Towards a digital library for language learning

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. Although the Internet at large is widely used for language education, it has critical disadvantages that can be overcome in a more controlled environment. This article describes a language learning digital library, and a new metadata set that characterizes linguistic features commonly taught in class as well as textual attributes used for selection of suitable exercise material. On the system is built a set of eight learning activities that together offer a classroom and self-study environment with a rich variety of interactive exercises, which are automatically generated from digital library content. The system has been evaluated by usability experts, language teachers, and students.

1 Introduction

The rise of computer-assisted language learning on the Internet has brought a new dimension to language classes. The Web offers learners a wealth of language material and gives students opportunities to learn in different ways. They can study by reading newspaper articles, listening to audio and viewing video clips; undertake online learning exercises; or join courses. Media such as email, chat and blogs enable them to communicate with other learners and with speakers of the target language all over the world. When preparing lessons, teachers benefit from the panoply of resources that the web provides. Automated tools can be used to build practice exercises and design lessons. Teachers construct language learning tasks based on the Internet because the language is real and the topics are contemporary, which motivates learners.

Despite all these advantages, the Internet has many drawbacks for language study. Although it offers innumerable language resources, learners and teachers alike face the challenge of discovering usable material. Search engines return an overwhelming amount of dross in response to any query, and locating suitable sources demands skill and judgment. When learners study on their own, it is hard for them to locate material that matches their language ability. Finally, students may accidentally encounter material with grossly unsuitable content.

Digital libraries, like traditional ones, can play a crucial role in education. Marchionini [1] identifies many advantages in using them for teaching and learning. As well as providing a safe and reliable educational environment, they have special advantages for language classes. Digital libraries are a great source of material that

teachers can turn into meaningful language exercises. They offer vast quantities of authentic text. Learners experience language in realistic and genuine contexts, which prepares them for what they will encounter in the real world. Searching and browsing facilities can be tailored to the special needs of language learners. Teachers can integrate digital libraries into classes that help students locate appropriate material, giving them the tools to study independently. Interpersonal communication media can be incorporated to create a socially engaging learning environment.

This project has built a language learning digital library called LLDL based on the Greenstone digital library software [2]. The goal is to explore the potential of digital libraries in this field by addressing issues intrinsic to language learning. We developed a language learning metadata set (LLM) that characterizes linguistic features commonly taught in class. By using it in searching and browsing, teachers and learners can locate appropriate material.

Eight learning activities are implemented that utilize LLDL's search and retrieval facilities. Together they offer a classroom and self-study environment with a rich variety of interactive exercises. Four features distinguish them from existing systems:

- They are student-centered
- They provide a communicative learning environment
- They provide a multilingual interface
- Exercises are automatically generated from digital library content.

While the present implementation of LLDL is for learning English, it is designed to provide a multilingual interface. English and Chinese versions exist; new languages can easily be added. We close the paper with some remarks on extending the interface and the language taught to other European languages.

2 DLs in Language Learning

Digital libraries can serve many roles in language education. First, they provide linguistic resources. In the classroom, text, pictures, models, audio, and video are used as material for teaching. Edge [3] summarizes three kinds of language resource, published, authentic and teacher-produced, and digital libraries allow teachers to build collections of each kind. Culturally situated learning helps students interpret the target language and master skills in communication and behavior within the target culture [4]. Teachers can build collections that introduce the people, history, environment, art, literature, music. The material can be presented in diverse media—text, images, audio, video, and maps. Students can experience the culture without leaving the classroom.

Second, digital libraries can bring teachers and learners together. Forums, discussion boards, electronic journals and chat programs can be incorporated to create a community where teachers share their thoughts, tips and lesson plans; learners meet their peers and exchange ideas; and teachers organize collaborative task-based, content-based projects. This community is especially meaningful for language learning because it embeds learners in an authentic social environment, and also integrates the various skills of learning and use [5]. As Vygotsky [6] points out, true learning involves socialization, where students internalize language by collaborating on common activities and sharing the means of communicating information.

