Browsers for Autonomous and Contextualized Language Learning: Tools and Theories

Oliver Streiter, Judith Knapp, Leonhard Voltmer, Daniel Zielinski
National University of Kaohsiung, Taiwan[1]; European Academy, Italy[2−3]; University of Saarland, Germany[4]
http://140.127.211.213/gymnazilla

***Abstract*— We present a new paradigm for Computer Assisted Language Learning (CALL). This paradigm aims at the development of linguistically and pedagogically competent Web-browsers for autonomous exploration of a second language (L2). We introduce Gymn@zilla, a browser-like application which converts Web-pages automatically into language lessons. Gymn@zilla combines annotated reading of Web-pages with exploration tools and the possibility to create personal wordlists and practice them in dynamically created exercises. The relevance and potentials of theses components for the acquisition of a second language are discussed. Possible usage scenarios range from individual language learning over school and university classes to daily working scenarios in non-native speaking environments.**

I. Introduction

A. CALL and NLP

The acquisition of a Second Language (L2)[1] occurs in diversified contexts and often outside the classroom. This learning process can be stimulated and systematized by tools for computer assisted language learning (CALL). As the development of CALL systems requires huge investments into pedagogical, linguistic and technical resources, we propose an automatic approach to CALL. Such an approach makes use of Natural Language Processing (NLP) to elaborate authentic documents[2] for L2-learning. It is thus possible, for example, to link an L2-document to an electronic dictionary. This allows the language learner to access translations of words and phrases in one click. This annotated reading within a Web-browser epitomizes the essentials of up-to-date L2-learning theories, such as autonomy [3], personalized learning [6], contextualized learning [26] and learning with authentic material [24].

B. CALL and Web-browsing

This paper explores the idea of a Web-browser as environment for L2-learning, drawing basically from our experience with the development of Gymn@zilla [27], [28], [11]. Gymn@zilla, which has been available until recently only as a unique Web-service, currently undergoes a major re-implementation to produce a free GNU GPL[3] software for the private and schools. While the second version is developed, a third version which foresees the complete integration in a browser (Firefox) is prepared. Our discussion of browser-based CALL, when citing examples of Gymn@zilla, will gloss over the different version and use example which seem most suitable to make our point.

Fig. 1. Annotated reading. When integrated into a browser, authentic Internet texts can be exploited autonomously for L2-learning.

C. Gymn@zilla and Web-browsing

Gymn@zilla supports browsing the Internet and a local document repository by dynamically annotating HTML and PDF documents with open dictionaries resources. By clicking onto new words the learner obtains information about translation and pronunciation, memory cues and options for active word exploration. In addition, the learner can construct wordlists which contain the personally chosen words with example sentences and images from the document the learner has seen. The learner can also convert the personal wordlist into a number of completion exercises to train new words.

II. Gymn@zilla, Starting the Tool

A. Language Selection

The learner can install Gymn@zilla on his or her private computer or visit a running Gymn@zilla web-service at his or her school. From the Gymn@zilla start-page the learner can select the language pair, e.g. eng-deu, for reading English web-pages with support in German.

The heart of Gymn@zilla is NLRDF [29], a huge XML-repository which contains the characterization of several hundred languages, their preferred encoding in the Internet, the countries or regions where a language is used and the URL of

[1]The language you are learning is called the *target language*. In second language acquisition, the target language is frequently referred to as L2. L1, then refers to the mother language or the language of instruction.

[2]"*materials can be considered authentic as they were originally intended to be used for non-educational purposes by native speakers of the target language*"[2] pg.24

[3]The GNU General Public License provides a legal means to guarantee the freedom to share and change free software and to guarantee that the software is free for all its users.

0-7803-8932-8/05/$20.00

dictionaries and other language resources which are used by Gymn@zilla when processing documents in that language. On the basis of the specification for each language (segmentation rules, morpheme lists, word lists), the tokenization into words is performed. Further directives in this file guide the stemming for a number of European languages.

