Content-based language learning in a digital library

Shaoqun Wu and Ian H. Witten

Department of Computer Science, University of Waikato
Hamilton, New Zealand
{shaoqun, ihw}@cs.waikato.ac.nz

Abstract. Digital libraries have untapped potential for supporting language teaching and learning. This paper describes a new scheme for automating topic-specific language learning using a specially built digital library. Three exercises of different types are generated automatically from the library content: one that learners undertake individually, one in which learners collaborate in pairs, and one in which a group of learners compete. The system aims to foster content-based language learning, which greatly increases students' motivation, fosters long-term recollection, and can be culturally situated in appropriate ways.

1 Introduction

Digital libraries have untapped potential for supporting language learning and teaching. They include an unprecedented supply of authentic linguistic material in the form of top-quality prose. They make language material easily accessible through purposeful searching and browsing. They include rich metadata that can support interesting linguistic exercises. They provide a safe and controlled learning environment. Socially-oriented library software can support collaborative activities that strengthen and enrich the students' learning experience. Exercise content can be focused on a particular subject. Last but not least, digital libraries can be distributed to people who lack the opportunity to attend traditional classroom lessons.

We are extending the Greenstone digital library software [8] and its metadata extraction tools to support language learning activities [9]. This paper describes a project in which articles on a chosen topic were harvested to create a collection that supports language learning activities. The material came from Wikipedia, and we derived metadata appropriate to language learning by mining its structured format and richly linked hypertext using standard natural language processing tools.

We implemented three vocabulary learning activities that offer challenging exercises within a particular domain utilizing the above metadata. The exercises are automatically generated from the digital library content; in some, learners select material using Greenstone's standard search and retrieval facilities. One exercise is done by individual learners; in another they collaborate in pairs; and in the third a group of learners compete. Together these exercises provide a learning environment in which students can improve their topic-specific vocabulary knowledge—we chose the example domain of *business*. The exercises have the following unique features:

- They draw students' attention to the salient vocabulary of a particular topic.

- They help students learn vocabulary from context.
- They increase the students’ encounters with relevant topic-related vocabulary.
- Collaborative learning helps sustain learning motivation and interests.

2 Digital libraries in language learning

Digital libraries have an important role to play in language education. They provide genre-specific, focused material that is carefully selected and organized. By exploring the authenticate material that digital libraries provide, learners are exposed to contemporary language usage. Subject-specific collections provide the opportunity to encounter key terms and grammatical constructions that rarely occur in general texts. For example, Fuentes [4] reports that students’ knowledge of business language is greatly enriched by basing learning on a corpus of business reports and product reviews. Digital libraries of multimedia can provide a rich and coherent learning context, which aids retention and reinforces learners’ knowledge of language. They can promote culturally situated learning by working with collections that introduce the target language’s people, history, environment, art, literature, and music.

Digital libraries can provide a safe learning community in which teachers share thoughts, tips and lesson plans, and organize collaborative task-based, content-based language projects; and learners meet their peers, exchange learning ideas, and engage in competitive or collaborative tasks. Pedagogically tuned search and browse facilities can meet the special needs of individual learners and teachers without bogging them down in fruitless tangential explorations. Earlier [9] we developed eight activities that are automatically generated from digital library content and utilize the search and retrieval facilities to illustrate new ways of supporting language study.

Supporting language learning with digital libraries is particularly relevant in developing countries where the ability to speak another language can make the difference between poverty and success. Language education traditionally takes place in classrooms, and many students are denied the opportunity because of scarce resources. Although the Internet is beginning to invade everyday life—and the classroom—even in developing countries, many people living in remote areas are still deprived of this learning facility. However, stand-alone digital libraries packed with unprecedented amounts of language material can reach out to those in areas that technology has forgotten and give them the means to escape the poverty trap.

3 Content-based language learning in digital libraries

Content-based language learning refers to two separate ideas: learning a subject (such as business or mathematics) through the medium of a foreign language, and learning a foreign language by studying a particular subject [3]. Although a relatively new research area it has promising educational implications. First, it makes language learning more interesting and motivational. For example, young children would enjoy reading science fiction or simple encyclopedia articles using vocabulary learned in the science classroom. Second, it helps those who need to upgrade their language knowledge in a particular domain but are hampered by time constraints or the inability to find suitable courses to further their career. Third, being subject-specific makes it natural and meaningful to introduce the culture of the target language.

