First Person Singular: A digital library collection that helps second language learners express themselves

Shaoqun Wu and Ian H. Witten
Department of Computer Science
University of Waikato
Hamilton, New Zealand
{shaoqun, ihw}@cs.waikato.ac.nz

Abstract: We are using digital library technology to help language learners express themselves by capitalizing on all the human-generated text available on the Web. From a massive collection of *n*-grams and their occurrence frequencies we extract sequences that begin with the word "I", sequences that begin a question, and sequences containing statistically significant collocations. These are preprocessed, filtered, and organized as a digital library collection using the Greenstone software. Users can search the collection to see how particular words are typically used, and browse by syntactic class. The digital library is richly interconnected to other resources. It includes links to external vocabularies and thesauri so that users can retrieve words related to any term of interest, and links the collection to the web by locating sample sentences containing these patterns and presenting them to the user. We have conducted an evaluation of how useful the system is in helping students, and the impact it has on their writing. Finally, language activities generated from the digital library content have been designed to help learners master important emotion related vocabulary and expressions. We predict that the application of digital library technology to assist language students will revolutionize second language learning.

Keywords: language learning, digital libraries, language teaching, web corpora, word *n*-grams

1. INTRODUCTION

Everybody wants to talk about themselves: their thoughts and feelings, what they have been doing and what they plan to do. In other words, we all aspire to become expert in the first person singular. But in a foreign language, it is not easy. Language learners often complain that they cannot express what they think, feel and do. You might answer a simple question like "How are you today?" factually ("My head aches"), perfunctorily ("OK"), or provocatively ("I'm feeling sexy"). But students find it hard to go beyond simple statements and talk about their feelings at greater depth. And the same applies to all forms of self-expression.

Part of the reason is that learners have not experienced enough of the language to express themselves in the first person in ways that sound natural. As Moskowitz (1978) notes, curricular material tends to focus on facts and everyday transactions, only rarely touching on vocabulary that is appropriate for communicating more subjective aspects of everyday life. To help remedy this she advocates integrating a humanistic approach to language teaching with a planned curriculum to promote self-actualization and self-esteem, so that students can express themselves meaningfully in the first person.

To be able to talk fluently about themselves, learners must command appropriate linguistic resources. This paper describes how to identify short sequences starting with (or, in some cases, containing) the word "I" and use them to help learners acquire important "*I*-vocabulary" and "*I*-expressions." Fluency does not blossom from a comprehensive lexicon of difficult words, nor even from familiarity with the most common ones. Instead, it requires an internalized repertoire of phrases and expressions composed of words used in everyday life (Lewis, 1993). Consequently our digital library focuses on the most commonly used English words and their associated expressions.

How can ordinary, everyday language be captured? Our approach is to capitalize on the text on the World-Wide Web, in particular the vast set of *n*-grams from the Web that Google has made available.[1] Only digital library technology can provide searching and browsing functions for such a massive body of text. Our system is based on the Greenstone software (Bainbridge et al., 2004). We have built a collection called "First Person Singular" that allows learners (and teachers) to locate phrases associated with a

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

Table 1. Number of units

(a) in the original n-gram collection

tokens	1,024,908,267,229	10^{12}
sentences	95,119,665,584	0.95×10^{11}
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}

(b) in the final collections

I-grams	*after initial filtering*	3,430,904	3.4×10^{6}
	in final collection	346,812	0.35×10^{6}
Wh-grams		33,842	0.03×10^{6}
Collocations		6,126,660 2-grams 15,464,201 3-grams	21.6×10^{6}

particular word, as well as synonyms, antonyms, and collocations. The digital library enables sentences containing these patterns to be retrieved from the Web and presented to the user as examples. We have conducted an evaluation with actual language students, and the results show the potential usefulness of the system in helping students correct grammar errors, generate text and expand text.

In this paper we first examine the *n*-grams Google has supplied and explain how to extract a subset that is useful for language learning. We then describe the design and implementation of the First Person Singular digital library collection: how it is built and the searching and browsing facilities it includes. Next we show how results obtained from the collection can be augmented by retrieving related material from the Web and the British National Corpus (BNC). Then we describe the findings from an evaluation with actual students.

We round out the paper by describing some language activities that we have designed to help students master important vocabulary and expressions. Although these have not been evaluated formally, they point the way to an exciting future. We believe that digital libraries in general—not just the First Person Singular collection described here—have the potential to revolutionize the area of second language learning by providing unlimited volumes of practice exercises that are generated automatically, directly from a library's contents. This general strategy will allow any digital library collection to be used as a basis for language learning exercises.

2. *N*-GRAMS FOR LANGUAGE LEARNING

Our starting point for this digital library is a corpus of word n-grams in English, ranging from unigrams or single words to 5-grams or sequences of five consecutive words, along with their frequency counts. These were generated by Google from approximately one trillion word tokens of text on publicly accessible web pages: a staggeringly large body of natural English. *N*-grams appearing 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 1a summarizes the extent of the corpus. The number of *n*-grams increases as *n* grows beyond 1, peaks at *n*=4, and then begins to decay.

Table 2a shows a few of these lines in the raw data files supplied by Google. They are very simple: each n-gram occupies a line:

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

Table 2b shows the ones that remain after the cleaning and selection operations described below.

Three subsets were extracted from the corpus and used as raw material for the First Person Singular digital library collection: 5-grams that begin with the word *I* (which we call *I*-grams), 5-grams that constitute the beginning of a question (*Wh*-grams), and 2- and 3-grams that contain statistically significant collocations. Restricting attention to 5-grams provides the greatest context, as well as reducing the collection to a manageable size.

