

Reuse of Lexicographic Examples in a Web-based Learners' Dictionary

Judith Knapp
European Academy Bolzano, Italy
jknapp@eurac.edu

Johann Gamper
Free University of Bolzano, Italy
jgamper@unibz.it

Peter Brusilovsky
University of Pittsburgh, USA
peterb+@pitt.edu

Abstract: Content creation for a language learning system is a time-consuming, laborious, and expensive task. It is therefore desirable to reuse content as much as possible in different learning situations. An important part of content is the illustrative content, i.e. example sentences. Such examples show a typical context in which a given rule can be used. We discuss a solution to reuse illustrative content in a Web-based learners' dictionary for the German and Italian languages. We explore fully the potential of hypertext, which allows building complex networks of small pieces of learning material. Natural language processing techniques are applied to facilitate the reuse of the manually created content. Our approach can be combined with other, more classical initiatives for sharing and reusing learning material. A first evaluation shows promising results.

Introduction

The use of new media and information technologies for language learning and teaching has become a distinct research discipline, known as *computer-assisted language learning (CALL)*. In (Gamper & Knapp 2002b) an extensive review of CALL systems that include Artificial Intelligence tools and technologies to increase the learning process (ICALL) is given. It was found, that in most cases AI tools are used on the interactive level: Student input is analyzed, either in written or in spoken form and appropriate feedback is given. However, since the technologies are still not mature, errors in the analysis may occur. Furthermore, for such an analysis to be efficient, a special error grammar has to be developed which considers typical student mistakes so that it is possible to provide sophisticated feedback, rather than simple right-wrong-answers.

An alternative approach could be to use natural language processing (NLP) techniques for data preparation. Educational materials could be annotated, language expert systems could be consulted in the annotation process, and their knowledge added to the educational material.

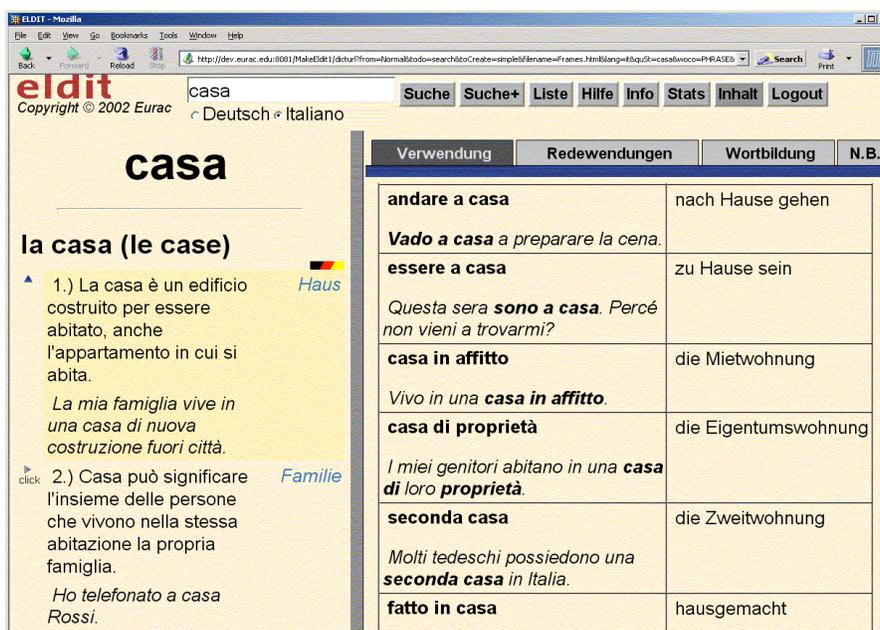
In this paper we present an approach that uses NLP for data preparation. We discuss a solution to reuse illustrative content in a Web-based learners' dictionary for the German and Italian languages. Content creation for a language learning system is a time-consuming, laborious, and expensive task. An important part of content in a language learning system is the illustrative content, namely example sentences. Our approach explores fully the potential of hypertext, which allows building complex networks of small pieces of learning material. NLP technologies such as "Lemmatization" and "Morphological Analysis" are used to prepare and enrich the content. In this way it can be reused more easily in different parts of the system.