Digital libraries, like traditional ones, can make a central contribution to education. The genre-specific nature of many collections matches the content-based language learning paradigm. Such collections provide a vast body of samples of authentic language use. Language is often topic-specific: "different genres may exhibit very different pattern in the use of both lexis and grammar." [7] Learners supported by a digital library can acquire knowledge of a subject and at the same time improve their linguistic ability. Activities can help them to notice linguistic features, give them the opportunity to acquire a core vocabulary and expressions relevant to that particular subject, and practice as much as they like.

4 Building a topic-related collection

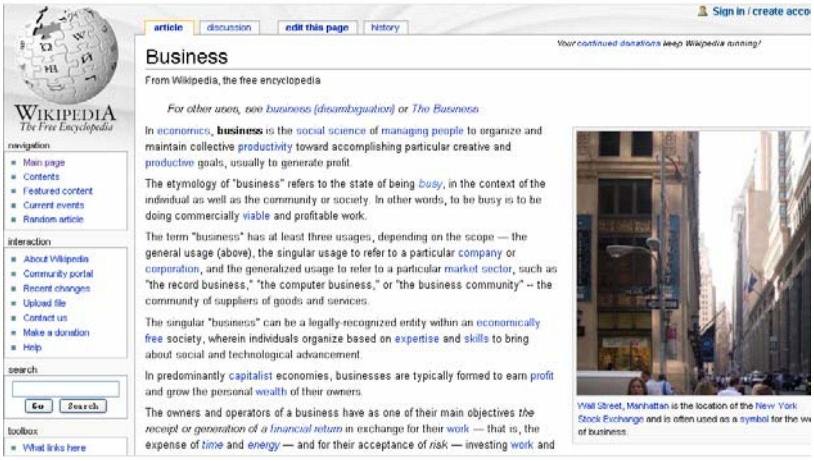
Wikipedia is a massive and constantly growing online encyclopedia with a unique "anyone can edit" philosophy. At the time of writing, its English version contains 1.8 million articles covering a very wide range of topics. Traditional encyclopedias and their new multimedia electronic counterparts are a valuable language learning resource that can enrich a learners' knowledge of the target language. However, their cost often discourages use. Wikipedia, which is online and completely free, opens up new opportunities for supporting language study in innovative and creative ways.

Figure 1 shows a typical article. It comprises the entry's title (in this case *business*), an accompanying picture, a definition, and a brief description. The sections that follow include detailed explanations and introduce related topics. One of the most striking features is the wealth of hyperlinks that are scattered through the articles, putting related material right at the learner‘s fingertips. Moreover, the hyperlinks are labeled by short phrases called the link's "anchor text." These manually assigned key-phrases have great pedagogical value.

Selecting the articles

We built a small digital library focused on the topic *business*. This term led directly to the Wikipedia page in Figure 1. All hyperlinks from this article to other Wikipedia articles were followed, resulting in about one hundred business-related articles.

In order to accomplish this we used the WikipediaMiner toolkit [10]. Wikipedia provides a standard procedure for exporting its entire database of articles, and WikipediaMiner connects to this database. The user starts by submitting a query term, in this case *business*. WikipediaMiner returns a list of matching articles. More than one article would be returned in the case of an ambiguous query. At this stage, the user must select one of more articles that correspond to the desired senses of the word.



Your continued donations keep Wikipedia running!

Business

From Wikipedia, the free encyclopedia

For other uses, see business (disambiguation) or The Business

In economics, **business** is the social science of managing people to organize and maintain collective productivity toward accomplishing particular creative and productive goals, usually to generate profit.

The etymology of "business" refers to the state of being *busy*, in the context of the individual as well as the community or society. In other words, to be busy is to be doing commercially viable and profitable work.

The term "business" has at least three usages, depending on the scope — the general usage (above), the singular usage to refer to a particular company or corporation, and the generalized usage to refer to a particular market sector, such as "the record business," "the computer business," or "the business community" -- the community of suppliers of goods and services.

The singular "business" can be a legally-recognized entity within an economically free society, wherein individuals organize based on expertise and skills to bring about social and technological advancement.

