocations. Cherries grow in twos and threes, which reinforces the idea of collocation, and the two words begin with the

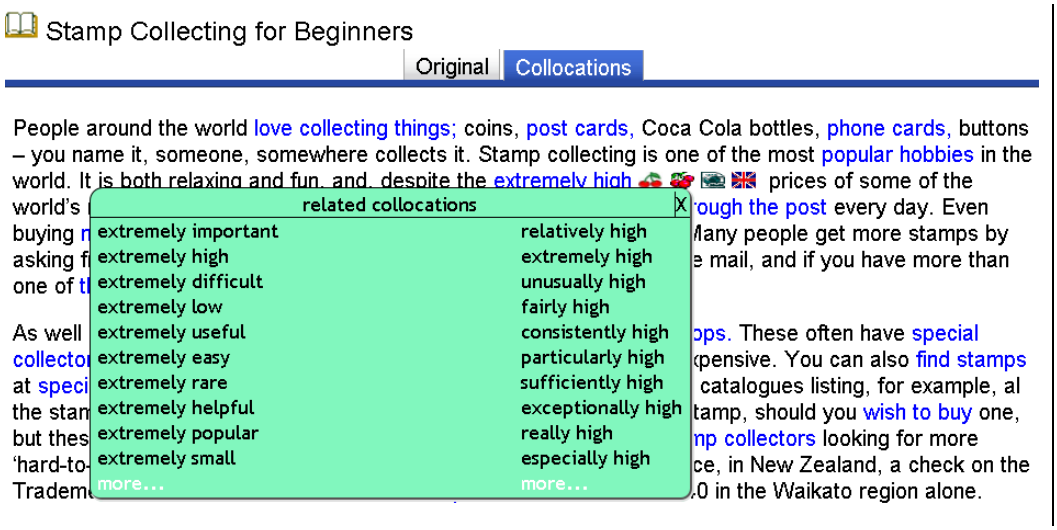

Figure 5. Collocations related to *extremely high*

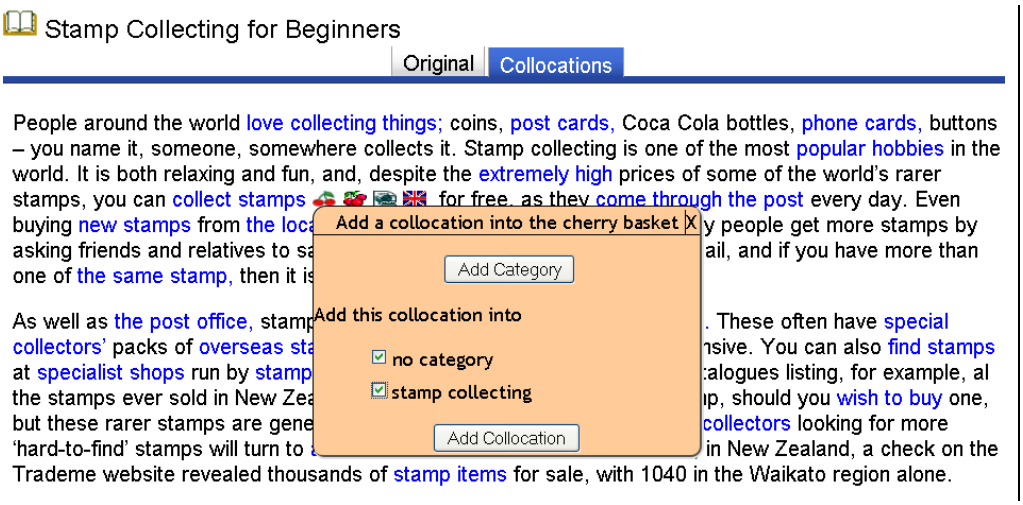

Figure 6. Picking cherries

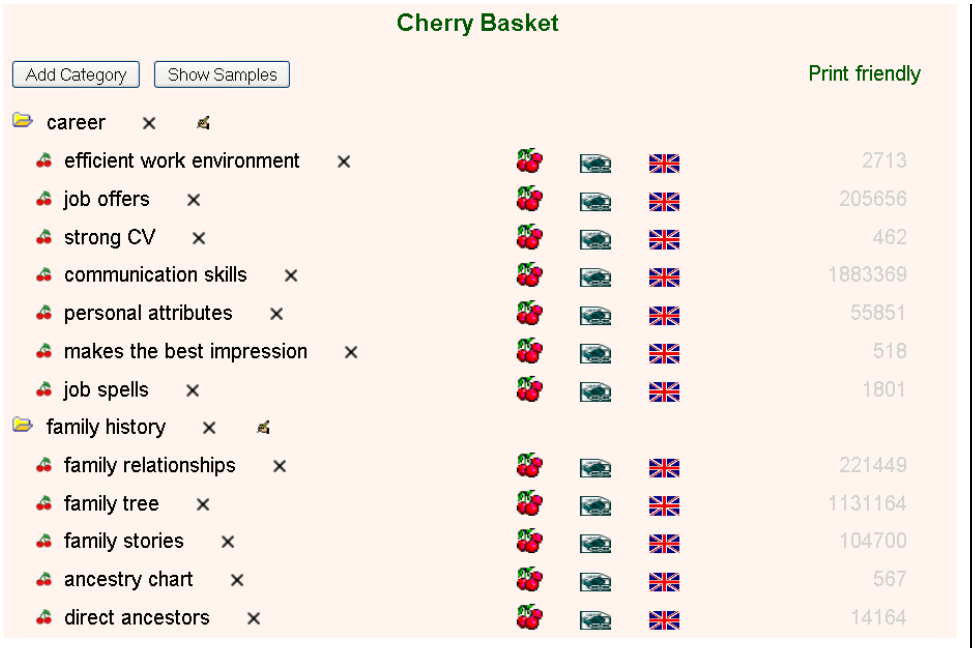

Figure 7. The cherry basket

same letter.

Figure 6 shows the cherry-picking interface that is launched by three-cherry icon that follows the collocation in Figure 2 (also seen in Figures 3 and 4), in this case, *collect stamps*. The selected collocation is added to the student's personal cherry basket. They can optionally assign it to a category or categories, or add a new category—say "stamp collecting"—for it, then assign the collocation to it. The