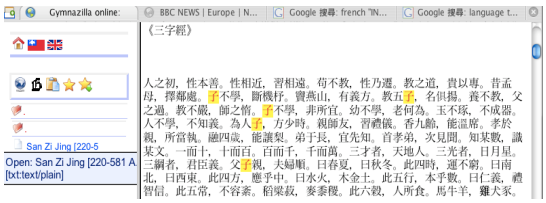
B. Document Selection

After selecting the languages, Gymn@zilla expects a text or document to be submitted. This is the join which might interacts with other programs. We strive for a maximal integration of Gymn@zilla into other activities such as Web-browsing, reading mails and word processing so that language learning is involved in all these task. Currently the learner can type in an URL, cut and past a text, browse the local directories or use pre-defined bookmarks.

C. Bookmarks and Project Gutenberg

The bookmarks have been assembled by the developers of Gymn@zilla and can be completed by language teachers. They point to high-quality web-sites such as BBC News for British English. In addition, Gymn@zilla's bookmarks include the entire Project Gutenberg's Document RDF, covering about 15.000 documents in 30 languages. Gymn@zilla thus integrates under the well-know form of bookmarks what otherwise is know in CALL as Internet Trails, e.g. collections of links to web-sites useful for L2 learning (e.g. http://www.fln.vcu.edu/default.html). With the bookmarks, the language, the region, the encoding and the document format are stored in RDF, to process documents correctly.

Fig. 2. Selection of a Chinese document from Project Gutenberg



Gymn@zilla means intercultural and interlingual learning. The cultural dimension in Gymn@zilla comes through the content of the Web-pages, the layout and design of the Web-pages which is not touched by Gymn@zilla and the overall structure imposed by the Gymn@zilla bookmarks. The cultural connection that traditional teaching materials tried to introduce artificially, thus comes to Gymn@zilla most naturally through the cultural texture of the resources.

D. Mirroring of Web pages

Mirroring of web pages is handled internally by WGET. Before downloading a web-page, it checks whether the web-page has been updated relative to a cached version. WGET creates appropriate directory structures to store pages locally. Irrespective of how WGET provides the most up-to-date file, Gymn@zilla takes the file from the WGET directory. Character encodings other than UTF-8 are converted to UTF-8. Document formats other than HTML such as PDF are converted to XHTML. Hyperlinks in a web page are transformed into CGI-parameters attached to Gymn@zilla's URL. This allows continuous browsing with Gymn@zilla. Links to multimedia documents such as audio, video and graphic files are preserved so that they can be integrated into the learning scenario.

E. From Cache to Background Corpus

Besides caching documents for reasons of efficiency, it might be interesting to build up a background corpus of documents seen by a specific learner. This background corpus could fulfil a number of functions, e.g. improve statistics about word n-grams, collocational forces and the specificity of a word for a given text, time, web-site, or simply to remind the learner of words and phrases he has read in the recent past.

III. Exploring the Web-page

When the learner queries a page, Gymn@zilla returns the almost original HTML-document. The annotations added by Gymn@zilla are first invisible. Before clicking on words to get a translation, the user may explore new L2 words.

A. KWIC and L2 Learning

Moving the mouse pointer onto a word highlightens all occurrences of the word in the given document (cf Fig. 1 and Fig. 2). Thus similar to a linguist who highlightens a word under scrutiny in a corps with the help of a KWIC-tool (Keyword in Context), the learner can highlighten all occurrences of a word in the document, keeping the original structure of the document. If only a few occurrences of a word show up, a background corpus might return additional examples from the documents the learner has seen before.

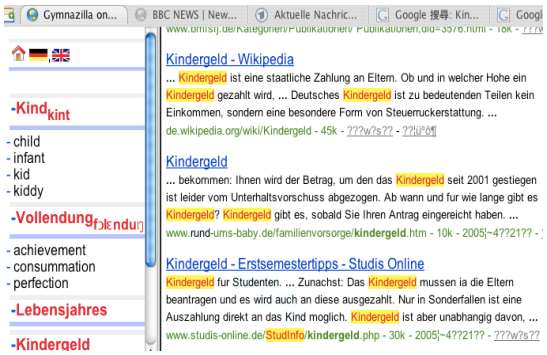
B. Term Extraction and L2 Learning

A personal background corpus might even be interesting for another function which encourages the learner to explore the target document. The very first version of Gymna@zilla included a term extraction tools which marked for each document the central terms. The main terms of a document might not only facilitate the understanding of the document, but the terms themselves are more likely to be learned as they are prominently marked.