In predominantly capitalist economies, businesses are typically formed to earn profit and grow the personal wealth of their owners.

The owners and operators of a business have as one of their main objectives *the receipt or generation of a financial return* in exchange for their work — that is, the expense of *time* and *energy* — and for their acceptance of *risk* — investing work and

Wall Street, Manhattan is the location of the New York Stock Exchange and is often used as a symbol for the w of business.

Fig. 1. Typical Wikipedia article

We chose the article shown in Figure 1, and specified that we are also interested all articles to which that article links. WikipediaMiner locates the articles and downloads their full text (which is not in the database) from the web.

Generating metadata

A set of pedagogically useful metadata is extracted from each article, to support the vocabulary exercises described below. It is extracted with the assistance of OpenNLP, a collection of software tools for natural language processing that perform sentence detection, tokenization, part-of-speech-tagging, chunking and parsing.

The metadata we extract from each article comprises:

- the term that describes the article, and its definition
- the illustrative picture (as shown in Figure 1), if any, and its caption
- key-phrases relating to the article
- the number of key-phrases and paragraphs in each section of the article.

The first sentence of the article usually contains the relevant *term and its definition*. The sentence is located using OpenNLP's sentence detector and parsed into linguistic phrases. Then the opening noun phrase and its following verb phrase are identified by string pattern matching, taking into account the possibility of an intervening adjectival phrase. The noun is marked as a term, and the remainder is marked as the definition.

For example, the first sentence of the page in Figure 1 reads

> *In economics, business is the social science of managing people to organize and maintain collective productivity toward accomplishing particular creative and productive goals, usually to generate profit.*

The net result is to remove the initial qualifier, identify *business* as the relevant term, and return the rest of the sentence as its definition:

business is the social science of managing people [continues as above].

This simple heuristic procedure works well. It works even when the term defined differs slightly from the article name, and when it is accompanied by a qualifier. In our example, terms and definitions were successfully extracted from 80 of the 100 articles. The procedure failed on the remaining 20 because the page structure was ill-formed. In one case out of 80 the extraction procedure yielded an ill-formed definition.

The page's *picture and caption* are extracted by seeking a particular configuration of HTML tags, including the <*img*> tag that signals an image.

As mentioned earlier, *key-phrases* are the anchor text of hyperlinks in the article, and thus are easily located. Care is taken to ignore the links used for navigation and special functions such as *search* and *edit*. Different sections of a Wikipedia article normally provide supplementary or complementary information about the topic: this makes them good sources of focused coherent text. The number of key-phrases and paragraphs in each section provide some indication of its pedagogical character and are calculated by counting the relevant HTML tags.

Building the collection

The Wikipedia articles extracted using WikipediaMiner were fed into Greenstone to build a digital library collection. The automatic metadata extraction described above was built into a special Greenstone plugin which processed each article and associated derived metadata with it.

The final collection seeded from the term *business* contained about 100 articles each with four kinds of extracted metadata: the term and its definition, a picture and caption, several key-phrases, and the number of paragraphs and key-phrases in each section. Of course, Wikipedia content is by no means definitive: it has been widely criticized as a reliable information source. However, it provides useful supplementary reading material for studying the subject.

5 Using the collection for language learning

Three learning activities, *Match-term-and-definition*, *Fill-in-the-blanks* and *Predict-words*, provide systematic study of vocabulary. They make use of the material in the digital library, augmented with pedagogically valuable metadata. They pull out the vocabulary that is important to the subject in question, namely *business*, and help students notice the salient language features of that particular subject.

The importance of explicit vocabulary learning has been widely recognized. "The brightest of students will not be able to recall and use new words without repeated meaningful contact with them."[2] Reading topic-related articles engenders some degree of familiarity with particular words. But follow-up learning activities that use the same material, presented through various kinds of individual and interactive exercises, maximize the chance of the word being retained over the long term.

The exercises we describe are automatically generated from the content of the articles and the extracted metadata, and do not place any extra burden on language example of *business*. Their design has been guided by the psychological conditions teachers. They work for any topic, and are not in any way specific to our running for the retention and ultimate mastery of a word [6]:

- Noticing: giving an attention to an item.
- Retrieval: giving a word form and retrieving its meaning or vice versus.
- Creative and generative use: using the word in a new context.