Third, digital libraries can provide students with activities, references and tools. Language activities include courses, practice exercises, and instructional programs. In traditional libraries students find reference works: dictionaries, thesauri, grammar tutorials, books of synonyms, antonyms and collocations, and so on. Equivalent resources in digital libraries can be used as the basis of stimulating educational games.

3 Language Learning Metadata

Metadata is a key component of any digital library. It is used to organize resources and locate material by searching and browsing. Metadata schemas developed specifically for education and training over the past few years have recently been formally standardized [7]. The two most prominent are LOM (Learning Object Metadata) and SCORM (Sharable Content Object Reference Model). LOM aims to specify the syntax and semantics of the attributes required to describe a learning object. It groups features into categories: general, life-cycle, meta-metadata, educational, technical, rights, annotation, and classification. SCORM aims to create flexible training options by ensuring that content is reusable, interoperable, durable, and accessible regardless of the delivery and management systems used. While LOM defines metadata for a learning object, SCORM references a set of technical specifications and guidelines designed to meet the needs of its developers, the US Department of Defense.

Neither of these standards proved particularly useful for our purpose. The aim of metadata is to help users find things. Although digital libraries make it easy to locate documents based on title, author, or content, they do not make it easy to find material for language lessons—such as texts written for a certain level of reading ability, or sentences that use the *present perfect* tense. To identify these users would have to sift through countless examples, most of which do not satisfy the search criteria.

The LLM metadata set is designed to help teachers and students locate material for particular learning activities. It has two levels: documents and sentences. All values are intended to be capable of being extracted automatically from full text: no human processing is required. Some LLM metadata are extracted with the help of tools from the OpenNLP package, which provides the underlying framework for linguistic analysis of the documents by tagging all words with their part of speech and identifying units such as prepositional phrases.

3.1 Document Metadata

Readability metadata can help both teachers and students locate material at an appropriate level. We have adopted two widely used measures recommended by practicing teachers: Flesch Reading Ease and the Flesch-Kincaid Grade Level [8]. The former is normally used to assess adult materials, and calculates an index between 0 and 100 from the average number of words per sentence and the average number of syllables per word. The latter is widely used for upper elementary and secondary material and scores text on a US grade-school scale ranging from 1 to 12.

LLM incorporates both these scores as separate pieces of metadata, and in addition computes LOM *Difficulty* metadata by quantizing the Grade Level into five bands.

3.2 Sentence Metadata

Readability metadata is associated both with the document as a whole and with individual sentences. Three further types of metadata are associated with sentences: sentence metadata, syntactic metadata, and usage metadata.

LLDL splits every document into individual sentences using a simple heuristic involving terminating punctuation, the case of initial words, common abbreviations, and HTML tags. Whereas sentences used as examples in the classroom or language teaching books have been carefully targeted, prefabricated, and honed into clean and polished examples, sentences extracted automatically from authentic text are often untidy and incomplete; some have inordinately complex structures.

LLM addresses this by defining the following metadata for each sentence:

- Processed version
- Tagged version
- State: clean or dirty
- Type: simple or complex.

The first two are variants of the original extracted sentence, which usually contains HTML mark-up. The *Processed* version contains plain text: mark-up has been stripped. The *Tagged* version has been annotated with linguistic tags that reflect the syntactic category of each word. Part-of-speech metadata is used by the language learning digital library to generate exercises, as described in Section 5.

Some extracted sentences are messy. *State* metadata is used to indicate whether a sentence is *clean*, comprising alphabetic characters and punctuation only, or *dirty*, including other extraneous characters. The *Type* of a sentence is *simple* if it has just one clause and *complex* otherwise, where a clause is a group of words containing a main verb. Teachers normally use simple sentences to explain grammar rules where possible.

The extraction process first detects sentence boundaries and strips HTML, yielding *Processed* sentence metadata. If sentences contain any characters other than alphabetic ones, space, and punctuation, their *State* metadata is *Dirty*. Clean sentences are analyzed by the OpenNLP tagger and chunker to yield *Tagged* sentence metadata. These contain syntactic tags that reflect the categories of individual words and reveal the sentence structure, facilitating the extraction of language metadata. Simple and complex sentences are differentiated by the number of verb phrases (*VP*) they contain.