Before creating these sets an algorithm was applied to regularize case, for as Table 2a illustrates the original corpus contains a haphazard mix of upper- and lower-case. We use OpenNLP's sentence tagger to identify proper nouns and the pronoun "I", and then capitalize them and make the remaining characters lower-case.[2]

I-grams

These selection steps were applied to form the subset of *I*-grams that are placed in the digital library:

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

Table 2. Sample n-grams

(a) from the original n-gram collection

I ASKED FOR ! </S>	53
I ASKED FOR A SO	67
I ASKED FOR I SAW	52
I Asked For , Inspirational	40
I Asked For It </S>	52
I Asked For Love </S>	66
I Asked For More Butter	318
I Asked For Reinforcements ,	77
I Asked For That robb06	926
I asked for ? </S>	1072
I asked for Anonymous --	339
I asked for Christmas .	61
I asked for Courage …	80
I asked for a 12	170
I asked for a 2	51
I asked for a </S>	237
I asked for a <UNK>	130
I asked for a >	71
I asked for a CD	83
I asked for a Coke	75
I asked for a Mgr	49
I asked for a river	163
I asked for a roll	43
I asked for a room	1395
I asked for a ruling	55
I asked for a sample	183

(b) after cleaning and selection

I asked for I saw	52
I asked for more butter	318
I asked for a CD	83
I asked for a Coke	75
I asked for a river	163
I asked for a roll	43
I asked for a room	1395
I asked for a ruling	55
I asked for a sample	183

1. select 5-grams that start with the word *I*
2. discard unless all words belong to a prespecified vocabulary
3. discard grammatically incorrect sequences.

Only a certain range of commonly accepted vocabulary is useful for language learning, and the second step checks each 5-gram against a standard word list and eliminates ones that include non-words or unusual words. We used the 47,224-word vocabulary of the million-word Brown corpus of natural English (Kucera & Francis 1967). This step removes most of the *n*-grams in Table 2a, leaving only the ones shown in Table 2b plus one other—*I asked for a so*, which is removed by the next step. Naturally, this means that certain legitimate phrases are removed from the corpus—such as ones that contain neologisms like (ironically) *Google*.

Surviving 5-grams are parsed into phrases by the OpenNLP chunker, and suspect ones discarded. OpenNLP uses the Penn Treebank tagset[3]—producing, for example, for the 5-gram *I asked for a room*

[NP **I**/PRP] [VP **asked**/VBD] [PP **for**/IN] [NP **a**/DT **room**/NN]

Square brackets indicate phrases, at the beginning of which is a phrase level tag that identifies the syntactic role of the phrase. This fragment contains the noun phrase (NP) *I*, the verb phrase (VP) *asked*, the prepositional phrase (PP) *for*, and the noun phrase (NP) *a room*. Word level tags follow each word and convey tense and number information: *I* is a proper pronoun (PRP), *asked* is a past-tense verb (VBD), *for* is a preposition (IN), *a* is a determiner (DT), and *room* is a singular noun (NN).

Tagged sentences are matched against a regular expression that specifies a noun phrase (NP), followed by a verb phrase (VP), optionally preceded by adverbial phrases (ADVP); and may optionally end with a noun, prepositional (PP), adverb, adjective (ADJP), particle (PRT) phrase or clause (SBAR). The effect is to discard ill-formed expressions such as *I asked for a so*. These selection steps reduce the number of 5-grams from 1.2×10^9 in the raw data to 3.4×10^6 (Table 1b).

Wh-grams

The same three steps were applied in slightly modified form to create the second subset. Here, 5-grams that begin with a question word (*When, Where, Why, How, What, Who, Whom* and *Which*) are selected in the first step, and the grammatical test in the third step checks that the question word is followed by an

[3] http://www.cis.upenn.edu/~treebank/

auxiliary verb (e.g. *do*, *does*, *have*, *am*, *are*) or modal verb (e.g. *can*, *will*, *would*), and then the word *I*. The process yields 34,000 *Wh*-grams (Table 1b).

Collocations

Collocations are short sequences of words that are commonly found together—more often than one might expect from their individual frequencies. For example, native speakers prefer the collocation *heavy rain* to the non-collocation *big rain*, or *totally convinced* to *absolutely convinced*. Native speakers carry in their heads hundreds of thousands—possibly millions—of collocations, ready to draw upon for fluent, precise and meaningful utterances (Lewis, 1997). This presents learners with a daunting challenge.

We adopt thirteen collocation patterns identified by Benson et al. (1986): two- and three-word syntactic fragments such as *adjective+noun* and *adverb+adjective*, and phrasal verbs of the form *verb+preposition*—for example, *make up* and *take off*. The collocation database is built in three steps:

1. select 2- and 3-grams that match the thirteen syntactic patterns
2. discard unless all words belong to a prespecified vocabulary
3. discard unless the word frequencies indicate a statistically significant pattern.

The OpenNLP tagger is again used to assign part-of-speech information, and the Brown vocabulary is used for step 2. Step 3 is an implementation of a standard method for collocation detection, which is to calculate the *t*-statistic and discard those whose *t*-value falls below a certain threshold. We used a value corresponding to a confidence of 99.5%, as recommended by Manning & Schütze (1999). The result is a set of 21.6×10^6 collocations; 20% of which are 2-grams and the remainder 3-grams (Table 1b).

3. BUILDING THE DIGITAL LIBRARY

We use the Greenstone digital library software, which allows librarians to build large collections of documents and metadata and serve them on the web.[4] In Greenstone, each collection offers readers different facilities, depending on the metadata available and the choices made by the collection designer. Full-text search is almost always included, possibly on different parts of the documents (or metadata), and browsing structures are built on particular metadata types at the discretion of the designer. First Person Singular is a collection of *n*-grams selected from the web as described above. We used Greenstone 3 (version 3.03), and took advantage of its flexible architecture to write specialized services through which users access it.

