A New Approach to CALL Content Authoring

Vom Fachbereich Informatik der Universität Hannover

zur Erlangung des Grades einer

Doktorin der Naturwissenschaften

Dr. rer. nat.

genehmigte Dissertation

von

Dipl.-Ing. JUDITH KNAPP

geboren am 21.03.1972 in Bruneck, Italien

2004

Referent: Prof. Dr. Wolfgang Nejdl

Korreferent: Prof. Dr. Bernardo Wagner

Tag der Promotion: 15. Oktober 2004

Preface

Nowadays science is becoming a working field in which knowledge of several disciplines is required from the researchers. The world cannot be divided into domains within which pure specialists could reach increasingly higher results. Reality shows that forgetting one aspect of a related field may significantly influence the entire outcome. Even a team of several specialists may not be able to cover an entire field, namely when exchanging knowledge between them does not work. Hence, more and more people with interdisciplinary knowledge are required. Such people can contribute to communication and network construction. If in a group of researchers the co-operation does not only consist of distributing tasks but of real openness and true mutual interest, it is much more likely that communication is a fruitful part of the common work, and that all facets of a given problem are discovered and considered within the development process.

This thesis is an interdisciplinary work which covers aspects of different disciplines: The main focus has been to find a way for developing a highly sophisticated electronic language learning system with realistic resources. For this task it was necessary to create a new method which facilitated the implementation and made it possible to join all the ideas. Work on the content side was initially carried out solely by linguists and language teachers. However, the techniques used allowed providing many more features than linguists and language teachers had demanded. Hence in order to be able to make meaningful suggestions about possible improvements of the system, it was necessary for the computer scientists to well understand the main principles of didactics, as well as the problems the linguists tried to resolve in their research.

Consequently, beyond technological issues, this thesis also contains a lot of linguistic, didactic, and sociological elements: the sociological background of the province of South Tyrol, which is home to the main target group of our work, has been examined; linguistic and didactic demands on language teaching have been studied; work on putting these ideas into practice in computer-assisted language learning has been reviewed; and modern technologies such as networking, computational linguistics, and Artificial Intelligence have been explored.

The overall outcome is a new approach to CALL content development. Its implementation and evaluation was realized within the ELDIT language learning project carried out at the European Academy Bozen/Bolzano.

Acknowledgments

First of all, I would like to thank my advisor Prof. Wolfgang Nejdl of the University of Hanover for his support. I am grateful for the time he had for me during his visits to Bolzano and for the invitation to the University of Hanover, where I spent three weeks in summer 2000. During this period, I had some helpful discussions with Nicola Henze and Martin Wolpers.

The closest collaboration was with Dr. Johann Gamper, who, as my local advisor, provided feedback and stimulating discussions on all different aspects of my work. When I started my work at the European Academy of Bolzano, he gave me the opportunity to become acquainted with a field that was very new to me at the time. Through all these years, he always had time to meet with me and to answer my questions.

Furthermore, I would like to thank my team colleagues Andrea Abel, Vanessa Weber, Claudia Richter, Stefania Campogianni, Julia Reichert, Sara Campi and Chiara Vettori for their openness and interest in an interdisciplinary collaboration. We had many vivacious discussions. I particularly want to thank Andrea for her optimism and never ending energy to search for funds, collaborations and possibilities to realize our ideas.

I remember also a lot of stimulating discussions with Oliver Streiter, Paolo Dongilli and Nikolaus Augsten. Their comments were helpful and stimulating for my work. Moreover, they always had time for me when I needed any kind of help.

Additionally, I would like to thank our project partners Pius ten Hacken and Sandro Pedrazzini for their first-rate collaboration. I am grateful to Pius for many interesting discussions we had, not only within the meeting rooms, but also during various walks and dinners in Bolzano and Lugano. I also thank Sandro for his constant availability and for answering my email questions almost always immediately.

Thank you also to Peter Warasin and Lorenzo Donati, who programmed some parts of the system and to the Institute for Computational Linguistics of the University of Stuttgart for POS-tagging our text corpus.

I am also very grateful to Christa Knapp and Albert Mairhofer for proofreading my English text. They did an invaluable service for me.

Last but not least, I would like to thank the European Academy of Bolzano for financing the ELDIT project, for the large amount of freedom we have, for the opportunities to participate in international workshops and conferences and for all kinds of support.

Abstract

The ELDIT project carried out at the European Academy Bozen/Bolzano aims at creating an innovative language learning system for the population of South Tyrol, allowing them to prepare for the exams in bilingualism. The system is furthermore sufficiently general, so that everybody interested in learning the German or Italian languages can use it. The project is carried out by linguists, language teachers, computational linguists, and computer scientists in interdisciplinary research work. The program can be freely used via the Internet (http://www.eurac.edu/eldit).

Efficient knowledge engineering is known to be a difficult task. On the one hand the linguists and language teachers demanded the implementation of very innovative but complex didactic ideas; on the other hand, approaching the realization in the usual way was judged as not feasible by them, since it would have meant the manual input and a detailed encoding of a large amount of data. Hence a new approach to content authoring was elaborated by the computer scientists which made it possible to develop the system with realistic resources.

The main solution we applied was (1) to support the linguists in the detailed data encoding process by electronically rewriting a hand made semi-structured version of the data into the final extensive version needed by the system and (2) to help them in the extensive data collection process by electronically adding as much information as possible to the manually collected educational data.

The approach we applied provided some more advantages: (3) the realization of some more innovative features than the linguists had originally demanded, became possible, (4) the inclusion of externally developed products that worked only on specific language features became feasible which considerably augmented the innovativeness of our system, and (5) the tools developed for data authoring were incorporated into the system and reused during runtime to provide online authoring for teachers.

ELDIT is an ongoing research project that started in autumn 1999. The approach was appreciated and judged as feasible by the linguists in our team. Hence it was applied throughout the development of the system. Its flexibility for the implementation of didactic demands has been confirmed each time we programmed a new module for the program.

Keywords Computer-assisted language learning, Computational Linguistics, Adaptive Hypermedia

Kurzfassung

Das Projekt ELDIT hat zum Ziel, ein innovatives Sprachlernprogramm für die Bevölkerung von Südtirol zur Vorbereitung auf die Zweisprachigkeitsprüfung zu entwickeln. Das System ist jedoch sehr allgemein und kann von jedem sinnvoll verwendet werden, der die deutsche oder italienische Sprache lernen möchte. Das Programm wird in interdisziplinärer Zusammenarbeit zwischen Linguisten, Sprachlehrern, Computerlinguisten und Softwareentwicklern durchgeführt und steht im Internet kostenlos zur Verfügung (http://www.eurac.edu/eldit).

Eine effiziente Wissensverarbeitung war eine schwierig zu lösende Aufgabe in diesem Projekt. Auf der einen Seite verlangten die Linguisten und Sprachlehrer in unserem Team die technische Umsetzung von zwar sehr innovativen, jedoch auch sehr komplexen didaktischen Ideen. Auf der anderen Seite wurde es von ihnen als für nicht machbar gehalten, die Umsetzung dieser Ideen mit traditionellen Methoden und Ansätzen zu realisieren, da dies bedeutet hätte, dass eine sehr große Menge an Daten manuell eingegeben und codiert werden hätte müssen. Aus diesem Grund entwickelten die Softwareentwickler einen neuen Ansatz zur Datenverarbeitung, der es möglich machte, das System mit realistischen Ressourcen zu entwickeln.

Die Hauptlösung war (1) die Linguisten im Kodierungsprozess zu unterstützen, indem eine manuell erarbeitete, teilstrukturierte Version der Daten elektronisch in die vom System benötigte ausführliche Version umgeschrieben worden ist und (2) sie beim Sammeln von Daten zu unterstützen, indem die manuell gesammelten Daten elektronisch mit so viel Information wie möglich angereichert worden sind.

Durch dieser Ansatz ergaben sich einige weitere Vorteile: (3) die Umsetzung neuer innovativer Elemente, welche die Linguisten ursprünglich nicht bedacht hatten, wurde möglich, (4) das Einbinden extern entwickelter Produkte, welche nur auf sehr spezielle Sprachelemente ansetzen, wurde machbar, und (5) die Module, die für den Datenerarbeitungsprozess programmiert worden sind, konnten in das System selber eingebunden und zur Laufzeit wiederverwendet werden.

ELDIT ist ein seit 1999 laufendes Forschungsprojekt. Der oben beschriebene Ansatz zur Datenerarbeitung wurde von den Linguisten in unserem Team geschätzt und als anwendbar bewertet. Seine Flexibilität für die Umsetzung didaktischer Anforderungen ist jedes Mal bestätigt worden, wenn ein neues Modul für ELDIT programmiert worden ist.

Schlagworte Computergestütztes Sprachenlernen, Computerlinguistik, Adaptive Hypermedia

Contents

Chapter 1

Introduction

In this chapter we focus on introductive issues: we first describe the sociological background of the local province of South Tyrol, since it is home to the main target group of our work (section 1.1). Then we describe some problems which should be considered when dealing with computer-assisted language learning (section 1.2). Afterwards we introduce ELDIT, an ongoing research project that aims at developing an innovative language learning system for the German and Italian languages (section 1.3). Next we will summarize the results of our work (section 1.4). Last we will give an overview of the organization of this thesis (section 1.5).

1.1 South Tyrol: a Bilingual Province

1.1.1 Sociological Background

South Tyrol is a bilingual (German and Italian) province located in the North-East of Italy (see Figure 1.1). Before World War One the province was a part of Austria and mainly German was spoken throughout the region. After the War the region became a part of Italy and ethnic Italians started to settle down. Today, although both languages, German and Italian, are official languages, still only few people consider themselves truly bilingual. For many people the German or Italian language is not even considered a second language but a foreign language [139]. Only 15% of the Italian-speaking population and 36% of the German-speaking population consider themselves able to hold a discourse in the second language without problems [58]. Many sociologists and linguists are even abandoning the definition of South Tyrol as a ”bilingual society”, preferring to talk instead of a society characterized by ”double monolingualism” [35].

The main reason for this apparently paradoxical phenomenon is a rather strict geographical separation between the two ethnic groups. For historical reasons the Italian population is living mainly in the two larger cities Bozen/Bolzano and Meran/Merano and in the very south of the province. In smaller towns and mountain villages the population is almost exclusively German-speaking [35, 63, 115, 138].

Not only geographically, but also socially the two groups live side by side and not together [115]. School instruction is in either German or in Italian. Although learning the second language is compulsory in primary and secondary schools, qualified language teachers were rare in the past. Moreover, language is often seen as a right

Figure 1.1: The bilingual region South Tyrol in the North of Italy

rather than as a resource [145]. Thus, many people use the other language only when they need to and not because of an explicit desire to speak the other language. Consequently, many students are not motivated to learn the second language [63, 115]. An additional problem for the Italian speaking population is the strong German dialect spoken in this region [35, 139]. Many Italians see this dialect as an obstacle as far as contact with the second language is concerned [145].

Although it is possible to watch TV and to read newspapers in the second language (even if newspapers with articles in both languages are still very rare), only few people take advantage of these opportunities. More than 75% of the German speaking population listen to news reports only in their own language. This number is even higher for Italian speakers, namely almost 90% [35]. Also cultural events are rarely shared. The German and Italian cultures are felt to be distinct, people are either unaware of, or do not attend, cultural events in the other language [35, 145].

Critics have stressed the fact that this situation is systematically enforced by the local government, which fears that widespread bilingualism could put at risk established rights and claims of both ethnic groups [35, 107, 145]. It is important, however, to notice that many people are not satisfied with the current situation at all. Italian parents for example have formed the association ”Parents for Bilingualism” with the objective of encouraging bilingualism in their children by means of adequate teaching programs [35]. Especially young people welcome opportunities to improve relations with the other group, declare themselves in favor of incentives to promote cooperation among the groups, and consider ethnic plurality a cultural asset [35, 115]. Ethno-linguistic boundaries are, however, still perceived to be fixed and difficult to cross [35].

1.1.2 Exams in Bilingualism

Since 1972 both languages, German and Italian, have equal status in South Tyrol [58]. Citizens are entitled to use their mother tongue in dealings with the public administration including judicial authorities. Therefore, passing the so-called exams in bilingualism is a prerequisite for employment in the public sector [35, 151]. The first version

of the exams consisted of rather strict translation exercises and an oral examination.

In 1998 the exams in bilingualism were reformed. The project order was given by the local *office of exams in bi- and trilingualism*[1] and carried out by the *European Academy of Bolzano*[2] in collaboration with the *Goethe Institute of Milan*[3]. A new, more modern mode of examination was elaborated. The new exams consist of reading two texts (one in German and one in Italian) and of answering some comprehension questions. Additionally an oral test is held to check the conversation and negotiation skills in the second language of a candidate. Within this project 400 texts for each language were elaborated, 200 for each level of difficulty. The texts are articles selected from various magazines and books which have been slightly modified and possibly shortened. They now contain approximately 150 words each. In addition to the new examination mode also learners' dictionaries in paper form were developed. All this material has been published and made available for learning and training [36].

Additionally, private language schools elaborated exercises and courses for preparation[4], and tutors are available for private training. Sometimes also Italian and German native speakers work together and correct each other. This mainly occurs in the towns of Bozen/Bolzano and Meran/Merano, where the Italian population is concentrated. The almost exclusively German speaking rural population has many more difficulties to establish such learning partnerships.

1.2 Computer-assisted Language Learning

CALL (computer-assisted language learning) is a research field which exploits the use of technology for language learning. A survey of CALL systems we carried out has revealed that nowadays language learning systems beyond hypermedia technologies (MCALL) increasingly adopt network technologies (NCALL) and Artificial Intelligence (ICALL) to improve the learning process. While MCALL and NCALL systems seem to be quite mature, many innovative ICALL systems and projects surveyed remained at a prototypical stage: their implementation usually concentrates on the exploitation of single technologies, with regard to content most of them concentrate on one single aspect of language learning, and didactic considerations are often neglected. Moreover, many systems have been implemented for a restricted domain (e.g. language for special purposes) and, if at all, for an easily controllable target group (e.g. the students of a specific lecture). This has been happening because the development of ICALL systems is still at an early stage, wherefore the implementation of ideas usually proves difficult and time-consuming.

Still, what is needed is a more integrated and comprehensive approach which considers didactic requirements and uses different technologies that are tailored for the training of specific skills. Clearly, it is not easy to develop such a system, a lot of expertise from different disciplines, as well as time and money, are needed. However, the development can be facilitated if methods are found to meet the following requirements:

[1] http://www.provinz.bz.it/praesidium/0101/01/index_d.asp
[2] http://www.eurac.edu
[3] http://www.goethe.de/it/mai/deindex.htm
[4] http://www.alphabeta.it, http://www.cedocs.it/versione_it/i_corsi03.php

- Data structure and software modules should be very flexible.
- Data and software modules should be reusable for several purposes and in different parts of the system.
- It should be possible to integrate externally developed products even if they work only on a very specific language feature.

1.3 The ELDIT Project

The main scope of the ELDIT project[5] is to create an innovative language learning system for the population of South Tyrol, allowing them to prepare for the exams in bilingualism. However, the system should also be sufficiently general, so that everybody interested in learning the German or Italian languages can use it.

ELDIT is developed at the *European Academy Bozen/Bolzano*[6] - *EURAC* in collaboration with the *Scuola Universitaria Professionale della Svizzera Italiana*[7] - *SUPSI* and the *Center for Education and Culture*[8] - *CEDOCS*, with financial support of the *European Union*[9] and the Offices of *Bilingualism and Foreign Languages*[10] and of *Education and Training*[11] of the *Autonomous Province of South Tyrol*[12]. An interdisciplinary team is working on the project. Currently at the main institution six linguists and two computer scientists are working on the system, further experts of the partner institutions are supporting the development process.

ELDIT is an ongoing research project. It started in autumn 1999 with a dictionary that gave its name to the entire project (ELDIT, "Elektronisches Lernerwörterbuch Deutsch ITalienisch", or "Dizionario ELettronico per apprendenti Italiano-Tedesco"). Since 2002 the dictionary has been extended to a language learning system that has a strong focus on vocabulary acquisition. A comprehensive overview of the system will be presented in chapter 4 and 5. The current state of the project will be given in section 10.4. Currently the system is freely accessible via the WWW at http://www.eurac.edu/eldit.

1.4 Achievements

The achievements of this thesis consist of the following aspects:

A New Approach to CALL Content Authoring Efficient knowledge engineering is known to be a difficult task. On the one hand the authors of the ELDIT system (linguists and language teachers) demanded the implementation of very innovative but

[5] http://www.eurac.edu/eldit
[6] http://www.eurac.edu/
[7] http://www.supsi.ch/index.html
[8] http://www.cedocs.it/
[9] http://europa.eu.int/comm/regional_policy/
[10] http://www.provincia.bz.it/cultura/bilinguismo/
[11] http://www.provinz.bz.it/kulturabteilung/Weiterbildung/
[12] http://www.provinz.bz.it/

complex didactic ideas; on the other hand, approaching the implementation in the usual way was judged as not feasible, since it would have meant the manual input and a detailed encoding of a large amount of data. Hence a new approach to content authoring had to be elaborated which made it possible to develop the system with realistic resources.

The main solution we applied was (1) to support the authors in the detailed data encoding process by *electronically rewriting* a hand made semi-structured version of the data into the final extensive version needed by the system and (2) using computational linguistics techniques to help the authors in the extensive data collection process by *electronically adding* as much information as possible.

The approach we applied provided some more advantages: (3) the realization of *many more innovative features* than the authors had originally demanded, became possible, (4) the *inclusion of externally developed products* became feasible which considerably augmented the innovativeness of our system, and (5) the tools developed for data authoring were incorporated into the system and *reused during runtime* to provide online authoring for teachers.

An Innovative Language Learning System Based on the findings of an extensive review of CALL systems [68, 70] and applying the previously outlined approach an innovative language learning system has been developed in interdisciplinary research work. The ELDIT system is uniquely rich in information, which is presented in an original way. It allows learning according to the most up-to-date didactic demands and technological possibilities.

A Real World System Not only a pure research prototype but a real world system has been developed. The data that has been elaborated up to now amounts to more than 8,000 word entries, 2,000 word groups, and 800 texts. According to our opinion many of the problems mentioned in section 1.1 can be addressed: thanks to network technologies included, geographical and social separations of the two ethnic groups can be overcome. Moreover, the unwillingness to study the other language might be reduced, since computer assistance has proved to be a successful motivator for language learning [34, 102, 109, 113, 177].

1.5 Organization

Chapter 2 In this chapter we will present the results of an extensive review of CALL systems. More than 60 projects have been analyzed and classified according to different dimensions: system type, supported language, technologies used, language skills taught, supported language elements, and the availability of the system. The results of the review indicate that many approaches are highly technology-driven. The implementation of a system, however, should follow an integrated approach and also include methodology and a curriculum in order to be effective and appreciated by learners and teachers.

Chapter 3 In this chapter we will present lexicographic, psycholinguistic and didactic findings we considered throughout the development process. We will list a detailed

analysis on the difficulties a learner usually encounters when dealing with a foreign language, in particular with regard to vocabulary and dictionary use. Then we will describe existing solutions. Such solutions are often demands that should be met when designing a course or a system. We will conclude by describing how we have implemented and improved the solutions in ELDIT with new media and technologies.

Chapter 4 In this chapter we will describe the basic ELDIT program, an innovative language learning system which focuses on vocabulary acquisition. The main modules are: a learners' dictionary, a pedagogical text corpus, interactive exercises, a tandem feature which enables collaboration between the learners, and an adaptive tutor which guides the learner individually through the learning material.

Chapter 5 ELDIT includes some special features that make it different from other educational systems. Applying computational techniques we reused the manually developed content on several places in the system. We included some externally developed products such as Word Manager and could also integrate WordNet. Also ELDIT itself or single modules can be reused and combined with external applications. Due to a user model included all information can be adapted to the individual learner.

Chapter 6 For the development of the system we applied a special approach to content authoring: To realize the didactic demands the educational data had to be encoded down to the level of single words and further. To help with this task we have developed special tools. Educational data is submitted by teachers or linguists in a semi-structured way, converted electronically into the extensive form, and enriched with additional information as much as possible.

Chapter 7 In this chapter we will first give an overview of the architecture of the system. Then we will present the data model of the ELDIT program. Its implementation has been realized in XML, which, according to our opinion, is an ideal technique to model such complicated data structures.

Chapter 8 The efficiency of our approach to data management and CALL system development has been confirmed for each new module we implemented for ELDIT. Moreover, we conducted an extensive evaluation of system use by examining several data sources: the server log files for daily access rates, the user models included for individual use of the system, and a questionnaire consisting of 38 questions. The results were very encouraging for our work, some interesting implications for our future plans emerged, too.

Chapter 9 The development of the ELDIT system has been inspired by a lot of other work. In this chapter we will refer to related work and try to emphasize what distinguishes our work from other work and how we have tried to improve the outcomes. We will furthermore discuss some possible generalizations, both of the system itself, as well as of the approach to content authoring.

Chapter 10 In this chapter we will summarize the innovative features ELDIT provides and the advantages of our methods of data management. Overall the approach has proved its effectiveness and flexibility in many different cases. The result of our work is a fully-fledged language learning system which is sophisticated and easily extendable.

Chapter 2

A Review of CALL Systems

The first step we carried out was an extensive examination of the state of the art in computer-assisted language learning (CALL) [68, 70]. In section 2.1 we describe the CALL research field by paying particular attention on ICALL. In section 2.2 we introduce a framework for the classification of CALL projects and give an overview of our analysis in table form. In the following sections (sections 2.3 to 2.7) we present the analysis of more than 60 advanced projects in detail. The projects have been investigated and classified with regard to several dimensions: system type, supported languages, techniques and technologies explored, the language skills trained, the language elements taught, and the availability of the system. Some outstanding systems will be described in more detail. In section 2.8 we will discuss some open research issues and remaining problems.

2.1 The Research Field

In this section we focus on the research field that deals with language learning and technology (see Figure 2.1). We start with the most general field, namely computer-assisted instruction (CAI) and summarize some common research paradigms. Then we go on with computer-assisted language learning (CALL) by referring to the main organizations and literature that deal with this field. Finally, we focus on intelligent computer-assisted language learning (ICALL) by sketching a brief history of the inclusion of Artificial Intelligence in language learning systems.