default is to leave it uncategorized. Students can pick collocations from the articles shown in Figures 2 and 5, or from the search results page (Figure 3), or from a page reached by searching or browsing collocations (Figure 4).

Figure 7 shows a student's personalized cherry basket. It displays collocations that the user has picked and placed into two categories: *family history* and *career*. Students can

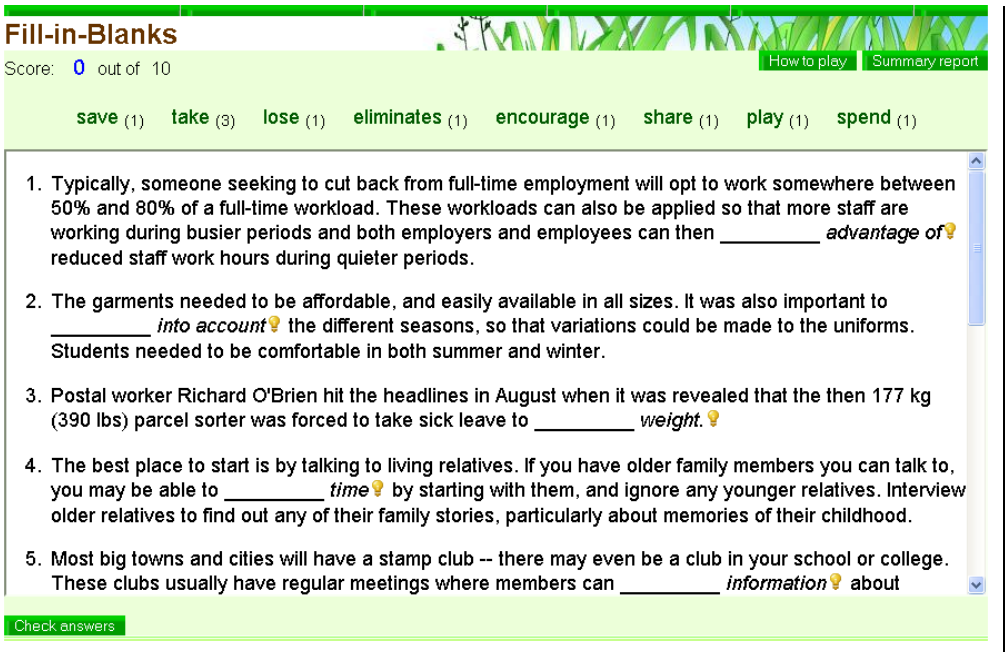

Figure 8. A Fill-in-Blanks exercise

study items in the basket using the three icons described in the previous section. They can also, of course, delete collocations, move them into different categories, create new categories and delete old ones, and print the basket to take home—or send it to friends (the "Print friendly" button).