Noticing plays a central role in every activity. *Match-term-and-definition* asks learners to associate a word form with its meaning: Nation's *retrieval* task. *Fill-in-the-blanks* promotes deeper reflection by asking learners to guess the meaning of a word from its context. *Predict-words* facilitates *creative and generative use* of a word by asking learners to predict what words might appear in a text.

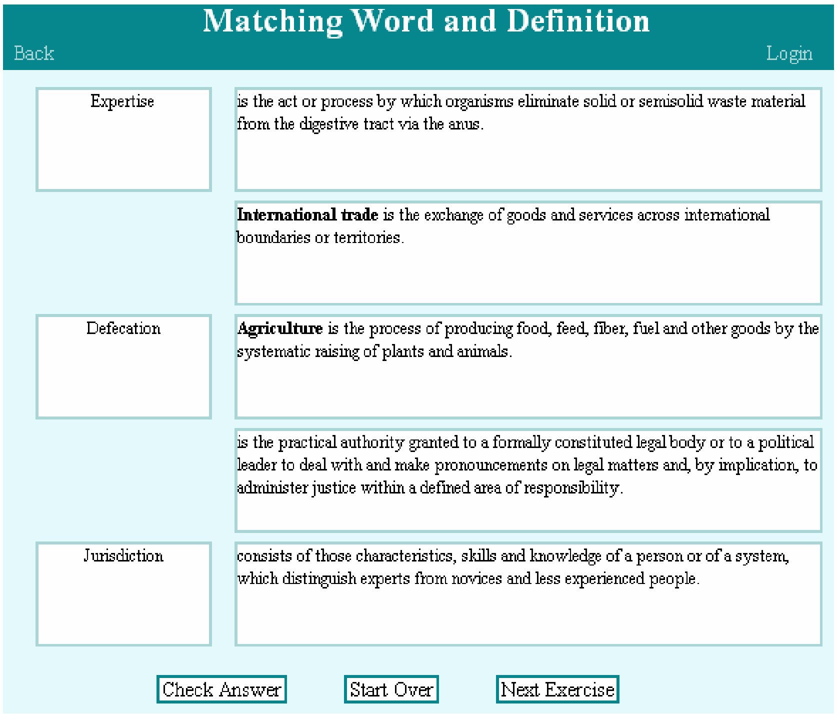
The exercises reinforce learning through repetition: subsequent encounters of a particular word help strengthen and enrich previous knowledge [5]. Moreover, the last two activities engage learners in collaborative activity, making the vocabulary learning less daunting and more enjoyable and effective. They embody a "chat" facility, creating an environment in which learners can practice communication skills by discussing with peers, seeking help, and negotiating tasks.

Exercise material for some activities comes from individual documents; for others it comes from the whole collection. *Match-term-and-definition* is a collection level activity. It uses the term and definition metadata associated with each article. Students reach the exercise from a link that is placed on the digital library collection's home page, which in Greenstone typically contains information about the provenance of the collection as a whole, and information on how to use it. In the case of these language learning collections, the home page briefly describes the searching and browsing functionality, and introduces the language activities. The other two activities, *Fill-in-the-blanks* and *Predict-words*, work at the document level. They use the content of a particular article, along with its metadata, as the exercise material. Students reach these exercises from links that are placed on the digital library page that displays each document. They reach the document in the normal way, by searching or browsing.

The language learning exercises are built on the top of the Greenstone run-time software. Like it, they follow the client-server model. The interface is explicitly designed to be multilingual—this is a particular strength of Greenstone. To add a new interface language it is necessary to create a "resource bundle" for that language and drop it into the appropriate folder. The Greenstone Translator's Interface [1] provides interactive tools for creating and maintaining language interfaces. The language learning activities are implemented using JavaScript and Ajax technology.

5.1 Matching terms with their definition

In "matching" activities, learners must find two items that are related in some way, such as having similar or opposite meanings. These activities can easily cater for learners at different levels, and are widely used for vocabulary study. The *Match-term-and-definition* exercise asks the learner to match terms with their definitions. Terms are words or phrases, and definitions are sentences whose subject is missing.



Expertise

is the act or process by which organisms eliminate solid or semisolid waste material from the digestive tract via the anus.