C. L2 Learning and the Credo of Constructivism

These are two examples, of how techniques borrowed from corpus linguistics or computational terminology might be included to support the exploration of a document without falling back onto a simple word to word translation. The learner is thus seen as active researcher who creates his or her own knowledge via contextual deduction, theory formation and falsification. Web-based CALL might thus combine the strand in CALL which explores techniques from corpus linguistics with the attempt to work on authentic documents. The corpus linguistic tools would be at hand for every document the learner reads with the browser.

Fig. 3. Google search result in Gymn@zilla creates KWIC from the Internet and suggest new, related words for further exploration.



From a psychological point of view, such meaningful activities around new items result in learning. Thus if one of the aims of CALL is to facilitate the acquisition of lexical material, a system should involve the learner in such meaningful activities which can be relate to already established knowledge fragments. We will return to this point when discussing *intentional*— and *unintentional learning* below.

Additional links for the explorations are easy to create: Links to online dictionaries of the target language as well as links to the google image and document search. Online dictionaries provide alternative translations, definitions and the possibility to check the back translation. The Google image searches may give important cues on possible polysemies, different denotational extensions in L1 and L2, cultural particularities and connotations. Google document search finally might function as a very simple KWIC-tool which uses the Internet as document. Interestingly, Gymn@zilla continues to browse in the annotation mode, even if dictionaries and research hits are rendered, taking the output of the query as starting point for new explorations.

IV. Annotated Reading

Clicking on an annotated word shows the word in its inflected form (e.g. *children*), in its base form (*child*), its possible translations in L1 and an image associated with L2 (see our discussion below). If provided by the dictionary, information on the pronunciation is given as well. This is of particular importance for languages where the writing does not or only partially reproduce the pronunciation (e.g. Chinese).

A. Annotating with NLP modules

In order to map the words of the document with high recall and precision onto dictionary entries, NLP modules for stemming (en:*children* ⇒ *child*), word-segmentation (de:*Importdefizit* ⇒ *Import/Defizit*, tagging (*man* ⇒ *noun*), and meaning disambiguation (*mouse* ⇒ *mouse pointer*) are used. Tagging and meaning disambiguation are particular difficult in Gymn@zilla, as the dictionaries come from different sites and often do not bear any information on the part of speech or meaning distinctions. If the inflected form or the base form of a word, or the concatenation of several words are found in a translation dictionary, the expression is annotated with the information coming from the free dictionaries.

B. Pedagogical aspects

The positive effect of annotated reading on L2-learning has been repeatedly shown ([18], [15]). As a tendency, learners check more lexical entries with hyperlinked texts than when using paper dictionaries. Accordingly, they remembered more words. The experiments have been done with European languages only. One might expect even more significant differences with an L2, such as Arabic, Chinese or Japanese, where a learner may occasionally need up to 1 hour to find a word in a paper dictionary.

C. Creating a Personal Wordlist

The learner can add new words to a personal wordlist by clicking a button. Following the link ”wordlist” in the menu bar, this personal wordlist is shown. The personal wordlist can be copied into a private word processor for further editing or it can be printed out for root learning. Given the five components of a wordlist entry, i.e. L2-lemma[4], pronunciation, image, context and L1-equivalent, learners may focus on specific relations (e.g. image - L2-relations).

D. Creating Exercises

The personal wordlist can be dynamically transformed into a series of exercises. These exercises may be used for training or testing. They are designed not to involve L1 and to reduce the interference of L1 in the processing of L2 (see our discussion below). Instead, the exercises train the bonding first of the image (representing the concept) and the L2-denomination, second, the lexical selection in an L2-context and third, the bonding of L2-lemma and the pronunciation (e.g. for Chinese). The learner may repeat the exercises with rising degrees of difficulties. The different degrees of difficulties are generated by scrambling characters in words or words in sentences. This feature has been introduced after first experiments showed that otherwise learners acquire only superficial knowledge of new words.