COLLOCATION ACTIVITIES

Into the collocation learning digital library described so far we have built four learning activities: Fill-in-Blanks, Common Alternatives, Multiple Choice and Correcting Errors. These are more accurately called activity *types*, for within each one a virtually unlimited number of exercises can be created by the teacher, using an interface described in the next section, from the content of the digital library collection. Some exercises are created from whole documents; others from sentences retrieved from the collection. In the latter case the following and preceding sentence of the target sentence are also provided, as context.

Below we describe each activity individually, focusing on interface and design issues. All exercises are generated automatically based on a set of predefined parameters, and the next section discusses how to create exercises and configure them.

Fill-in-Blanks

Fill-in-Blanks exercises involve a set of collocations and their associated sentences. Under control of the exercise designer, one or more words of collocations are removed from the text, and the learner is asked to choose the word that completes each collocation.

Figure 8 shows one such exercise, which focuses on finding the right verb for a noun. The missing verbs are given at the top of the exercise panel. When chosen, they disappear from this list—except for words that occur more than once, in which case the occurrence

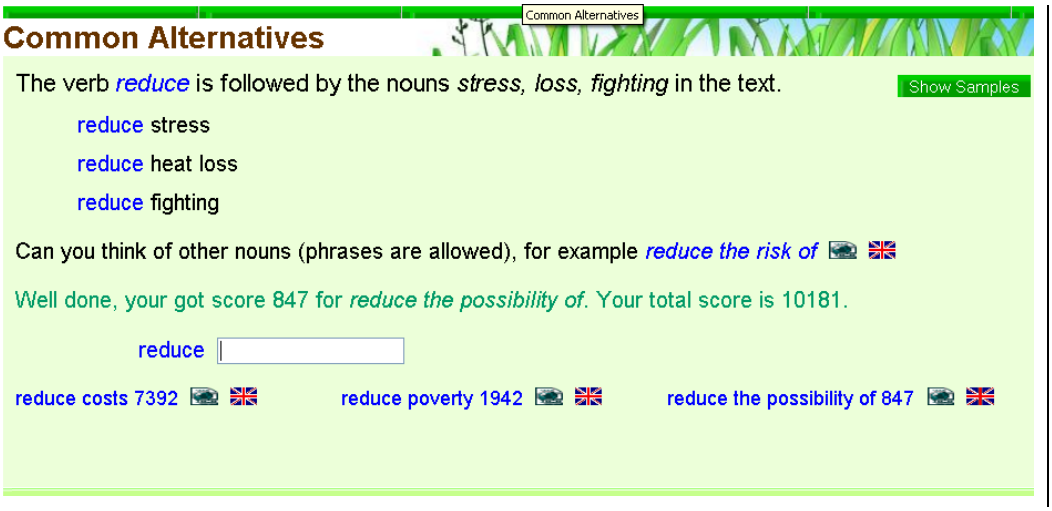

Figure 9. A Common Alternatives exercise

count (in parentheses) is decremented. Below is a list of items with target verbs omitted and the remainder of the collocation rendered in italics. The learner completes a collocation by dragging a word from the top and dropping it in place, where it appears in blue; the move can be undone by clicking the word. When the Check Answer button at the lower left is clicked, correctly formed collocations remain, but the offending word is removed from incorrect ones and reinstated at the top of the panel. The light bulb beside each collocation signifies a hint, and clicking it retrieves relevant items from the *Web collocations* collection. For example, the hint for *advantage of* includes *added advantage of*, *gain a competitive advantage*, *create a competitive advantage*, *offer a tremendous advantage*, *get the advantage of*, and *see the advantage of*.

This activity works well for sets of words that share similar meanings but have different usage. Learners are frequently confused by common words—*make and do*, *speak* and *tell*, *see* and *look*—and find it difficult to understand their differences by consulting dictionaries. Studying their collocations is an effective way to help learners distinguish their various shades of meaning. As presented in Figure 8, it reinforces receptive rather than productive knowledge, but the teacher can select a version in which the missing verbs are not shown at all but must be typed in by the learner. This reinforces productive knowledge, and is far more challenging.

Common Alternatives

To add strength to adjectives learners tend to use the word *very*, but in specific contexts there are usually more precise qualifiers that perform the same function. When describing someone as very beautiful, alternatives like *really*, *truly*, *stunningly* and *incredibly* spring quickly to the mind of a native speaker, and are usually preferred. These alternatives can be found in the *Web collocations* collection—in this case a quick search finds 100 adverbs with frequency exceeding 1000. The Common Alternatives activity helps elicit and expand this knowledge. Given a target word along with some collocation examples, learners are asked to enter as many collocations as possible—and their choices are scored.