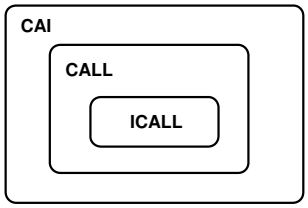


Figure 2.1: The research field

2.1.1 Computer-assisted Instruction

The Foldoc Computing Dictionary[1] defines *computer-assisted instruction (CAI)* as "the research field that deals with the use of (personal) computers for education and training". Considering newer subfields such as mobile learning, which is learning through hand-held devices such as PDAs and digital cell phones, we should better generalize the definition to "the research field that deals with the use of electronic devices for education and training".

Since the beginning of the information technology age computers have seemed appealing for learning. Learning "what, where and when" were the keywords that made learners, teachers and researchers dream of a simple, effective, life-long learning experience. But where has the learning society got to? What are the achievements reached? Some of the reviews we will list now have been designed to give an insight into this topic.

In [130] an introductory overview of several *common issues* nowadays closely related to computer-assisted learning (CAI) is given, namely to learning theories related to technology, to CAI software classifications according to subject matters and educational paradigms, and to different software types. In [157] the *effectiveness* of online learning has been reviewed in the relevant literature. The author concludes that we now have good evidence that students generally learn at least as much online as they do in traditional classroom environments. However, unique characteristics of the medium may promote or constrain particular kinds of learning, which in turn suggests certain approaches that might enhance the learning effectiveness. In [29] an analysis of *course delivery tools* and technologies for building a virtual university is given. Various aspects of currently existing tools are analyzed, and a kind of grading is introduced by which the tools and systems are subdivided into three levels of advancement. It is concluded that very few tools exist that support an author in more than simply uploading content. In [166] it has been reviewed how *mobile learning* could help reach the goals of better learning. The nature of mobile devices can make them very efficient, namely when the design of the content succeeds in supporting new learning situations, for instance small fragments of waiting or idle time. Design guidelines for the achievement of this task are listed.

In a modern CAI environment the roles of the student and the teacher usually change. Learning becomes constructivist, the student acts as a researcher, the teacher serves as a facilitator. System features more and more take care of the learning process. In [167] a historic review of the development of *Artificial Intelligence and Educational Psychology* is discussed. It is concluded that an interdisciplinary collaboration between the two fields might be a fruitful way of teaching and helpful for the evolution of both. In [117] a general overview of the use and role of *Intelligent Tutoring Systems* is given. The key message in this paper is that new technologies usually initialize new ways of learning and new curricula, wherefore a good preparation of learners and teachers is needed to make the new approaches efficient. Finally in [27] *adaptivity and adaptability* is proposed to increase the efficiency of educational hypermedia systems.

[1] http://wombat.doc.ic.ac.uk/foldoc/

2.1.2 Computer-assisted Language Learning

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)*. Other, more general expressions such as *software-assisted language learning (SALL)* have sometimes been proposed, but CALL was the expression agreed upon at the TESOL convention in Toronto in 1983 [41].

There are several relevant organizations, conferences, journals, and books which have dedicated themselves to language learning and technology.

The *European Association for Computer-assisted Language Learning (EUROCALL)*[2] wants to "provide a European focus for all aspects of the use of technology for language learning" and organizes the informative annual *EUROCALL conference*. Its *ReCALL journal* is issued twice a year in May and November, back issues are freely accessible online[3].

The *Computer-assisted Language Instruction Consortium (CALICO)*[4] is an American organization which "has an emphasis on modern language teaching and learning, but reaches out to all areas that employ the languages of the world to instruct and to learn". It publishes the *CALICO journal*. The annual *CALICO symposium* "features uses of cutting edge technologies in foreign language teaching and learning".

CALL has had a long tradition in Japan and other Asian countries. The *Japan Association for Language Teaching (JALT)*[5] is a nonprofit organization dedicated to "the improvement of language teaching and learning both within Japan and internationally". The annual *JALT CALL conference* as well as the two journals *TLT* and *JALT* provide useful information about language teaching and research, particularly in an Asian context.

Some more journals should be pointed out: the *CALL journal*[6], an international journal published by Swets & Zeitlinger, claims to "lead the field in its dedication to all matters associated with the use of computers in language learning (L1 and L2), teaching and testing". The freely accessible online journal *Language Learning and Technology (LL&T)*[7] seeks "to disseminate research to foreign and second language educators in issues related to technology and language education". The also freely accessible online journal *CALL-EJ-ONLINE*[8] has resulted from a merge of two formerly independent journals. Its editors are mainly from the Australian and Japanese region. The *International Journal of Artificial Intelligence in Education (IJAIED)*[9] is not particularly focused on language learning, but generally "publishes papers and other items concerned with the application of Artificial Intelligence techniques and concepts to the design of systems to support learning"; however, from time to time some interesting ICALL papers can be found there.

From 1992 on several books on the topic of CALL have been published. Some important ones are: *Intelligent Tutoring Systems for Foreign Language Learning* [158],

[2] http://www.eurocall-languages.org/about/ecabout.htm
[3] http://www.hull.ac.uk/cti/eurocall/recall/r_online.htm
[4] http://calico.org/
[5] http://jalt.org/
[6] http://www.szp.swets.nl/szp/journals/ca.htm
[7] http://llt.msu.edu/
[8] http://www.clec.ritsumei.ac.jp/english/callejonline/callej.html
[9] http://cbl.leeds.ac.uk/ijaied/

Intelligent Language Tutors - Theory Shaping Technology [92], *Computer-assisted Language Learning - Context and Conceptualization* [111], *CALL - Media, Design and Applications* [32], and others.

2.1.3 Intelligent Computer-assisted Language Learning

ICALL started a distinct research field about a decade ago, when some Artificial Intelligence (AI) technologies were mature enough to be included in language learning systems, at least in experimental settings.

At the CATH90 (Computers and Teaching in the Humanities) conference in 1990 Chris Bowerman published an overview of the state of the art in CALL [20]. He also included a short review of the ICALL systems then in use. He mentioned six systems which were using syntactic or semantic knowledge to check a foreign language text. Only one of these systems also maintains a student model. In those days all of the systems in question were research prototypes and never in everyday use.

The first seminal publication is the book *Intelligent Tutoring Systems for Foreign Language Learning*, edited by M. L. Swartz and M. Yazdani in 1992 [158]. This book forms a systematic basis for this multidisciplinary research discipline and tries to combine ITS and NLP to build an instructional framework for language learning. Methods such as grammar checking, error analysis, user modeling, and tutoring are discussed, and the question, how they can be adapted and combined to be useful for language learning systems, is dealt with.

Two years later, in 1994, a special issue of the *International Journal of Artificial Intelligence in Education* was dedicated to the topic of language learning [37]. Besides the technological aspects, researchers began to include results from the educational and cognitive sciences. They also started considering pragmatics and socio-linguistic competences. System development was still at a prototypical stage. At that time, none of the systems presented in the issue was used in real learning situations.

In 1995 the book *Intelligent Language Tutors – Theory Shaping Technology* was published [92]. While the main focus was still on ITS augmented with NLP, many systems were presented which provided a more comprehensive language teaching environment including negotiations and discourse between the learner and the system. NLP technologies were developed to train fluency. Systems became available on the personal computer, and huge computational lexicons as well as large linguistic databases for different languages have been added.

In the following years, research in Automated Speech Recognition (ASR) matured, and the technology became powerful enough to be incorporated into ICALL systems to support the training of pronunciation and communication skills. In 1999 the *CALICO journal* devoted a special issue to speech recognition techniques for language learning [90]. Most of the systems presented implemented discrete speech recognition. This technology is more reliable than continuous speech recognition, but its use is limited, since the user has to reproduce one of several predefined patterns. Continuous speech recognition allows processing freely produced sentences. However, only a few systems have implemented this technology.

The book *CALL – Media, Design & Applications*, published in 1999, devotes some chapters to intelligent technologies used in CALL [32]. The use of ASR and NLP were, once again, the main topics. At that time, both technologies were successfully imple-

mented in some fully working systems, but they had to be supported by alternative input modalities or limitations of the domain of discourse. The inclusion of semantics and pragmatics in the recognition process was still an open problem.

In 2002 the *CALL journal* devoted a special issue to ICALL [162]. Again ASR and NLP projects were presented, but also some new approaches, such as the use of machine translation for learning how to translate were discussed.

2.2 Classification Framework

In this section we will introduce a classification framework which allows an analysis of CALL projects with regard to the following dimensions: project type, languages supported, technologies applied, language skills trained, language elements that can be learned, and the availability of the systems. Table 2.1 summarizes the results.

With respect to learners' autonomy or collaboration, we distinguish several *types of language learning projects*: ICALL projects (I), MCALL projects (M), NCALL projects (N), and the use of additional aids (D). The category of additional aids covers resources and tools which are used as a complement in a language learning environment. The category of MCALL refers to the use of more traditional software which explores hypertext and multimedia features for language learning. In NCALL projects network technologies are used for collaboration. In ICALL projects Artificial Intelligence techniques are explored to improve interaction.

The next dimension concerns the *target languages* which can be learned through the system: English (e), Japanese (j), French (f), German (g), Spanish (s), Italian (i), Dutch (d), Russian (r), Greek (k), Mandarin Chinese (c), Arabic (a), Hebrew (h), Thai (t), Quechua (m), an Indigenous language in the Andean region spoken by approximately 13 million people in Bolivia, and Warlpiri (w), an Australian Aboriginal language spoken in the Tanami desert in the North West of Alice. An asterisk (*) means that the system has been developed for a larger group of languages or is language-independent.

In the third dimension we have a look at the *tools, techniques, and technologies* that are used in the single projects: dictionaries (dict) and Internet resources (inter) are used as stand-alone products to support the classroom activities. In older MCALL systems only hypertext (hyper), in newer ones also multimedia (mult) and text corpora together with concordance tools (cor) are used to organize and illustrate information in an innovative way. In NCALL projects, e-mail (e-mail), discussion forums (forum), MUD and MOO like chat tools (chat), real time audio (audio) and video conferencing systems (video) are explored. The core part of our analysis is focused on ICALL and concerns the use of AI systems and techniques in computer-assisted language learning. Some approaches follow the line of expert systems (EX) and store domain knowledge. This approach allows detailed feedback for the learner. Intelligent tutoring systems (ITS) guide the user through the learning space individually. Many of these systems store information about each individual learner in a so-called user model. In some cases the user model is aimed at being inspected by the learner or the teacher (UM). Other systems use this information to adapt to the needs of the individual user (A). Natural language processing (NLP) and natural language generation (NLG) are two other technologies which are mainly applied in systems for training

writing skills. Systems for training speaking skills often include an automated speech recognition (ASR) component which allows control over the learner's pronunciation. Finally, machine translation (MT) is used in some projects to improve communication and translation skills.

The fourth dimension classifies CALL projects according to different *language skills* which can be trained. A common, high-level distinction is between reading (R), writing (W), listening (L), and speaking (S) skills. In our analysis we will consider translating (T) and interpreting (I) as two additional high-level language skills. For the latter, however, no corresponding project has been found. Speaking skills are further subdivided into pronunciation (p), fluency (f), and social skills (s).

In the fifth dimension we classify projects according to different *language elements* which can be trained. The analysis distinguishes between grammar (G), vocabulary (V), and dialogue elements (D). Grammar may be approached in two different ways: deductively (d), where students are given a rule which they practice, or inductively (i), where students infer the rules by themselves. Vocabulary acquisition can occur intentionally (t), for example by studying a list of words, or incidentally (c), for example by reading.

Finally, we will have a look at the *availability of the systems* and indicate whether a system is Web-based (w), whether a demo version is available (a), or whether the product can be purchased (p).

CLASSIFICATION OF PROJECTS

Name	Ty.	Lang.	Technology	Language skills					Lang. elem.			Av.
				R	W	L	S	T	G	V	D	
Cambr.[83]	D	e	dict	*	*				i	c		w
Cobuild[150]	D	e	dict	*	*				i	c		p
Longman[114]	D	e	dict	*	*				i	c		w
Oxford[94]	D	e	dict	*	*				i	c		w
Langen.[168]	D	d	dict	*	*				i	c		p
VcuTrail[74]	D	*	inter,forum	*	*				i	c	*	w
Alexia[38]	M	f	mult	*	*					t		
Cross Talk[112]	M	e	mult				s				*	
Harben[81]	M	e	mult			*					*	
Cresh.[31]	M	i	mult			*	s				*	p
Ucuchi[46]	M	q	mult			*	s				*	p
CyberBuch[134]	M	g	mult	*					i	c	*	p
Key4phil[15]	M	r	hyper					*	i	c		
Cavoca[78]	M	d	mult	*	*					t		a
TransIt.[141]	M	*	hyper					*	i	c		p
Gertie[98]	M	*	mult	*	*					t		
Lexica[76, 106]	N	e	cor,forum	*	*			s	i	c		w
Web learn[34]	N	e	e-mail,chat	*	*		s		i	c	*	w
MOO 1[101]	N	j	chat	*	*		s		i	c	*	w
MOO 2[170]	N	j	chat,video	*	*		s		i	c	*	w
email 1[109]	N	g,s	e-mail	*	*				i	c		w
email 2[177]	N	e,s	e-mail,forum	*	*		s		i	c	*	w
email 3[113]	N	e,g	e-mail	*	*				i	c	*	w
email 4[102]	N	g,f	e-mail,audio	*	*	*	f		i	c	*	w
email 5[25]	N	*	e-mail	*	*	*	f		i	c	*	w
Leverage[57]	N	e,s,f	video	*	*	*	s		i	c	*	w
KirrKirr[97]	I	w	UM,mult	*	*					c		w

Name	Ty.	Lang.	Technology	Language skills					Lang. elem.			Av.
				R	W	L	S	T	G	V	D	
Pet2000[43]	I	e	UM,cor	*	*					c		
W.P.Voice [172]	I	e	NLP		*				d			w
RECALL [104]	I	e	ITS,NLP		*				d		*	
MTSystran [165]	I	e	MT					*	i	c		w
ISLE [118]	I	e	A,ASR				p				*	a
FollowYou! [148]	I	e	A	*						t		
ArtCheck [146]	I	e	NLP		*				d			
L2tutor [136]	I	e	UM,NLP	*	*		f				*	
Inst.Dict. [123]	I	e	ASR,MT,				p			c		
LISTEN [119]	I	e	ASR	*							*	
CAPIT [116]	I	e	ITS		*				d			
Spengels [19]	I	e	EX,UM		*				d			
ILTS [163]	I	e	A,NLP,MT		*			*	d	c		w
MT.system [149]	I	e	NLP,MT					*	i	c		
Fluent-2 [79]	I	e	ITS,NLP,NLG	*	*						*	
J.T.Texts [178]	I	j	NLP		*				i			w
CoCoaJ [129]	I	j	NLP		*				i			
Subarashii [16]	I	j	ASR			*	f,s				*	
Nih.Cali [124]	I	j	NLP		*				d			
CompLex [85]	I	j	EX,UM	*						t		
FreeText [169]	I	f	NLP		*				d			
GLOSSER [126]	I	f	NLP	*					i	c		w
PIL*É*FACE [110]	I	f	NLP		*		s				*	
McGill [65]	I	f	UM,NLP	*					i	c		
Hr. Komm. [52]	I	g	UM,NLP,NLG	*	*		f				*	
Lice [21]	I	g	ITS		*				i			a
Span.Verb [152]	I	s	EX,NLG		*				d			
Voc.Tutor [45]	I	c	A	*						t		
Conversim [82]	I	a	ASR	*		*	p,f,s				*	p
TLS/CATL [49]	I	t	UM,NLP		*				d			w
Targumatic [9]	I	h	MT		*			*	i	c		
Tait [173]	I	g,s	A,ASR				p				*	
LingWorlds [55]	I	j,e	ITS,NLG			*					*	
Glossaries [44]	I	i,e	NLP	*					i	c		
Pronto [48]	I	s,e	A,ASR			*	p				*	
German Tutor [87]	I	g,e,k	ITS,NLP		*				d			w
WordMang. [161]	I	g,i,e	NLP	*					i	c		w
Fluency [62]	I	f,e,k	ASR				p				*	
MILT [91]	I	e,s,a	A,NLP,ASR	*	*		p,f				*	
Athena [121]	I	*	NLP,NLG,ASR	*	*		p				*	p

Table 2.1: Classification of CALL projects.

2.3 Project Types

2.3.1 Use of Aids

Aids are defined as all resources which provide useful complements in a language learning environment. Mostly dictionaries (dict) and Internet trails (inter) are used for such tasks.

Learners' Dictionaries Special learners' dictionaries have been developed with the objective of both, supporting the language learner to decode what cannot be understood and of helping in text production. Many paper-based learners' dictionaries offer an online version of their data [83, 168, 114, 94, 150]. They are characterized by a carefully selected vocabulary, illustrations and lexicographic examples. Other dictionaries are not intentionally designed as learners' dictionaries, but they nevertheless apply psycholinguistic and didactic theories to support the construction of mental models. Both KirrKirr [97] and Alexia [38] show a graphical representation of the so-called mental map and thus provide content-based access to semantically related words. The words are organized in networks. Nodes represent word entries, and edges represent semantic relationships between the words. The nodes are hyperlinked to the corresponding dictionary entries. Learners' dictionaries are available as stand-alone applications [81, 168, 114, 94, 150] as well as an integrated part of a larger CALL system [38, 97]. They are available on CD-ROM [168, 127, 150] and some of them online as well [83, 97, 114, 94].

Internet Trails Internet trails are used by teachers to collect authentic language material and to provide their students with resources for reading and practicing the target language. As an example, we would like to mention the "VCU Trail Guide to International Sites and Language Resources" [74] which links Internet resources like clickable maps, country and regional information, as well as information on discussion groups for English and for other languages.

2.3.2 Multimedia Systems

The possibilities of the electronic medium, namely working with hypertext and hyperlinks (hyper), exploring large text corpora with concordance tools (cor), and providing multimedia elements such as sound files, digital images, or video films (mult), has attracted students, teachers and researchers for decades. All in all the meaningful combination of hypertext, hyperlinks and different multimedia elements makes for an effective and interesting learning environment.

Hypertext and Hyperlinks Systems which mainly exploit hypertext and hyperlink possibilities to organize the information, and to provide rapid information access, are described in [15, 141]. On the computer mere linear presentation can be avoided. It is possible to organize information in small modules, and to hyperlink these modules to each other in a networking way. Electronic search possibilities and interactive quizzes further distinguish this medium from the traditional paper-based approach. Explorative learning and personalization of the information are facilitated. The KEY4PHIL system [15], for example, is intended to provide maximal flexibility in choosing translation exercises. Texts to be translated are organized in a strongly hierarchical structure and can be accessed by applying different selection filters. A list of keywords indicates the linguistic features the text is designed to train, and a number indicates the level of difficulty. The user can also choose between five different learning modes: learning with or without help, training with or without the repeating of errors, and the examination mode.

Multimedia In many other systems all kinds of multimedia are additionally used to illustrate the learning material and to provide an interesting and effective learning environment [40, 78, 81, 98, 134]. Pictures are used to illustrate concepts or complex contexts, sound files make it possible to listen to, and practice, the pronunciation of words. Using video films to show traditions of a country or typical communicational situations of the target language, is a popular method to convey cultural knowledge and to teach social skills [31, 46, 112].

Corpora and Concordance Tools Concordance tools designed to rapidly and conveniently access the rich information that is provided by large text corpora are sometimes used by teachers and students. In the projects described in [43, 106] the users were reading texts and learning new vocabulary by exploring collocations of an unknown word with the support of a concordance tool. In [43] a user model was included so that students could create their personal dictionary of learned words. In [106] students furthermore used a discussion forum to reflect upon the language, the expressions they searched, and the solutions they came up with.

2.3.3 Networking Systems

When network technologies arose, teachers soon had the idea to provide their students with the possibility of discussing online with their teachers, each others, peers from other countries and native speakers. Asynchronous technologies such as discussion forums (forum) and e-mail (e-mail) as well as synchronous technologies such as chat tools (chat), telephone conferencing (audio) and video conferencing (video), were used for this task.

Discussion forums Discussion forums are used to provide possibilities to pose questions also outside the classroom. More often, however, they are used to motivate students to train their communication skills, to encourage them to discuss their opinions and to be faced with culturally diverse thinking [74, 106, 177]. In [106] the target language itself was used as topic of reflective conversations. Computer-mediated asynchronous discussion around language topics and language learning issues were examined, and their value for learning was analyzed.

E-mail E-mailing with native speakers has proved to be a very motivating language learning method. In all projects examined [25, 34, 102, 109, 113, 177], evaluations showed that both learners and teachers were in favor of using this tool. Usually e-mail is used in a tandem way, i.e. partners were on an equal level of language competence and supported one another in acquiring the others' native language. During the last ten years, entire networks of tandem communities have arisen: since 1994 universities and other educational institutions from many different countries have worked together in the "International Tandem Network" to help their students learn languages via tandem, primarily using the Internet [25]. Since students have to correct their peers, they start reflecting and asking questions about vocabulary and grammar rules, i.e. their critical thinking abilities improve [177]. In [109] an experiment was described in which one group of US students communicated in Spanish via e-mail with peers in Mexico, and

another control group read and summarized newspaper articles. Even if the results of the study revealed no significant difference in language performance and the students' confidence, both, students and teachers, were in favor of using e-mail in language learning.

Chat Tools Chat tools allow for communication in real time, and are hence often used in CALL systems. Two experiments on using chat in language learning are reported in [101] and [170]. In both projects advanced students of the Japanese language had to carry out a task and to plan and discuss it with native Japanese speakers using MUD/MOO-like chat environments. The students were very interested and motivated, and the development of language skills as well as skills in decision making, planning, or negotiating was evident.

Audio Conferencing In addition to e-mail, real time audio was used to train conversational fluency in the project described in [102]. However, due to a certain shyness of the participants and, possibly, the lack of eye contact, some problems were encountered. In the telephone conversations considerable pauses between students' utterances were observed, because "no one knew who should speak". During the entire study the subjects remained reserved and did not try to compensate for their insecurity. On the other hand, also positive outcomes were reported: students became more aware of the pitfalls of the foreign language and of the errors they made.

Video Conferencing The projects in [57, 170] explore video conferencing for language learning. In [57] students of the English, French, and Spanish languages met over a computer network and studied each others' native language. The students had to carry out different tasks. Their work was evaluated by applying criteria like "arguments employed" or "specialized language correctly acquired". Evaluations showed that the students were very motivated, group effort was evident in 11 of 14 groups.

2.3.4 Artificial Intelligence Systems

Over the last decade, it has become increasingly common to adopt AI techniques in CALL systems. While the first systems mainly focused on expert systems and NLP techniques, nowadays also ASR and MT, or a combination of different techniques, are often used.