International trade is the exchange of goods and services across international boundaries or territories.

Defecation

Agriculture is the process of producing food, feed, fiber, fuel and other goods by the systematic raising of plants and animals.

is the practical authority granted to a formally constituted legal body or to a political leader to deal with and make pronouncements on legal matters and, by implication, to administer justice within a defined area of responsibility.

Jurisdiction

consists of those characteristics, skills and knowledge of a person or of a system, which distinguish experts from novices and less experienced people.

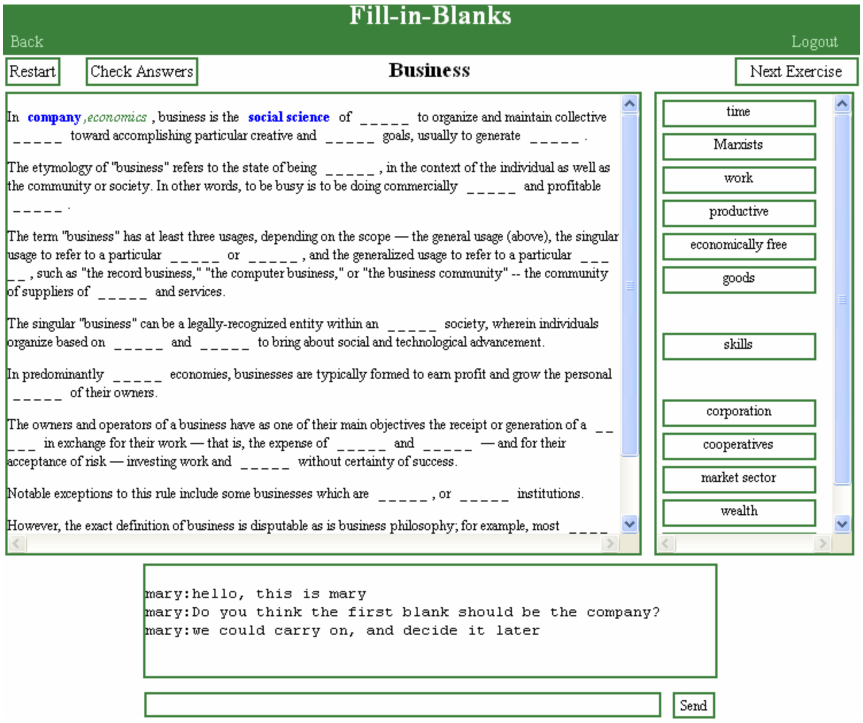
Fig. 2. The Match-*term-and-definition* activity

Figure 2 illustrates the interface to this activity. This exercise comprises five pairs of terms and definitions, scrambled. The learner drags a term on the left side into a definition box on the right. The system formats the result by highlighting the term. In Figure 2, the learner has matched *International trade* (second row) and *Agriculture* (third row) with their definitions, and has yet to match *Expertise*, *Defecation* and *Jurisdiction*.

At any stage the learner can use the buttons at the bottom to check the answers (this causes incorrect ones to be colored red), start this exercise over again (this reinitializes the boxes), or move to the next exercise. The scheme is designed for motivated learners, and for exercising rather than testing: it does not enforce the completion of one exercise before allowing the next to begin. At the beginning of the exercise the activity server retrieves five phrases and their associated definitions from the collection at random, using the metadata computed as described earlier.

5.2 Filling in the blanks

Fill-in-the-blanks exercises are created by cutting target words or phrases out of a sentence or article and having students fill them in. Human tutors can judge free word choices, but in computer-assisted environments the excised words or phrases are invariably displayed so that a student’s choice can be checked automatically. Our implementation of this activity is novel because pairs of learners must work together



In company ,*economics* , business is the social science of _____ to organize and maintain collective _____ toward accomplishing particular creative and _____ goals, usually to generate _____ .

The etymology of "business" refers to the state of being _____ , in the context of the individual as well as the community or society. In other words, to be busy is to be doing commercially _____ and profitable _____ .

The term "business" has at least three usages, depending on the scope — the general usage (above), the singular usage to refer to a particular _____ or _____ , and the generalized usage to refer to a particular _____ , such as "the record business," "the computer business," or "the business community" -- the community of suppliers of _____ and services.