The paper is structured as follows: First we present the project ELDIT. We outline the didactic background and explain the concept of "illustrative content" in more details. Then we focus on the reuse of the illustrative content by analyzing the problems we encountered and indicating the corresponding steps to resolve them. Next we describe some evaluations we carried out that showed promising results. Last we refer to related work about sharing and reusing teaching material. Finally we conclude our discussion.

Illustrative Content in ELDIT

At the European Academy of Bolzano (<http://www.eurac.edu>) an interdisciplinary research team is currently developing an innovative, Web-based language learning system for the German and Italian languages, called ELDIT (<http://www.eurac.edu/eldit>). Its core module consists of a German and an Italian learners' dictionary, each one containing approximately 3,000 word entries.

Figure 1 shows a screenshot of the Italian dictionary entry “casa” (house). In the left-hand frame the lemma (casa) together with morphological information and different word meanings are shown. Each word meaning is described by a definition, a translation, and an example sentence. In the right-hand frame additional information regarding the correct usage of the word is shown. A tab metaphor is used to illustrate typical usage patterns, word relations, etc. For example, the tab "Verwendung" (usage), which is selected in figure 1, shows collocations and word combinations. Each item is described by a pattern, a translation equivalent, and a lexicographic example. More details can be found in (Abel & Weber 2000, Gamper & Knapp 2003a).



The screenshot displays the ELDIT dictionary entry for the Italian word "casa". The left panel shows the lemma "casa" and its plural forms "la casa (le case)". Below this, two numbered definitions are provided, each with an example sentence and a German translation. The right panel features a tabbed interface with "Verwendung" selected, showing a table of collocations and their German equivalents.

Verwendung	Redewendungen	Wortbildung	N.B.
andare a casa	nach Hause gehen		
Vado a casa a preparare la cena.			
essere a casa	zu Hause sein		
Questa sera sono a casa . Perché non vieni a trovarmi?			
casa in affitto	die Mietwohnung		
Vivo in una casa in affitto .			
casa di proprietà	die Eigentumswohnung		
I miei genitori abitano in una casa di loro proprietà .			
seconda casa	die Zweitwohnung		
Molti tedeschi possiedono una seconda casa in Italia.			
fatto in casa	hausgemacht		

Figure 1: Dictionary entry of the Italian word "casa" (house) in ELDIT

Psycholinguistic findings suggest that words should not be learned in an isolated way (e.g. as lists of words and respective translations), but in relation to each others and in context, i.e. applied in example sentences and explored in texts (Aitchison 1994, Kielhöfer 1996). Therefore, ELDIT contains a large number of text sentences and lexicographic examples which are used to illustrate language in use, i.e. how the rules can be applied to form correct and meaningful sentences. We call this example sentences *illustrative content*.

Our linguists created manually an illustrative example sentence for the following pieces of information in the dictionary: definitions, collocations, verb valency, and idiomatic expressions. Other pieces of information were listed without examples: adverbs, compound words and derivations. Altogether we have more than 60,000 example sentences which are already included in ELDIT. This is one of the innovative aspects of the system.

The following guidelines are applied when creating example sentences (Abel & Weber 2000):

- the example must show a typical context in which the pattern can be used;
- the example gives a first choice of words that typically occur with it;
- it must be simple, since it illustrates information and should not provide new unknown items.

To overcome the problem with new words in illustrative sentences, each word is linked to the corresponding entry in the dictionary (Gamper & Knapp 2003a).

While ELDIT already contains a large number of illustrative examples, there is still a need for even more example sentences. Especially when learners have a specific context or sentence in mind, they might want to see an example sentence which is rather close to their specific needs. This is an important finding resulting from a first evaluation of ELDIT.