Figure 9 shows an exercise that focuses on nouns commonly associated with the verb *reduce*. To get learners started, they are given some sample collocations: three from the

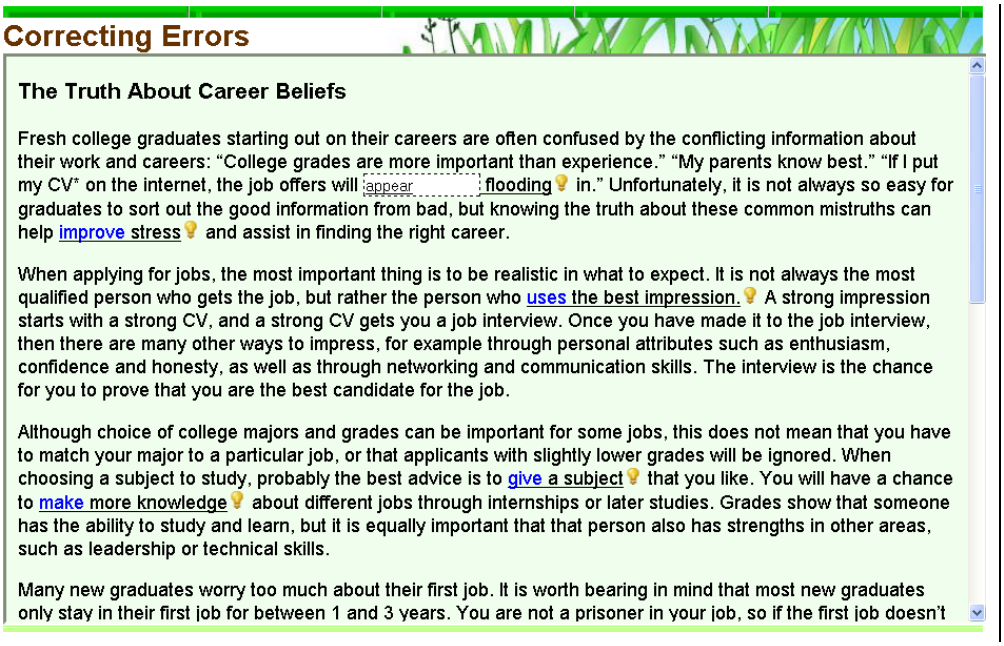

Figure 10. A Correcting Errors exercise

original text—in this case *reduce stress*, *reduce heat loss* and *reduce fighting*—and one from the *Web collocations* collection—here, *reduce the risk of*. The first three, from an article in the library that the teacher may already have asked students to read, refreshes their memory of this word. The other is the most frequent *reduce* + noun item in *Web collocations*, and is intended to help students think of other common ones. The icons that follow each collocation allow students to retrieve text samples from the Web and the British National Corpus.

Learners type a word or phrase into the text box and press the Enter key, at which point the system checks it. For example, *reduce more* would be invalid because this exercise requires nouns, or a phrase that contains a noun. If it is valid, the input text, preceded by the word *reduce*, is sought amongst *n*-grams of the same length in the *Web phrases* collection. If it is found, the associated frequency is retrieved from that collection and used as a score. The learner is notified if the collocation is invalid or the phrase does not appear amongst the *Web phrases*; otherwise it is displayed along with its score and the standard two icons for further exploration (Web and British National Corpus). In Figure 9 the user has already entered *reduce costs*, *reduce poverty*, and *reduce the possibility of*, for a total score of 10,181.

Competitive factors make this activity compelling. Learners can be connected to work on the same exercise and see each other’s scores. This challenges them to outwit one another, and encourages them to discover more collocations.

Correcting Errors

Unlike the preceding activities, Correcting Errors exercises are created from full documents rather than excerpts. Correcting language errors is relatively difficult task because of the ambiguity of language, so the entire document is used to provide as much context as possible. The teacher first chooses a document and several target collocation