The First Person Singular collection

The primary collection contains phrases beginning with the word "I". Being a general-purpose digital library system, Greenstone works with a basic unit of *document*.[5] Documents consist of *sections*, and Greenstone accommodates hierarchies of sections—typically chapters, sections, subsections, etc.—of arbitrary depth. Searching can be at both the document and section level.

Making each *I*-gram a separate document yields a collection with 3.4 million documents; putting them as separate sections of the same document yields a document with 3.4 million sections. Both are undesirable for performance reasons. As a compromise, the *I*-grams were grouped based on the first adjective and verb encountered. For example, *I was a little disappointed* and *I was disappointed in the* are placed in the same file, along with all other *I*-grams that have *disappointed* as the first adjective. The smallest documents correspond to rare words and contain just one section. The largest have many thousands of sections, which again impacts search performance, but we decided to truncate them to the 100 most frequent *I*-grams containing that adjective and verb. This yielded 35,000 documents with an average of about 10 *I*-grams each. Following this selection procedure, the final collection contained 347,000 *I*-grams, about 10% of the original figure (Table 1b).

Greenstone has a scheme of "plugins" that allows it to deal with different document formats in an extensible manner. We developed a custom plugin to process files that contain lists of *I*-grams, treating each one as an independent document. It extracts metadata corresponding to *frequency*, *word type* and *tense*. For the last two, the plugin identifies the nouns, verbs, adjectives, adverbs and prepositions in each *I*-gram and associates them with that document as metadata. In the case of verbs, the root form is determined. (For example, the verbs *enjoyed*, *enjoying* and *enjoys* all share the same root form, namely *enjoy*.) Automated morphological decomposition has an inevitable risk of error, a risk that is unacceptable

[4] http://www.greenstone.org
[5] Documents may contain text or multimedia, though the latter does not concern us here.

in a system designed for language learners. Consequently to find root forms we simply consult three word family lists that have been downloaded from the Complete Lexical Tutor website:[6]

1. the most frequent 1000 headwords
2. the most frequent 2000 headwords
3. Coxhead's Academic Words (Coxhead, 1998).

Despite being incomplete—together they cover only 2,500 headwords—these lists are adequate for our purpose because the aim is to help learners master the most common words.

An important step in the design of any Greenstone collection is to determine its searching and browsing facilities. This collection has four full-text indexes, described below, and a hierarchical browser that allows users to browse by wordlist and see the *I*-grams in which any particular word appears.

Subsidiary indexes

Alongside the main collection, specialized search facilities are provided by two subsidiary indexes. One contains *Wh*-grams, divided into documents (one per *Wh*-word), and is equipped with a standard Greenstone full-text index. The other has forward and reverse indexes of the two- and three-word collocations, and is consulted using a simple Java program.

Retrieval services

Greenstone's full-text search can only respond to ranked, Boolean, and phrase queries. However, it has an extensible service-oriented architecture in which specialized services can be created to fulfill non-standard requirements (Bainbridge et al., 2004) . We implemented five new Greenstone services:

- find *I*-grams that start or end with a given query term or terms
- find *Wh*-grams that contain the query terms
- find collocations that contain the term
- retrieve examples from the web and the British National Corpus
- find synonyms, antonyms, related words, and associated words from auxiliary resources.

In addition, we wrote XSLT statements to display the information retrieved by the services in appropriate ways, as illustrated below—another extensibility feature of the Greenstone system.

4. USING THE DIGITAL LIBRARY

Learners are overwhelmed by choice when trying to construct a sentence using a newly acquired vocabulary item. What sentence structure should be used? What tense? Would native speakers say it that way? What linguistic hedges might reduce the impact of the utterance? How could its impact be strengthened? Our system allows learners to study *I*-grams that contain particular terms. We have implemented four ways for learners to examine the usage of a word: phrases that contain it, its prefix and suffix patterns, and questions that use it. These are described below. We next describe the browsing operations that are built into the digital library collection, and then look at the facilities provided for consulting external auxiliary resources, including the web and the British National Corpus.

Phrases containing a particular word

I-grams are 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 1 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. Each phrase is assigned tense metadata during the collection building process. For example, *I have been happy* is determined to have present perfect tense because it matches the "*have been* + adjective" pattern.

Clicking the phrase or the image icon that follows the frequency retrieves samples from the Web and the British National Corpus respectively (see the Section below on "Using auxiliary resources"). 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.

More information on the query term appears above the search results: links to synonyms, antonyms, and related words, each grouped by part of speech, and to associated words and collocations. We discuss these further below, under "Using auxiliary resources." If more than one term is typed into the search box,

[6] www.lextutor.ca/lists_download

phrases containing each one are presented under the various categories in the search results. Quotation marks can be used to signify that the query should be treated as a phrase. It is interesting and often instructive to lengthen a chosen phrase word by word and see how the popular contexts change.

Figure 1. Search facilities that the collection provides

Table 3. Patterns that follow the words *love* and *hate*

love		hate	
you		**you**	
... love you ...	246236	... hate you ...	20825
... love you I love ...	106610	... hate you I hate ...	8165
... love you so much ...	97106	... hate you so much ...	6675
... love you because I ...	43553	**the**	
... love you more than ...	41664	... hate the fact that ...	20010
... love you and I ...	37987	... hate the ...	17283
... love you forever ...	34593	... hate the idea of ...	11524
the		... hate the thought of ...	6754
... love the fact that ...	69154	... hate the way you ...	6502
... love the way you ...	62343	**myself**	
... love the idea of ...	49511	... hate myself and want ...	16722
... love the smell of ...	40600	... hate myself for losing ...	6322
these		**him**	
... love these shoes so ...	35925	... hate him ...	6893

Phrases preceding a particular word

Given a word, learners can study language patterns that frequently precede it. In the pull-down menu near the top of Figure 1b *Phrases preceding* has been selected, and in this case the search results are grouped by words that appear in the preceding context. They show that the most common sentence structure with *disappointed* takes the form *be* + *disappointed*, and again the past tense is most common. The hedges *very, really, so, extremely, quite, somewhat rather, pretty* are often used in this context.