3.3 Syntactic metadata

English grammar is relatively simple because it has fixed rules. On other hand, the number of rules is large and there are many exceptions. Based on recommendations from language teachers, we identified nine syntactic metadata elements that can be extracted automatically by natural language processing tools. While these do not cover all aspects of English grammar, they form the basis of a useful digital library.

The syntactic metadata elements are Adjective, Preposition, Possessive pronoun and determiner, Modal, Tense, Voice, Coordinating conjunction, Subordinate conjunction, That-clause and wh-clause. For each one a regular expression is defined—for example, *\\w+/JJ* is the expression for *Adjective* metadata: it indicates a string that contains one or more word characters (*\\w+*) followed by */JJ*, the syntactic tag for adjective. *Tense* and *Voice* metadata are also extracted using tagged sentences. They comprise both the tense or voice and the verbs or verb groups that are so marked.

The extraction process for the remaining syntactic metadata is similar. Understanding the grammatical implications of the tags is the key to successful extraction. *Preposition* metadata is extracted by searching for prepositional phrases, tagged *PP*. *Subordinate conjunction* and *that-clause* metadata are extracted by seeking subordinating clauses tagged as *SBAR*. *Wh-clauses* are not indicated by a clause-level tag, and must be identified by combining phrase tags and *wh-word* tags.

3.4 Usage Metadata

LLM contains a single usage metadata element: Collocation. This is a group of two or more words that are commonly found together or in close proximity. For example, native speakers usually prefer the collocation *heavy rain* to the non-collocation *big rain*, or *totally convinced* to *absolutely convinced*. Lewis [9] points out that native speakers carry hundreds of thousands, possibly millions, of collocations in their heads ready to draw upon in order to produce fluent, accurate and meaningful language, and this presents great challenges to language learners.

We define collocations in terms of 9 two- and three-word syntactic patterns such as *adjective+noun*, *adverb+adjective*, and phrasal verbs in the form *verb+preposition*—for example, *make up* and *take off*. They are identified by looking for particular tags and matching them with the nine syntactic collocation patterns. Following common practice [10] we use the *t*-statistic to rank potential collocations. This uses the number of occurrences of words individually and in combination, and the total number of tokens in the corpus. Its accuracy depends on the size of the corpus: good collocations that occur just once do not receive high scores.

4 Searching the Digital Library

LLM metadata captures linguistic aspects of the documents in a digital library. It allows users to search and browse language learning materials. This section demonstrates the use of the extracted metadata in LLDL. In this project, we have built five demonstration collections for use in the activities described in the next section:

- Documents from the UN FAO *Better farming series*
- Children's short stories from *East of the web*
- News articles from the *BBC World Service*
- Sample articles from *Password*, a magazine for new English speakers
- Collection of plant and animal images downloaded from the Internet.

The first collection includes practical articles intentionally written in a simple style, but not targeted at children. The second contains material specifically for children. The third and fourth are made from material that is intended to be particularly suitable for second language learners. These four collections exhibit a wide variety of styles and difficulty levels.

LLDL uses standard Greenstone facilities [2] to present options for browsing and searching on entry to the library. When users browse, they can select *Titles*, *Difficulty*, and other metadata elements. Clicking *Titles* presents an alphabetical list of titles of the documents in the collection, broken down into alphabetic ranges; the full text of the documents is available by clicking beside the appropriate title. *Difficulty* also applies to documents, and allows the reader to browse titles in each of the five difficulty levels mentioned above.