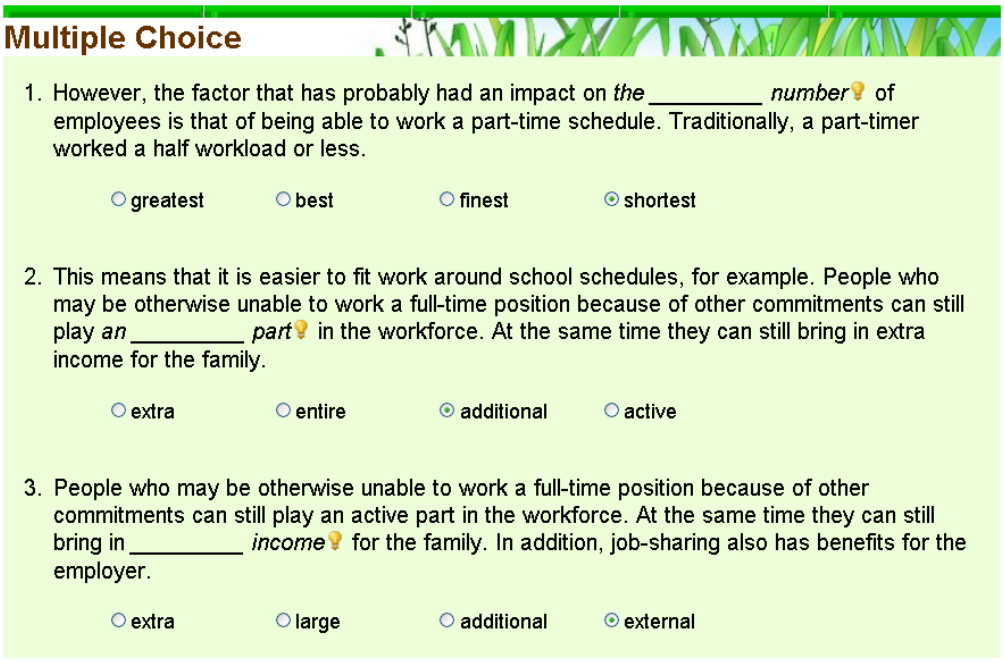

Figure 11. A Multiple Choice exercise

types, and then decides whether learners will work on the first or last constituent word. The system replaces these words are replaced with infelicitous choices that learners must correct.

Figure 10 shows an example, *The Truth About Career Beliefs*, which focuses on collocations of the verb + noun type and asks learners to find the right verb for the noun. Target collocations are underlined, and incorrect words colored in blue. Clicking a blue word brings up a box in which the student types in a new word. The answer is checked when the learner presses the Enter key or moves to another word. Correct entries are changed to black, while incorrect ones remain blue. The hint icon (light bulb) shows more collocations, retrieved using the target collocation's first and last words respectively. For example, the first set of hints for *improve stress* include *improve the accuracy of*, *improve performance*, and *improve the lives of*; while the second set includes *reduce stress*, *cope with stress*, and *handle stress*. To make them more relevant, the collocations adapt to what the learner has entered—if the learner changes *improve stress* to *decrease stress*, the collocations of *improve* are replaced by those of *decrease*.

Multiple Choice

Multiple Choice exercises, comprising a question and a set of choices—typically four—from which the correct answer must be selected, are widely used language drills for learning vocabulary. We tailor this activity to collocation learning by using sentences containing particular collocations as questions, with one collocation part missing. Four choices, including the correct one, are shown to students, who must select one that forms a valid collocation.

Figure 11 shows an exercise that asks students to complete adjective + noun collocations. The collocation is rendered in italics, and one part is missing: learners must select the

correct choice. When the Check answer button at the bottom of the screen (not shown) is clicked, the learner's correct choices are inserted into the blanks, while incorrect ones are left so that they can continue working on them. As with other activities, clicking the light bulb brings up further related collocations.

CREATING EXERCISES

Exercises are created by teachers, who select content and provide answers where necessary. First they must determine the purpose of the exercise and select material accordingly. Then they review the questions that the system provides, and remove unsuitable ones. For some activities—for example, Fill-in-blanks—answers are taken from the original text, but for others—Correcting errors and Multiple Choice—they are generated by the system. This is cheap but potentially unreliable, and teachers may wish to correct the system's suggested answers before the exercise is used.

All activities share the same principles and use similar algorithms, and we describe their parameters below. Then we introduce the interactive user interface through which teachers configure and review exercises.

Setting up parameters

Exercise content is selected by determining a few parameters that control the material retrieved by the system. All have default values, and if no configuration is necessary a complete exercise can be generated with a couple of clicks of the mouse. Here are the principal parameters.

Collocation type determines what types of collocation are to be used, selected from a drop-down list that shows the ten types in Table 2 (multiple selections are possible). Learning can be enhanced by tailoring collocation types to the teacher's goals and the students' ability.