Phrases following a particular word

This allows users to explore what words and phrases follow a particular word. Figure 1c shows that the prepositions *with* and *in* commonly follow *disappointed*, and that *disappointed* is often followed by *that*- and *when*-clauses. These indicate useful sentence structures that learners can employ in conversation (or writing) when they want to express disappointment about something. In fact, learning the common clause patterns—that is, the basic sentence structures—associated with particular verbs helps improve learners' communicative fluency enormously.

Table 3 contrasts the patterns that follow the words *love* and *hate* (obtained by the same method but, for succinctness, displayed in tabular form rather than as screenshots). This not only reveals what people commonly love or hate, but also helps learners choose appropriate words when they want to express similar feelings. It is disturbing that people often say they hate themselves! Also, there is evidence here that women tend to talk more about their feelings than men (*hate him*, and maybe *love these shoes*).

Questions

Users can search for questions that contain particular words to learn how to formulate their own questions in the first person singular. Behind the scenes, the system searches *Wh*-grams, one of the two subsidiary indexes described above. Figure 1d shows a question search for the word *happy*: the results are grouped by *Wh*-word. Existential angst pervades the output shown.

Browsing

For *I*-grams, four wordlists were generated and sorted into inverse frequency order:

1. all words regardless of type;
2. main verbs;
3. main adjectives;
4. modal words.

The digital library collection was configured with browsing facilities that allow users to examine these wordlists. Figure 2a 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* and *like*. Figure 2b gives the language patterns that are associated with *think* in the first person context.

For the 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

Table 4. Collocations for *sad*

very sad	731	sad that	626
so sad	679	sad to	604
really sad	433	sad day	452
pretty sad	267	sad thing	423
little sad	266	sad story	347
bit sad	238	sad fact	317
just sad	233	sad news	311
quite sad	208	sad because	296
kinda sad	201	sad about	279
rather sad	197	sad part	275

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 phrases *I think that* and *I think he*. However, neither study exposes the surprising fact that the pattern *I* + *think* occurs most frequently as a negative statement.

Using auxiliary resources

Learners find feelings difficult to articulate, particularly in rich and appropriate ways. We use external databases—WordNet, Roget's thesaurus, and the Edinburgh Word Association thesaurus—to retrieve words related to or associated with a particular term. In WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct concept.[7] Roget is a widely used thesaurus; the online version contains 15,000 words.[8] The Edinburgh Word Association thesaurus contains word association strengths derived experimentally from human subjects.[9] We downloaded these resources and developed computer programs to incorporate them into the web pronoun phrases collection.

Each resource is filtered to remove words and phrases that do not appear in the collection. This eliminates usage that rarely occurs in self-expressions today, and prevents learners from becoming overwhelmed with choice. For example, WordNet contains these synonyms for the adjective *sad*:

> ***bad***, *bittersweet,* ***depressing****, depressive,* ***gloomy****, saddening, doleful, mournful, heavyhearted,* ***melancholy****, melancholic,* ***pensive****, wistful, tragic, tragical, tragicomic, tragicomical, sorrowful, deplorable, distressing, lamentable,* ***pitiful****,* ***sorry***

Only those in boldface, a small minority, actually occur in the First Person Singular collection. On the other hand, all three WordNet antonyms—*glad, joyful* and *good*—occur. Related words from Roget that appear in the collection include *unpleasant, unacceptable, touching, troublesome, fearful, hard* and *cutting*, while associated words in the Edinburgh database include *happy, unhappy, bad, cry, death, girl, glad* and *me*.

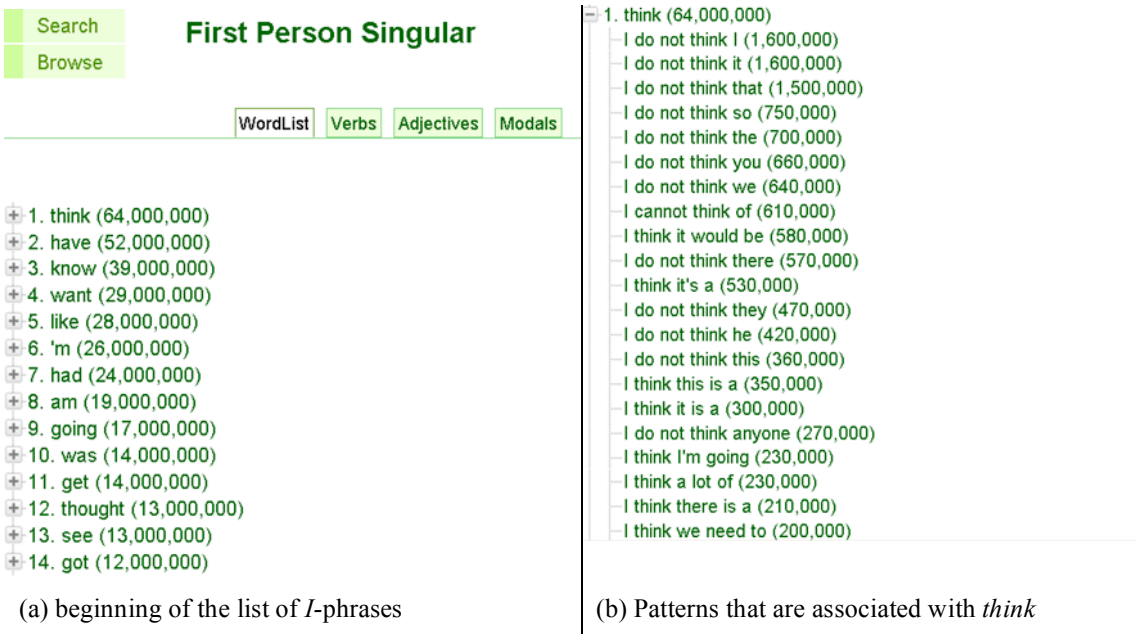
Preceding the query results in Figures 1a–c are links to information extracted from each database: synonyms and antonyms from WordNet, related words from Roget, and associated words from the Edinburgh thesaurus. Synonyms, antonyms and related words are further grouped by syntactic class. The words themselves appear on a separate page, and are sorted by frequency in the n-gram corpus.