The other browsing options refer to individual sentences: they are *Tense*, *Preposition*, *Clause*, *Difficulty* (which differs from the document-level *Difficulty* above because it refers to individual sentences), and *Type*. Sentences are the essential units in language communication. Students study vocabulary and learn grammars in order to construct sentences. Conversely, studying good sentences helps master word usage or grammar rules in context. LLDL allows readers to browse for particular grammatical constructions or identify particular parts of speech. For example, selecting *Preposition* shows the sentences of the collection, with the prepositions that each one contains listed in parentheses after it. The sentences are presented in alphabetic groups according to preposition: those under the *A–B* section of the hierarchy contain *about*, *at*, *above*, *as*, *between*, *before*, *by*, *beside*, … These sample sentences help students learn the usage of particular prepositions and study what words commonly appear before and after them—for example, *above all*, *ask about*.

Searching is more highly targeted than browsing. Users can perform an ordinary full-text search to locate documents that treat particular topics; the search results show the title and difficulty level of matching documents. Advanced search allows users to specify metadata as well as content. For example, one might search for particular full-text content but confine the search to documents that are *easy* (in terms of difficulty level). Or search for individual sentences rather than documents, whose type is *simple* (i.e., no compound sentences), or whose state is *clean* (i.e., no non-alphabetic characters). Users can combine these criteria in a search form to find *simple* and *clean* sentences from *easy* documents whose text contains specified words or phrases.

Users can also search for sentences that contain particular words. New learners are often confused about word usage—for example, distinguishing the different implications of *look*, *see* and *watch*. One way to help is to provide many authentic samples that show these words in context. LLDL can retrieve sentences that include a specified word or phrase, and are restricted by the above-mentioned sentence-level metadata. Students can also search for sentences that exhibit any of the grammatical constructs that are identified by metadata, for example passive voice sentences, modal sentences or sentences in a particular tense.

5 Language Learning Activities

LLDL facilitates the creation of language learning activities. To demonstrate this we have developed eight activities: *Image Guessing*, *Collocation Matching*, *Quiz*, *Scrambled Sentences*, *Collocation Identifying*, *Predicting Words*, *Fill-in-blanks*, and *Scrambled Documents*; unfortunately space permits a description of the first four activities only. They share the common components *login*, *chat*, *scoring* and *feedback*.

5.1 Common Components

Learners are not required to register, but must **login** by providing a user name and select a difficulty level (easy, medium or hard). This parameter is used to select sentences or documents for each activity, to determine which image collections are used to generate exercises, and to group students for activities in which they work in pairs. For these activities the system maintains a queue of users waiting at each level. When a student logs in, the queue is checked and they are either paired up with a waiting student at the same level, or queued to await a new opponent.

LLDL makes a **chat** facility available in all activities, in order to create an environment in which students can practice communication skills by discussing with peers, seeking help, and negotiating tasks. The chat panel resides either in the activity interface or a window that is launched by clicking a *Chat* button.

Each activity contains a **scoring** system intended to maintain a high level of motivation by encouraging students to compete with each other informally. Students can view the accumulated scores of all participants, sorted so that the high scorers appear at the top. Additional statistical information is provided such as the number of identified collocations in the *Collocation* activity or the number of predicted words in the *Predicting Words* activity. The implementation of the scoring mechanism varies from one activity to another, depending on whether students do the exercise individually, or collaborate in pairs, or compete in pairs.

Students are provided with **feedback** on whether the response is correct or incorrect, and in the latter case they are invited to try again, perhaps with a hint that leads to the correct response. In general, feedback is given item by item, at logical content breaks, at the end of the unit or session, or when requested by the student. Students also see their accumulated scores. Some activities provide an exercise-based summary that includes questions, correct answers, and answers by the student's partner.

Hints provide direct help without giving away the answer. They can be offered through text, pictures, audio or video clips, or by directing students to reference articles or relevant tutorials. Some exercises give hints that use text from the digital library. For example, the *Quiz* activity allows students to ask for other sentences containing the same words; *Collocation Matching* provides more surrounding text so that students can study the question in context.