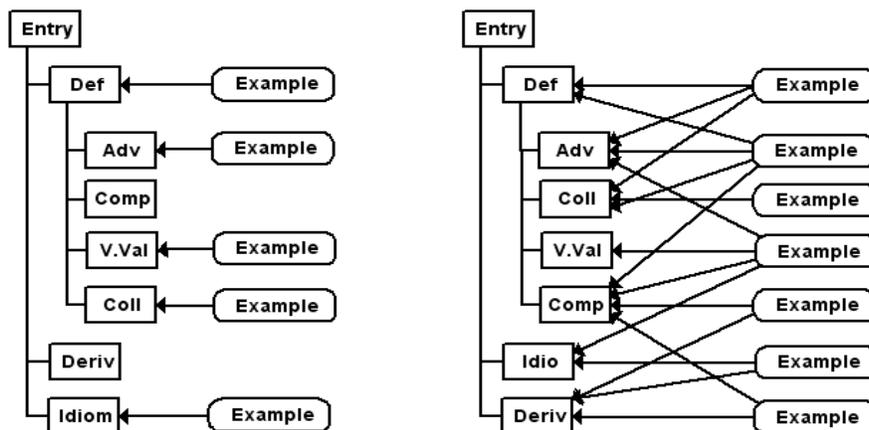


Figure 2: Traditional and new way of providing information and example sentences in ELDIT.

The authoring of didactic material is difficult and time-consuming. Modern hypermedia technologies, however, make it possible to use the existing content in different parts of the system. Hence we can neglect traditional paper based approaches which apply a 1:1 relation between information and illustration. In electronic systems the definition of an n:m relation between information and illustration is possible, which magnifies the amount of example sentences that can be provided for each piece of information (see Figure 2).

We have programmed a module that allows reusing the existing illustrative content in ELDIT at several places. Now for each piece of information not only the main example added by the linguists, but also additional examples created for other contexts but showing the same information can be retrieved. The additional examples can be accessed via a button placed next to the standard example.

Figure 3 shows the result of our work for some collocations of the Italian word *casa*. The first collocation is *andare a casa* (to go home) with the main example *Vado a casa a preparare la cena* (I will go home to prepare dinner). In the additional examples the expressions “I will go home to sleep...” and “...let us go home” can be found. For the word combination *casa in affitto* (rented flat) not only the main example *Vivo in una casa in affitto* (I am living in a flat), but also the additional example *Mio fratello ha preso una casa in affitto ed è andato a vivere per conto suo* (My brother has rent a flat and is now living on his own) can be found, etc. In the first list of additional examples the word testa (head) was unknown to the user. Since all words in the example sentences are linked to the dictionary entry a click on the unknown word shows its definition and translation.

Reusing the Illustrative Content

The basic idea is to retrieve all example sentences which are useful to further illustrate the concept under consideration.

Problems

The retrieval of additional example sentences is not trivial. A simple search in our database does not lead to the desired results. The following problems occur:

1. The patterns under consideration might be unstructured, for instance they may contain meta information such as slashes to indicate variations. For instance, the pattern *gli occhi, la bocca, il viso, ..., belli/bella/bello* indicates several patterns: *gli occhi belli* (beautiful eyes), *la bocca bella* (a beautiful mouth), and *il viso bello* (a beautiful face). All these patterns should be considered when searching additional examples.
2. Words occur in declined or conjugated form both in the patterns and in the example sentences. For example, the collocation *to go home* occurs in the sentence *Yesterday I went home very late*, and therefore this sentence should be matched.
3. It is not sufficient that all words of a pattern occur in a lexicographic example, but they must occur as a collocation. For instance, the word combination *to go home* occurs as a collocation in the sentence *I went home very late*, but not in the sentence *I went out and came home very late*.
4. Words usually have several meanings. For instance, the word *house* may be a building but also a dynasty. Hence, the sentence *The royal house of Norway is a branch of the princely family of Glücksburg* is not a good illustration for the definition *A house is a place to live and to work*.