The singular "business" can be a legally-recognized entity within an _____ society, wherein individuals organize based on _____ and _____ to bring about social and technological advancement.

In predominantly _____ economies, businesses are typically formed to earn profit and grow the personal _____ of their owners.

The owners and operators of a business have as one of their main objectives the receipt or generation of a _____ in exchange for their work — that is, the expense of _____ and _____ — and for their acceptance of risk — investing work and _____ without certainty of success.

Notable exceptions to this rule include some businesses which are _____ , or _____ institutions.

However, the exact definition of business is disputable as is business philosophy; for example, most ____

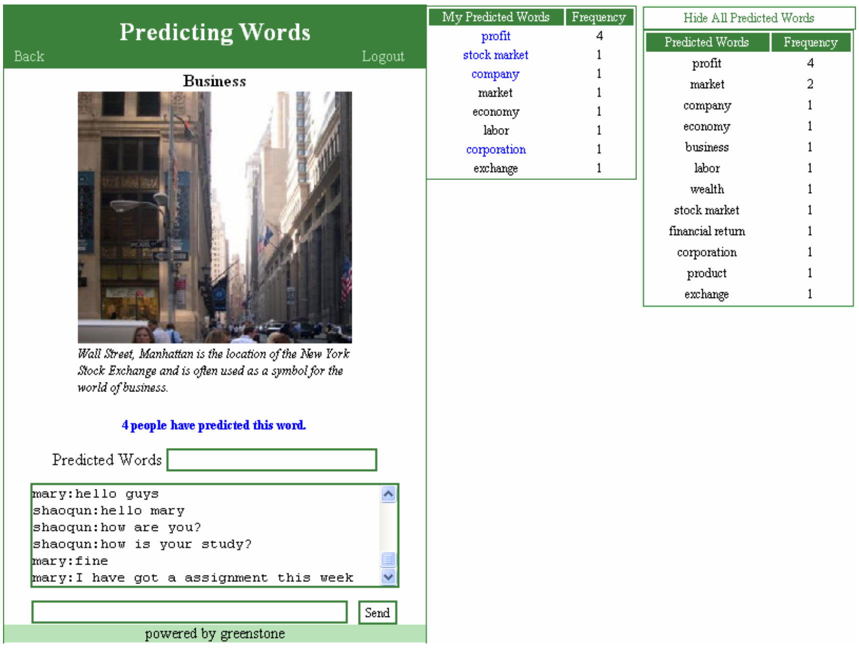
```
mary:hello, this is mary
mary:Do you think the first blank should be the company?
mary:we could carry on, and decide it later
```

Fig. 3. The fill-in-the-blanks activity

to complete the task. The necessary information is split: each student holds part of it and in order to succeed they must cooperate. Such tasks are known to have great pedagogical value: they encourage negotiation of meaning, promote implicit or explicit corrective feedback between learners, and improve performance by stimulating modifications to the output [6]. In this activity, each learner has only half the missing words, and in order to achieve the goal of filling them in they must exchange information verbally using a built-in “chat” system.

Learners are paired based on the article they have selected. An exercise is constructed using the text of a particular section and the key-phrases that occur within it. Not all sections provide suitable material: some have too few key-phrases; others have too many paragraphs. The key-phrase and paragraph count metadata is used to select sections of modest length that contain a high density of key-phrases.

Figure 3 shows the interface. The key-phrases in the text are replaced with dashed lines. Each learner is given half of them and can drag them and drop them into the gaps. A move can be undone by clicking the dropped word. Each learner sees the moves their partner makes, in real time. In Figure 3, this learner has filled in the first blank with the words *company* and *social science*. Of the word that follows, *economics* appear in a different color and font style because it was filled in by the partner. The buttons at the top are only displayed to one of the two learners. That learner can check the answers at any time and control the progress of the activity, restarting it or passing on to the next exercise.



Business

Wall Street, Manhattan is the location of the New York Stock Exchange and is often used as a symbol for the world of business.

4 people have predicted this word.