Collocation position specifies either the first or the last word of collocations. For example, in Fill-in-Blanks learners may be asked to specify *make* in ____ *an effort*, or *effort* in *make an* _____. Based on their objectives, teachers set either component as the target. Here, the first word would be an appropriate choice if the focus is on learning verbs associated with the noun *effort*.

Hint determines whether learners can receive extra help while doing the exercise. The *Web Collocations* collection is used as the source of hints. Given the example ____ *an effort*, a hint displays the 20 most frequent verbs that collocate with the noun *effort*.[9]

Number of sentences determines the size of the exercise, in terms of how many questions are posed to learners. For the Correcting Errors activity, which does not use individual sentences, the teacher instead specifies **Document title** to determine which document to use. Document metadata includes language level, which teachers can use to help make their choice.[10]

[9] We are implementing further hint options, such as giving the first letter, last letter, or dictionary definition of the target word.

[10] Language level metadata can be specified explicitly for each document when the collection is built; if it is not, the Flesch-Kincaid grade level (http://en.wikipedia.org/wiki/Flesch-E2_Kincaid_readability_test) is used.

Contains words, specific to the Fill-in-Blanks activity allows teachers to design exercises focusing on particular words. If specified, only collocations that contain those target words are used. For example, teachers can create an exercise specifically to help students differentiate the commonly confused words *do* and *make*.

Providing answers

In Correcting Errors the original words are replaced with incorrect ones, and in Multiple Choice there are three incorrect choices for each question. It is not easy to find words that are incorrect yet plausible! Here we examine how the system reduces the teacher's burden by providing a list of candidates. When creating an exercise, teachers can determine which of these to use, or provide their own.

For each collocation, 20 candidates are generated during the collection building process. They are not randomly chosen. Rather, they must (1) somehow fit the context, (2) be of the correct form, and (3) not form a valid collocation. As an example of the second criterion, if a past tense verb or plural noun is used in the original text, the same must be true of each candidate. For the third, if the target collocation is *make a complaint*, candidates such as *file, lodge, resolve, investigate* are not selected because they collocate strongly with *complaint*.

The process involves three steps, corresponding to the three criteria described above. We explain it using the example sentence

> Some of these communities have made a great effort to improve this situation by running special classes …

where *improve this situation* is the target collocation and *improve* the target word. First the preceding text, *effort to*, is used to locate verbs that somehow fit the context. The system consults the *Web phrases* collection and retrieves verbs that follow *effort to.* Using just two words as context generally yields a satisfactory list of candidates. Next the candidates are tagged and discarded if their tag does not match that of the target word—in this case, *improve* is a verb in base form (recall that words of collocations are tagged when the collection is built). Finally, to remove candidates that form good collocations with *this situation*, the five-word phrase that encloses *improve this situation* is extracted from the original text, yielding *to improve this situation by*. Then verbs extracted in the second step are used to replace *improve*, and discarded if the resulting phrase does occur in the *Web phrases* collection.