5.2 The *Image guessing* exercise

In *Image Guessing*, the system pairs students according to their self-selected difficulty level. One plays the role of describer, while the other is the guesser. An image is chosen randomly from a digital library collection of images and shown to the describer alone; the guesser must identify that exact image. The describer describes the picture in words that are automatically used by the system as a query term, and also decides how many of the search results the guesser will see. The guesser does not see the description; the describer does not see the search results. The guesser and describer can communicate using the chat facility. The activity moves to the next image when the guesser successfully identifies the image, chooses the wrong one, or the timer expires. The students use the search and chat facility to identify as many images as possible in a given time. They can pass on a particular image, or switch roles.

The difficulty level is determined by the image collection, which teachers build for their student population. They select simple images—e.g. animal images or cartoons—for lower level students, and more complex ones—e.g. landscapes—for advanced ones. For searching, image collections use metadata provided by the teacher, which they tailor to the students' linguistic ability. The more specifically the metadata describes the images, the easier the game.

5.3 The *Collocation matching* exercise

Collocations are the key to language fluency and competence. Lewis [9] believes that fluency is based on the acquisition of a large store of fixed or semi-fixed prefabricated items. Hill [11] 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. Today, teachers spend more time helping students develop a large stock of collocations; less on grammar rules.

LLDL is particularly useful for learning collocations because it contains a large amount of genuine text and provides useful search facilities. In the *Collocation matching* activity, students compete in pairs to match parts of a collocation pattern. This is a traditional gap filling exercise in which one part of a collocation is removed and the students fill the gap with an appropriate word. For example, for *verb+preposition* collocations, verbs or prepositions are deleted. Students select the collocation type they want to practice on, and decide which component will be excised. The exercises use complete sentences retrieved from the library as question text.

Students are paired up and one is chosen to control the activity by selecting collocation types. The other one can see what is going on and negotiate using chat. Then complete sentences are presented one by one, with the target collocation colored blue and missing words replaced by a line. The students select the most appropriate word from four choices before the count-down timer expires. When the exercise is complete the pair view their performance in a summary window that shows the question text with collocations highlighted, and the students' answers and scores.

Exercises are generated from collocation metadata. Sentences at the appropriate difficulty level and collocation type are retrieved. The words that appear in the collo-

cations are grouped according to their syntactic tags and used as choices for the exercise. For each sentence, four choices, including the correct one, are picked randomly.

5.4 The *Quiz* exercise

Quizzes comprising a question and a few choices from which the correct answer must be selected are widely used language drills for learning grammar and vocabulary. Traditionally, teachers construct quizzes and use them for practice exercises, tests or exams. Our system offers a unique feature that makes quizzes far more motivational: students can create their own.

The teacher begins by defining a list of topics and perhaps creating a few initial quizzes. Students can select a topic and construct a new quiz by entering up to four words or phrases; limiting the maximum number of questions; choosing whether or not to stem the terms; and specifying sentence types—simple, complex or both.

Once the learner has defined a new quiz or selected an existing one, the system presents the questions. Each has two to five possible answers. When the student selects one, the system indicates its correctness and moves to the next question. Students can get help by initiating a digital library search for sentences that contain the correct word or words, without being told which one it is. When the quiz is finished a summary is shown of all questions, along with the correct answer and the student's incorrect ones. Students then re-take the questions on which they performed poorly.

This activity uses a simple quiz-generation mechanism that constructs questions and answers using words or phrases specified by students. For example, a question might be *What did you think ___ the film?* with possible answers *of*, *at*, *about*, and *over*. The question text comprises a single sentence retrieved from the digital library using words or phrases specified by the student. These are excised from the question text and used as the correct answer. Sentence retrieval employs full text search on the sentence text and metadata. For example, to construct questions on prepositions, teachers retrieve sentences by searching on *Preposition* metadata. To avoid students having to understand the metadata structure, they are only asked to provide the words or phrases of interest when creating new quizzes.

Stemming is a key parameter for quiz generation that significantly affects the number of available questions and choices. Without stemming, the question text for a *make* and *do* quiz would be restricted to sentences that contain *make* or *do*, and students would have only two answer choices. With stemming, different forms such as *making*, *makes*, *doing* and *does* are also provided as alternatives.