E. Incidental vs. Intentional Learning

Depending on the web-page, Gymn@zilla exposes learners abundantly to new vocabulary in personal contexts. Gymn@zilla activates the mental lexicon in several steps and on several levels. On one hand, words are learned incidentally by reading web-pages annotated with dictionary information ([4], [21]). On the other hand, words are learned intentionally by editing and studying the personal wordlist and working through completion exercises. By the combination of incidental and intentional learning, we hope to overcome the respective disadvantages of each approach.

Intentional vocabulary acquisition results from purposely learning words and their respective translations with the aid of wordlists. This method is usually appreciated by learners.

[4]The lemma of a word (e.g. *child*) is the *base form* or *citation form* of a word that represents a paradigm of words, which is also called the lemma, e.g. *child, children*.

It is fast and direct. The disadvantage is its superficiality. The learner encounters isolated words in their base form. Therefore, the learner may rapidly forget learned words or may not be able to use them in context. Instead, extensive word exposition is necessary to ensure a deep and solid embedding of vocabulary in the mental lexicon ([1], [10]). Moreover, vocabulary acquisition should be personalized and should occur in context with authentic text ([9], [17], [25]).

Incidental vocabulary acquisition refers to vocabulary acquisition by contextual deduction, e.g. when reading text in a foreign language. This method ensures that words are encountered in several semantic contexts and morphological forms. Hence new words are embedded much deeper within the mental lexicon. Through the patterns of syntactic contexts the right usage of a word is acquired together with the word.

However, as [5] points out, this method is not fast. It would require a considerable amount of time to master the vocabulary needed for fluent and correct conversations. Another problem with this approach is the possibly large amount of new words around the words to be learned [8]. Thus learners at beginners levels may be overwhelmed with this approach. A strategy according to which a learner switches between the two learning approaches thus seems to be promising. Many words may be learned quickly with intentional learning and then rooted with incidental learning.

V. Image Processing

A. Image-Lemma Mapping

When learners solicit pages through Gymn@zilla, the addresses of images (the SRC-attribute) and the alternative text (the ALT-attribute) of images are compared to the text. If the string of a certain word – e.g. the word ”Pakistan” – appears in one of the attributes (e.g. ”.../Pakistan/....jpg) and in the text on the page, Gymn@zilla stores the image’s address with ’Pakistan’ as a search key.

In this and future Web-pages, Gymn@zilla is thus in state not only to provide a translation of foreign words, but to show a fairly up-to-date image related to that word. The shelf life of an image may be quite short as any newer image provided by this Web-domain will overwrite the older one. In order to guarantee a thematic consistency of text and image, images have been limited to the domain from which they originated.

Fig. 4. Terms associated with images: *new, Pakistan, gay marriage*

B. Image and the Part of Speech

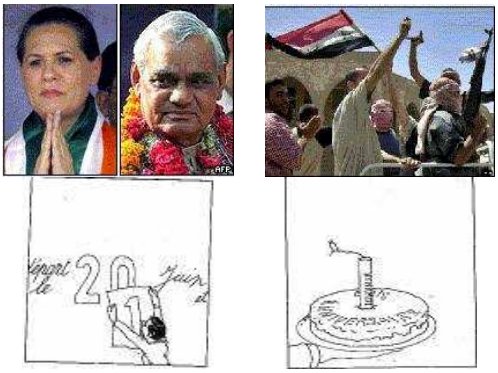
Image processing seems to be especially successful with proper nouns (names of persons, cities, countries). Although such words may not be included in a dictionary, they provide a learner with valuable cultural knowledge which facilitates the assimilation of additional information. Proper nouns and concrete nouns of the basic level ([23]) can be identified from their image. All other words, including more abstract or more specific nouns, adjectives or verbs have to be understood, in order to understand why a given image has been selected. We hypothesize that finding a match between meaning and the image requires a cognitively demanding activity which becomes wonderfully rewarded when the match is found.

Fig. 5. Annotation of proper names with images in Gymn@zilla.