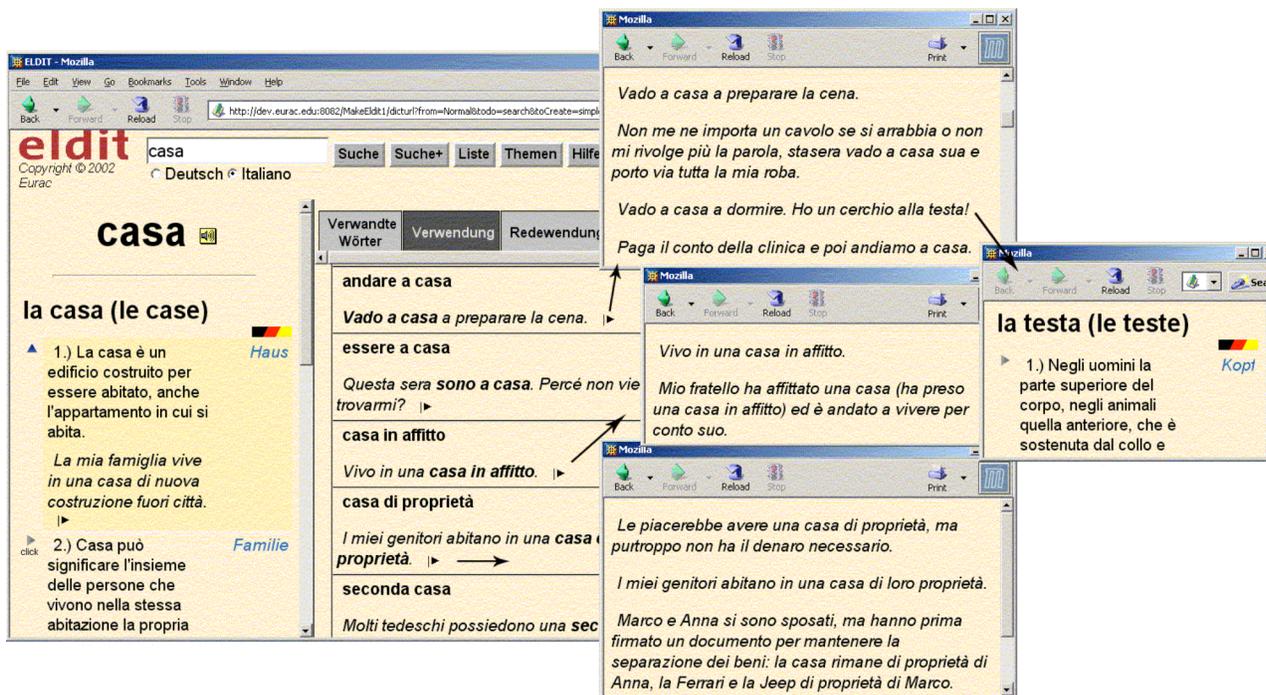


Figure 3: Reusing examples in ELDIT. All words are linked to the corresponding dictionary entry.

Implementation

With regards to the previously mentioned problems the retrieval of these additional examples is a four-step process:

1. Extracting “clean” patterns
2. Retrieving all example sentences
3. Recognition of collocations
4. Disambiguation of meaning

Extracting “Clean” Patterns

The first step is to construct new, “clean” patterns which can be passed to our search engine. Three situations have to be distinguished:

- a) Derivations, compound words, adverbs

- b) Collocations and idiomatic expressions
- c) Definitions and verb valency

Case a) is about words which are generated from a dictionary entry by word formation rules (adverbs, derivations, and compound words). These words are listed within the entry and the word formation rules are highlighted in order to communicate them to the learner. In this case it is easy to get such a “clean” pattern, since the given pattern consists only of the generated word, it occurs in the citation form, and different meanings are not considered. Hence, the word can directly be passed to the search engine.

In case b) we have multiple word expressions which are given in an unstructured form. The first step is to unfold all meta-symbols. For example, unfolding the pattern *ein Auto fährt schnell/langsam* (a car runs fast/slowly) yields the patterns *ein Auto fährt schnell* and *ein Auto fährt langsam*. Then some very general words (such as articles) are removed and the remaining words are connected with an AND function and marked as obligatory. The result for our example is the following set of search patterns: *+Auto AND +fährt AND +schnell* as well as *+Auto AND +fährt AND +langsam*.