Students can use stemming to explore the variants of a word. When teaching a new word, teachers often encourage students to check its variants in a dictionary. This activity enables students to find variants and practice them by creating an appropriate quiz. For example, students use a quiz to learn more about the variants of *effect*, namely *effects*, *effective*, and *effectively*.

5.5 The *Scrambled Sentence* exercise

The words of sentences are permuted and students must sort them into their original order, to help study sentence structure. Students can select suitable material to practice on.

LLDL retrieves sentences from the digital library, according to selected criteria specified by the student:

- Word or phrases that must appear
- Corpus that the sentences come from
- Difficulty level
- Sentence type (simple, complex, or both)
- Number of sentences retrieved
- Whether to sort in ascending or descending length order.

Once the sentences have been retrieved, they are permuted and presented one after another. The search terms are put in their correct position, highlighted in blue. Students can view the title of the document containing the sentence, and the sentences preceding and following it, by clicking the *help* icon.

In this activity, students can see what other students are doing, in order to encourage them to help each other and learn from their peers' mistakes. Their names are shown (the list is updated when students log in and out); clicking a name allows you to observe how that student unscrambles a sentence by observing his word moves. Students can use chat to discuss the exercise or help each other. Teachers can also log in and observe what the students are doing, and identify and analyze their errors.

6 Evaluation

LLDL demonstrates the roles that digital libraries can play in language study. It has been extensively evaluated, although we have not attempted to assess effectiveness—whether it results in efficient learning—because this paper addresses digital library issues rather than educational ones. We have also drawn a line between evaluating the system itself and evaluating the language material that teachers have put into it.

We conducted four kinds of evaluation: metadata extraction, usability, and activity evaluation with both teachers and learners. We recruited three different kinds of evaluator: usability experts, teachers, and students. The teachers also contributed to the system throughout its development, and helped recruit language students as evaluators. The evaluation is anecdotal rather than quantitative.

6.1 Evaluating Metadata Extraction

Extracted metadata provides the underlying framework for LLDL by facilitating automatic exercise generation for the various language activities. However, they are not always accurate. Sample documents were used to assess the accuracy of sentence boundary detection and identify language constructions and collocations. We identified several tags that had been incorrectly assigned by OpenNLP, causing errors in both the *Tagged sentence* metadata and the values associated with the syntactic metadata types. Four factors affect the accuracy of *collocation* metadata. First, errors in tagging produce incorrect matches against the underlying syntactic pattern. Second, the numbers used to calculate the *t*-values are not exact. Third, the choice of the rejec-

tion threshold is arbitrary. Fourth, groups of words that commonly come together more often than chance are not necessarily good collocations.

6.2 Evaluating Usability

Evaluators examined the interface and judged its compliance with recognized usability principles. They focused on:

- Explicitness: users understand how to use the system
- Compatibility: operations meet expectations formed from previous experience
- Consistency: similar tasks are performed in similar ways
- Learnability: users can learn about the system's capability
- Feedback: actions are acknowledged and responses are meaningful.

Three rounds of usability evaluation were conducted, by usability experts, students, and language teachers. This feedback was used to improve the interface before embarking on the next stage of evaluation.

6.3 Evaluating Activities by Language Teachers

We showed the system to teachers at an early stage, and they proposed several activities that were incorporated into the system we have described. We also made other modifications based on their feedback, giving more search options for the scrambled sentence exercises, excising only nouns and verbs in the *Predicting words* activity, and showing students extracted collocations for the *Collocation identification* activity.

Later we performed a further evaluation, focusing on:

- Do the activities meet the teachers' original expectation?
- What do they think of the feedback provided to students?
- Which ability levels are the activities suitable for?
- What do they think of the exercise material that is used?

On the whole, the teachers thought the activities exceeded their original expectations. They especially liked the use of authentic reading material. They also liked the feedback provided to students, particularly the summaries provided at the end of exercises. They made many constructive and detailed comments on the individual exercises which were used for further improvements such as providing help and hints, and in some cases to enhance the functionality of the exercises.