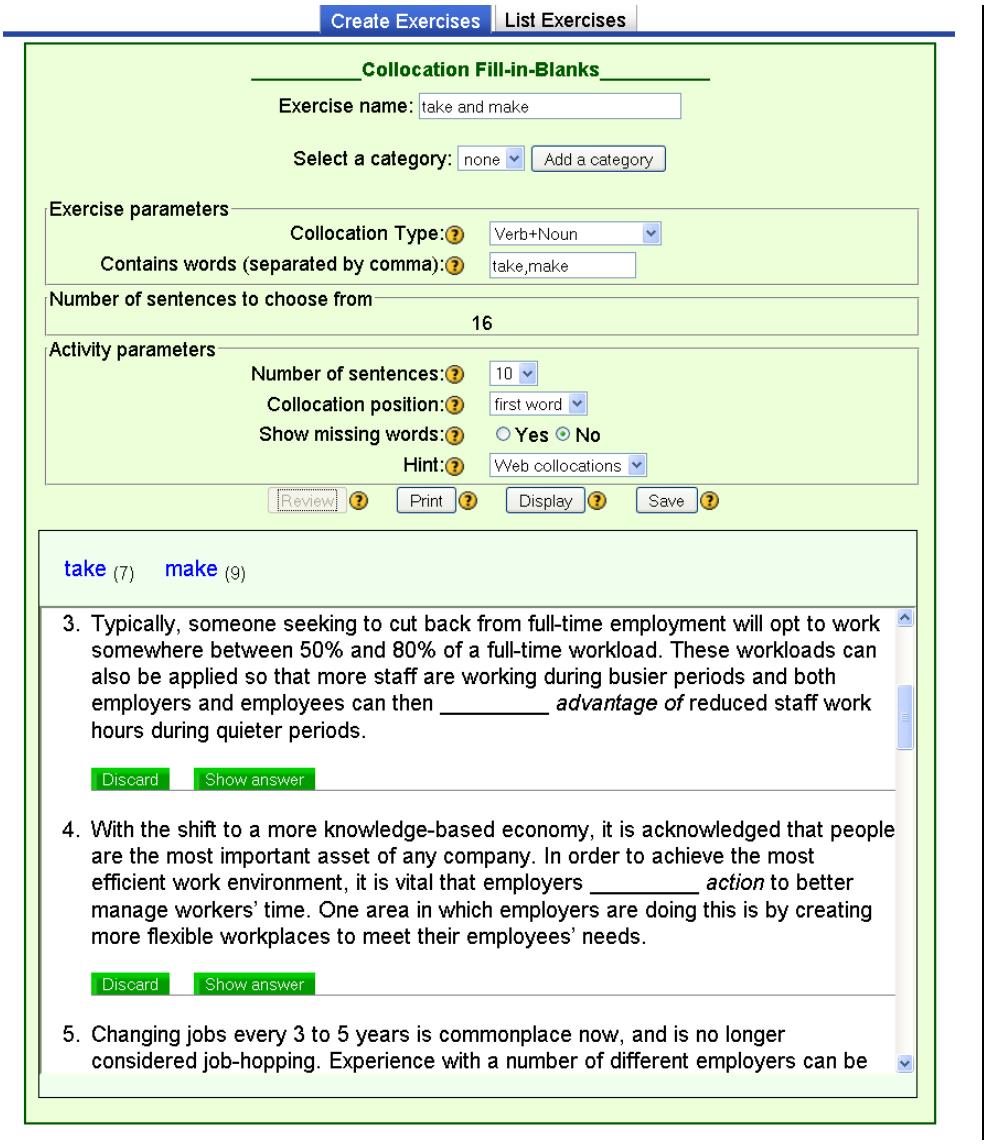

Figure 12. The design interface of the Fill-in-Blanks activity

Exercise design interface

The exercise design interface allows teachers to select materials for their students, create exercises at different levels of linguistic difficulty, make exercises collaborative or competitive, and apply quality control to the automatically-generated exercise content.

The collection's "About" page (Figure 1) displays a list of available activities, including a brief description of each one, and three related buttons—*exercises*, *create an exercise*, and a button depicting a person. The first button presents a list of exercises that have already been created; newly created ones are added automatically when the teacher saves them. The second button allows students or casual visitors to create (and use) temporary exercises with all the functionality of ones supplied by teachers, but does not appear in the exercise list; for this they use precisely the same interface as teachers, described

below. Only registered users—typically teachers—can create exercises that persist, and they must first log in using the third icon.

All activities share the same design interface, although the configuration parameters are slightly different. We illustrate it using the Fill-in-Blanks activity. Figure 12 shows its interface, which has five parts. At the top, teachers enter a name for the exercise, and, optionally, select a category. Categories can be used to create exercises at different levels of difficulty, and new ones added if desired. The next panel is for exercise parameters, where the teacher selects a collocation type and, if desired, enters a word or words that must appear in all collocations—*take* and *make*, in this case.

The next panel gives the number of sentences to choose from, and is automatically updated following any parameter change. For example, this collection includes 180 sentences that contain verb + noun collocations, but this changed to 10 in the interface when the words *make* and *take* were entered, because this is the number that includes those words. In the next panel the teacher decides how many sentences to use in the exercise, whether learners have to guess the first or last word of collocations, whether the missing words are shown and whether hints are allowed. The buttons underneath, Review, Display, Print and Save, allow teachers to review the sentences and collocations that have been chosen, try out the exercise just as a student would, print it on paper, and save it for students to use. The last three are self-explanatory; we look at the first in more detail.

All exercise content is determined automatically based on the parameters specified. However, teachers may not be satisfied with what they see because (1) the question text may contain complicated structures or difficult vocabulary items that could hinder learning; (2) students have already mastered some collocations that have been retrieved; (3) there are errors in collocations (e.g., a noun + noun type may be marked as verb + noun); or (4) the items are unsuitable for other reasons. During the review process teachers apply quality control, discarding unsatisfactory questions and modifying the automatically generated answers or replacing them with their own.