In case c) different word meanings have to be considered which are described by a definition such as *Eine Glocke ist ein Gegenstand aus Metall, der irgendwo hängt* (a bell is a metal device that hangs somewhere) or verb valency patterns such as *jemand baut etwas* (somebody builds something). In both cases quite general words are frequently used, for instance the words *Gegenstand* (thing), *irgendwo* (somewhere), *jemand* (somebody), or *etwas* (something). Moreover, in both cases we describe a specific word meaning, which should be matched in the retrieved example. The general words are definitely not suitable for building a good search expression and to find examples that are appropriate for a specific word meaning.

The solution we applied was to build our search expression based on some specific words from the main lexicographic example. From these words we assume that they describe the meaning of the word in consideration: We first extract the clause which contains the word under consideration. From this clause we extract all verbs as well as the first nouns occurring on the left-hand side and on the right-hand side of the word in consideration.

Let us have a look at an example sentence provided for a definition, namely the definition of *house* as the group of people in a theater:

Als die Vorstellung zu Ende war und der Vorhang zu ging, klatschte das ganze Haus begeistert Beifall. (When the performance was finished, the whole house applauded enthusiastically.)

The main clause is “klatschte das ganze Haus begeistert Beifall”. The verb in this clause is *klatscht* (to applaud), there is no noun on the left-hand side of “Haus”, but a noun on the right-hand side of “Haus”, namely *Beifall* (applause). From this sentence we generate the search patterns “+Haus AND +klatscht” and “+Haus AND +Beifall”.

Retrieving Examples

The second step is to retrieve all possible examples from the ELDIT dictionary by passing the search pattern obtained in step 1 to the search engine.

Our search engine applies NLP techniques both on the search pattern and on the learning material in order to retrieve also results that might occur in inflected form. The task is carried out by components derived from Word Manager¹ (Domenig & ten Hacken 1992, Pedrazzini 1998). One of these components, the “WMTrans Lemmatizer”, returns the citation form of any valid word for a specified language. The result of a query is a list of corresponding citation forms followed by the corresponding category (i.e. word class):

[1] The collaboration has been funded by the European Union, Interreg IIIA, Italia-Swizzera.

query → ging
result → gehen (Cat V)

We have first transformed the ELDIT data to its lemmatized form. For instance, the following example sentence is changed to the second version:

I) *Die Vögel bauten ihre Nester aus kleinen Ästen, Blättern und Erde (the birds have built their nests out of branches, leafs, and soil)*

II) *Der Vogel bauen mein Nest aus klein Ast, Blatt und Erde (The bird build my nest out of small branch, leaf, and soil)*

Then we have generated two indices. One index contains the data in the original form, the other one contains all data in the lemmatized form (i.e. the citation form). The two indices include cross references to each others.

The search engine first lemmatizes the obtained search pattern, for instance *Nest AND baut* becomes *Nest AND bauen*, then the lemmatized pattern is searched in the lemmatized index, where the example sentence *Der Vogel bauen mein Nest aus Ast, Blatt und Erde* is found. Last, by using the cross reference system the corresponding original example sentence is retrieved, namely *Die Vögel bauten ihre Nester aus kleinen Ästen, Blättern und Erde*, and given back to the module.

Recognizing Collocations

The third step is to recognize collocations. A sophisticated concordance tool is required to check whether a word combination forms a collocation in a sentence or not. Currently, we are using a rather simple approach based on a set of rules to identify real collocations. Two typical rules are the following ones: (i) the words of a pattern have to occur within one main sentence or within a subordinate clause; (ii) there should not be more than a determined number of other words between the words of the pattern. We are currently working with three words in the case of noun-verb combinations and zero words in the case of noun-adjective and verb-adjective combinations. The application of these rules allows eliminating invalid sentences for a given pattern. For example, let us consider the pattern *ein großes Haus* (a big house). Among others, the search engine retrieves *Meine Eltern zogen in ein Haus mit einem großen Garten* (my parents moved into a house with a big garden). In this sentence the garden is big, not the house. Applying the zero-words rule identifies this sentence as invalid for our specific pattern.