Expert Systems Expert systems (EX) store a large body of knowledge about language learning such as typical mistakes, learning strategies, questions and answers, etc. [19, 85, 152]. This knowledge is then used for the analysis of the students' interaction with the system. It allows providing a more detailed feedback than in traditional systems. For example, the Spengels system [19] is equipped with knowledge of the spelling of verb forms. The student can either learn spelling and conjugation rules from scratch or practice them by completing gapped sentences. Error-specific feedback can be given at each stage of the learning process.

Intelligent Tutoring Systems Intelligent Tutoring Systems (ITS) usually consist of the following core modules [158]: (1) an expert module which stores the domain knowledge, (2) a tutor module which represents the tutoring strategies and learning goals, (3) a learner module which describes the learners' knowledge about the domain and allows the tutor module to plan the interaction between the student and the system, and (4) a graphical user interface. There are several classical ITSs which have been developed for language learning [21, 55, 79, 87, 104, 116]. As an example we mention the RECALL system [104] in which an electronic tutor guides the student through communicative role-play scenarios. If the student makes a grammatical error, the tutor provides user-specific feedback on different levels. Grammatical rules can be studied systematically and in an individual way.

User Modeling and Adaptivity Many language-learning systems include a so-called inspectable or viewable user model (UM) which records the user's steps and mistakes. The mistakes are classified by the system and can be examined by the student and the teacher after the learning session [19, 43, 49, 52, 65, 85, 97, 136]. The system in [136] additionally uses NLP techniques to analyze writing errors in more detail.

A more advanced approach is the combination of user modeling and adaptivity. Traditional teachware is designed for a prototypical learner. This approach is not very useful in many learning situations. New media in combination with adaptation technologies (A) allow the development of teachware which - based on the information in the user model - adapts content and presentation to the individual learner. For example, remedial exercises are offered in [118, 163, 173], and, depending on the error frequency, adaptive sequencing is used in [45, 48, 91, 148]. The FollowYou! system [148] automatically generates a language lesson adapted to the specific needs of the learner. Since language input should be comprehensible and a little further beyond the learner's current level of competence, the system checks the user model to determine the information which should be included in the generated lesson.

Natural Language Processing and Natural Language Generation Natural language processing (NLP) and natural language generation (NLG) are among the earliest AI techniques which have been explored in ICALL. Although these techniques have been studied for more than 20 years, we have only recently well working tools [80]. NLP and NLG are very promising for language learning. Grammar checkers in combination with a lexicon can be used to check written input by the user for spelling errors and grammatical correctness. Some grammar checkers apply constraint-relaxation techniques to deal better with erroneous user input [52, 169]. Another method is to restrict the possible answers to a limited number of patterns and to apply simple pattern matching techniques. This approach considerably improves speed and reliability of the system [136]. Finally, some systems explicitly model an error grammar to provide meaningful feedback [104, 149, 172].

A fully-fledged analysis of a written text in all its complexity is a very difficult task and exceeds the current state-of-the-art technology in NLP. In general, four different levels of complexity can be distinguished. At the first level only morphology is considered, which includes errors on the word level like gender, number, and conjugation [44, 161, 126, 129, 152]. The second level is that of syntax and includes errors

on the sentence level like noun, verb, and prepositional agreement [87, 104, 136, 146, 163, 169, 172]. The third level includes semantics: it requires a large body of domain knowledge for a system to "understand" a conversation [49, 52, 65, 79, 91, 124, 121, 136, 178]. The fourth level is that of pragmatics. Only very few projects deal with the delicate relationship between speaker and listener. In the project described in [110] a large body of knowledge related to greeting situations was collected to deal with pragmatic aspects of the language.

The NLP/NLG system developed in the Athena language learning project [121] is a very sophisticated system which includes morphological, syntactic, semantic, and pragmatic components. Several applications including the Fluent-2 program [79] incorporate this NLP system.

Automated Speech Recognition Recently automated speech recognition (ASR) technologies have become mature enough to deliver reliable results. Since only a limited number of answers has to be expected from the learner, speech recognition in a CALL context is quite fast and reliable. The major drawback is the big difference between native and non-native accents which makes it difficult for the software to provide a fine grained analysis.

Two ASR technologies have to be distinguished: discrete and continuous speech recognition. Discrete speech recognition allows for an analysis of single patterns which are known to the system. This technique is often used to train pronunciation [48, 62, 121, 123]. It is also used for the training of fluency, where the user can choose from a predefined set of patterns [82, 91]. Continuous speech recognition aims at analyzing free and fluently spoken input [16, 173]. While accurate recognition of spontaneous speech is still beyond the state of the art [7], it works to some extent if the system can expect a certain input [118, 119].

The Conversim ^{TM} system [82], for example, allows users to have extensive "face to face" dialogues in real time with virtual characters. The system continuously prompts three relevant questions which the user can ask. A video instructor helps to pronounce the questions and phrases correctly. The user must rely on language skills, experience, and intuition to judge whether the virtual character says the truth or not.

Machine Translation Machine translation (MT) technologies entered into the field of CALL only a few years ago. Translations provided by a system are used as a preliminary version of a text in the target language on which the student should continue to work [9, 165]. Instead of using other MT systems, some researchers have developed ICALL systems with MT-like capabilities which are easier to integrate into the learning process [149, 163]. The ILTS system [163] shows to the learner texts with typical key English patterns which have to be translated. The system compares these translations with correct and erroneous model sentences and provides error comments and adaptive remediation.

A rather unusual approach was used in a study which aimed at investigating whether a hand-held translation machine used in daily life could be helpful for language learning [123]. Both incidental and intentional vocabulary acquisition were observed, as well as the improvement of pronunciation due to the speech processing facilities included.

2.4 Supported Languages

The vast majority of language learning programs have been developed for English, followed by Japanese, French, and German (see Table 2.1). Almost none have been developed for Italian. One of the reasons is that for English CALL already has quite a long tradition and thus a lot of resources are already existing and freely available (corpora, NLP-tools, etc.).

As the mistakes in foreign language learning very often depend on the mother tongue of the student, many systems in fact have been developed for a language pair in the sense that a specific source language has been considered. This allows, among other things, the use of error grammars which model typical mistakes of students with that specific native language. Moreover, the menu and explanations can be shown in the source language.

2.5 Supported Language Skills

This dimension classifies ICALL systems according to different language skills which can be trained. We distinguish between reading (R), writing (W), listening (L), speaking (S), and translating (T) skills.

Reading Some systems combine reading with vocabulary or grammar acquisition [45, 65, 85, 165, 148]. In other systems the training of reading skills is embedded in conversations [52, 79, 82, 91, 121, 136]. In many NCALL projects reading and correctly understanding e-mail or chat messages is an integrated part of the project [34, 101, 102, 109, 113, 170, 177].

Highlighting unknown words and adding electronic glosses or translations which appear at a mouse click, have been explored in a number of applications [44, 161, 126, 134]. Experiments reported in [137] showed that students appreciated the speed of the electronic medium and the easy access to the information, and hence checked unknown words more often. Moreover, incidental vocabulary acquisition occurred, sometimes even in a more efficient way than without electronic aids [108]. The CyberBuch [134] aims at training reading in the German language. Assistance is provided in both top-down processing (the learner uses global knowledge about the domain) and bottom-up processing (the reader first decodes the meaning of single words and then works out the meaning of sentences and the whole text). Different cognitive styles are supported by allowing the students to choose the type of explanations that best fit their needs.

The system LISTEN described in [119] uses continuous speech recognition to listen to children while they are reading aloud in their native language. The effects of ASR errors are minimized by never saying that the student is right or wrong. Instead, the system responds to a possibly incorrect word by communicating the correct word or by indicating ”mmm?”. Effort is praised instead of performance.

Writing In MCALL reading and writing skills are usually trained by teaching grammar and vocabulary rules. Training writing skills up to the conversational level is the main focus of most NCALL projects. Asynchronous technologies such as e-mail and discussion forums are used [25, 34, 102, 109, 113, 177] as well as synchronous chat

and MUD/MOO like environments [34, 101, 170]. Correction is provided either by a teacher or by peers who are native speakers of the target language.

Most of the analyzed ICALL systems allow for the training of writing skills by NLP techniques. Many systems are limited to single linguistic aspects such as spelling, punctuation, or writing sentences in passive form [19, 116, 146, 152, 172]. Other systems are more general and include a sophisticated parser. The user can freely type in sentences and gets detailed feedback, e.g. about grammatical aspects or collocations [9, 87, 104, 129, 149, 163]. A third class of systems even concentrate on communication skills and the content of the learner's answers [49, 110, 121, 124, 169], in some cases embedded in stimulating conversations [52, 79, 91, 136]. More sophisticated systems provide support for the overall text-composition process including the structuring of texts according to different schemes [21, 178]. The LICE system [21] is an example which provides support at all stages of the writing process, from the initial elaboration of ideas to the generation and revision of the text. Diagnosed errors are recorded in a student model, which enables the tutor module to determine the appropriate reaction.

Listening Understanding spoken language is a crucial aspect of successfully using a second language. Most MCALL systems provide listening tasks implicitly with static sound files or video clips [31, 81, 46]. The only MCALL system in our review with an explicit focus on listening skills is described in [81]. The learner can listen to a piece of text and has to answer comprehension questions such as "What has the character said?". The answer has to be given in written form by exactly repeating a heard sentence. The system design follows psycholinguistic theories about top-down and bottom-up processing as well as learning theories concerning learner independence and autonomy.

In NCALL systems, conferencing tools and direct conversations with native speakers are used to bring students into contact with spoken language and foreign accents [57, 102].

In ICALL system one common way to provide language comprehension is to use automated speech recognition techniques [48]. Another approach has been implemented in the system LingWorlds [55]. A micro-world is presented and the user is asked to move objects in this world (e.g. provisioning a lifeboat before an ocean liner sinks). The reactions of the user are analyzed by the system. If the user misunderstands a command, the relevant part of the sentence is repeated. In [16, 82] understanding pragmatic aspects has additionally been considered. This requires cultural knowledge and is very crucial for a correct language comprehension.

Speaking Many teachers and researchers consider the ability to speak and communicate in a foreign language one of the main tasks in language learning. In the analysis of speaking skills we distinguish between pronunciation, fluency, and social abilities.

In MCALL systems speaking abilities are generally trained in a passive way, i.e. by learning rules and by being faced with prefabricated situations of communication. A very interesting project, however, is described in [112]. CrossTalk is a multimedia program on CD-ROM which contains short filmed dialogs. The aim of the program is to train socio-linguistic skills of advanced learners by making them aware of cultural

differences between Australia and Asia. During the film sequences the student's attention is drawn to cultural characteristics (opening and closing sequences of a discourse, who is leading a conversation, impolite questions, etc.) by more or less intense lights depending on how strong the respective characteristics occur. In this way the student becomes aware of cultural differences and can discuss and device possible alternative approaches.

In NCALL projects teachers have made use of the possibilities to electronically collaborate with each other: developing the ability to express oneself, negotiating meanings, and mutually culturally understanding has been the aim of many e-mail tandem projects [34, 109, 113, 177]. Furthermore, fluency was meant to be trained by using chat tools [101, 170]. Audio and video conferencing tools were used to further train pronunciation and communication abilities [57, 102, 170].

Most of the ICALL systems for speaking concentrate on training pronunciation with ASR. The systems differ in the way they give feedback to the learner: transforming the utterances into visual cues [62, 82, 121], commenting orally on the user's performance [62, 173], highlighting the actual position in the orthographic representation of the word [118], and rejecting answers if they are not understood by the system [48, 91, 123]. Training fluency requires very efficient speech analyzers. While the approach with a set of prefabricated answers is faster and more reliable [82, 91], the recognition of free answers is also possible [16]. The systems in [52, 136] attempt to train fluency via written tasks. The training of social skills is realized by teaching cultural competence and checking the vocabulary used in specific situations [16, 82, 110]. The Subarashii system [16] offers to beginners of Japanese the opportunity to solve simple problems through spoken dialogues with Japanese native speakers. The vocabulary needed can be learned by hearing and repeating in the stand-alone mode. In addition to pronunciation and fluency, also cultural competence like tactfulness is trained, e.g. the student has to politely refuse an invitation.

Translating Two systems with a strong focus on teaching translation skills are described in [15, 141]. The TransIt-Tiger system [141] had originally been conceived as a course in translation. Later, it was shown that the system assists students in extending their range of linguistic competence in both their target and their source language. The assistance during the translation process consists of glossaries and hints as well as two prefabricated alternative translations.

Training translation skills can naturally be integrated into CALL systems by including a machine translation component. The translation system is either used as supportive tool [165] or as a starting point. The student can analyze the output of the system and try to understand and eliminate the translation errors and to refine the text [9]. Instead of using a translation system, the approaches in [149, 163] developed components which were specifically designed to support the training of translation skills. The system in [149] is meant to help student translators who are simultaneously learning English as a foreign language, and translating from and into that language: a generator for lexicalized sentence stems (i.e. typical phrases a native speaker uses) and a collocation matcher help students generate more easily a native-like text in the target language.

2.6 Supported Language Elements

For the purpose of this analysis we will distinguish between the following three language elements: grammar (G), vocabulary (V), and dialogue elements (D).

Grammar Learning grammar was considered one of the main objectives in most language learning curricula until some years ago.

Inductive grammar acquisition, where the students work out the grammar rules by themselves, has been the aim of several systems which all, in different ways, focus on communication [31, 34, 101, 102, 46, 109, 113, 134, 141, 170, 177], reading [44, 65, 161, 126], writing [21, 129, 178], and translation tasks [9, 149, 165].

In deductive grammar acquisition the students first learn the grammar rules and then practice these rules on examples. Systems which support this approach have to perform a grammatical analysis of the student's input. As the development of an efficient and reliable grammar checker for correct and wrong input is very difficult, many systems are limited to single aspects, e.g. usage of articles [146], conjugation [152], punctuation and capitalization [116], gender, number and person [169], passive sentences [172], or morphological aspects [19]. Other systems include a sentence parser combined with an error grammar and are able to detect a certain class of grammatical errors [49, 87, 104, 124, 163]. For example, with the *German Tutor* [87] the students can build sentences with some indicated words. Several grammar and parser modules analyze the input (spelling, word usage, grammar, punctuation). If a module detects an error, further processing is blocked until the student corrects the mistake. Similar systems have been implemented by the same authors for the Greek and English languages.

Vocabulary Another important part of a language is its vocabulary. Many systems which train reading in a foreign language focus on incidental vocabulary acquisition [108, 134]. Other systems have been developed for intentional vocabulary acquisition. In CAVOCA [78] vocabulary is taught by guiding the learner through three stages corresponding to the sequence of mental operations which make up the word acquisition process.

In the MCALL system Gertie [98] the key principles in the design process were the possibilities to contextualize and to personalize the learning process. Vocabulary is studied by moving the mouse pointer over the items of a picture and at the same time reading and hearing the appropriate words. Students can record their own voice and compare their pronunciation with a native model.

There are several systems which apply intelligent techniques for vocabulary acquisition. The systems in [45, 85, 148] for intentional vocabulary acquisition apply an adaptive approach and consider the vocabulary learned previously by e.g. generating a lesson which is just little beyond the learner's current knowledge. Most systems pursue incidental vocabulary acquisition in combination with translating [9, 123, 149, 163, 165] or reading [44, 65, 161, 126]. Knowledge bases or NLP techniques are applied to provide extensive information on unknown words.

Dialogue Elements Many systems do not explicitly address grammar or vocabulary acquisition, but support the training of dialogue elements. MCALL systems concentrate on conveying cultural knowledge and social skills, e.g. by supplying culturally authentic texts or short video films [31, 81, 46, 112, 134]. In NCALL systems dialogue elements are trained through written communication with peers [34, 57, 101, 102, 113, 170, 177].

There are also some very advanced ICALL systems which allow the training of dialogue elements like pragmatics [110], phonology [48, 62, 118, 173], or an entire conversation [16, 82]. Several ICALL systems focus on the training of comprehension and fluency using written dialogues [52, 55, 79, 91, 104, 121, 136]. Such systems provide learning environments such as a micro-world or a virtual interlocutor. The user typically has to interview the characters or to negotiate a goal. The systems include a grammar and robust parsing abilities to analyze errors up to the semantic level and to give meaningful feedback.

2.7 Availability of the Systems

Depending on the course content or on the intended audience, different media have been used in the development of CALL systems. Depending on the language skills that are to be taught and on how much learner autonomy is supported, the systems include a more or less sophisticated user interface, high-level input processing and output generating components, and intelligent tutoring as well as help features. Nowadays many projects are Web-based or have been realized with the help of Web-based tools (w). Others, mostly memory-intensive systems, have been implemented as stand-alone applications in CD-ROM-form or as device on its own, and can either be freely downloaded and tested (a) or purchased (p).

Web-based Systems The Internet and in particular the WWW open new doors to the development of CALL software. Various services can be integrated, such as e-mail, news, and chat forums. Almost half of the projects examined are Web-based and explore various features offered by this medium including the possibility, on the part of teachers and learners, of communicating through the net.

Systems on CD-ROM and stand-alone applications A large number of CD-ROMs for language learning is available. Developers very often use this medium to provide large amounts of authentic material in the form of spoken language and video clips [31, 78, 46, 134, 141]. An enhanced study of differing cultural behavior and communicational abilities is supported by adding multimedia features [112, 134]. Many of these CD-ROMs are designed for autonomous language learning, and hence include extensive help facilities or intelligent tutoring features.

Devices Hand-held devices are sometimes recommended by a teacher to enhance foreign language knowledge [123, 173]. TAIT [173] is an adaptive authoring device which can be used for training pronunciation skills. Language patterns are recited and the student is supposed to repeat them. The system uses speech recognition software to record the user's voice and to check the input for errors. A user model is maintained

and different users get different feedback messages at different stages of their work. TAIT also provides facilities for a systematic repetition of previously studied concepts.

2.8 Discussion

Most of the MCALL systems we analyzed have reached a very high standard. They follow an integrated approach, include the training of communication skills, and provide information about cultural and social circumstances. Many of the systems subsumed in this category are now either in everyday use at universities or commercially available.

Also NCALL projects seem to be carried out quite successfully. The main problems encountered are of a formal nature, i.e. teachers or students were not well prepared for the medium, network connections did not work, and lab times had to be organized and coordinated. The contact with native language partners, and hence an unknown culture, sometimes made the communication challenging for both students and teachers. Nevertheless, such approaches were considered effective and motivating and were usually appreciated by all participants.

Artificial Intelligence technologies are thought to be very useful to support authentic language learning environments and to simulate real communication situations. On the other hand, a certain neglect of ICALL is lamented, even within the research community. One high-level commercialized system ("Herr Kommissar"), a robust conversator for German, was dropped of the marked after some time because the demand for it was too low[10]. Many other very interesting and promising systems have remained at a prototypical stage. They were abandoned after project termination and have never been used in a real language-learning environment — often due to lack of funding[11].

One reason for this skepticism may be the fact that, because of the difficult implementation process, many systems just focus on single aspects of language learning. Most of the ICALL systems that use NLP techniques concentrate on syntax, few of them include semantic components, and even fewer address the problem of pragmatics. The systematic training of cultural knowledge and social abilities is only rarely considered in the training process. Sometimes technology is used quite experimentally. There are several systems which apply NLP techniques to foster conversational fluency. But can oral capabilities be learned by written tasks (sometimes student input consists of single sentences)? It should be analyzed and evaluated which technology best supports which language skills. The machine-translation approach is basically to learn from mistakes, which is no doubt valuable if the mistakes are so called "false friends" and other mistakes commonly made by students. It is, however, questionable whether a student benefits from artificial mistakes made by a machine.

In almost all cases AI tools have been used on the interactive level. Student input is analyzed, either in written (NLP) or in spoken (ASR) 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 able to provide sophisticated feedback, rather than simple right-wrong-answers. A new and maybe

[10] e-mail conversation with Bill DeSmedt, author of "Herr Kommissar [52]"

[11] e-mail conversation with Farzad Ehsani, author of "Subarashii [16]"

less complicated approach could be to use NLP for data preparation. The use of annotated corpora for language learning exists, but also educational materials could be annotated, language expert systems could be consulted in the annotation process and their knowledge added to the educational material.

It is furthermore striking that we could classify all the projects analyzed rather easily into distinct sections MCALL, NCALL or ICALL, and that few systems try to integrate more kinds of technologies. MCALL technologies could focus on pragmatic aspects of a language, while ICALL technologies could concentrate on answer checking. In the same system collaboration could be provided by NCALL technologies. The development of integrated systems and their wise incorporation into a CALL environment is certainly very important for the acceptance and applicability of ICALL systems in the real language learning lab.

As it is usually claimed and nowadays also widely applied in MCALL and NCALL, also in ICALL a more comprehensive approach is needed which takes into account didactic requirements, and uses different technologies tailored for the training of specific skills. Clearly, it is not easy to develop such a system, and much expertise from different disciplines is needed. More interdisciplinary co-operation between experts from different fields could positively influence this process.

Chapter 3

Problem Analysis and Design Goals of ELDIT

In order to avoid that technology would drive our approach to system development, research paradigms out of different disciplines considering language learning, lexicon acquisition, and dictionary use are considered throughout the development process. We carried out an analysis of learners' difficulties and needs in foreign language learning and potential problems in the use of a straightforward aid, namely a dictionary (see section 3.1). We reviewed existing solutions found in lexicography, didactics, (psycho)linguistics and other fields (see section 3.2). Many of the solutions we found are in fact demands. We not only implemented but also tried to improve them in the ELDIT language learning system (see section 3.3). The analysis is illustrated in Figure 3.1.

3.1 Analysis of Problems in Language Learning

An analysis of learners' difficulties and needs when dealing with a foreign language and a straightforward aid, namely a dictionary, revealed a list of problems they usually encounter.

3.1.1 Problems with Foreign Language Use

This part of the analysis is illustrated in the upper left half of Figure 3.1.

Students may encounter problems when reading or listening to a text (decoding) and when producing a written or oral text in a foreign language (encoding). In text decoding students may encounter difficulties on the semantic level, namely if they do not understand the meaning of certain words. With regard to text encoding, we can again distinguish between two main categories of difficulties: one on the paradigmatic level and one on the syntagmatic level.