The final link that precedes the query results is to collocations that are generated using 2- and 3-grams. They are grouped according to whether they occur on the left or right of the target word and ordered by statistical score. Table 4 shows collocations for the word *sad.* On the left are common modifiers used to boost (*very, so really, pretty*, *quite*) or hedge (*little, bit, just, kinda*) the expression of sadness. On the right, s*ad story, sad news* and *sad day* are strong collocations of type *adjective+noun*. Collocations are highlighted if they occur in the web pronoun phrases collection; users can click the link to find phrases containing the collocation.

Retrieving samples from the web and from the British National Corpus

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.

[7] See http://wordnet.princeton.edu for more information.
[8] See http://www.gutenberg.org/etext/10681 for more information.
[9] See http://www.eat.rl.ac.uk for more information.



(a) beginning of the list of *I*-phrases

(b) Patterns that are associated with *think*

Figure 2. Browse facilities that the collection provides

The first 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. 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 disables automatic queries from computer programs other than Web.

Figure 3 shows 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 3a 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 3b. They tend to be more coherent than the Web snippets, and are presented in 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.

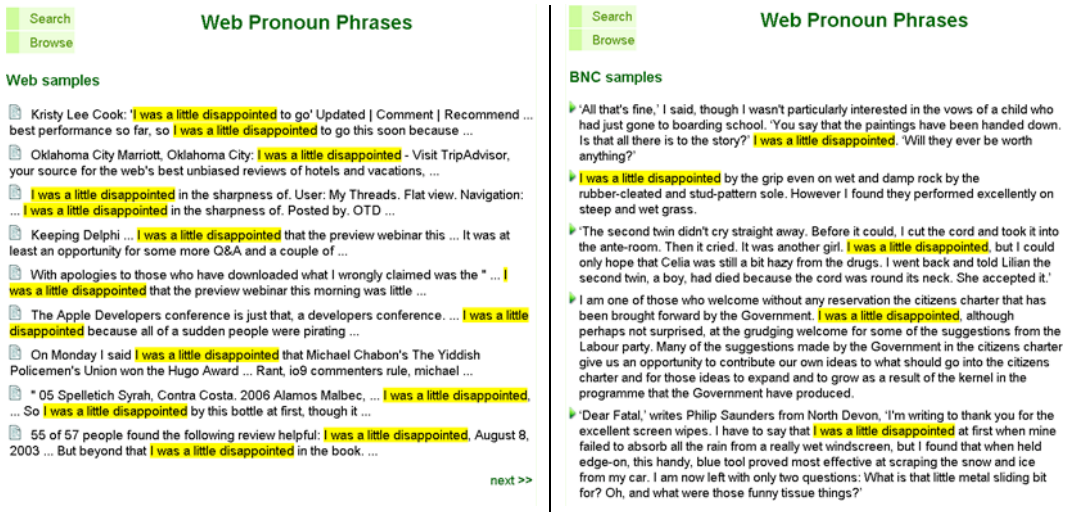
5. EVALUATION

We conducted an evaluation on the usefulness of the digital library system for supporting writing in the context of self-expression, focusing on how students use the system and the impact it had on their work.

Participants and procedure

Twelve language students were recruited, six females and six males aged from 19 to 40 years. They were native speakers of six different languages. Their abilities in grammar, reading, speaking, and writing had been graded by the college at which they were studying. Grammar and reading were tested by the Oxford entry test, which yields two scores for each skill. Writing and speaking were tested by a writing task and interviews with teachers, who gave scores for each. The four scores were combined in order to allocate students to different classes.

All our participants were from the same intermediate class. However, their abilities varied greatly—for example, some excelled in speaking but performed poorly in writing and vice versa. Despite our best



(a) from the Web (b) from the British National Corpus

Figure 3. Samples retrieved for *I was a little disappointed*

efforts to ensure uniformity, we still ended up with participants who had a range of different writing ability. To compare their ability before and after using the system we asked them to write a 150–200 word description of themselves the day before the evaluation.

The evaluation was conducted during a 2-hour session in the computer lab of a local language school. In the first half hour we explained how the *I*-grams were gathered and introduced the functionality of the system. Then subjects were asked to prepare a personal profile of themselves for a home-stay, including their background, interests, likes and dislikes, and any other things that they thought would make them look interesting. They wrote on paper, and in order to track their changes they were instructed not to erase errors but to cross them out or rewrite above the text. They were encouraged to bring dictionaries and use them, because the system did not check spelling. They could request help from the teacher on how to use the system, or to explain unknown words. Finally, they were asked to circle any text fragments that the system had helped them generate or improve.

Each student was given an anonymous identifier, and their use of the system was recorded in detail and written to a log file. The log data includes (1) the search terms entered, (2) synonyms or collocations that were looked up, and (3) the retrieved samples, whether from the web or the British National Corpus. Data were recorded sequentially, with a timestamp to make it easy to trace the course of each student's work.

Results

Students using the system adopted one of two strategies. Most finished their writing first and then used it to check text they were uncertain of. Some students (three) used the system to help generate text by finding the correct usage of a word and suggesting suitable sentence structures. Most searches used content words as queries to find phrases containing words they were interested in. For example, they would search for *student*, *study* and *university* to describe their student status, or *like, love* and *hobby* to talk about their personal interests. In a few cases students searched on function words such as *been*, *will*, *why* and *when*.