C. Images in Gymn@zilla, are they strange?

Images have been widely used for L2-learning. We notice, however, that the fully automatic selection of images in Gymn@zilla contrasts with editorial habits in the elaboration of learner’s dictionaries, encyclopedias, and children’s books. While Gymn@zilla features colored photos of utmost contextualization, the traditional usage of images is based on drawings which show a considerable degree of abstraction and de-contextualization. Figure 6 illustrates this contrast, opposing images in Gymn@zilla to those in [7].

Fig. 6. Automatically selected images (14.05.2004) and images in a paper dictionary for French as foreign language ([7]): *to change* and *to celebrate*.



The difference is even greater if we consider the short period in which a contextualized image is understandable. Therefore, images are constantly updated when newer images become available. The images in paper dictionary last for 40 years.

D. Images and L2-Learning

When looking for psycholinguistic motivations for the usage of images, we find that already Paivio showed that words with

Fig. 7. Image annotations for "to celebrate" as they are found on 3 consecutive days in Gymn@zilla (13.5.2004-15.5.2004)



high imagery are more easily processed ([19]). According to his Dual Coding Theory, there are two cognitive subsystems. One is specialized for the representation and processing of nonverbal objects and the other is specialized for language. A word that symbolizes a concrete object can be encoded twice in memory. As a consequence, there should also be two ways the word can be retrieved from memory. An increasing body of research supported this conjecture. Learning is affected positively by presenting text and images together ([16]). Dynamic visualizations have been shown to be superior to static visualizations ([22]).

Another strand of research shows that proficiency in L2 is related to the degree to which the L2-production is conceptually mediated. In beginners, the L2-production is lexically mediated through L1: The translation of words from L1 into L2 is faster than picture naming in L2. For proficient bilinguals however, these activities are almost equally fast, suggesting that both activities are conceptually mediated ([12], [20]).
A lexical mediation can still be found with proficient bilinguals in a L2 to L1 translation task. This disappears however, when words to be translated are preceded by a image context ([14]). The presence of images thus reduces lexical mediation.

What helps memorizing are "unique environmental cues for L2 or unique concepts of nuances of meaning that are distinctly associated with L2" ([13]). Unusual images attenuated the regress to L1 and privileged the bonding between L2 and the concept, identified with the non-canonical image. Focusing of particular features of L2-denoted entities may thus be particularly rewarding. The features of difference which can be represented in drawings to a limited extend only, are omnipresent in photos, even if they are due to incidental differences or the association with different events or episodes.

VI. Conclusion

In this paper we presented Gymn@zilla, an open source browser-like application for L2-learning. Gymn@zilla combines annotated reading with KWIC and the creation of personal wordlists and dynamic exercises. Possible usage scenarios range from individual language learning over school and university classes to daily work scenarios in a non-native speaking environment.
Gymn@zilla follows a learner centered approach. In principle any Internet or Intranet page can be chosen. While the use of Gymn@zilla remains the same, the resources can change from extremely technical text with content explanations by professors to rather simple texts for beginners, helped by a language teacher. With changing resources in the Internet, Gymn@zilla's resources will automatically be updated.