6.4 Evaluating Activities by Language Learners

Ten language learners, from 18 to 67 years old and native speakers of Arabic, Chinese, Italian and Japanese, participated in an experiment aimed at assessing student satisfaction with the activities. They were grouped into beginner (2), intermediate (4) and advanced (4), and paired up with like partners. In each session they tried out three activities. They filled out a questionnaire and answered verbal questions. The eight activities were allocated to the different levels in accordance with the teachers' advice.

On the whole, we did not learn a great deal from the language learners themselves. It was gratifying to find that the participants liked the activities, and appreciated any opportunity to do exercises outside the classroom. They could understand the feedback provided, but would have liked explanations of answers to be provided. It is easier for younger people with better computer skills to adjust to this learning environment and make the most of it. The evaluation also showed that the competitive activities are more attractive to younger students, and (predictably) male ones.

7 Conclusion

Digital libraries have stunning potential for improving language teaching and learning. But while there are thousands of language learning systems on the web, the potential of digital libraries in this domain remains virtually untapped. Digital libraries contain authentic text, have comprehensive search capabilities, and can automatically generate precisely-targeted exercise material. They can also provide a social environment for students to work in. Teachers can build their own collections—such as the image collections used for the image guessing activity. The library paradigm of assigning metadata to documents serves to separate the structure of exercises from their content. The digital library paradigm of automatic metadata extraction frees teachers from the onerous task of producing exercise material by hand.

We have demonstrated that stimulating educational activities can be build on top of digital libraries that have been augmented with metadata designed specifically to support language teaching. The activities are novel, and incorporate elements of cooperation, competition, and communication. All use authentic material from the digital library instead of artificial made-up examples. They have analogies in traditional classroom activities used for language teaching, but most go much farther than it is feasible to do in the classroom environment, particularly under the inevitable constraints of material that has to be carefully prepared in advance.

The activities we have devised can be used in a classroom setting or for private study. In exercises that involve pairs of students, the system matches them up automatically. In many cases students can create their own exercises. The chat facility provides a social environment that is integrated into the educational setting.

The interface to the LLDL system is explicitly designed to be multilingual (it is available in Chinese as well as English). Resource bundles for different languages have the same set of keys, which are used internally by the program, and different value strings for different languages. They are named following certain conventions that make them easy to locate. To add a new interface language all you need to do is create a bundle for that language and drop it into the folder where the resources are stored.

The system is not restricted to teaching English, the current target language. To extend it to other languages, difficulty metrics and open source implementations of rudimentary parsing techniques are needed in the target language. Second language learning is one of society's greatest challenges, and one that is particularly relevant to Europe. We believe that language learning will prove to be a key application of the European digital library.

References

1. Marchionini, G. and Maurer, H. (1995) "The role of digital libraries in teaching and learning." *Comm ACM*, 38(4), 67-75.
2. Witten, I.H. and Bainbridge, D. (2003) *How to build a digital library*. Morgan Kaufmann.
3. Edge, J. (1993) *Essentials of English language teaching*. Addison Wesley Longman.
4. Clouston, M.L. (1997). Towards an understanding of culture in L2/FL education. *Ronko: K.G. Studies in English*, 25, 131-150.
5. Warschauer, M. and Healey, D. (1998) "Computers and language learning: An overview." *Language Teaching* 31: 57-71.
6. Vygotsky, L.S. (1978). *Mind and society*. Cambridge, MA: Harvard University Press.
7. Friesen, N. Mason, J. and Ward, N. (2003) "Building educational metadata application profiles." *Proc Int Conf on Dublin Core and Metadata for e-Communities*, pp. 63-69.
8. Flesch, R. (1948). A new readability yardstick. *J. Applied Psychology, 32*, 211–233.
9. Lewis, M. (1997) *Implementing the Lexical Approach*. Language Teaching Publications.
10. Manning, C. and Schütze, H. (1999) *Foundations of Statistical NLP*. MIT Press.
11. Hill, J. (1999) "Collocational competence." *English Teaching Professional*, V. 11, pp. 3-6.