On the paradigmatic level, learners may find it difficult to discriminate distinctions between words. For example, should the learner say "ein Haus bauen" (to build a house) or "eine Hütte bauen" (to build a cabin), does one say "die Sonne geht auf" (the sun rises) or "die Sonne geht unter" (the sun goes down), etc.? Grammatically speaking all these phrases are correct, as regards content, however, there are big differences between the expressions. General dictionaries or specific dictionaries of synonyms

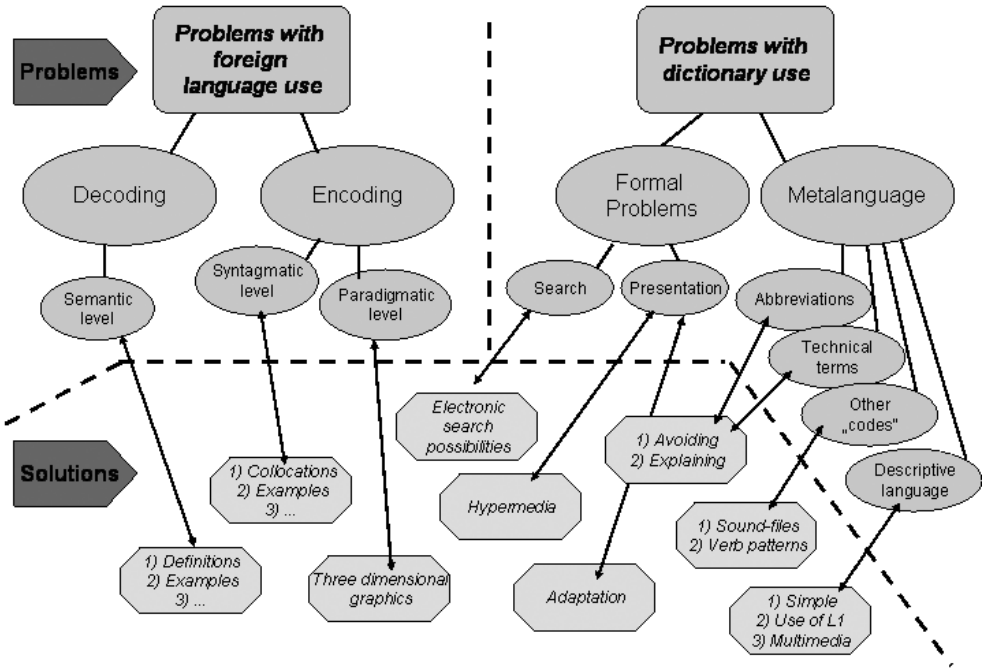


Figure 3.1: An analysis of learners' difficulties and needs

provide a set of synonyms or antonyms for a word, but they do not specify the differences between the synonyms, or explain in which context they can be used.

On the syntagmatic level, students often have problems to cope with collocations (i.e. correct word combinations), because they are unpredictable. For example, in English you say "to brush one's teeth", in Italian "lavarsi i denti" (to wash one's teeth), and in German "sich die Zähne putzen" (to clean one's teeth). Also verb valency, i.e. how to construct valid sentences around a verb, is a common source of errors due to the fact that verb constructions are, more often than not, arbitrary and unpredictable.

3.1.2 Problems with Dictionary Use

This part of the analysis is illustrated in the upper right part of Figure 3.1.

Generally, a dictionary should help students to cope with the previously mentioned difficulties. But when they look for an answer in dictionaries they come across a whole series of new difficulties.

We can divide the problems with dictionary use into two main categories: formal problems and meta language.

Formal problems result from the way information is stored: it is rather time-consuming to look up a word in a dictionary, especially if you are looking for a particular collocation or construction. The main questions are: is a collocation "x does y" to be found under the lemma x or under the lemma y? If one cannot find a particular piece of information, is it because the information is not available or because one has not succeeded in finding it?

Meta language problems concern the language used to convey information: it is not self-evident that users understand the descriptive language used to define the entry word of e.g. monolingual dictionaries. In fact, it is not unusual to find the definition of a word more difficult than the word looked up.

Then, students will find many abbreviations whose meaning they will have to work out. These are used to refer to the word class or gender, to indicate restrictions of use, and to give various kinds of labels (temporal labels, labels for marking style and situation or labels for special fields of activities).

Once students have been able to work out the meaning of the abbreviations, they will still have to understand the meaning of the underlying indication, as they are technical terms which refer to grammar or to language variations. This implies that students need a certain level of grammar knowledge and a general understanding of the concept of language variations in terms of style and register.

The most difficult problem are probably special codes that are used to describe particular kinds of information. For example, most dictionaries use the International Phonetic Alphabet to describe how a word is pronounced.

Also information on verb valency (i.e. sentence construction) is given in coded form. The following problems occur (see Table 3.1): the German monolingual dictionary *Duden*, for instance, contains implicit information on verb valency combined with information on collocations. With regard to Italian lexicography, we can note a tendency to discriminate above all transitivity and intransitivity. Only the *Disc* includes the number of obligatory elements (e.g. ”2 argom.”). Special verb valency dictionaries contain very extensive, complex and condensed information on valency. However, they usually use a notation that is difficult to decode and which requires consulting and studying the special explanatory charts. Alphanumeric codes are used which allow giving a rather precise description, but are by no means intuitive.

WORD	VERB VALENCY PATTERN	DICTIONARY
General monolingual dictionaries		
fragen	[jemdn.] unvermittelt, ... etw. fragen	Duden [56]
chiedere	v.tr. (2 argom.)	Disc [143]
chiedere	v.tr.	Devoto/Oli [53]
Special mono- and bilingual verb valency dictionaries		
fragen	01a v 1b C	Bianco [17]
chiedere	N-V	Blumenthal/Rovere [18]
(Monolingual) learner’s dictionaries		
fragen	Vt/i (j-n) (etw.) f.	Langenscheidt [168]
fragen	tr K jd fragt jdn [nach etw dat]	Pons Basiswörterbuch [86]
chiedere	tr.	DIB [50]

Table 3.1: The description of verb valency for the verbs “fragen” and “chiedere” (to ask) in different dictionaries.

Using, remembering, or even learning words with the help of a reference tool provides further problems: words are encountered in an isolated way and usually only in the citation form. The learner may not be able to use the word correctly in context and – although special effort may have been made against it [179] – forget it quite soon.

3.2 Existing Solutions

This part of the analysis is not represented in Figure 3.1.

A combination of existing solutions from several disciplines turned out to be a good approach to the problems mentioned in the last section. It served as a starting point for the elaboration of solutions in the ELDIT system.

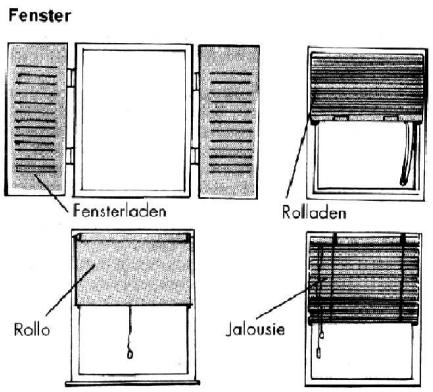
3.2.1 Lexicography

Lexicographers designed a special kind of dictionary, so-called *learners' dictionaries*. A learners' dictionary serves both as a reference book decoding what the learner does not understand and as an instrument which supports text production [1].

Research on learners' dictionaries started in the 1920s and 1930s with the so-called "vocabulary control" movement [67]. The basic idea was to ease the burden of foreign language learning by limiting the vocabulary to a core part which would suffice for everyday communication. Since then much research has been done, not only in terms of coverage but also in the way meanings are defined and illustrated. Several learners' dictionaries have been published. The four major dictionaries for the English language are: *Oxford Advanced Learners' Dictionary* [94], *Longman Dictionary of Contemporary English* [114], *Collins Cobuild English Dictionary* [150], and *Cambridge International Dictionary of English* [83]. Langenscheidt's *Großwörterbuch Deutsch als Fremdsprache* [168] is one of the mayor German learners' dictionaries. To our knowledge, except the ELDIT dictionary currently no special dictionary for learners of the Italian language exists. Most learners' dictionaries are monolingual, and until recently most of them were only available as textbooks.

Figure 3.2 shows the entry of the word "Fenster" (window) in Langenscheidt's learners' dictionary "Deutsch als Fremdsprache" [168]. The entry contains pictures of different windows and window parts, a definition, collocations, an example, compound words, an idiomatic expression, and an adjective. In a printed dictionary space is limited, information is structured in a linear way, and many abbreviations are used. It is therefore difficult to get an overview of the information provided. Moreover, it is impossible to directly access semantically related words, except derivations and compound words which, according to their usual lexicographic order, are physically close to the word under consideration. And even such words are difficult to find if the user is not familiar with the global organization of the dictionary: note for example the compound word "Fenstersims" (windowsill) which is listed within the entry "Fenster". "Fensterbrett", a synonym to "Fenstersims", could also be expected to be listed within "Fenster" and alphabetically ordered *before* "Fenstersims". However, it is listed as an own entry at the same level as "Fenster" further below.

Learners' dictionaries are no doubt very useful, but, in paper form, they do not tackle the problems mentioned above. Nowadays also an electronic version is often available. However, for simplicity reasons these are often one-to-one conversions of the printed form and do not exploit the vast possibilities offered by new hypermedia systems. The dictionary of the Third Millennium is no doubt electronic, but new approaches have to be developed in order to tackle problems regarding an efficient exploitation of the medium and an underlying didactic theory [51].



verwendet ‖ K-: **Fenchel-, -gemüse, -salat, -tee**
Fẹn·ster *das*; -s, -; **1** das Glas (mit e-m Rahmen aus Holz, Metall od. Kunststoff), das in e-m Haus od. Fahrzeug e-e Öffnung schließt, durch die das Licht hineinkommt ⟨aus dem F. sehen, zum F. hinausschauen; das F. steht offen / auf, ist geschlossen / zu; das F. klemmt; das F. öffnen, schließen⟩: *Er kurbelte das F. herunter u. fragte e-n Passanten nach dem Weg* ‖ K-: **Fenster-, -glas, -griff, -öffnung, -rahmen, -scheibe, -sims** ‖ -K: **Auto-, Dach-, Keller-, Kirchen-, Küchen-, Wohnzimmer-** **2** *gespr*, *Kurzw* ↑ **Schaufenster** ⟨ein F. dekorieren⟩ ‖ ID **weg vom F. sein** *gespr*; in der Öffentlichkeit nicht mehr beachtet werden od. keine Chance mehr haben ‖ *zu* **1 fẹn·ster·los** *Adj*; *nicht adv*
Fẹn·ster·bank *die* ≈ Fensterbrett
Fẹn·ster·brett *das*; e-e schmale Platte aus Holz, Metall od. Stein am unteren Ende des Fensters (1): *Blumentöpfe auf das F. stellen*
Fẹn·ster·flü·gel *der*; eine Hälfte e-s Fensters (1), die man öffnen kann: *Die Fensterflügel sind weit geöffnet* ‖ ↑ Abb. unter **Flügel**
Fẹn·ster·la·den *der*; e-e Vorrichtung (*mst* aus Holz) außen am Fenster, die man morgens aufklappt u.

Figure 3.2: The entry for the German word “Fenster” (window) in Langenscheidt’s learners’ dictionary “Deutsch als Fremdsprache”.

3.2.2 Psycholinguistics

According to psycholinguistic insights, extensive word exposure is necessary to ensure a deep and solid embedding of new words in the mental lexicon. Moreover, people do not only remember words in a possible context, but they group the words into multidimensional word nets in their minds. Several word groups should be visualized for the learner to enhance the learning process and to facilitate the memorization of the information [8, 100]:

- *Paradigmatic relations* (i.e. synonyms, antonyms, etc.) between words should be shown to help with the decision which word to use in a given context. They should help with the decision whether to say “ein Haus bauen” (to build a house) or “eine Hütte bauen” (to build a cabin).

- *Syntagmatic relations* should be shown to help with the decision whether some words can be connected grammatically correctly in a sentence. They should help with the decision whether to say “die Zähne putzen” (to clean one’s teeth) or “die Zähne waschen” (to wash one’s teeth).

- The *associative field* groups semantically related words, e.g. “Dach” (roof), “Fenster” (window), “Tür” (door), “wohnen” (live), etc. for the word “Haus” (house). This allows acquiring all important words of a domain and makes the learner able to communicate in specific situations.

- Words that *sound similarly*, such as “appartamento” (appartment), “dipartimento” (department) and “compartimento” (compartment) should also be shown and explained together to the learner. This avoids possible misconceptions and supports an immediate correct embedding of them in the mental lexicon.

3.2.3 Didactics

Since CALL is a quite young research field, many didactic suggestions about how to design a CALL system exist, but the field seems to be explored in quite an experimental way. CALL researchers complain that it is hard to find some generally accepted research paradigms [39]. Carol A. Chapelle suggests that some design features for CALL might be developed based on hypotheses about second language acquisition [40]. She outlines a relevant theory of second language acquisition, and, based on this model, states seven demands on a good CALL system. We are listing these demands in an abbreviated form:

- The linguistic characteristics of target language input need to be made salient, since emphasizing information in materials, which prompts learners to notice particular syntactic forms, positively influences their acquisition.
- Learners should receive help in comprehending semantic and syntactic aspects of linguistic input. Often only a subset of the data the learner receives, is actually useful, because the information contains words or linguistic forms that the learner does not know.
- Learners need to have opportunities to produce target language output. Moreover, it is important that learners have an audience, so that they attempt to construct meaningful utterances.
- Learners need to notice errors in their own output and need opportunities to correct their linguistic output.
- Learners need to engage in target-language interaction whose structure can be modified for negotiation of meaning. They should furthermore engage in L2 tasks designed to maximize opportunities for good interaction.

3.2.4 Vocabulary Acquisition

Two ways of learning new words are commonly known, namely “intentional” and “incidental” vocabulary acquisition [43, 78].

“Intentional” vocabulary acquisition occurs when intentionally learning words and translations with the help of a word list. This method is very often used in traditional learning environments and, usually, also appreciated by learners, since it is quite fast.

However, there are also the following recommendations as far as vocabulary acquisition is concerned: extensive word exposition is necessary to ensure a deep and solid embedding of acquired vocabulary in the mental lexicon [8, 100], moreover, vocabulary acquisition should be personalized and occur in context with authentic text [78, 98, 120, 148].

These requirements would be met by “incidental” vocabulary acquisition, namely vocabulary acquisition by contextual deduction, which happens when reading a text in a foreign language. This method provides two advantages: first, it ensures that words are encountered in several forms and contexts, wherefore they will be embedded much more deeply in the mental lexicon. Second, words are encountered together with

syntactic information which helps to learn how to use them correctly. However, also this method is not devoid of problems: as stated in [43, 76], the method is often not fast enough for the learner to master within limited time the vocabulary needed for fluent conversations and correct formulations. Another problem with incidental vocabulary acquisition by reading an authentic text has been summarized in [78]: it is the large number of unknown words around the items to be learned.

3.2.5 Computational Linguistics

We now start analyzing some more technologically oriented research fields and what kind of inspiration and help they could provide for the development of a language learning system.

Computational Linguistics (CL) is the field that deals with the use of computers for linguistic research and applications. We are outlining some well known techniques out of this field:

- A *Stemmer* is a program or algorithm which determines the morphological root of a given inflected (or sometimes derived) word form. A Stemmer for German, for example, identifies the string ”kaufte” (bought) as based on the root ”kauf” (buy).

- A *Lemmatizer* is a program or algorithm which determines the lemma of a given inflected word form. Usually also information on the word class is provided. For instance, for the previously mentioned word “kaufte” the information “kaufen - verb” would be obtained [33].

- A *Morphological Analyzer* is able to analyze an inflected word form in more detail [33]. For instance for the word “Hause” the information “Haus, noun, dative, singular” is obtained.

- Lemmatizing and a morphological analysis create ambiguities. For instance, the Italian word “posto” can be “posto - noun” (the place) or “porre - verb” (to put). A *part-of-speech tagger* eliminates ambiguities in a lemmatized or morphologically analyzed text [144].

- A *Parser* consults a grammar model of a natural language. It parses a linear text and maps it to a tree-like representation of it. The structure of the tree is determined by the grammar [159].

- A *Term Extractor* is a program able to extract new terms of a text [155].

- A *Meaning Disambiguator* can distinguish different word meanings within a text: for instance “a house to live” (i.e. house as a building) is not the same as “the royal house of England” (i.e. house as a dynasty).

- A *Concordance Tool* is a program that generates a concordance, i.e. a list of all occurrences of a given word or expression in a given corpus, with the context in which it occurs [12]. Usually a frequency analysis is carried out as well and the result is ordered according to these frequency values.

- A *Text Classification Program* is able to determine the domain (sports, music, etc.) a text belongs to [154].

Corpus annotation is the electronic adding of information to a given electronic text. A computer program scans the text by applying the above mentioned so-called natural language processing (NLP) techniques to obtain new information that can be added to the given text.

The use of annotated text corpora for language learning is very common. Form focused activities are in the foreground, i.e. the students should establish and practice the use of particular linguistic features. Data-driven learning (DDL) is the main keyword used in this field [11, 12, 43, 106, 148]. Teachers and students who practice DDL use different concordances and authoring tools to research and create language materials. Learners have the opportunity to discover language rules by themselves. Usually students learn through problem-solving activities rather than being instructed directly by the teacher [142].

3.2.6 Adaptation

Learners may have difficulties when being faced with the large amount of information a hypermedia system provides and the new, usually non-linear way of structuring and presenting this information. Moreover, each learner is different, and a personalized learning process may increase the learners' motivation and the effectiveness of the learning activity. In order to provide personalization and individual guidance, adaptive features are often proposed in the relevant literature. Adaptive hypermedia systems are a combination of user-model-based Intelligent Tutoring Systems and user-neutral Hypermedia systems [27]. The first ones of them appeared about 10 years ago. They cover the aspect that the system automatically adapts content presentation and navigation possibilities to the individual user. This is achieved by explicit indications about the user, assumptions about the user, and observations about the user's interaction with the system. The information is collected in a user model and interpreted with intelligent reasoning and machine learning algorithms [23, 30, 26, 61, 89, 125]. In language learning user input is often additionally interpreted by applying computational-linguistic and Artificial Intelligence (NLP and ASR) techniques [55, 79, 87, 104, 163]. Based on these interpretations, content presentation and navigation features are adapted to individual user preferences and needs. Such features have proved to be effective if the user collaborates with the system [28], and if the user is allowed to keep control over content, structure, and organization of the learning material [28, 59, 93]. Also in a domain with almost no structure, for instance in an unrelated collection of learning items, adaptation for structuring the hyperspace may work very well [93].

3.3 Design Goals of ELDIT

The ELDIT program is a Web-based system. Its core module is an extensive learners' dictionary which is currently extended to a complete language learning platform. The system has a strong focus on vocabulary acquisition.

For each section discussed in chapter 3.2 we will now explain the corresponding features in the ELDIT program. This part of the analysis corresponds to the lower half of Figure 3.1.

3.3.1 The ELDIT Learners' dictionary

Basic Vocabulary

For the selection of the vocabulary the following considerations have been taken into account: ELDIT has been conceived to contain about 3,000 entries for each language. For both, the Italian and the German languages, an intersection of different so-called "basic vocabularies" has been built up. The two resulting vocabularies have been adjusted to each other to avoid big differences between the two languages. Finally, the two vocabularies have been enlarged by some words which are frequently used in South Tyrol such as terms related to farming or viniculture ("farmer", "wine", etc.). The resulting vocabularies cover standard German and Italian including some language variants spoken in South Tyrol.

Exploiting New Media

The ELDIT dictionary is solely designed for computer use and tries to fully exploit the facilities offered by new technologies. The information is stored in hypertext-form and organized in independent modules which can be accessed separately by clicking on hyperlinked tabs.

Multimedia allows the simultaneous use of different media such as text, images and sounds. This makes it possible to avoid complicated codes such as the phonetic alphabet and offers a more illustrative and multi-faceted way of presenting the lexical material. It furthermore encourages a learning process that integrates different modes of presentation and is adaptable to different cognitive styles.

Finally network technologies (collaboration via e-mail), computational linguistics approaches (corpus annotation), and Artificial Intelligence techniques (NLP and adaptation) are included or planned for the future.

A New Type of Dictionary

From a lexicographic point of view, ELDIT is a completely new type of dictionary [6, 3], we call it a "crosslingual" dictionary: it is on one hand designed as two separate monolingual dictionaries in that the meaning of each word is explained by a definition in the same language. This approach fulfills pedagogical demands which claim that it is better for the learner to remain in the target language.

On the other hand, the definitions are extended with translations, a typical element of bilingual dictionaries. This add-on fulfills the demands of learners who usually prefer bilingual dictionaries [13, 128]. Moreover, the translation equivalent serves as an entry point to the corresponding part of the other module. In this way, with a simple mouse click, a user can easily switch between the two dictionary modules.

The crosslingual version is especially useful for German and Italian native speakers who wish to acquire the other language. However, the monolingual versions can be used by anybody wishing to acquire the German or Italian languages.

Dictionary Elements

Abbreviations are not strictly necessary in a hypermedia system, since there is enough space available. In ELDIT the first rule for handling abbreviations and technical terms is to simply avoid them. In those cases in which the linguists really could not do without them, they are explained in an extra window which appears when the user clicks on the expression. Also linguistic difficulties and possible explanations of grammatical rules are handled in this way.

Complicated codes such as the phonetic alphabet are also avoided. Sound files help the user with the pronunciation of a word. The presentation of verb valency is realized by exploiting semiotic didactics (colors and movable elements).

Further Improvements

An electronic search engine able to deal with erroneous or inflected input provides efficient and direct access to different kinds of information. Data reuse allows showing a large amount of applications and example sentences. An extensive link structure helps with unknown words and word forms. Interaction is possible due to quizzes that are corrected by the system. In order not to overload the learner with information, personalization features are used to adapt the information to different users.

3.3.2 Psycholinguistics in ELDIT

Paradigmatic Level On the paradigmatic level, learners find it difficult to make semantic distinctions between more or less similar words and word meanings. In ELDIT word fields explaining the meaning of related words and distinctions between them are represented by interactive graphs. The words in the graphs are hyperlinked to the respective entry words, so that the learner can easily consult definitions and explanations about word meanings and word relations.

Syntagmatic Level On a syntagmatic level, students have difficulties coping with word combinations. We give long lists of free combinations, restricted collocations, and idiomatic expressions. Each word combination or collocation is treated as a separate item supplied with a translation, possibly an explanation, and at least one example that shows a typical phrase in which the expression may occur.

Associative Level Semantically related words are planned to be shown in groups and by an image in which each object is linked to the corresponding dictionary entry.