References

[1] J. Aitchison. *Words in the Mind: An Introduction to the Mental Lexicon*. Blackwell Publishers Ltd, Oxford, UK, 2nd edition, 1994.
[2] K. Beatty. *Teaching and Researching Comp.-ass'd Lang. Learning*. Applied Linguistics in Action. Longman, London, 2003.
[3] P. Benson. *Teaching and Researching Autonomy in Lang. Learning*. Applied Linguistics in Action. Longman, London, 2001.
[4] D. M. Chun. L2 reading on the Web: Strategies for accessing information in hypermedia. *Comp. Ass'd Lang. Learning*, 14(5):367–403, 2001.
[5] T. Cobb. Breadth and depth of lexical acquisition with hands-on concordancing. *Comp. Ass'd Lang. Learning*, 12(4):345–360, Oct. 1999.
[6] J. DiMartin (ed). *Personalized Learning*. Scarecrow Press, 2003.
[7] J. Dubois. *Dictionnaire du français langue étrangère, Niveau 1*. Larousse, Paris, 1987.
[8] P. J. Groot. Comp. ass'd second lang. vocabulary acquisition. *Lang. Learning & Technology*, 4(1):60–81, May 2000.
[9] C. Jones. Contextualise & personalise: Key strategies for vocabulary acquisition. *ReCALL*, 11(3):34–40, Nov. 1999.
[10] B. Kielhöfer. Psycholinguistische Grundlagen der Wortschatzarbeit. *Babylonia*, 1996.
[11] J. Knapp, C. Vettori, L. Voltmer, O. Streiter, and D. Zielinski. Image Ass'd L2 Acquisition with Internet. In *Proc. the IADIS-CELDA 2004: Cognition and Exploratory Learning in Digital Age*, Lisbon, Dec. 2004.
[12] J. Kroll and J. Curley. Lexical memory in novice bilinguals: The role of concepts in retrieving second-lang. words. In *Practical aspects of memory*, volume 2, pages 389–395. Wiley, 1988.
[13] J. F. Kroll and N. Tokowicz. The development of conceptual representation for words in a second lang.. In *One Mind, Two Lang.s. Bilingual Lang. Processing*. Blackwell Publishers, Malden–Oxford, 2001.
[14] W. la Heij, R. Kerling, and E. van der Velden. Nonverbal context effects in forward and backward translation: Evidence for concept mediation. *J. Memory and Lang.*, (35):648–665, 1996.
[15] B. Laufer. Electronic dictionaries and incidental vocabulary acquisition: Does technology make a difference? In *Proc. the 9th EURALEX International Congress on Lexicography*, pages 849–854, 2000.
[16] R. Mayer and V. Sims. For whom is a picture worth a thousand words? extensions of a dual coding theory of multimedia learning. *J. Educational Psychology*, 86(3):389–401, 1994.
[17] M. Müller and L. Wertenschlag. Wortschatz-lernen ganzheitlich: effektiv und effizient. *Babylonia*, 2:25–31, 1996.
[18] J. Nerbonne, D. Dokter, and P. Smit. Morphological processing and comp.-ass'd lang. learning. *Comp. Ass'd Lang. Learning*, 11(5), 1998.
[19] A. Paivio and E. Rowe. Noun imagery, frequency, and meaningfulness in verbal discrimination. *J. Experimental Psychology*, 85:264–269, 1970.
[20] M. Potter, K. So, B. von Eckardt, and L. Feldman. Lexical and conceptual representation in beginning and more proficient bilinguals. *J. Verbal Learning and Verbal Behavior*, 23:23–38, 1984.
[21] I. D. Ridder. Are we conditioned to follow links? Highlights in CALL materials and their impacts on the reading process. *Comp. Ass'd Lang. Learning*, 13(2):183–195, Apr. 2000.
[22] L. Rieber. Using comp. animated graphics with science instruction with children. *J. Educational Psychology*, 82(1):135–140, 1990.
[23] E. Rosch, C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–440, 1975.
[24] B. Rüschoff. Authentische Materialien. *Medienbrief*, (2), 2003.
[25] C.-C. Shei. FollowYou! an autom. lang. lesson generation system. *Comp. Ass'd Lang. Learning*, 14(2), 2001.
[26] J. L. Shrum and E. Glisan. *Teacher's Handbook: Contextualized Lang. Instruction*. Heinle & Heinle, 2004.
[27] O. Streiter, J. Knapp, and L. Voltmer. GYMN@ZILLA: a browser-like repository for open learning resources. In *Proc. World Conf. on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA 2003)*, pages 1371–1379, Honolulu, July 23-28 2003.
[28] O. Streiter, J. Knapp, and L. Voltmer. GYMN@ZILLA: Lang. learning with the Internet. In *Proc. T.A.L.C 2004 The sixth Teaching and Lang. Corpora conference*, Granada, July, 6-9 2004.
[29] O. Streiter and M. Stuflesser. XNLRDF, A Framework for the Description of Natural Lang. Resources. – A proposal and first implementation. In *Lesser Used Lang.s & Computational Linguistics*, Bolzano/Bozen, Italy, Oct. 27-28 2005.