CONCLUSION

This paper has described a scheme for supporting collocation learning with a digital library. The design is guided by the psychological conditions that facilitate acquisition: noticing, retrieval and generation. Articles such as those that teachers have prepared for their students are built into a digital library collection and augmented with automatically identified collocations organized by syntactic composition and ranked by frequency. Once the collection has been constructed, students interact through an interface specially designed for learners to seek, study, and collect collocations. While reading the articles, their attention is drawn to highlighted examples. They expand and enrich their knowledge by examining related items retrieved from a vast corpus of naturally occurring collocations, and by studying exemplary text in the British National Corpus and live samples from today's Web. They select and collect collocations for their own use later.

We have developed four activity types. For each one, teachers can generate unlimited numbers of exercises, tailored to their classes, from the content of the digital library, using a specially created interactive exercise design interface. Common Alternatives is a game-like activity that help learners maintain high motivation. Fill-in-blanks, Correcting

Errors and Multiple Choice are traditional collocation learning activities with proven pedagogical value.

Evaluation of the system is ongoing, and will lead to refinements in both the collocations it generates and the interfaces through which teachers and learners use it. But preliminary experience with student users indicates that this digital library already provides a new and engaging way of enriching their knowledge of collocations.

ACKNOWLEDGEMENTS

We gratefully acknowledge the stimulating environment provided by the digital library laboratory at the University of Waikato. This research is funded by the Royal Society of New Zealand Marsden fund.

NOTES ON CONTRIBUTIORS

The contributors come from the University of Waikato, Hamilton, New Zealand and work together on a project funded by the Royal Society of New Zealand Marsden fund. Shaoqun Wu is a doctoral candidate in the Computer Science Department. Ian H. Witten is a professor in the Computer Science Department and has published extensively in the area of digital libraries. Margaret Franken is an applied linguist in the School of Education and has an interest in data-driven learning.

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.

Bates, M.J. (1989) "The design of browsing and berrypicking techniques for the online search interface." *Online Review*, 13, 407-424.

Benson, M., Benson, E. & Ilson, R. (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* (pp. 227-244). Philadelphia, PA, USA: John Benjamins Publishing Company.

Brown, D.F. (1974). "Advanced vocabulary teaching: The problem of collocation." *RELC Journal*, 5(2), 1-11.

Channell, J. (1981). "Applying semantic theory to vocabulary teaching." *English Language Teaching Journal*, 35, 115-122.

Conzett, J. (2000). "Integrating collocation into a reading and writing course." In *Teaching Collocation*, edited by Lewis Michael. 70–87, LTP, England.

Farghal, M. & Obidedat, H. (1995). "Collocations: A neglected variable in EFL." *IRAL*, 33(4), 315-331.

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

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.

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.

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

Hill, J. (1999). "Collocational competence." *ETP* 11.

Hoey, M. (2000). "A world beyond collocation: new perspectives on vocabulary teaching." In *Teaching Collocation*, edited by Lewis Michael. 224-243, LTP, England.

Hornby, A.S. (1974). *Oxford Advanced Learners' Dictionary*. Oxford: Oxford University Press.

Lewis, M. (1993). *The lexical approach*. Language Teaching Publication, England.

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

Lewis, M. (Ed.) (2000). *Teaching Collocations*. Hove: Language Teaching Publications.

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

Nation, P. (2001) *Learning vocabulary in another language*. 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.

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

Palmer H.E. (1933). *Second interim report in 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.

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

Stevens, V. (2001). "Calico software review. Concordance ver 2.0." *CALICO Journal*, 19(3), 690-708. Retrieved October 22, 2009, from https://www.calico.org/p-180-Concordance%20(62001).html

Swan, M. (1996). *Language teaching is teaching language.* Plenary IATEFL.

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. (ERIC Document Reproduction Service No. ED457690).

Witten, I.H., Bainbridge, D. and Nichols, D.M. (2010). *How to Build a Digital Library*. Morgan Kaufmann, Burlington, MA (second edition).

Wray, A. 2002. *Formulaic Language and the Lexicon*. New York: Oxford University Press.

Wu, S. & Witten, I.H. (2006). "Towards a digital library for language learning." *Proc European Conference on Digital Libraries*, 341–352, Alicante, Spain.

Wu, S. & Witten, I.H. (2007). “Content-Based Language Learning in a Digital Library.” *Proc. International Conference on Asian Digital Libraries*, 424-433, Hanoi, Vietnam.

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.

Yorio, C.A. (1980). “Conventionalized language forms and the development of communicative competence.” *TESOL Quarterly*, 14(4), 433-442.