Soundly Related Words For each word the user can request words that sound similarly, such as “compartimento”, “dipartimento”, or “appartamento”, by using the wildcard feature of the ELDIT search engine. Searching for instance “*mento” reveals the previously mentioned list.

Grammatical Level In ELDIT we have included some more groups of words: grammatically related words and word forms are shown in groups to a language learner, namely words related by *word formation* (i.e. derivations and compound words) and

word forms related by *word inflection* (i.e. conjugated and declined word forms). The goal is to convey the complex subject matter in quite a simple and transparent way: We provide the entire inflection paradigm for each dictionary entry in ELDIT. In the future stem and ending will be highlighted. We also show a group of derivations for each word entry and a group of compound words for each word meaning by emphasizing prefixes, basis and suffixes of the term. In both cases, as a result the user should be able to analyze and learn the grammatical structure of words and to draw conclusions about new forms in respective linguistic situations.

3.3.3 Didactics in ELDIT

We now deal with Chapelle's demands on a didactically meaningful CALL system:

- *The linguistic characteristics of target language input need to be made salient*. As outlined in the last two sections and described in more details in later sections hypermedia technologies are used throughout the system to explain, illustrate and emphasize linguistic rules, exceptions and difficulties. Colors and movable elements, for example, are used to emphasize verb valency. Linked footnotes are used to indicate linguistic difficulties such as exceptions to a rule or false friends.

- *Learners should receive help in comprehending semantic and syntactic aspects of linguistic input*. The transmission of complicated syntactic and semantic information has already been described in section 3.3.2. ELDIT is furthermore equipped with a glossary, and each single word used in the system is linked to the corresponding dictionary entry. Hence with one simple mouse click the learner can inspect the meaning of an unknown word. The educational value of such an approach is widely accepted [108, 137].

- *Learners need to have opportunities to produce target language output*. In ELDIT, quizzes can be created which allow, as a first step, the practicing of acquired information in a simple drill and practice way. The learner also has the opportunity to freely produce target language output, namely by reading texts and answering questions. These answers are corrected either by peers or by a human tutor.

- *Learners need to notice errors in their own output* and *Learners need to correct their linguistic output*. Due to computational-linguistics technologies included in ELDIT, advanced error correction by the system up to the morphological level, in some cases even up to the syntactic level, is possible for quizzes. Correction of the answers by a human tutor can be provided by included network technologies. Personalization features can refine the interaction process by adapting feedback and correction possibilities to different steps within a course, or to different language levels.

- *Learners need to engage in target language interaction whose structure can be modified for negotiation of meaning* and *Learners should engage in L2 tasks designed to maximize opportunities for good interaction*. These demands are difficult to realize if not by human interaction. To realize this demand ELDIT is

used in language schools[1] which provide "blended learning", i.e. learning which combines online and face-to-face approaches.

3.3.4 Vocabulary Acquisition in ELDIT

According to our opinion, a combination of several methods may provide a solution to the problems mentioned in section 3.2.4: words should be systematically learned by a word list, interactively practiced, and read in authentic texts. The process should be personalized to meet individual students' interests and needs.

In ELDIT, depending on user preferences, a smaller or larger group of words from a user specific interest domain is first offered to be studied. The words should be acquired in several steps:

1) *Perception:* Being able to identify a word, to distinguish different word meanings and to differentiate words from related words, is necessary for the learner to understand spoken and written language correctly. In ELDIT the learner can explore the various descriptions of a word meaning by checking definitions, translations, pictures, and example sentences, and by examining in the same way related words.

2) *Usage:* The next step is to learn how to use a word. Grammatical and collocational knowledge is necessary for this task. The learner should study typical patterns of word usage listed as collocations and idiomatic expressions, as well as the conjugation and declension of a word.

3) *Characteristics:* Finally, the learner should enlarge and complete the knowledge about a word. For that purpose, word characteristics, exceptions to a rule, false friends, etc. listed in the footnotes of a word, should be studied.

4) *Context:* After the words have been learned in all their facets, the user should explore them in an authentic text. The system can select a text from the text corpus included in the system, which contains the words studied before and, preferably, only a few new words.

The first three steps are supported by simple quizzes and questions. Since quizzes and questions arose from the experience of decades of traditional classroom teaching activities, their *usefulness* is generally accepted regardless of the teaching methodology applied. In the electronic medium *interactivity* and *immediate feedback* are moreover possible, which is considered the main positive feature that distinguishes traditional paper-based material from electronic material [157]. The *effectiveness* of such quizzes especially for weaker students has moreover been shown in [96].

The fourth step is supported by authentic texts with comprehension questions. In the text, all the words are linked to the corresponding dictionary entry, so that unknown words can easily be inspected by the learner, an approach which is considered a very valuable feature in language learning [42, 108, 137]. A text can either be read as a whole, or practiced as a gap-filling exercise.

[1] http://www.cedocs.it/versione_it/i_corsi03.php

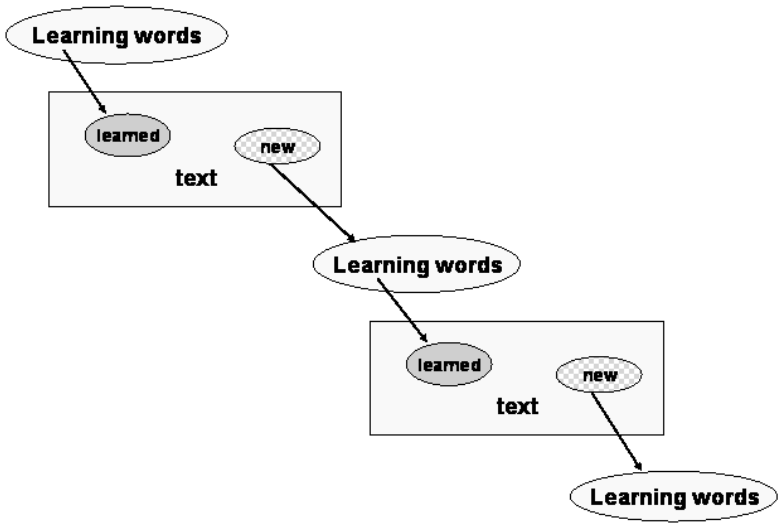


Figure 3.3: Contextualized, adaptive vocabulary acquisition

A combination of these steps is supported by an electronic tutor. Each text will contain some new words that the user will check. These new words are summarized by the tutor and at the end offered to be studied systematically according to steps 1 to 3. After the user has studied these new items, step 4 is applied again and a new unknown and interesting text, in which these words occur, is proposed to be read, and so on. The method is illustrated in Figure 3.3. We call this method of learning "adaptive, contextualized vocabulary acquisition".

3.3.5 Computational Linguistics in ELDIT

Annotating the manually developed text material and in this way supporting the linguists in their content creation work was the main principle applied in our approach to CALL content development. In ELDIT no raw text material but carefully designed educational material has been annotated and enriched with information that can be explored. All smaller or larger text pieces (explanations, definitions, patterns, etc) have been encoded down to the level of single words and annotated with lemma and word class. These and other annotations revealed a large amount of possibilities, out of which we will mention only a few:

- Advanced search possibilities became possible that allow finding inflected word forms and expressions.

- Reusing the manually created content in several places of the system became possible.

- Frequency analyses have been carried out in order to establish "importance" values for words and texts.
- Links were established between explanatory information and dictionary entries. In this way every unknown word can be checked by a simple mouse click.
- A meaning disambiguator using "context vectors" extracted from the collocations will be implemented in the near future to refine the links.

A special authoring tool supports annotating the educational material and adding arbitrary other information to the material, and of course, the more information is added the more possibilities for exploration and an innovative presentation are given.

3.3.6 Adaptation in ELDIT

Not only does the ELDIT system provide a large amount of information, but it also offers a great variety of activities. Different language levels are supported as well as different interest domains.

But it is questionable whether all users will be happy with the default design of the system. In fact, first evaluations have revealed that different users are interested in different aspects of the system and need different contents [4]. These problems can be tackled by adapting the content and the presentation of the system to the individual user [69]. ELDIT exploits common adaptation techniques such as adaptive content presentation, adaptive link annotation, or hiding by slightly modifying them for language learning. In [93] it is furthermore suggested that in a domain with almost no structure, for instance in an unrelated collection of learning items, adaptation for structuring the space may work very well. This suggestion is particularly important for our system, since what we have is such a collection of quite unrelated words and texts. The solution we implemented is an adaptive tutor (realized by an adaptive "next" link) which uses information stored in a user model, advanced search possibilities, our fine-grained data model, morphological knowledge, and some simple rule bases to coordinate the learning process.

Chapter 4

Basics of the ELDIT System

We now describe in detail the conceptualization and implementation of the ELDIT system.

Figure 4.1 shows the overall architecture of the program. The query engine analyzes and dispatches the user requests. The most advanced module is an extensive learners' dictionary (see section 4.1). Furthermore, a text corpus has been developed which allows the learner to read and answer questions in the target language (see section 4.2). Sophisticated quizzes can be generated automatically from the data elaborated for the dictionary and the text corpus (see section 4.3). Collaboration between peers and teachers via a tandem module is possible (see section 4.4). The core module is an adaptive tutor which coordinates the learning material, dynamically selects next best items to be studied, and proposes individual paths through the learning material (see section 4.5). Since the user model influences all features of the ELDIT program, it will be described at the end of the next chapter (see section 5.4).

4.1 Dictionary

The dictionary is the most advanced module of the system. An extensive description of the linguistic and didactic background is given in [3], the main characteristics are explained in [5] as well.

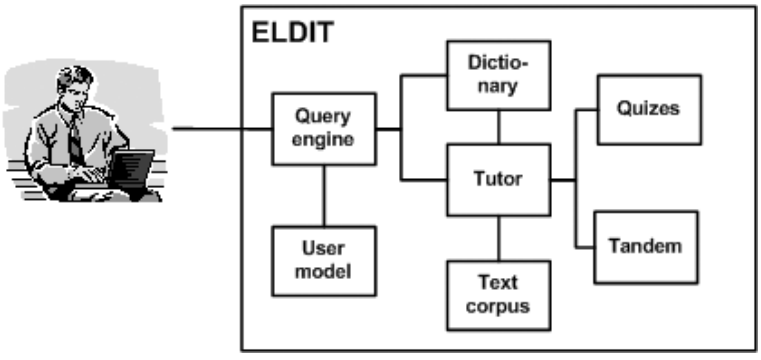


Figure 4.1: Core modules of the ELDIT vocabulary acquisition system.

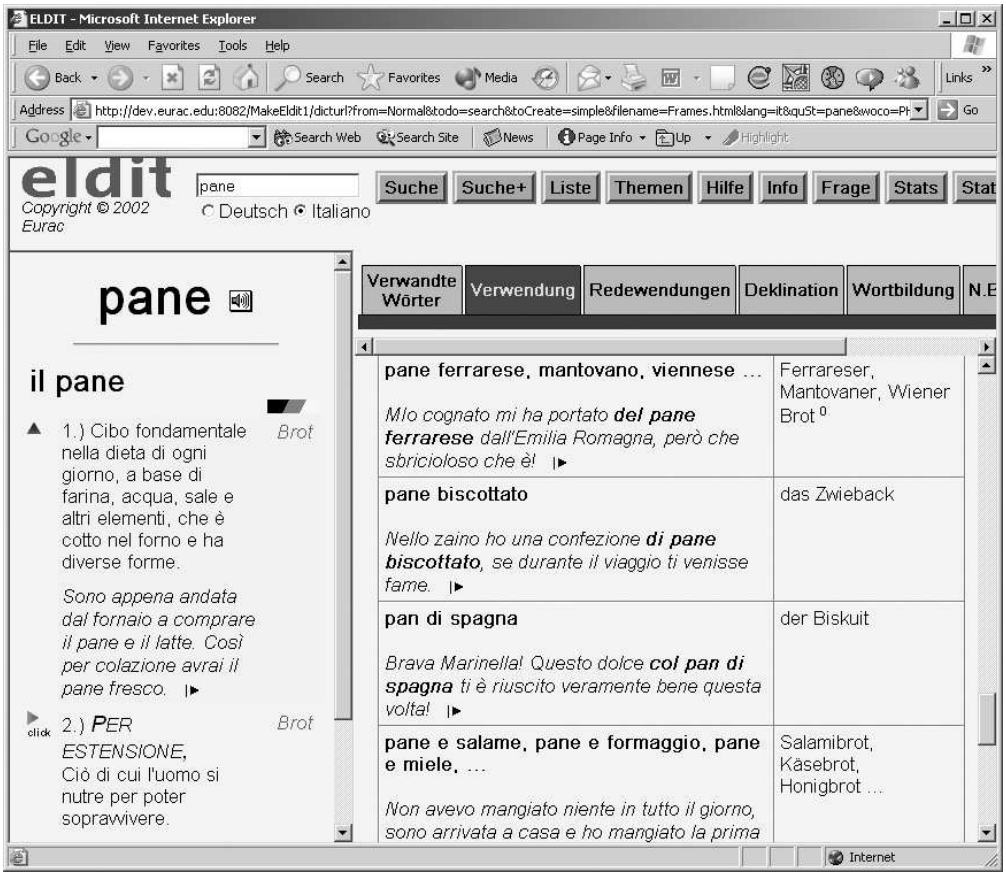


Figure 4.2: Dictionary screenshot for the Italian word *pane* (bread): the tab "Verwendung" (usage) is activated

4.1.1 User Interface

A screenshot of the dictionary is shown in Figure 4.2. A navigation bar at the top of the screen allows accessing the single modules of the system. General features such as search and help features, statistics, or online evaluation questionnaires are listed there.

The information of a dictionary entry is presented in two frames and with the help of a tab metaphor. Smaller blocks within the larger modules are organized in clear chunks or in a network metaphor.

The left-hand frame shows first the lemma (i.e. citation form). Clicking on the loudspeaker button next to the lemma activates a sound file with the pronunciation of the word. Below, morphological information is shown (article and plural forms, if they exist). Further below different meanings are listed, each of which is described by a definition, a translation, an optional comment, and a lexicographic example.

When clicking on the small triangles next to the lexicographic examples, further example sentences from the large example pot in ELDIT can be retrieved which contain the word or expression under consideration.

The right-hand frame is divided into several tabs. In Figure 4.2 the tab "Verwendung" (usage) is activated. It shows typical collocations and word combinations which are described by the following information: a pattern to represent the rule or general

model, a translation equivalent, a lexicographic example, and possibly style and register information (e.g. “colloquial”).

Note the small footnote next to the translation of the first collocation. When clicking on this number, a short explanation appears in an extra window in which the user is informed that “pane ferrarese, mantovano, viennese” are typical Italian types of bread. Such footnotes can appear everywhere. They inform the user about linguistic difficulties such as exceptions to a rule or false friends, cultural particularities, etc. The sum of all footnotes, and hence of all difficulties related to a particular word entry, can be inspected within the tab “N.B.”.

When the user drives over an arbitrary piece of text with the mouse cursor, the single words appear in blue and underlined, because they are linked to the corresponding dictionary entry. When clicking on the word, the description of its meaning appears in a separate window.

The system includes personalization features which, however, are not enabled yet. Different information can be shown to users with different backgrounds.

4.1.2 Description of the Content

We now describe the information packages that have been collected for each dictionary entry (called “word entry” as well), and explain how they have been prepared for the user. We mention concrete words as examples and refer the interested reader to the system itself to check out these examples[1].

Lemma and Morphology

The lemma (called “base form” or “citation form” as well) is the core description of a word, e.g. the word “pane” on the left-hand side in Figure 4.2. In cases where there are a male and a female version, such as for “Einwohner/Einwohnerin” (inhabitant), both forms are stated. Below the lemma morphological information is given, namely the article and plural forms. This information may also include comments on restrictions and particularities, for example when a word is used only in singular or when it is a compound word (see “Ausverkauf” (sale) and “aeroporto” (airport)).

Meaning Description

Each lemma may have several meanings, for instance “a *house* to live” is not the same as “the royal *house* of Scotland”. Since we follow a crosslingual approach, each meaning is explained by a definition and one or more translation equivalents which are linked to the corresponding lemma in the other dictionary.

A lexicographic example illustrates the meaning in context. The lexicographic examples consist of several short sentences. They have been created manually by the linguists according to specific criteria in order to illustrate the word.

Where verbs are concerned, also a short pattern, a so-called “minimal sentence”, is indicated to illustrate the meaning (for instance “jemand baut” means that somebody builds his own home). A small triangle on the left-hand side of each definition allows

[1] http://www.eurac.edu/eldit/

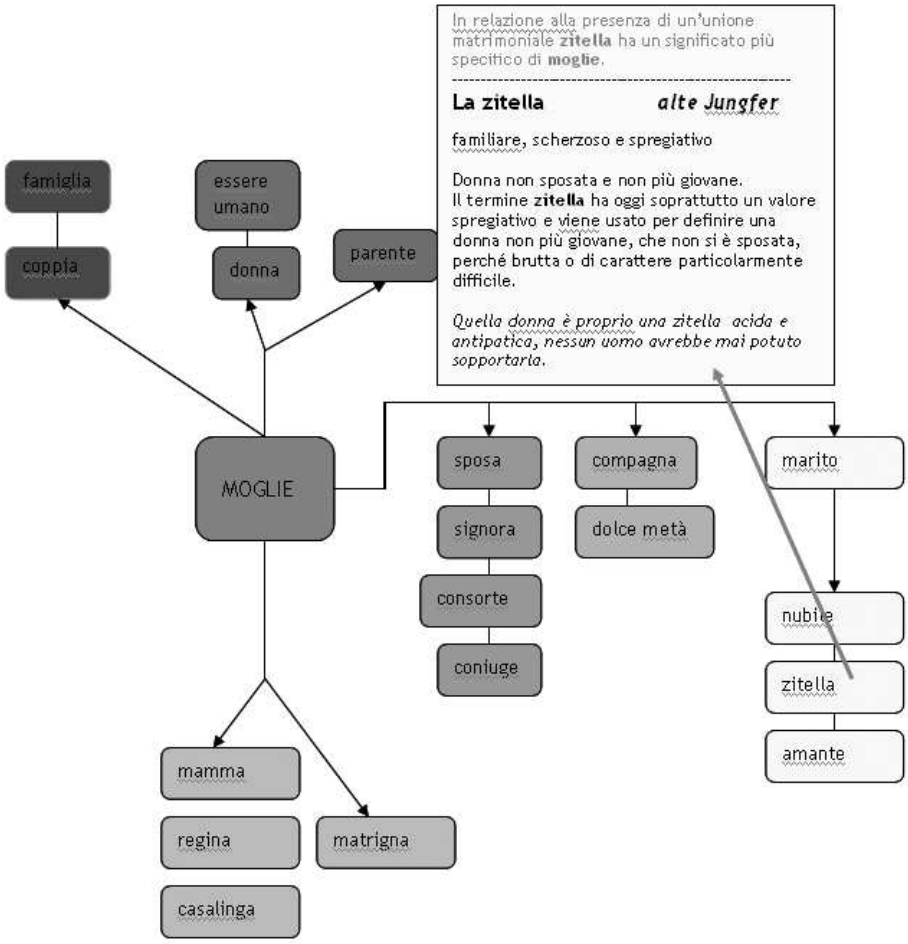


Figure 4.3: The semantic field of the Italian word "moglie" (woman)

activating one particular meaning. This gives access to the information provided in the right-hand frame, some of which is different for different meanings.

Semantic Fields

Paradigmatic relations, i.e. semantically related words or so-called "semantic fields" or "word fields", are shown within the first tab on the right-hand side, namely the tab "Verwandte Wörter/campi semantici". They are illustrated in two-dimensional graphs in ELDIT[2].

We are currently elaborating about 250 word fields for each language (see Figure 4.3). Since we are not interested in elaborating complete word nets of our vocabulary, but rather in describing slight differences between important words for language learners, we have not elaborated one connected graph of all the words in ELDIT, but several small graphs. In each graph we distinguish three levels: (1) relations of the lemma to more general words comprise hypernymy and holonymy, (2) relations

[2]With financial support of the Autonomous Province of South Tyrol.

to words at the same level comprise synonymy, quasi synonymy, antonymy, entailment, and causation, and (3) relations to more special words comprise hyponymy, meronymy, troponymy and particle verbs. The relations are indicated by special colors and explained by definitions, comments, translations, examples, and - most important - differences to the lemma in question. This descriptive information can be inspected by clicking on the nodes in the graph.

Collocations and Idiomatic Expressions

Syntagmatic relations are shown within the second and third tab. For a language learner it is very difficult to infer valid word combinations, thus detailed information has to be given in a good reference system. The following information is presented within the tabs "Verwendung/combinazioni" (usage) and "Redewendungen/fraseologia" (idiomatic expressions) in our dictionary (see Figure 4.2):

- Each word combination or collocation is treated as a separate item supplied with a translation and an example sentence. Also information about the stylistic level of an expression (colloquial, literary, etc.) may be included.
- Furthermore, all kinds of expressions are listed which present a different degree of idiomaticity but are perceived as stable units. These items are accompanied by a translation whenever it is possible to find an equivalent idiomatic expression in the other language. In other cases an explanation is used to illustrate the meaning of opaque expressions and the origin of their figurative sense. Last, an example shows a typical situation in which the idiomatic expression may occur.
- Furthermore, several adjectives which are typically used together with the lemma under consideration are listed together with respective translations and a few examples.

Verb Valency

Within the tab "Verwendung" or "costruzioni" (usage) we also show information on verb valency[3]. This information is very important for the learners, since it indicates how a sentence is constructed in a grammatically correct way. This module is an outstanding example how hypermedia features are used in ELDIT to describe complex information in an innovative and comprehensible way to the learner [2].

The core verb valency information is represented in a simple way by a table (see Figure 4.4). We indicate some short patterns, for instance "qualcuno chiede qualcosa a qualcuno" (someone asks somebody something) which looks like a minimal sentence. The patterns indicate the obligatory parts of a sentence which depend on the verb. Each of these parts is represented in a single cell. The facultative parts of the sentence are shown in brackets. The selection of the indicated parts of the sentence is carried out according to a model proposed in [88, 176]. Below the pattern one or more lexicographic examples are listed which show the use of the verb in an authentic context.

[3]With financial support of the European Union - Interreg IIIA.