Predicted Words

```
mary:hello guys
shaoqun:hello mary
shaoqun:how are you?
shaoqun:how is your study?
mary:fine
mary:I have got a assignment this week
```

My Predicted Words	Frequency
profit	4
stock market	1
company	1
market	1
economy	1
labor	1
corporation	1
exchange	1

Predicted Words	Frequency
profit	4
market	2
company	1
economy	1
business	1
labor	1
wealth	1
stock market	1
financial return	1
corporation	1
product	1
exchange	1

Fig. 4. *The* Predicting *Words* activity

5.3 Predicting words

Given an article's topic, students compete to predict words they think will occur in it. This traditional pre-reading activity is often played in a classroom to stimulate interest and facilitate comprehension before students begin reading. It can also be used to brainstorm suitable vocabulary for a forthcoming essay, or serve as a retrospective activity where learners recall and review a list of expressions and collocations that are important for accurately expressing the ideas relevant to the topic learners.

We have implemented the exercise in its traditional form to provide a collaborative learning environment in which learners help each other by sharing information and exchanging ideas. This is a document level activity: learners who have chosen the same article are grouped together.

Figure 4 shows the interface. The title, picture, and caption are presented to convey the context of the original article. Each learner enters predicted words in the *Predicted words* field and can "chat" on the system to other learners. On the right hand side, the *My Predicted Words* table shows the words that this learner has predicted, colored blue if they occur in the key-phrase list. The table is ordered by the number of participants who have chosen that word. In the Figure, this student has predicted eight words, four of which appear in the article's key-phrase list.

The positions and counts change dynamically to reflect work by other learners. For example, *profit* has been predicted by four students. The *All Predicted Words* table, which learners can hide, lists the predictions of other participants.

6 Conclusions

This paper has described a scheme for supporting content-based language learning with a digital library. First, an instructor decides on a topic and interacts with the WikipediaMiner system to retrieve a suitable selection of related articles. This is done by issuing a command-line statement which could easily be automated within a web-based form. Next, the articles are built into a digital library collection, using the Greenstone software augmented with a purpose-built plugin that extracts language learning metadata. These operations are independent of whatever topic was chosen for the collection.

Once the collection has been constructed, language learners interact through the ordinary Greenstone interface. Special facilities have been built into the run-time system to present three types of exercise. Learners access these by clicking on links in the collection. For the *Match-term-and-definition* exercise, which involves the whole collection, a link is placed on the collection's home page. Learners access the other exercises, *Fill-in-the-blanks* and *Predict-words*, once they have reached a particular document in the collection through links that are presented alongside the document.

The three exercise types illustrate different modes of interaction. *Match-term-and-definition* is done by individual students, working independently. *Fill-in-the-blanks* shares the necessary information between pairs of students who must cooperate in order to solve the exercise. In *Predict-words* all students who choose the same article see the same information, and compete in an informal manner to guess as many words as possible. In the last two cases the different displays are updated simultaneously, and users are able to communicate with each other through a "chat" panel.

These exercises have been devised and implemented, but not yet field tested. Interested readers can access them at *http://www.nzdl.org/language_learning*. Our next step is to test them in an actual classroom environment and a self-study setting.

References

1. Bainbridge, D., Edgar, K.D., McPherson, J.R. and Witten, I.H. (2003) "Managing change in a digital library system with many interface languages." *Proc European Conference on Digital Libraries ECDL2003*, pp. 350-361, Trondheim, Norway, August.
2. Conzett, J. (2000) "Integrating collocation into a reading and writing course." In *Teaching Collocation*, edited by Lewis Michael. pp. 70–87, LTP, England.
3. Darn, S. (2006). "Content and Language Integrated Learning." *British Council Teaching English*. http://www.teachingenglish.org.uk/think/methodology/clil.shtml (May 23, 2007)
4. 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.
5. Nation, I.S.P. (2001) *Learning vocabulary in another language*. Cambridge Univ. Press.
6. Pica, T., Kang, H.S., and Sauro, S. (2006) "Information Gap tasks." *SSLA* 28, pp. 301–338.
7. Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press, Oxford.
8. Witten, I.H. and Bainbridge, D. (2003) *How to build a digital library*. Morgan Kaufmann.
9. Wu, S. and Witten, I.H. (2006) "Towards a digital library for language learning." *Proc European Conf on Digital Libraries*, pp. 341–352, Alicante, Spain.
10. Milne, D. (2007) WikipediaMiner. *http://wikipedia-miner.sourceforge.net*