The students produced fairly short texts, averaging 20 sentences per essay and 11 words per sentence. Grammatical errors, incorrect sentence structures, and incomplete sentences were scattered throughout their work. Because of the constraints of the topic—themselves—and their limited language ability, their writing exhibited a narrow range of vocabulary and few idiomatic expressions. For example, the four most common words used were *like*, *come*, *want* and *live*. Sentence structure was simple and basic. Most sentences began with a pronoun, followed by the main verb and a noun or prepositional phrase. Feelings and emotions were expressed in a rather plain way; linguistic boosters or hedges were rarely used.

Table 5 summarizes the log data. For each of the 12 students it shows the number of sentences in their text, the number of searches they launched, the number of times sample text on the web or the British National Corpus was viewed, and the number of lexical resources, i.e., synonyms and collocations, viewed. The last two columns give the positive or negative uses the students made in the text when using the system. We elaborate on these shortly.

Table 5. Summary of the log data

	sentences	searching	samples (web or BNC)	lexical resources (synonyms/collocations)	positive uses	negative uses
1	40	14	6	3	3	0
2	29	32	32	5	7	0
3	26	15	0	0	5	1
4	25	32	24	2	8	2
5	21	29	24	5	8	2
6	19	39	9	25	3	0
7	18	45	29	20	12	2
8	16	14	19	1	6	0
9	15	12	5	2	4	0
10	9	13	13	12	4	1
11	9	8	5	0	2	0
12	8	14	12	3	3	0
total	235	267	178	78	65	8

A total of 267 searches were conducted, ranging from 8 to 45 per student with an average of 22. Students evidently used the system actively, for searches outnumbered the sentences generated. Except for the first student, the number of searches correlates well with the amount of text produced, and also, with rare exceptions, with the number of look-ups on the web or the British National Corpus. It is encouraging to see that the students tried to understand samples in context before rushing to use them. Surprisingly, most samples viewed came from the web rather than the British National Corpus—perhaps because the latter snippets tend to be lengthy paragraphs, and students were under time pressure to finish their essay.

The number of times lexical resources were consulted—in most cases five or fewer—paints a different picture. The logs reveal unexpected searches for words such as *and* and *will*, which suggests that some students did not understand the nature of these resources. However, students 6 and 7, whose writing skills were the best amongst all participants, used them extensively. This indicates that more advanced learners are more likely to explore alternative language usage.

What impact did the system have on the students' work in terms of text generation and revision? We went through their text manually and identified 73 uses. A "use" is identified based on:

1. the student indicated use of the system by circling the text;
2. there was no evidence of such language usage in the text the student produced the previous day;
3. log data confirmed that the altered text was suggested by the system.

The first criterion provides strong evidence of use, but in many cases students forgot to circle the text and consequently the second criterion was used as well. (For the second criterion, recall that students were asked to write two pieces of text: the first without using the system and the second during the evaluation the following day.) Here it is important to differentiate errors from mistakes. Students make language *errors* because they have no knowledge, or limited knowledge, of the relevant linguistic feature—for example, one wrote *we want do something* because he did not know the correct usage of the verb *want*. Students make language *mistakes* when they write the wrong thing despite knowing the rules: in this case they are capable of recognizing the mistake and fixing it themselves. Mistakes were discarded if there was evidence of correct use elsewhere in the text.

It is important to note that a "use" of the system does not necessarily guarantee that the result is correct. The student might misinterpret the samples the system provides, resulting in a negative use. For example, one student changed *I like eat Taiwan's snack* to *I would like to eat Taiwan's snack* after searching for the word *like*. Unfortunately, in the original context the first version, although grammatically incorrect, is nevertheless more appropriate. This negative use is attributed to the student misunderstanding the pragmatic meaning of *I would like to*. Moreover, *I would like to* is the dominant usage of the verb *like* and therefore accounts for most of the search results, which confused that student. On the other hand, a positive use is a correct use of the search result in a text, leading to correct grammar, better sentence structure, and idiomatic, natural expressions such as *it would be better to*, *I enjoyed it a lot* and *I wish I could*.

There are 65 positive and 8 negative uses, which means that every 3½ searches resulted in a use, 90% of which were positive. Most negative uses were due to inadequate pragmatic knowledge of a language

Table 6. Samples extracted from student text

category	original	new
checking grammar	I was born Seoul I went performance hall for singing I've been in NZ since four month ago I want find a good job	I was born in Seoul I went to performance hall for singing I've been in NZ since April I want to find a good job
generating text		I graduate from the music school My sister is very good at cooking I wish I could become a social worker I have developed interest in movies I think it is important to learn English I can travel all over the world
expanding text	I am close to them It is a beautiful place It is hard to speak English I thought to find another home-stay	I was born and raised in Taiwan I am very close to them It is a absolutely beautiful place It is really hard to speak English I thought it would be better to find another home stay
confirming text	I did my best to study English I cannot afford to lose more time	

expression, for example, the difference between *my friend was performing* and *my friend was going to perform*, or *I was singing* and *I have been singing*.

Now let us look at what the students used the system for. We grouped uses into four categories:

1. checking grammar
2. generating text
3. expanding text
4. confirming text.

Table 6 gives some samples extracted from student text for each category.

In the first category, students used the system to help correct grammar errors, find the right prepositions, correct verb forms and use conjunctions correctly. The system provides a wealth of examples of usage of common verbs such as *go*, *want*, *continue* and *live*, which resulted in many corrections. One student even changed *I've been in NZ since four month ago* to *I've been in NZ since April* on searching for *been*. We did not point out errors in student text, so some uncorrected errors remained.