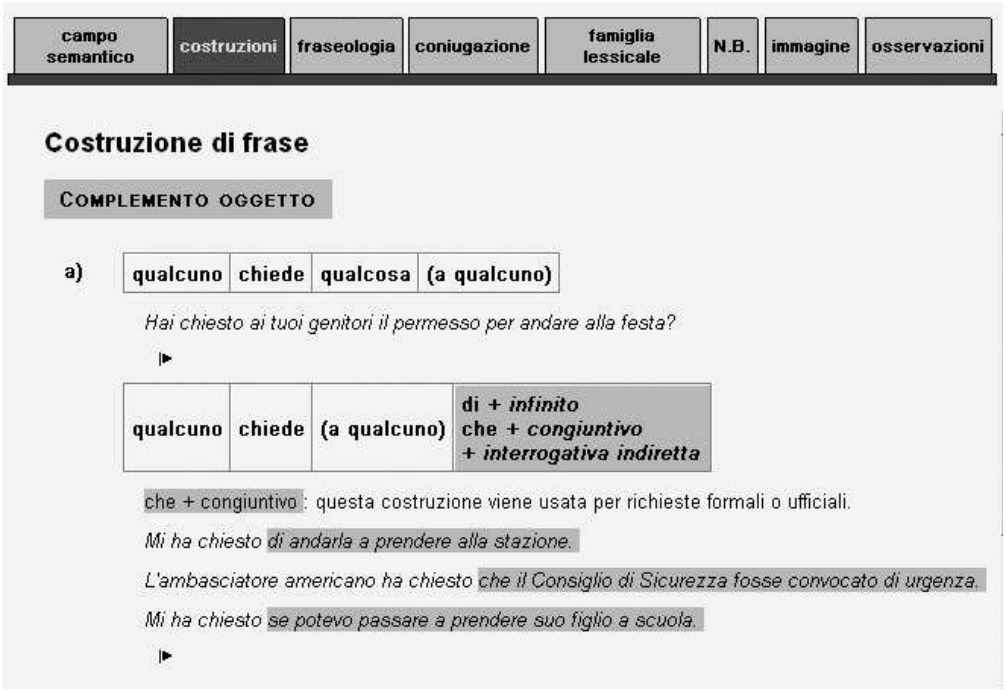


Figure 4.4: Description of verb valency for the Italian word “chiedere” (to ask) in ELDIT

We are exploiting semiotic principles (colors and animations) to describe a complex linguistic phenomenon: the user can pass over the single elements in the table with the mouse and each element is highlighted in a different color. At the same time the corresponding element in the example below the pattern table lights up. The specific use of colors serves to facilitate comprehension and communication processes.

The colors fulfill different functions: (1) they show which elements of the pattern correspond to which elements in the concrete examples. (2) They show which parts of the verbs belong together since such parts always light up at the same moment (e.g. verbs with an auxiliary verb or separable verbs in German). (3) Last but not least the same parts of the sentence always appear in the same color, which allows for a quick and easy recognition of identical parts (e.g. the verb always appears in green, the subject always in red).

Inflection

Inflection is the declension of nouns or adjectives and the conjugation of verbs. In co-operation with the “Scuola Universitaria Professionale della Svizzera Italiana - SUPSI”[4] we implemented the following outstanding feature: we provide the entire conjugation or declension for each single word in ELDIT and show this information within the tab “Deklination/declinazione” (declension) and “Konjugation/coniugazione” (conjugation), respectively (see Figure 4.5).The word forms are organized in tables according to modes (indicative, subjunctive), tenses (present, perfect, etc), persons (first,

[4]This co-operation has been financially supported by the European Union - Interreg IIIA

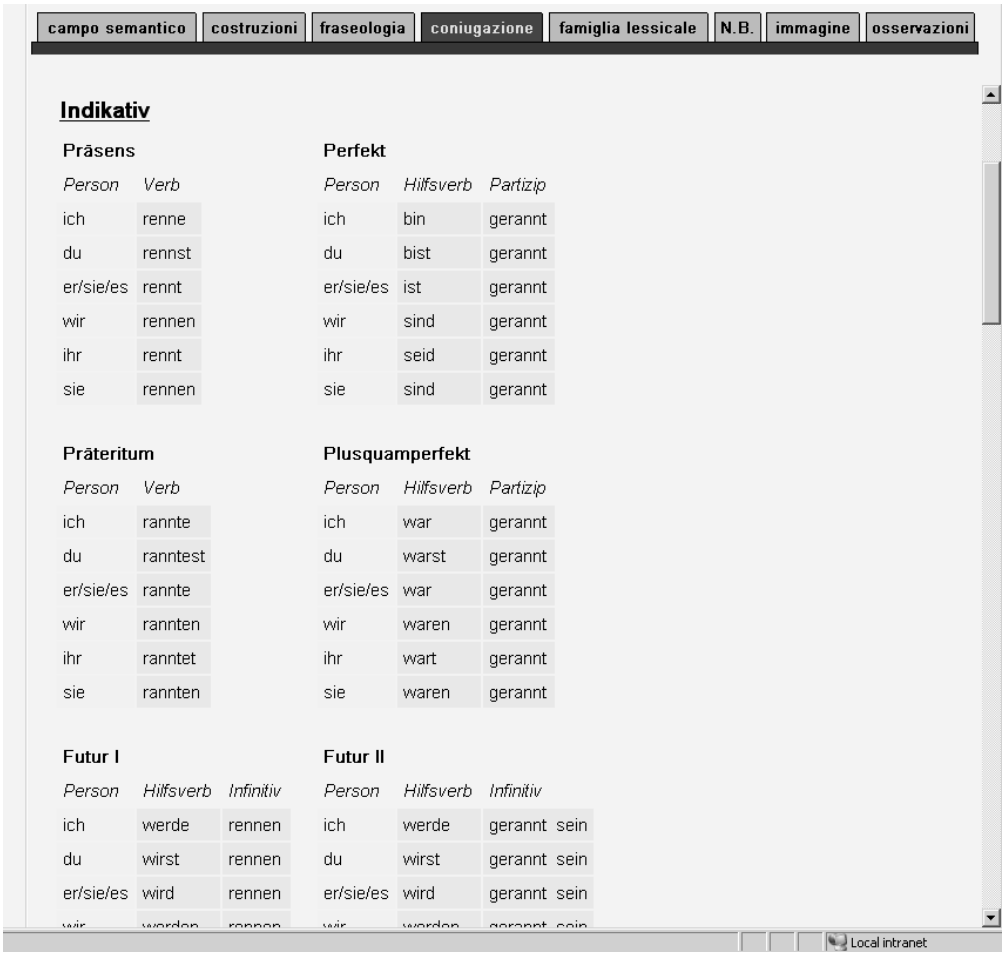


Figure 4.5: The conjugation of the German word “rennen” (to run) in ELDIT

second and third), cases (nominative, genitive, etc.) and number (singular, plural). We are also planning to show the generation rules by emphasizing stem and ending of each word form.

Word Formation

Word formation is about derivation and composition of words. We show a group of derivations for each word, and a group of compound words for each word meaning, within the tab “Wortbildung/famiglia lessicale” (word formation) by emphasizing prefixes, basis and suffixes of each word (see Figure 4.6). By clicking on the prefixes (suffixes) of the derivations, an extra window opens which leads to a section where word formation rules for this specific prefix (suffix) are explained. In this way the user can see the underlying construction rule. Next to the words, small triangles indicate the possibility of searching the ELDIT example pot for lexicographic examples containing the word in question.

Figure 4.6: Word formation of the German word “Wurzel” (root) in ELDIT

Footnotes

One of the didactic demands listed in section 3.2.3 is that linguistic characteristics of the target language should be made salient. ELDIT uses footnotes next to the corresponding content piece to inform learners of potential linguistic difficulties. Footnotes may appear anywhere in the text. By activating (clicking) a footnote, a small window opens where the linguistic difficulty (particularities, general remarks, stylistic nuances, false friends, etc.) is explained. These remarks are seen both in an intralingual and a contrastive perspective and serve the function of error prevention.

Within the tab “N.B.” we list all the footnotes relevant to the entry word so that the learner gets an overview of all the particularities that characterize the word in question.

Pictures

Sometimes “a picture says more than 1,000 words” (for instance it is difficult to define an “apple” unambiguously). Thus we have decided to add pictures to concrete nouns in ELDIT. They are shown in the tab “Bild/imagine”. The pictures are selected according to psycholinguistic criteria [103, 175]. Psycholinguistics suggests the use of prototypical representations of an object (see Figure 4.7), e.g. simple drawings and no photographs, because the latter are not culturally independent and might cause misconceptions, or even not be understood at all. Movements, certain actions, or scenarios

Figure 4.7: Pictures in ELDIT

will be visualized by animated graphics in ELDIT.

Sound Files

It is very important for a language learner to learn the correct pronunciation of a word. In an electronic dictionary the phonetic alphabet can be avoided and replaced by sound files. In ELDIT a sound file with the pronunciation of the lemma can be activated by clicking on the loudspeaker button next to the lemma (see Figure 4.2). The pronunciation of longer text passages could be provided by streaming audio or speech generation[5].

Annotations

Activating the annotation tab “Anmerkung/annotazione” (observation) results in showing an editable window in which the user can save personal comments on each word. These annotations are saved in the user model and retrieved only for this particular user whenever the user accesses the system.

[5] http://www.webspeech.de/index2.php

4.2 The Text Corpus

One of the didactic demands listed in section 3.2.3 is that learners need to have opportunities to produce target language output. In order to provide these opportunities, ELDIT includes a text corpus consisting of 800 texts (400 for each language, within them 200 for each level of difficulty) which have been especially developed for language learners[6] [72, 73] (see Figure 4.8). The texts have been classified manually into 20 different, possibly overlapping, interest domains. Each text consists of about 150 words. Every word is linked to the corresponding dictionary entry so that learners can easily check unknown words.Each text is furthermore enlarged with six comprehension questions.

Figure 4.8: A text in ELDIT

4.2.1 Learning Activities

The following learning activities can be carried out by a learner:

[6]With financial support of the Autonomous Province of South Tyrol.

- The texts can be read, words are linked with the dictionary. Unknown vocabulary can be checked with a simple mouse click, the corresponding meaning description is shown in an extra window. If desired, the learner can access the entire dictionary entry directly from this extra window.
- Each text includes six questions which the learner has to answer in complete sentences in the target language. Hints for the correct answers to the questions are provided in the following way: when the learner clicks on the help link, the sentences that could provide the answer are highlighted. Moreover, the dictionary and its large amount of information on the construction of semantically and syntactically correct sentences is always accessible simultaneously with the texts.
- The answers can be printed or saved in the user model.
- The answers can be sent to a learning partner or to a human tutor for correction.
- In order to increase the incidentally obtained vocabulary knowledge, a text can also be practiced as a fill-in-the-blanks exercise.
- An electronic tutor will make suggestions of "next best texts" to be practiced.
- Text sentences can be re-encountered in the dictionary when calling the "more examples" feature of the system.
- Users can submit their own texts and prepare them in the same way the system texts have been elaborated.

4.2.2 Corpus Generation

The texts were reused from a former project carried out at the European Acdaemy Bozen/Bolzano. For the inclusion of them into the ELDIT language learning system several conversions of the original material was carried out.

From MSWord to HTML The text corpus was provided by the *Göthe Institute of Milan* in MSWord format. It was converted electronically into HTML format. Some manual adaptations were carried out in cases where the conversion did not work properly.

From HTML to XML For the links the texts had to be encoded down to the level of single words. Since all the material in ELDIT is stored in XML format (see section 7.3), we first elaborated a DTD which describes the texts. Then we programmed some Perl scripts which converted the HTML files into valid XML files in which each single word is encoded with a <w>-tag. Linking the words to the corresponding dictionary entry electronically would have been possible now, but only up to a certain degree: it would not have been possible yet to link inflected words such as "Häuser" (houses) to "Haus" (house), and many ambiguities would have occurred (e.g. the Italian word "essere" represents a noun and a verb ("to be", and "the human being")).

POS-tagging the XML files Part-of-speech tagging is a computational linguistic technique to add morphological and syntactic information to each word of a text. Our XML files were POS-tagged at the *Institute for Computational Linguistics* of the University of Stuttgart by using the TreeTagger [144] developed by Helmut Schmid. Up to now the TreeTagger has been successfully used to tag German, English, French, Italian, Greek and old French texts. In the ELDIT texts the information added consists of the lemma ("Haus" for "Häuser") and the word class ("noun" for "Haus") for each word.

Manual Error Correction No POS-tagger works without errors. When electronically linking the words to the corresponding dictionary entries, we systematically collected and analyzed the words for which the program did not find a link. In this way we found some typing mistakes in the corpus and some tagging mistakes. These mistakes were corrected manually.

Adding Meta Information Then meta information was added to the texts: a unique ID for each single element for electronic searching purposes, the language of the text, the difficulty level of the exam the text is intended for, and domain and frequency information.

For the domain information, the texts were classified manually into 20 different interest domains. In this way particularly interesting texts can be recommended to each user if they match user interests recorded in the user model.

A frequency list has been calculated to have a kind of importance value for each text. The importance value of a text is the mean value of the frequency values of the significant words of the text:

Let $\mathcal{T}$ be the text corpus in ELDIT:

$$\mathcal{T} := \{T_j \mid T_j \text{ is a text of ELDIT}\}$$

Let w_i be a significant word (i.e. a noun, verb, adjective, or adverb) in a text T_j, and let $f(w_i)$ be the frequency of w_i in $\mathcal{T}$:

$$f_i := f(w_i) := \text{frequency of } w_i \text{ in } \mathcal{T}$$

Let n be the number of significant words in a text T:

$$n := \#\text{ significant words in } T$$

Then the frequency value g_j of a text T_j is defined to be the mean value of the frequency values f_i of the significant words w_i in the text T_j:

$$g_j := \frac{\sum_{i=1}^{n} f_i}{n}$$

The number g_j indicates whether a text is a general text or a text for special purposes: a text about family life will contain many words that are used very frequently

(such as “Mutter” (mother), “Schule” (school), etc.). Highly frequent words are easier to remember for a user and at the same time often needed. Hence such words and a text that contains such words can be considered important for a language learner. A text about aircraft carriers will contain much language for special purposes. These words are difficult to remember (because possibly encountered only in this text) and probably also less often needed. Hence such words and a corresponding text are possibly less important for a language learner.

Due to the calculation of the mean value the number g_j obtained is independent of the length of a text (otherwise a very long text about aircraft carriers would have a higher value than a very short text about family life).

Linking the Words After that each word has been equipped with unique information about its lemma and word form, links to the ELDIT dictionary can be created electronically. Links are not only created to the lemmas, but also to derivations, compound words, thematically related words and to structure words. Moreover proper names were linked to the lemma “Name/nome” (name), and words with a hyphen, such as “Ex-Freund” (ex-boyfriend) were linked to one of the words involved, in this case to the word “Freund” (friend/boyfriend). In this way, it was possible to achieve a very high number of links, namely for more than 90% of the text words.

Hints to Questions We are planning to electronically create hints for each question, namely by electronically searching significant words (i.e. nouns, verbs, adjectives, or adverbs) of the questions in the sentences of the text body and by marking corresponding questions and sentences. In cases, where this approach does not work, a manually supplied list of key words is used for searching. During runtime the sentences helping the learner answer a specific question are highlighted if the user requests them.

4.3 Quizzes and Questions

As stated in section 3.3.4 interactivity is considered the main positive feature that distinguishes traditional paper-based material from electronic material. In order to provide opportunities to interactively practice the information in ELDIT, the data collected for the dictionary and the text corpus are reused to create quizzes and questions. We want to emphasize that our detailed data model allows for the automatic generation of these quizzes without manual authoring. Moreover, by reusing some analysis modules that have been used for the data preparation process highly sophisticated correction possibilities and individual feedback can be provided. A small prototype of the quiz module has been implemented to test the ideas [72, 73].

4.3.1 Quizzes, Quiz Types and Quiz Groups

The following possibilities include only some straightforward examples that came to our mind, a detailed examination of the data and software tools combined with much creativity on the part of the developers would result in a more exhaustive variety of possibilities to allow the user to interactively practice the information provided by the system. An overview of the following analysis is given in Table 4.1.

CLASSIFICATION OF QUIZZES								
Quiz	Quiz Types				Quiz Groups			
	Gap.	Mul.	Msq.	Crs.	Perc.	Usg.	Cha.	Cont.
Matching	*	*	*	*	*		*	*
Direct questions	*	*	*	*	*			*
Morphology quizzes	*			*		*		
Syntax quizzes	*	*				*		

Table 4.1: Quizzes, quiz types, and quiz groups. The meaning of the abbreviations is as follows: Gap. = Gap-filling, Mul. = multiple choice, Msq. = magic square, Crs. = crossword puzzle; Perc. = training perception, Usg. = training usage, Cha. = training characteristics, Cont. = training context.

Quizzes

Different kinds of quizzes can be created of the given data. *Matching quizzes* are quizzes where the user has to match two representations of a word or word meaning, for instance, the lemma with a picture, the definition with a translation, a picture with a sound file, and so on. *Direct questions* such as "What is the opposite of...?", "Which of the following words is part of an apartment?" can be generated from the relations between words stored in the semantic fields. *Morphology quizzes* can be provided by asking the user to write a specific word form or by asking them to write the entire inflection paradigm into edit fields. *Syntax quizzes* can be provided by reusing the example sentences of verb valency: the user has to indicate the name of the corresponding sentence part (subject, verb, ...).

Quiz Types

The interaction process between user and system varies according to different quiz types: *Gap-filling* or fill-in-the-blank quizzes are text pieces in which some words have been replaced by an edit field. The user is asked to write the missing words into edit fields. *Multiple choice* quizzes are groups of options from which one or several of them have to be selected by clicking a check mark or radio button. *Magic squares* are squares consisting of letters within which the user has to search some given words. *Crossword puzzles* are puzzles which ask the user to fill in a magic square by finding the words described.

Quiz Groups

The quizzes can be grouped according to the support they provide for the single steps in the word acquisition process (described in section 3.2.4).

Quizzes for Perception Translation quizzes can be generated by providing the lemma, a derivation, a compound word, a definition, a pattern, a picture, a sound file, etc. and asking for a translation.

A gap-filling quiz can be created of a definition or example sentence by removing the lemma and asking the user to complete the sentence.

Multiple choice quizzes based on the semantic fields can be created. A group of words can be taken, then the user is asked to choose the word which does not fit into the group, or to complete the group with a word from a separate word list.

Direct questions generated from the semantic fields can be used to check whether the user understood relations and similarities between words.

Words that sound similarly such as “appartamento” (flat), “compartimento” (compartment), and “dipartimento” (department) can be searched by using the “wildcard feature” of the search engine and presented as a matching quiz.

Magic squares or crossword puzzles can be used to test the learners’ knowledge about a group of vocabulary. Definitions or translations can be provided, and the user has to find (magic squares) or fill in (crossword puzzles) the corresponding word.

Quizzes for Usage Gap-filling quizzes to train the inflection of single words or collocations can be created by removing the corresponding words and asking the user to complete the sentence.

The syntax of a sentence can be practiced on the example sentences of verb valency information (since we explicitly encoded the syntax for these example sentences).

Gap-filling inflection tables or crossword puzzles can be provided to practice the conjugation or declension of a word.

Quizzes for Characteristics “Characteristics” refers to the content of the footnotes in the ELDIT system. This information is semantically important for the learner, wherefore it is difficult to generate meaningful quizzes automatically. In some cases, however, it is possible to identify some emphasized words (e.g. in the case of a false friend or of given related words). In such cases showing the footnote together with a magic square that involves the emphasized words, might lead to a deeper awareness of the underlying linguistic characteristics.

Quizzes for Context An example sentence or a text of the learner corpus can be presented as a gap filling quiz.

4.3.2 Parameterized Quizzes

Many of the quizzes mentioned above can be parameterized, which further augments the number of possibilities. Let us suppose we have five (in general n) synonyms of the word *family*: “Familie, Klan, Sippe, Stamm, Verwandtschaft”. We would like to generate a multiple choice quiz showing three of these words (in general k) together with one randomly assigned word to the learner, and we would like to ask the learner to indicate the one that does not fit:

∘*Famile* ∘ *Klan* ∘ *Fenster* ∘ *Sippe*

It is possible to generate

$$\binom{n}{k} = \binom{5}{3} = 10$$

possibilities to choose a group of three words out of the five synonyms. There are four (in general $k + 1$) different places where the word that does not fit can be shown: at

the beginning of the list, somewhere in the middle, or at the end. Since we have about 20,000 different words for each language in ELDIT we have 20,000 possibilities to randomly assign a word that does not fit. Hence altogether we have

$$\binom{n}{k} \cdot (k+1) \cdot 20,000 = 10 \cdot 4 \cdot 20,000 = 800,000$$

different possibilities to generate this quiz.

In other cases we have many example sentences to illustrate a word, namely the original hand-made one and additional ones taken from other contexts (see section 5.1.1). In fact, for frequent words (e.g. “Schule”, “Jahr”, “gehen”, ...) several hundreds of example sentences exist in which the word is used.

4.3.3 Levels of Difficulty

Depending on the demands of a teacher or depending on adaptation parameters, the quizzes can be devised for different levels of difficulty. The level could also be changed if the user asks the system for help. We will now explain the generation of different levels of difficulty for the different quiz types:

Gap-filling Quizzes The removed words can be indicated below the sentence or not. If indicated below the sentence, they can be given in their original inflected form or in the citation form. The gap itself can be totally blank, show the first letter of the missing word, or show the translation in the mother tongue.

Multiple Choice Quizzes For the wrong possibilities a totally different word, a similar word (e.g. a word that starts or ends with the same letters as one of the original words), or even a quasi synonym given in the semantic fields may be chosen by the system. If the user asks for help, one of the items can be removed so that the number of possible solutions decreases.

Crossword Puzzles and Magic Squares The size of the square in which the words under consideration have to be searched for can vary. The description of the words can be given by translations (easiest), synonyms (more difficult), or definitions (most difficult). The first letter of the word in question can be indicated or not, etc.

4.3.4 Corrections

Since the system includes a large body of knowledge (definitions of words, information on inflection and spelling errors, synonyms, etc), it is possible to do much more than giving simple “correct/wrong” answers. In fact, it is possible to electronically identify several classes of errors and to give the corresponding feedback. We now list only the error classes for the gap-filling quizzes, since they can be adapted in a similar way to the other quiz types. Some of the correction possibilities are shown in Figure 4.9.

- The supplied word can be the original one, hence the solution is correct.