A more sophisticated disambiguation could be done by including a tool such as “Phrase Manager” (Pedrazzini 1994). The inclusion of this system into ELDIT is part of our future work.

Meaning Disambiguation

The last and most difficult step is the disambiguation of word meanings. Currently this step is compiled into the search patterns which include nouns and verbs from the original example sentences (see case c) more above). This is a first step in the creation of a program able to perform meaning disambiguation. Further steps would be possible: Meaning disambiguation programs include context vectors, which are lists of words that are semantically related to the specific meaning of a word. Such context vectors could be obtained e.g. by listing the nouns, verbs, and adjectives of the collocations collected for a specific word meaning. In ELDIT the general context vector of the word *house* as *a place to live and work* would be [*bauen, renovieren, wohnen, kaufen, mieten, vermieten...*], and the general context vector of the word *house* as *a group of people in a theater* would be [*Theater, Vorstellung, Beifall, toben, klatschen, Begeisterung, ...*]. The context vector of the example sentence

Als die Vorstellung zu Ende war und der Vorhang zu ging, klatschte das ganze Haus begeistert Beifall.

could be the list of words used in this example sentence, namely [*Vorstellung, Ende, Vorhang, zugehen, klatschen, Haus, begeistert, Beifall*]. The context vectors of the obtained example sentences could be compared with the general context vector of the word meaning in consideration, and in this way a better indication could be obtained, whether an example sentence really matches this meaning or not.

Performance

In this section we present the results of a short performance analysis in which we examined the didactic validity of the feature discussed in this paper. We systematically inspected the generation of additional examples for some words. We analyzed how many additional examples are found and how many of these items are valid from the didactic point of view. “Validity” means that the example applies the pattern in a correct way and is hence useful for the learner.

Case a): Finding examples for *derivations, compound words, and adverbs* is easy and works fine, since only one word has to be searched, and different meanings are not considered in the presentation. Hence, in our analysis we got a precision of 100%. This result is very important, since we have no manually created examples for these words.

Case b): Finding examples for *noun-adjective combinations* and *verb-adjective combinations* is also quite easy. We got a precision of 100%, since a large number of examples contain such combinations, and hence, we can apply very restrictive rules to determine the final set. For *noun-verb combinations* the first difficulties with invalid results arise (precision is approximately 90%) since our rule system is not a very precise concordance tool. For the *idiomatic expressions* the results are not yet satisfactory. Very few examples are found and the found ones have mostly been considered as invalid. The reason is that the examples in ELDIT are kept simple, and idiomatic expressions are hardly used within them.

Case c) In the part of examples for *verb valency* and different *word meanings* we used the original example sentence to obtain a search pattern. A lot of (mostly valid) examples are found for common, frequently used meanings and patterns, whereas only very few (and mostly invalid) examples are found for the less frequently used meanings and patterns.

We decided the following policy: In those cases where the feature performs well it might be a useful tool for learners working with the ELDIT system. For the unsatisfactory cases, however, the feature is actually disabled in the production version until we have appropriate tools to improve the results.

Related Work

Recently, much research has been carried out to develop systems, data models, and standards for Web-based learning, specifically for sharing and reusing teaching material over the Web. In (Henze & Nejd1 1999) a data model to support constructivist learning in the KBS-Hyperbook system is described. In (Süß et. al. 1999) a meta-modeling approach to adaptive hypermedia-based electronic teachware is described. SCORM (Sharable Content Object Reference Model) is a suite of technical standards that enable web-based learning systems to find, import, share, reuse, and export learning content in a standardized way (Dodds & Thropp 2004). LOM (Learning Object Metadata) (Hodgins & Duval 2002) and DC (Dublin Core) (DCMI Usage Board 2004) are standards for management of educational content metadata. All these approaches focus on document structures and navigation services. The main difference of these projects to our approach is the level of granularity. While in general the basic building blocks are learning objects which represent a domain concept, we need a more fine-grained approach which considers the learning material down to the level of single words and further.