In the second category, some students constructed sentences based on the samples they found in the collection. They either used them directly or modified them to suit their need. For example, the sentence *I enjoyed spending time with my close friend* stemmed from the *I*-gram *I enjoyed spending time with*. In one particular text we found seven idiomatic expressions such as *I wish I could*, *I think it is important to* and *is very good at*. The original version of this text was mostly made up of simply structured sentences and showed no evidence that the student knew these idioms. This student told us that she could write a text in different ways by using the phrases found in the system.

Some students found it difficult to make their writing interesting and colorful because of their limited stock of vocabulary and idiomatic expressions. In the third category, many students made efforts to expand the text using samples provided by the system. A common strategy involves the use of language boosters and hedges, including adverbials such as *very*, *really*, *so much*, *a lot*; expressions such as *I thought it would be better*; and collocations such as *born and raised*, *absolutely beautiful*.

The fourth category is use of the system to confirm text that has been written. A student's original text may show that they know the language features in question, but they may nevertheless consult the system for confirmation. For example, one student searched for the word *best*, and then checked the sample *I did my best to*—despite the fact that he has already used it correctly.

In summary, the results of this evaluation suggest that the First Person Singular digital library is a valuable resource for language learning, particularly in helping students to express themselves in richer and more natural ways. Proficient learners can use the collection to generate text as well as revise it, but the limited vocabulary knowledge of less proficient learners restricts them to revisions. However, most

(a)

Match beginnings with endings	
Beginnings	**Endings**
1. I am just wondering if	a. should be doing
2. I am trying hard to	b. you have to be very good in math to become a pyrotechnition?
3.I get the feeling that	c. be nice but cheney and bush are really pushing my buttons what do I do about it?
4. I am doing what I	d. everyone needs a next big thing, and if there is not one, they create it.

(b)

Make the expression stronger by choosing a word (or words) in brackets and putting it in the right place:
1. I am upset because this might be the last time he is in the films (really, all).
2. I was disappointed but there was nothing I could do about it. (bitterly, absolutely)
Make the expression weaker by choosing a word (or words) in brackets and putting it in the right place:
1. I was upset when they told me. (extremely, a bit)
2. I was annoyed that he was late. (somewhat, rather)

(c)

Use one of the words *speak, tell, say, talk* to complete these sentences
1. I got to _____ to him a bit and he was such a friendly guy.
2. First of all I can not _____ for him, Love was one subject matter that he knew a lot about, the good and the bad parts of it.
3. I like this guy but I can not _____ if he likes me back.
4. I am sorry to _____ I don't find that very helpful

(d)

Put the verb *get* in its proper place
1. I a chance to talk to Bun B today and do a little interview **got**.
2. I a kick out **get**.
3. Perhaps when I finally around to watching Alias, it will change my point of view **get**.
4. I always something from my mom, despite my age **get**.

(e)

Complete the sentence in a way that makes it true for you:
1. I get upset when I
2. I get very impatient if

(f)

Use one of these word pairs to complete the sentence: ***proud and glad, willing and able, small and despised, confident and sure,*** **and** ***fine and dandy.***
1. I am ____ and ___ of myself, or so I seem.
2. I am ____and ___ to be a Singaporean.
3. I am ____and ___ to work with the school children's needs.
4. I am ____ and ___.
5. I am ____ and ___, but I do not forget your firm advice.

Figure 4. Language activities automatically generated using the content of the collection

student text demonstrated positive effects at the lexical, grammatical and perhaps most saliently the pragmatic level.

The system has several limitations. The collection only contains *I*-grams, but self-expressions do not necessarily begin with the word *I*. One possibility is to include *my*-grams as well. Some students found the lexical resources unsatisfactory because they could not find the words they wanted. One complained that the samples shown are similar; he wanted different phrases on the first page of search results. We acknowledge this problem: results are sometimes flooded with phrases dominated by particular structures. One remedy is to remove phrases with similar structures, but the extent to which this should be done is unclear. Last but not least, students must know the word before they can use the system. What if they only have a vague idea of what they are seeking? A wordlist could be compiled grouped into different categories of feeling or emotion—*happy*, *sad*, *like*, *dislike*, *angry*, …—and made available to students. However, this would have to be manually selected and categorized by language instructors.

6. USING DIGITAL LIBRARY CONTENT FOR LEARNING ACTIVITIES

We have described how language learners can search and browse the content of the First Person Singular collection, and obtain from the web or the British National Corpus sample text containing the language

Table 7. Common five-grams of *speak, talk, say, tell*

speak	I can not speak for
	I did not speak out
	I am speaking of the
talk	I am not talking about
	I am talking to you
	I got to talk to
say	I can not say I
	I am sorry to say
	I am not saying it
tell	I can not tell you
	I can not tell if
	I am here to tell

fragments they find in the collection. We have also explained how students can use these facilities to help improve their writing.

Another practical application for this content is the automatic construction of language learning exercises. There are many web sites that provide practice exercises for language students. However, they tend to be constructed in an *ad hoc* way, and use language examples that have been chosen manually by language teachers. Digital libraries offer the potential for far more extensive text corpora to be tapped automatically to produce a wide range of exercises that draw on a huge body of material. By way of example, we show in this section how the material in the First Person Singular collection can be used for exercises that are generated automatically by the system. To illustrate the potential of this approach we describe five possible language activities.

Sentence heads

This activity requires learners to match the first part of a sentence with its ending. It focuses on common, and therefore important, sentence heads. The exercises are generated in three steps:

1. select the target sentence heads
2. retrieve sample sentences from the Web
3. split these sentences into two parts.