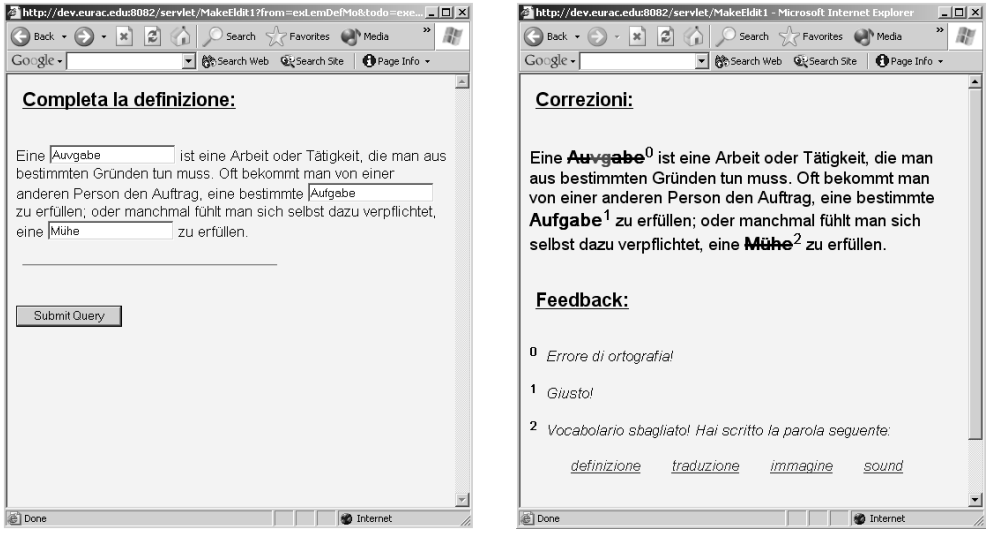


Figure 4.9: Screenshot of a quiz

- The supplied word can be a synonym to the original one, hence the solution is correct. Synonyms and quasi-synonyms can be identified thanks to the semantic fields added to ELDIT. An even better analysis could be carried out if the knowledge collected in different WordNet projects were to be included in the system (see section 5.2.3).

- The supplied word could be a valid word, but neither the original one nor a synonym. Thanks to the WMTransducers (see section 5.2.2) ELDIT can identify valid words. In this way a “wrong word” mistake is indicated. If the word is furthermore included in the dictionary, the system can give some information which describes the meaning (for instance: “You have written the following word: ...”). If the word is furthermore related to the meaning of the original word (which can be obtained from the semantic fields), the system can give indications of this relationship (for instance: “You used the opposite word!”).

- Due to the spell checker included (see section 5.2.1), spelling errors can be identified. By comparing the word used by the learner with the original word, the location of the error within the word can be indicated by the system.

- Due to the WMTransducers (see section 5.2.2), ELDIT can identify inflection mistakes.

- The last class of mistakes contains non-words (words that do not exist). Again the WMTransducers are able to identify such mistakes (see section 5.2.2).

4.3.5 Feedback

Feedback can be given on several levels. In this way a user can practice a quiz several times and approach the solution step by step:

- The system can indicate that there is a mistake somewhere.
- The system can indicate the location of the mistake.
- The system can correct the mistake.

4.3.6 Remediation

After a mistake, remedial information is fundamental to ensure that no wrong knowledge is embedded in the mental lexicon, and to eliminate misconceptions. Remedial quizzes can consist of getting the user to read the correct information as well as to practice new quizzes:

- A quiz similar to the original one can be generated by changing one or more of the parameters.
- The dictionary information giving detailed feedback on a mistake (e.g. vocabulary definitions or the declension of a word) can be shown to the user, and the user can be asked to study the information.
- The dictionary information giving detailed feedback on a mistake can be given in the form of a remedial quiz, and the user can be asked to practice this information.

4.3.7 Learning Activities

There are different ways of guiding a user through the practicing process. There could be users who only want to have a look at our system. Such users might wish to see all possibilities at once without being forced to go through the entire learning process. Also in autonomous learning, freedom may be appropriate, since we imagine the user to be rather motivated when working voluntarily with the system. However, in learning sessions in a course or a classroom some restrictions could be wanted by the teacher and a step-by-step guidance through the quiz might be welcome. Information overload could also be a problem which might make it necessary to follow an individual path through the learning material.

Hence we distinguish between different guidance modes: *Step-by-step guidance* is for concrete learning situations: the user first gets the plain quiz, then the quiz plus some hints, and, only then, the solution. *All-at-once guidance* is for situations where impatient users want to have all information at a glance. In this case the user gets the quiz together with buttons to request hints and the correct solution. For *adaptive guidance* the system considers parameters stored in the user model and possibly a concrete curriculum to generate an individual guidance mode.

4.4 Tandem

In section 4.3 we have described how we provide possibilities to interactively practice the information provided in ELDIT, and in section 4.2 how we provide opportunities

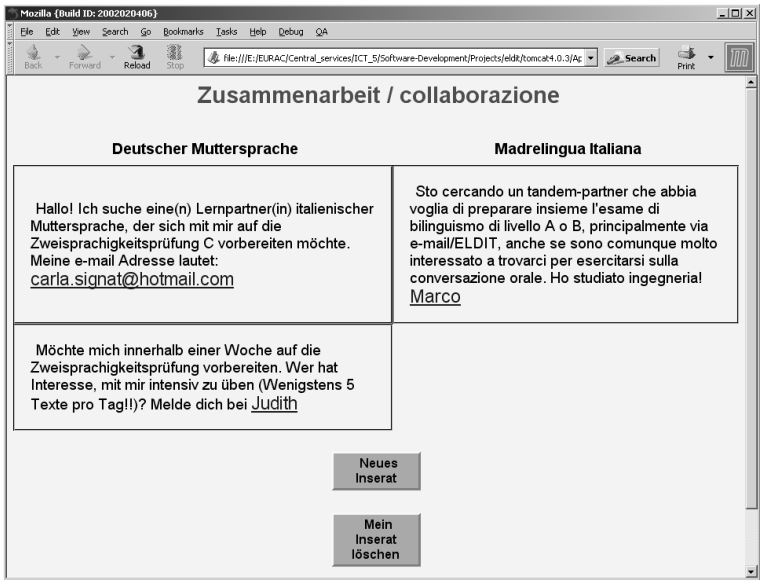


Figure 4.10: The advertisement page of the eTandem module

to apply language by writing entire sentences and small pieces of text. As the current version of ELDIT does not include a Parser and is hence not able to correct freely typed text, we are implementing an e-mail tandem module (eTandem) which is designed to bring Italian and German native speakers into contact with each other, so that they can form learning partnerships and correct each other. This module could also provide a great opportunity for the population in South Tyrol to overcome many of the problems mentioned in section 1.1.

The conceptualization of the eTandem module is concluded. The implementation was started in 2001 but had to be stopped after some time, since funds which had been promised were then withdrawn by the South Tyrolian local government. Hence the final implementation as well as the evaluation of the module will need to be carried out at some point in the future.

4.4.1 Tandem Learning

Learning a language via tandem means that two learners with different native languages try to learn the language of their partner by holding controlled conversations with each other. The learning progress is based on reciprocal dependence and mutual support. The partners teach each other, correct each other, and evaluate each other's progress. Usually, a teacher monitors the contact and provides guidance if required. Such contacts and conversations can also occur via electronic media such as e-mail, chat, or video conferencing. These media offer new possibilities to overcome problems like spatial and temporal distances.

4.4.2 The eTandem Module

The eTandem module in ELDIT offers a communication forum on the Web which allows the establishment of learning partnerships and exchanging information between already existing partnerships.

The very first step for each user is the registration. When the user logs in for the first time, an introductory page is shown which explains the use of the system. At the end of this introduction the user is asked to examine the advertisement board.

This board serves either to communicate interest in a learning partnership, or to contact somebody else who has already announced interest in such a partnership. The screenshot in Figure 4.10 shows an advertisement page. The advertisements are sorted according to the native language of the advertiser and the exam level. Furthermore, information about the preferred date of the exam or personal interests can be indicated.

Then the user can start to work by selecting a text, read it carefully, and answer the questions. As long as no partnership has been established, the texts can be saved as drafts and sent later.

A mail page provides a communication forum between learning partners. The answers are supposed to be sent to the learning partner for correction. Similarly, the messages from the learning partner are shown on this page. Furthermore, short messages, such as comments on concluded work or organizational aspects, can be exchanged in this forum. If the work is accompanied by a supervisor, communication with the supervisor is also carried out via this page.

4.5 Tutor

In order to guide the learner through the large amount of learning material, an electronic tutor might be advantageous which presents new items to be studied individually for each learner [71, 72].

As described in section 3.2.4, two different approaches to vocabulary acquisition can be distinguished. *Intentional* learning refers to an approach where the student explicitly learns a list of words and their translations. This form is faster, but can be superficial. *Incidental* learning refers to an approach where the student learns new words by extensive reading which generally results in a deeper embedding of the information in the mental lexicon but is slower. In ELDIT we want to explore a combination of intentional and incidental vocabulary learning and additionally exploit adaptation techniques. The method has been described in detail in section 3.3.4.

4.5.1 A Learning Scenario

The adaptive tutor described in the following has been conceptualized but not yet implemented in ELDIT. We now describe a possible learning scenario to explain our ideas in more detail.

The learner will first encounter the interface shown in Figure 4.11 which lists different word groups, different text groups, and a “next” link (“Ich möchte weiterarbeiten...”). The learner can now decide to study either words from a word group or a text from a text group, or let the system decide the best next step.

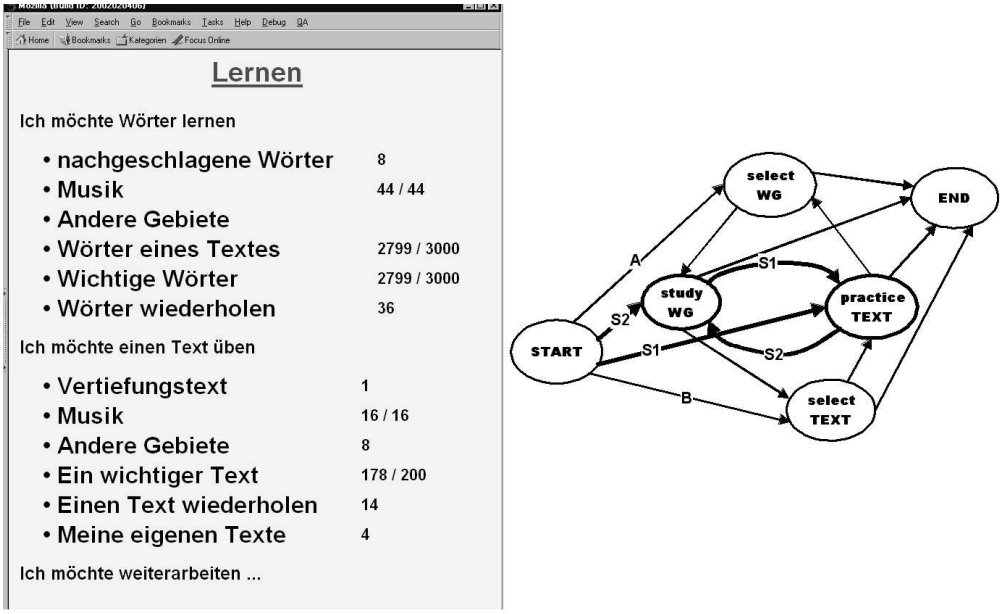


Figure 4.11: The learning scenario

Let us suppose the learner decides to study new words from a specific word group. They are studied by consulting the corresponding dictionary entries and carrying out the corresponding quizzes. The learner reads the word definitions, compares different word meanings and their synonyms, examines collocations and idiomatic expressions, and notices linguistic differences between the two languages. During this task the user model will be continuously updated with information about how well the user has done the quizzes.

Once these new words are sufficiently known, the learner returns to the main interface. Following the link “Ich möchte weiterarbeiten...”, the system now selects a suitable text to be practiced. The text questions contain as much as possible of the vocabulary just studied. In this way it is likely that the user will have to use these words when formulating the answers.

The user works through this text, i.e. reads it, checks unknown vocabulary, answers the questions, and sends the answers to the e-mail partner. The user model is updated to reflect the fact that this text has been practiced and will be refined when the learning partner gives a score for the answers of the corresponding questions.

When the questions of the text have been answered, the user returns to the main interface. Following again the link “Ich möchte weiterarbeiten...”, the system will now propose the next group of words. In general, the new group will contain those words from the text practiced in the last learning session which the learner looked up in the dictionary. Again the user should study these words and do the corresponding exercises, then practice a suitable new text, send it to the partner, etc.

In such a way vocabulary acquisition and application occurs alternately in a cyclic way. Figure 4.12 illustrates the approach. The cycle can always be interrupted and restarted. The learner can ignore the prepositions of the system and select new word

groups or texts – maybe to change the topic – and afterwards again take advantage of the system's capabilities to select next items.

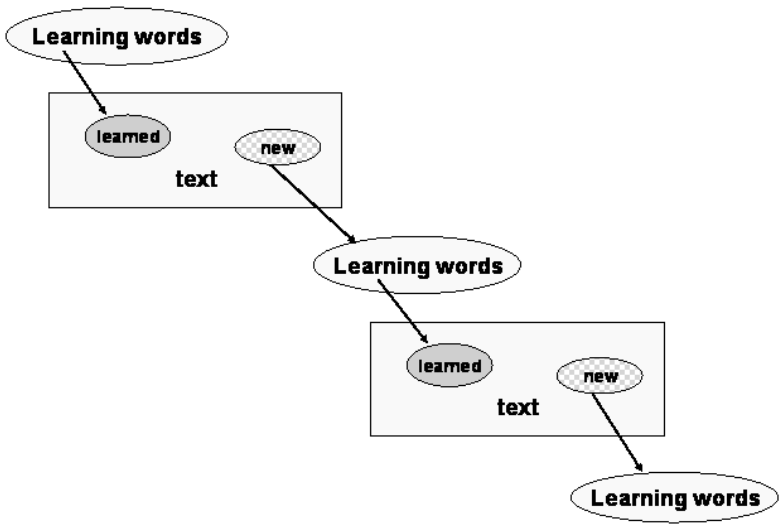


Figure 4.12: Contextualized, adaptive vocabulary acquisition

4.5.2 Ordering Words and Texts

Words and texts are quite independent of each other. Therefore, it is difficult to define prerequisites or "next best" items to be learned if the tutor finds equally well suited texts or words to be practiced next. One indication could be the frequency of words in a language or, as in our case, in the given text corpus. Our policy is to offer highly frequent words before low frequent words, and thus general vocabulary before more specialized vocabulary. We assume that the higher the frequency of a word is, the more common the word is and the easier to learn. Words with a low frequency are considered specialized vocabulary and less important. They are more difficult to learn, as they are rarely encountered. Similarly, we assume that a text with many frequent words is more useful for the general language knowledge and easier to practice than a text with a specialized and low frequency vocabulary. Thus, we have ordered words and texts from "more important" to "less important".

To implement this policy, we have ordered the words in the dictionary according to their frequency of occurrence in our text corpus, taking into consideration the citation form for counting. Based on this overall frequency list of words, we have been able to calculate an importance (frequency) value for the texts (see section 4.2.2).

Apart from these two frequency lists, words and texts are arranged in different, possibly overlapping groups. As for the words, the most comprehensive group contains all words of the ELDIT dictionary. Other groups only contain the words of a specific domain, e.g. vocabulary related to sports, music, or traveling. At the lowest level

of granularity we have a word group for each text which contains all words in the corresponding text. The texts are grouped in a similar way as the words. For example, we have groups with different levels of difficulty, or groups which contain all texts belonging to a specific domain.

4.5.3 The Tutoring Process

All interaction between the adaptive tutor and the learner occurs through the interface shown in Figure 4.11. At this point the learner has three possibilities: studying words, practicing a text, or asking the system for the next step. The state diagram in Figure 4.11 visualizes the different possibilities and their interaction, all of which is described in more detail below.

Studying Words

This option ("Ich möchte Wörter lernen") corresponds to the path A in the state diagram. Different groups are listed with adaptive annotations about how many of the words have already been worked through and about how well this work has been done. The user can select a word group, and the system helps to choose important (i.e. frequent and unknown) and interesting vocabulary (i.e. words of user specific interest domains). The following word groups are offered:

1. *Nachgeschlagene Wörter*: the words which the learner checked the last time when practicing a text;

2. *Sport, Musik, Reisen*: the words organized in different interest domains. Domains which correspond to the indicated user interests are highlighted and listed before the other ones.

3. *Wörter eines Textes*: some users might prefer to systematically study the words of a specific text before practicing this text. Hence all unknown words of the next best text are offered in this group;

4. *Wichtige Wörter*: the next group of important words which are still unknown to the user. The size of the group may depend on users learning speed or diligence.

5. *Wörter wiederholen*: repeating formerly studied items is very important for language learning and especially for vocabulary acquisition. If a user selects this option, the system will propose words to be repeated.

Practicing a Text

This option ("Ich möchte einen Text üben") corresponds to path B in the state diagram, where the user wants to start by practicing a text. Different groups of and options for texts are listed with indications about the number of not yet practiced texts in a text group. The user can select an option or a text group, and the system provides an interesting (i.e. a user specific interest domain) and relevant (i.e. high frequency value and unknown) text for this specific user. The following text groups are offered:

1. *Vertiefungstext*: a suitable text to practice the items of vocabulary acquired last, i.e. a text which contains as much as possible of the words last studied.

2. *Sport, Musik, Reisen*: the texts organized into interest domains like sports, music, traveling. Domains which correspond to the indicated user interests can again be highlighted or listed before the other ones.

3. *Ein wichtiger Text*: an important text with many frequent words which has not been practiced yet.

4. *Einen Text wiederholen*: if the user wants to repeat a text, some important texts are proposed in which the user performed badly or which was done long time ago.

5. *Meine eigenen Texte*: we also provide the user with the opportunity to practice an own text in the same way as the system texts.

Asking the System

This option ("Ich möchte weiterarbeiten") corresponds to the paths S1 and S2 in the state diagram and activates the adaptive, contextualized vocabulary acquisition process which alternates between learning new words and practicing a text. The process can be described as follows:

First, the last activity recorded in the user model is consulted: if the learner has last learned some words, the tutor suggests practicing a text. The decision is made as in *transition S1* described below. If the learner has last practiced a text, the tutor will recommend to learn some new words. The decision is made as in *transition S2* described below. If no last step can be determined (e.g. when the learner works with the system for the first time), the tutor proposes both possibilities and encourages the learner to choose the preferred starting activity.

In the *transition S1* the electronic tutor selects a text for the individual learner. The following decision strategy is applied:

- If the user has previously studied a group of words belonging to one specific text, the system selects this text.

- If the user has studied other words, a best matching text is selected which contains as many as possible of the words just studied (the words are primarily searched in the text questions, because it is likely that the user has to apply them when writing the answer) and many other well-known words, i.e. the option "Vertiefungstext" is applied.

- If no vocabulary has been studied previously (perhaps because the user is working with the system for the first time or because this user prefers to work with texts only), a new text or one which should be repeated is suggested, i.e. the options "Ein wichtiger Text" and "Einen Text wiederholen" are applied.

In the *transition S2* the system selects some new words to be studied next. The following decision strategy is applied:

- If the user checked some words the last time a text was worked through, these words are proposed for a more detailed study, i.e. the option “Nachgeschlagene Wörter” is applied.
- If no vocabulary was checked the last time (either because everything was known to the user or because the user is working with the system for the first time), some new words or words which should be repeated are proposed by the tutor, i.e. the options “Wichtige Wörter” and “Wörter wiederholen” are applied.

Learner Control

The learner always has access to all possibilities and is allowed to interrupt a cycle. This may occur if the user wants to change topic or wishes to carry out another activity. The system records these new actions, and, based on them, starts a new cycle the next time the user requests it.

Chapter 5

Extensions of the ELDIT System

ELDIT includes some special features that affect the entire system and make it innovative and outstanding from other educational systems. In section 5.1 we describe how we reused the manually developed content on several places in the system. In section 5.2 we report on the successful inclusion of some externally developed products such as Word Manager. Also ELDIT itself can be reused and combined with external applications, in section 5.3 we report on a project of this kind. The ELDIT system is able to adapt the information to each learner, hence, in section 5.4 we describe the user model which stores personal interests, goals and background information for each user.

5.1 Content Reuse

The content of the ELDIT system can be divided into two classes: educational content and illustrative content. The educational content defines the rules of a language, i.e. how to combine words in order to form correct sentences. It includes definitions, word usage patterns, collocations, idiomatic expressions, and translations. Illustrative content comprises the text sentences and lexicographic examples which are used to show language in use, i.e. how the rules can be applied to form correct and meaningful sentences.

All in all the ELDIT system contains a large amount of educational and illustrative content chunks: there are about 4,000 word entries for each language. Each word entry contains an average of five to ten derivations, compound words, and semantically related words. This amounts to approximately 20,000 words with translations for each language. Moreover, each word entry contains an average of ten to twenty collocations and idiomatic expressions. This yields approximately 45,000 usage patterns. Hence, all in all we have about *65,000 pieces of educational content chunks* in ELDIT.

For each dictionary entry about fifteen to twenty example sentences are listed, this amounts to about 50,000 lexicographic examples for each language. Text sentences (about 12,000 sentences) can also be used as example sentences. Hence, all in all there are about *62,000 pieces of illustrative content chunks* in the dictionary.

Both types of content can be searched in a highly sophisticated way and reused on several places throughout the system. First, we will describe the possibility of reusing the illustrative content to provide a large amount of example sentences for

each piece of information (see section 5.1.1). Then we will describe how we reused also the educational content to link each single word used anywhere in the system to the corresponding dictionary entry (see section 5.1.2). Moreover, in ELDIT these two modules are combined: even in the additional examples all words are linked to the corresponding dictionary entry, and for each entry description more example sentences can be requested, etc. The combination of these two modules will be described in section 5.1.3.

5.1.1 More Examples - Reusing the Illustrative Content

Approach

Since illustrative examples are of paramount importance for language learning and since it is very time-consuming to generate them, it is desirable to reuse the illustrative content as much as possible in different learning situations. ELDIT contains for each pattern, collocation, idiomatic expression, etc. an illustrative example. However, during an exhibition, where we presented and evaluated our system[1], we observed a strong need for more example sentences. Some users wished to see further examples which show the word in different contexts. Especially when learners had a specific context or sentence in mind, they wanted to see an example sentence which was rather close to their specific needs.

Therefore we now offer the possibility of accessing additional example sentences which are not explicitly provided for the collocation or definition under consideration but are retrieved from other dictionary entries.