Searching text corpora with concordance tools is not new in language learning. For instance, the aim of the PET 2000 (Cobb 1999) and Lexica (Goodfellow 1999) projects, was to help students to acquire the necessary vocabulary for the English language. Students can search a text corpus using concordance tools. In this way they can explore words in context and build their own vocabulary database consisting of words with definitions and examples. ELDIT takes a different approach by combining hand-crafted examples with automatically retrieved examples. The

detailed data model, the hypertext nature of our system, and the application of computational linguistics technologies allow searching and reusing these pieces on different places, hence we can offer a large number of carefully prepared text pieces on many different places which magnifies the amount of information provided in the system. Moreover, in ELDIT all words are linked to the corresponding dictionary entry, hence unknown words found in the example sentences can be checked by a simple mouse click.

Conclusions

We presented a solution to content reuse in the ELDIT dictionary for learners of German and Italian. Since the authoring of learning material is very time-consuming, it is desirable to reuse it for different learning situations. Our approach to reuse illustrative example sentences explores the potential of hypertext and NLP technologies, which allows building complex networks of small pieces of learning material and identifying correctly small pieces of information within larger text pieces. Our approach can be combined with other, more classical initiatives for sharing and reusing educational contents. Using NLP for the identification of specific words and collocations could also work for the semiautomatic adding of metadata to existing educational content.

A first evaluation shows promising results of our approach and the didactic usefulness of this additional information for the learner. Still, the automatic generation of the hypertext links using shallow reasoning and pattern matching techniques can and has to be improved.

References

- Abel, A., & Weber, V. (2000). ELDIT, prototype of an innovative dictionary. In *Proceedings of the 9th EURALEX International Congress on Lexicography (EURALEX'00)*, Stuttgart, Germany.
- Aitchison, J. (1994). *Words in the Mind: An Introduction to the Mental Lexicon*. Blackwell Publishers Ltd, Oxford, UK, 2nd edition.
- Cobb, T. (1999). Breath and depth of lexical acquisition with hands-on concordancing. *Computer Assisted Language Learning*, 12(4):345–360.
- DCMI Usage Board (2004). *DCMI Metadata Terms*, <http://dublincore.org/documents/dcmi-terms/>
- Dodds, P & Thropp, S. (2004). *Sharable Content Object Reference Model (SCORM) 2004 Overview*, Advanced Distributed Learning (ADL).
- Domenig, M. & ten Hacken, P. (1992). *Word Manager: A System for Morphological Dictionaries*, volume 1 of *Informatik und Sprache*. Olms Verlag, Hildesheim.
- Henze, N. & Nejdil, W. (1999). Adaptivity in the KBS hyperbook system. In *Proceedings of the 2nd Workshop on Adaptive Systems and User Modeling on the WWW*.
- Gamper, J. & Knapp, J. (2003a). A data model and its implementation for a Web-based language learning system. In *Proceedings of the 12th International World Wide Web Conference (WWW2003)*, pages 217–225.
- Gamper, J. & Knapp, J. (2002b). A review of intelligent CALL systems. *Computer Assisted Language Learning*, 15(4):329–342.
- Goodfellow, R. (1999). Evaluating performance, approach and outcome in the design of CALL. In *CALL: Media, Design & Applications*, pages 109–140. Swets & Zeitlinger, The Netherlands.
- Hodgins, W. & Duval, E. (2002). *Draft Standard for Learning Object Metadata*, Institute of Electrical and Electronics Engineers, Inc (IEEE).
- Kielhöfer, B. (1996). Psycholinguistische Grundlagen der Wortschatzarbeit. *Babylonia*.
- Pedrazzini, S. (1994). *Phrase Manager: A System for Phrasal and Idiomatic Dictionaries*, volume 3 of *Informatik und Sprache*. Olms Verlag, Hildesheim.
- Pedrazzini, S. (1998). The finite-state automata's design patterns. In *Proceedings of Third International Workshop on Implementing Automata (IWA'98)*.
- Süß, C., Freitag, B. & Brössler, P. (1999). Meta-modeling for Webbased teachware management. In *Advances in Conceptual modeling – Workshop on the World-Wide Web and Conceptual Modeling (ER'99)*, Springer Verlag.