The person constructing the exercise—whether teacher or student—submits a list of words to provide a focus for it. For example, they might choose *wondering*, *trying*, *feeling*, *doing*. The system retrieves items from the *I*-gram collection that contains those words. The teacher can then determine which of these head patterns should be used in the exercise; alternatively, this step can be skipped, in which case the system simply uses the most frequent patterns. Then the system retrieves documents from the Web, locates target sentences containing the *I*-gram, splits them into two parts, and scrambles them. The learner's job is to match which beginning goes with which ending. Figure 4a shows an exercise that uses *wondering*, *trying*, *feeling*, and *doing* in turn.

Semi-fixed expressions with qualifiers

Native speakers use qualifiers such as *quite, really, so* and *just* to strengthen or weaken the feeling that they are conveying. However, to achieve this, language learners tend to formulate more complicated expressions—ones that could easily be replaced by a simple modifier. This activity helps learners master these common and useful qualifiers to make what they are expressing sound strong, weak or negotiable in a simple and natural way. The teacher prepares a list of feeling-related words such as *annoyed, grateful, upset* and *disappointed*. The system uses these to retrieve related *I*-grams, along with the most frequently associated qualifiers. For example, common qualifiers for *disappointed* include *so, very, quite, pretty* and *rather*. Then the teacher manually sorts qualifiers into categories according to their degree of strength. Figure 4b shows an exercise that asks learners to use identifiers to strengthen and weaken expressions in sentences retrieved from the web.

Related verbs

It is hard for learners to differentiate between words with similar meanings, such as *speak, talk, say* and *tell*. This activity helps them make these distinctions by studying related collocation and sentence structures. The teacher gives a list of words and the system retrieves their most frequently used 5-grams, illustrated in Table 7 for *speak, talk, say* and *tell*. Then sample text containing these fragments is retrieved from the Web and used to construct the fill-in-the-blanks exercise shown in Figure 4c.

De-lexicalized verbs

One of the best ways to make one's spoken English more natural is to learn expressions that use the verb *get*. This is generally a far more productive way for learners to spend their time and energy than studying unusual new words. We are working on automatically constructing different kinds of *get*-related exercises. Figure 4d shows an exercise that asks learners to put the verb into its proper place, while Figure 4e shows an exercise that asks learners to complete a sentence.

Double gapping and common adjectives

Students can enrich their vocabulary knowledge by learning pairs of adjectives that commonly appear together. Given two adjectives, exercises can be constructed by using *I*-grams that contain them joined by the word *and*. Adjectives and sample sentences can be used to create "double gapping" exercises such as the ones shown in Figure 4f.

CONCLUSION

This paper has described a way to capitalize on the vast amount of human-generated text readily available on the Web to help language learners express themselves in the first person. We have built a digital library collection, First Person Singular, containing fragments of text that are useful for self-expression, along with searching and browsing facilities specially designed for language learners. For pedagogical reasons it is essential to avoid errors, idiosyncrasies, and other dross: this is done using various language and grammar filters. Words and phrases are also sorted by frequency of use to exclude all but very common usages. The key is to use a huge collection of *n*-grams, along with their occurrence frequencies.

We contend that "*I*-grams"—common word sequences starting with the word I—have real value in enabling learners to study the usage of the most common words and expressions, thereby helping them to express themselves articulately. The digital library collection reveals many fascinating aspects of human life: what most people are worried about or afraid of, what they enjoy, love, and hate, what disgusts them, and so on. These are of value not only to sociologists and psychologists, but also to help language learners enrich their expressive repertoire.

We believe that digital libraries have enormous untapped potential for language education. In general, digital libraries contain the world's best prose, which makes them an ideal source for instructive examples of real language. As every language teacher knows, it now takes inordinate time to prepare and administer high-quality student exercises. The use of digital libraries will dramatically improve efficiency and effectiveness by utilizing automated techniques of language parsing, word collocation detection, identification of reading level, exercise construction, etc., to empower teachers to capitalize on top-quality prose already present in the world's libraries, and simultaneously give students an unprecedented choice of linguistic material. In the case of the First Person Singular collection, we have described five language activities that draw on this material to help students master important vocabulary and expressions through interesting and compelling online activities. Although these have not been evaluated formally, they point the way to an exciting new future. We predict that the application of digital library technology to assist language students will revolutionize second language learning.

Many fictional robots are able to express themselves "as a person does." But despite 40 years of research in artificial intelligence, scant progress has been made on realistic self-expression. Surprisingly, although they may never have feelings of their own, computers can nevertheless utilize the World-Wide Web to help users express themselves precisely and eloquently, and thereby participate more meaningfully in society.

ACKNOWLEGEMENTS

We acknowledge the entire New Zealand Digital Library Project team for their unstinting work in providing an environment that makes this kind of research meaningful—and fun. This work is funded by the Royal Society of NZ Marsden Fund and the NZ Foundation for Research, Science and Technology.

REFERENCES

Bainbridge, D., Don, K.J., Buchanan, G.R., Witten, I.H., Jones, S.R., Jones, M. and Barr, M.I. (2004) "Dynamic digital library construction and configuration." Proc European Digital Library Conference ECDL2004, edited by R. Heery, et al., pp. 1-13, Bath, England, September.

Benson, M., Benson E., and Ilson, R. (1986). Lexicographic description of English. Amsterdam, The Netherlands: John Benjamins Publishing Company.

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

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.

Coxhead, A. (1998) *An academic word list.* Occasional Publication Number 18, LALS, Victoria University of Wellington, New Zealand.

Kucera, H., and Francis, W.N. (1967) *Computational analysis of present-day American English*. Brown University Press, Providence, R.I.

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

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

Manning, C. and Schütze, H. (1999) *Foundations of statistical natural language processing*. MIT Press.

Moskowitz, G. (1978) *Caring and sharing in the Foreign Language class*. Heinle & Heinle.

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

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

Witten, I.H. and Bainbridge, D. (2007) "A retrospective look at Greenstone: Lessons from the first decade." *Proc Joint Conference of Digital Libraries*, Vancouver, Canada, pp. 147-156, June.