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

- The patterns under consideration might be unstructured, i.e. they may contain meta information such as slashes and commas to indicate variations of a pattern, abbreviations like “etc.”, brackets and “...” signs. 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.

- 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 with the pattern.

- The recognition of collocations is another problem. It is not sufficient that all words of a pattern occur in a lexicographic example, but the words 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”. Hence the second example is not valid for the pattern “to go home”.

[1] http://www.fieralingue.it/cgi-bin/WebObjects/Fiera.woa/wa/Main

- 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”.

Implementation

For the reuse of illustrative content we place a small triangular button next to the standard example sentence of definitions, collocations, etc. These buttons can be used to dynamically retrieve additional illustrative examples in the current learning situation.

The dynamic retrieval of 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 fit the needs of our search engine. Three situations have to be distinguished:

a) Derivations, compounds, adverbs

b) Collocations and idiomatic expressions

c) Definitions and verb valency

In case a) it is easy to get such a “clean” pattern, since the given pattern consists of one single word which even occurs in the base form. This word can directly be passed to the search engine.

In case b) we have multiple word expressions. Here, the first step is to remove all meta-symbols, abbreviations, certain structure words, etc. from the pattern under consideration. Furthermore, a pattern may contain slashes or commas to indicate variations which can be unfolded into several patterns. For example, unfolding the pattern *ein Auto fährt schnell/langsam* (a car runs quickly/slowly) yields the patterns *ein Auto fährt schnell* (a car runs quickly) and *ein Auto fährt langsam* (a car runs slowly). Then the remaining words are connected with an AND function and marked as obligatory. The result for the car example would be 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). Neither the definitions nor the verb valency patterns are suitable for building a search expression. In both cases quite general words

are frequently used, for instance the words ”Gegenstand” (thing), ”irgendwo” (somewhere), ”jemand” (somebody), or ”something” (etwas). 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 to find examples that are appropriate for a specific word meaning.

Therefore, we build our search expression from some words which are taken from the main lexicographic example. We first extract the clause which contains the lemma 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 lemma. The resulting words are combined with the lemma by an AND operator.

Let us consider the following example for the definition of “house” as the group of people in a theater:

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

From this example sentence we generate the following search patterns: “Haus AND klatschen” and “Haus AND Beifall”.

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

Recognizing Collocations The next step is to recognize collocations. In the case of collocations the search pattern consists of several words and a kind of concordance tool is required to check whether a word combination forms a collocation in a sentence or not. Again we are using a rather simple approach based on some rules to identify real collocations. Typical rules are: the words of a pattern have to occur within one main sentence or within a subordinate clause. 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 the case of noun-adjective combinations. A more sophisticated disambiguation could be done by using a program such as “Phrase Manager” described in [131]. The inclusion of this system is one of our ideas for the future.

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.

This is a first step in the creation of a program able to perform meaning disambiguation. We are currently experimenting with the following approach: 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 zuging, klatschte das ganze Haus begeistert Beifall.* (When the performance was finished, the whole house applauded enthusiastically.)

could be the list of significant words (nouns, verbs, adjectives, adverbs) 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.

5.1.2 The Glossary - Reusing the Educational Content

Approach

Since the example sentences are given in the target language, they may contain new words which are unknown to the learner. This also holds for all other text pieces in the dictionary such as definitions and explanations. Hence we also want to reuse the educational content by linking each word used in the system with a short description of the word meaning.

For the automatic generation of these links we have to consider the following problems:

- We have encoded all our data in XML. Since each word will be linked to its explanatory information, the XML-files have to be broken down to the word level.
- Since a word in an example sentence usually occurs in a conjugated or declined form, we have to annotate each word with its lemma (base form). For example, the word *ging* (went) will be annotated with its lemma *gehen* (to go).
- Ambiguous word forms are another problem. For instance, the word *posto* can either be the noun "place" with the lemma *posto* or the verb "placed" with the citation form *porre*.
- A more difficult kind of ambiguity occurs at the semantic level. Our linguists respected lexicographic traditions for homonymous words and in some cases they created two dictionary entries for the same word form. For instance, there are two dictionary entries for the German word *Bank*, one for "bench" and one for "bank".

Implementation

For encoding the XML-files down to the word level we use a simple tokenizer. For adding the base form we used the WMTrans Lemmatizer described in section 5.2.2. The next step would be to use a POS-tagger (part-of-speech tagger) which disambiguates the POS (i.e. the word class) to each word. Currently we do not have a

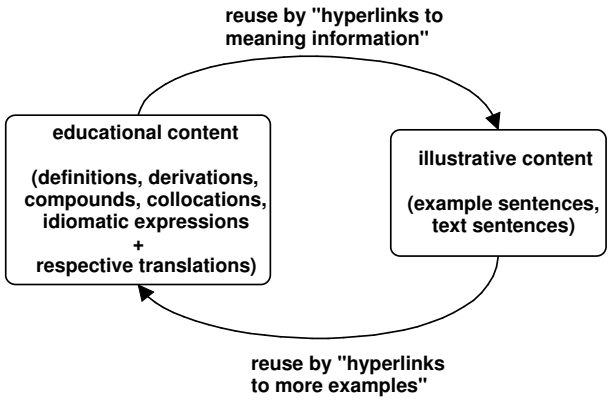


Figure 5.1: Reuse of educational and illustrative content in ELDIT.

sophisticated POS-tagger at our disposal yet. Instead, we use a simple set of rules like: there must be a noun or an adjective after an article, there must be a verb after an auxiliary verb, etc. Quite a lot of disambiguities can be solved in this way.

The last problem is at the semantic level, namely with homonymous words (words with more dictionary entries) and polysemous words (words with several meanings). A program able to perform meaning disambiguation is needed for this task. At the moment only the manual authoring of such links eliminates these ambiguities. When no manual authoring has been done we rely on learner autonomy and present several links in the first case and several meanings in the second case.

5.1.3 Combining these Modules in ELDIT

The previously described modules are combined in ELDIT. Additional examples can be retrieved for each piece of information. In the examples as well all words are linked to the corresponding dictionary entry. For each entry description again more examples can be requested, etc. Figure 5.1 illustrates the process.

Figure 5.2 shows a possible learning scenario in ELDIT. Suppose a user is searching the collocation *ein Haus kaufen* (to buy a house) and types the two words *Haus* and *kaufen* into the search field. The dictionary entry for the word *Haus* appears. The tab “Verwendung” (usage) is activated and the collocation is highlighted in red. The user reads the collocation, the translation, and the lexicographic example.

In our scenario the example sentence is not sufficient for the learner, therefore a click on the icon for more examples follows. A new window with additional example sentences opens. When studying these examples, our learner reads the unknown word *Schulden* (debts). A click on this word opens a new window which shows the word *Schuld* (lemma of *Schulden*), its Italian translation, and the indication that the word occurs within the dictionary entry *Entschuldigung*. Since this information is not sufficient for our learner, again a click on the icon for examples follows. A new window with many examples for the word *Schuld* opens.

The learner reads the example sentences and again encounters an unknown word, namely “Krankenhaus” (hospital). This word is in the basic vocabulary of ELDIT and hence a corresponding dictionary entry exists. Therefore the explanatory information of this word is more comprehensive and consists of the definition, the translation and

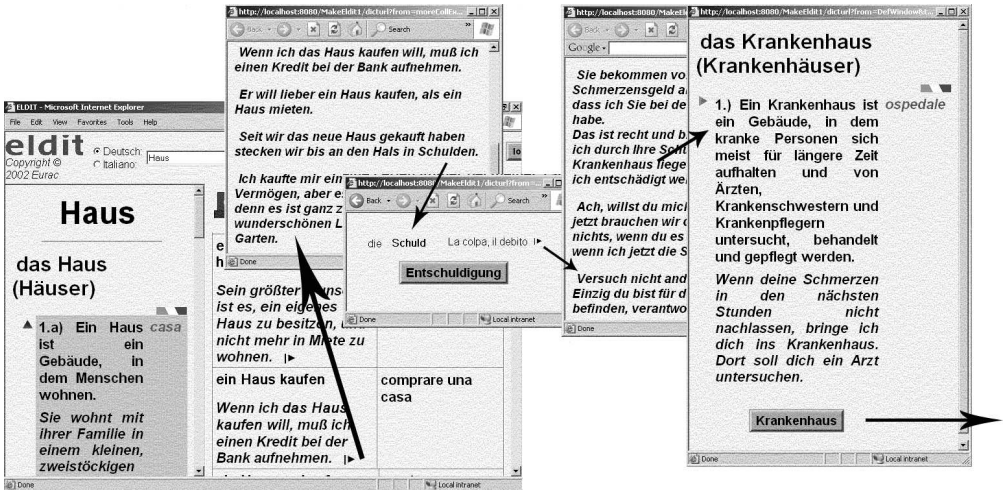


Figure 5.2: A possible learning scenario.

an example sentence. Now the learner wants to know more about the word *Krankenhaus* and clicks on the corresponding entry button. All windows with examples and explanatory information close and the original window shows the dictionary entry for the word *Krankenhaus*.

5.2 Inclusion of External Software

Some authors of non educational systems claim that it could be useful to integrated their systems into a language learning package [66, 105, 161]. In this section we will describe the possibilities of including externally developed products into ELDIT as we demanded in section 1.2. The first module we included is a self-developed search engine which has especially been designed for language learners (see section 5.2.1). Secondly, we will report on how we have successfully included Word Manager [160], a system for morphological dictionaries (see section 5.2.2). Thirdly, we will outline our ideas about including WordNet [64] as well into the ELDIT language learning system (see section 5.2.3).

5.2.1 The Search Engine

The search engine we implemented tries to address some problems language learners commonly encounter: it makes it possible to find single words and multi word expressions (collocations and idiomatic expressions), conjugated and declined word forms, and misspelled words. In the last case a hint is given by the system so that learners can discover their mistakes and eliminate them from the very beginning.

Searching Single and Multiple Words Single words can easily be found in ELDIT. The user simply types the word into a search field at top of the screen, indicates the language, presses the enter key, and gets the result. Speed, however, is not the only advantage electronic systems have over paper based dictionaries. This becomes clear when trying to find an expression consisting of more than one word, e.g. ”nach Hause

gehen" (to go home). In paper dictionaries the expression could be listed in the entry for the verb "gehen" (to go), for the noun "Haus" (house), or even for the preposition "nach". In the worst case the user has to check all three entries. In ELDIT the user simply types in the entire expression or just part of it.

Wildcards Sometimes a language learner might not know the correct spelling of a word. Let us consider the expression "ein Gebäude errichten" (to put up a building). There are several difficulties: "Gebäude" could be spelled with *eu* or with *äu*, "errichten" with one or with two *r*. In ELDIT a wildcard search allows the user to ignore problematic parts of the expression, e.g. searching "ein Geb*de e*ichten" results in "ein Gebäude errichten".

Spell Checking What happens if users are not even aware of their own spelling difficulties? ELDIT is able to indicate spelling errors, too. For example, "Schwiegermuter" (mother in law) typed with one t is wrong. A warning about the spelling error appears followed by the correct word "Schwiegermutter". ELDIT is able to detect up to two - sometimes even more - spelling mistakes per word.

Lemmatizing Further problems can arise if a user is sure about the spelling of a word, but has problems with grammar, e.g. the learner does not know the citation form of a word or has problems with inflected word forms within collocations. In ELDIT advanced help is offered. Searching for the verb "ging" yields the citation form "gehen" (to go). Searching for the words "Haus" and "kommen" yields the collocation "nach Hause kommen".

Structured Full-Text Search When searching the Internet usually a large number of useless results are encountered. One reason for this problem is the fact that most search engines perform a full-text search over the entire document. In ELDIT we have tried to avoid such useless search results by implementing a so-called structured full-text search. Each ELDIT entry consists of several fields (definitions, examples, idiomatic expressions, text titles, questions, ...). The search operations are restricted to single fields and the results are provided for each field separately.

Default Search and Extended Search We implemented two search modes: default search and extended search. In the default search mode the user types the desired expression into a text field and presses enter. The default search process is as follows: 1) the given words are searched in the indicated language, 2) if no results are obtained stemming of the words is performed and the search process is started again, 3) if there are still no results ELDIT supposes a mistake (language indication or spelling) and searches the words in the other language and checks its spelling.

However, the user does not depend on the system's decisions about how and where to search. Using the extended search mode explicit indications about the desired search features can be given: different word-connections, searching with or without wildcards, stemming, and simple or extended spell-check. Moreover, the user can indicate the fields to be searched, while the possibility of cross-field searching and a simple full-text search are provided as well.

Presenting Search Results Since all pages are generated on demand, it is possible to adapt the pages to different learning situations. When accessing a page after a search, the desired result is shown in red to attract the learner's attention to the corresponding part of the interface. The next time the user accesses this page e.g. by simply browsing, the same information is shown in black.

Implementation For the implementation of the search engine we have used the java-API Lucene[2]. Lucene makes it possible to index arbitrary documents.

The "normal" index contains the data in its original form. We have encoded our data in XML (see section 7.3), each XML element is added to the index as a separate document and can be referenced by its ID. In this way restricting the search request to single elements is easily feasible.

The "lemmas" index contains the data in the same form as the "normal" index, but transformed to their lemmatized form. A pattern such as "er baute Häuser" (he constructed houses) is first converted to "ich bauen Haus" (I construct house) and then stored in the index. On the one hand this makes it possible to find inflected queries (for instance "ging" (went) leads to "gehen" (to go)). On the other hand queries in the citation form which actually occur in an inflected form in the text can also be found (for instance "kommen Haus" (come home) leads to the collocation "nach Hause kommen" (coming home)).

Last, the "wildcard" index was created. This index contains the data in a transformed form. The word "Gebäude" (building) is transformed to "G e b ä u d e" which is a sentence consisting of seven words, each of which consists of just one letter. In this way it is possible to find the word "Geb*de" by searching the phrase "G e b" followed by the phrase "d e".

For a spell-check-search several wildcard-searches are performed. In each search cycle two consecutive letters are substituted by a wildcard and the resulting expression is searched for. For example, for the word "Gebeude" the following six expressions are searched for: (1) *beude, (2) G*eude, (3) Ge*ude, (4) Geb*de, (5) Gebe*e, (6) Gebeu*. In case (3) and (4) the error is eliminated and a valid result is found. In order to find more than one error several wildcards are inserted at all possible places. This is an efficient way to search our data, however, in most cases too many results are found. Thus, a post-processing is performed, in which the set of proposed solutions is diminished by comparing the numbers of letters of the original search expression and of the result. The difference must not be more than two letters. If this rule holds, the result is proposed to the user.

Use of the Search Engine The search engine is used both in the CALL system and in the authoring tool for the linguists and language teachers. Search operations in the authoring tool include the elaboration of the link structure. Search operations in the CALL system are carried out when users request word entries, or when the "more example feature" is invoked.

[2] http://jakarta.apache.org/lucene/docs/index.html

5.2.2 ELDIT and Word Manager

In the ELDIT dictionary we present the entire inflection paradigm (declension and conjugation) for each single word. This is possible thanks to the co-operation between the European Academy Bozen/Bolzano and the Scuola Universitaria Professionale della Svizzera Italiana involving Word Manager and ELDIT[3] [99, 133]. Word Manager is a system for reusable morphological dictionaries[4]. The underlying technology is described in [54]. Word Manager's lexical resources include morphological and orthographic knowledge for three languages, namely English, German and Italian[5]. ELDIT could profit from the Word Manager components in several ways. Hence a co-operation was started in which the two systems were joined. The flexibility of ELDIT, the reusability of the Word Manager modules (WMTrans modules), and the resulting creation of an innovative language learning feature were achieved. This work is an outstanding example for the fact that it is possible to include externally developed modules into ELDIT to enlarge the system in a rapid and effective way [99, 133].

Word Manager - a System for Reusable Lexical Databases

In Word Manager lexical resources are developed at two stages. At the first stage the morphological system of a language is described. This results in a morphological rule database, in a description of the inflectional classes, and in the word formation processes of the language. At the second stage lexemes are entered by classifying them in terms of the morphological rule database. The result is a reusable lexicon database. On the basis of this lexicon database specialized tools for any individual task can be developed using the knowledge made available in the database. The general architecture and the rule types involved are described in [160]. So far large lexicon databases have been developed for German, English, and Italian.

One of the most straightforward applications of Word Manager based tools is the recognition of words in unseen text. In this context the inflection component is used for the analysis of word forms for which the lexemes are in the database. The word formation component is used for unseen words formed on the basis of lexemes in the database. In [161] another application of Word Manager databases is sketched. In that application they are used to learn inflection and word formation rules in second language acquisition. Thanks to the flexibility of the presentation of information Word Manager databases are ideally suited for the exploration of morphology by an independent learner. Without a considerable amount of guidance, however, the richness of information can only be exploited by expert learners. This means that only the development of specialized tools and their embedding in a more general learning environment can bring out the full potential of Word Manager databases for language learning.

Word Manager for ELDIT

Due to the inclusion of the Word Manager databases and their operational modules in ELDIT we were able to provide several new features: an additional tab "declension"

[3] This co-operation has been financially supported by the European Union - Interreg IIIA

[4] http://www.unibas.ch/LIlab/projects/wordmanager/wordmanager.html

[5] For a demo in German see http://www.canoo.net

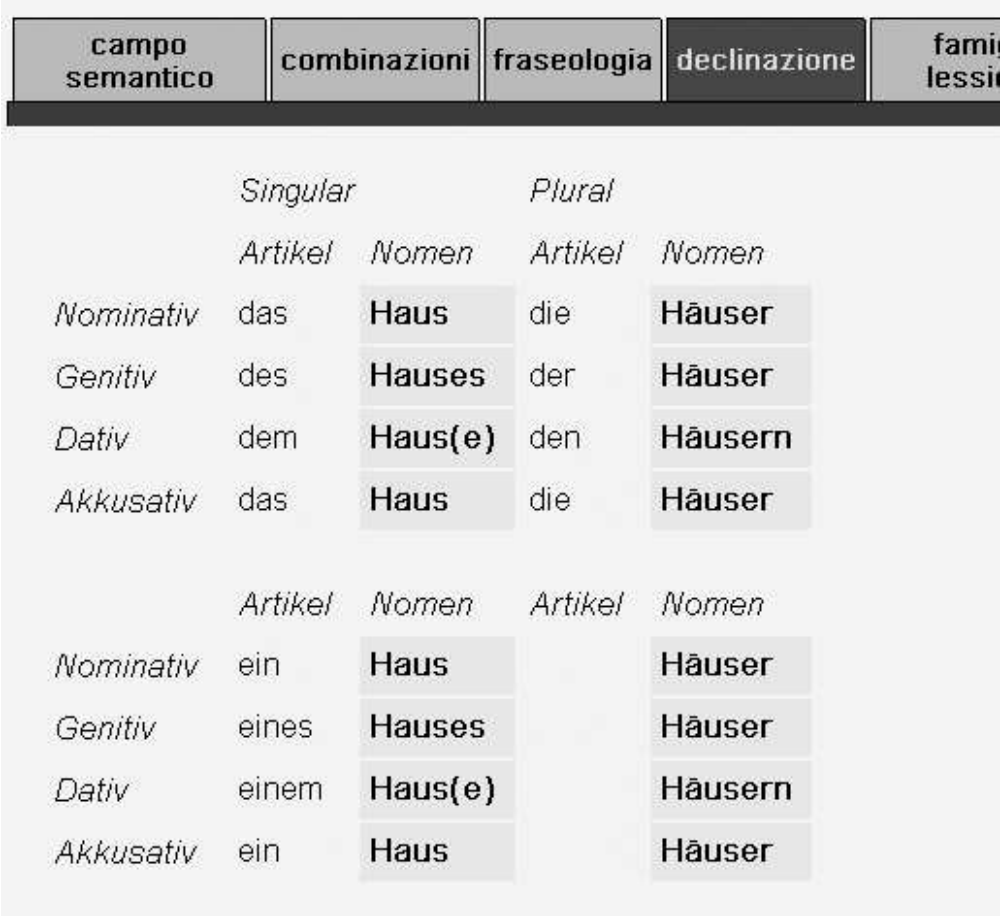


Figure 5.3: Describing inflection in ELDIT

or "conjugation" has been added to each word (see Figure 5.3). Word families (compounds and derivations) are shown to teach word formation processes. Hyperlinks from each word used in the system to the corresponding dictionary entry have been added electronically, the stemmer and spell checker of the search engine have been improved, and quiz and exercise data submitted by the user can be morphologically analyzed.

WMTrans Finite State Transducers

The previously mentioned tasks were achieved by integrating so-called WMTrans inflection modules into ELDIT. In order to deploy Word Manager dictionary databases in the context of complex applications the Word Manager lexicon entries have been compiled into finite-state transducers (WMTrans). The production of the Transducers is carried out by means of the WMTrans Finite State Transducer Framework [132]. On the basis of a specification of the target transducer's input and output the corresponding transducer is generated automatically. Up to now a whole range of WMTrans products have been developed[6]: WMTrans Word Recognizers, Lemmatizers, Analyzers and

[6]Demo versions are available at http://www.canoo.com/wmtrans

Generators for inflection and word formation [33]. Each product type is available for different vocabulary sizes and processing volumes. Typical uses of WMTransducers include their integration into search engines, text indexing and text mining applications, language learning systems, word stemming and hyperlink generation programs, spell and grammar checking tools, and machine translation applications.

The software currently integrated into ELDIT contains the first four products, the other ones will be integrated in the future:

- The *WMTrans Recognizer* is a program which is able to recognize any valid word, be it inflected or in citation form. The result of a query is yes/no, in form of 1/0 or true/false.

```
query   -> ging
result  -> 1

query   -> gingxyz
result  -> 0
```

- The *WMTrans Lemmatizer* is a program that 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):

```
query  -> ging
result -> gehen (Cat V)
```

- The *WMTrans Inflection Analyzer* returns the citation form and morphosyntactic classification of any valid word. A query result provides a list of citation forms followed by a list of morphosyntactic features related to the analyzed word form.

```
query  -> ging
result -> gehen
          (Cat V)(Aux sein)(Mod Ind)(Temp Impf)(Pers 1st)(Num SG),
          (Cat V)(Aux sein)(Mod Ind)(Temp Impf)(Pers 3rd)(Num SG)
```

- The *WMTrans Inflection Generator* is a program that returns all inflection paradigm word forms, starting from a valid citation form. The result of a query is a list of word forms followed by a list of morphosyntactic features related to each single word form.

```
query   -> haus
result  -> häuser
             (Cat N)(Gender N)(Num PL)(Case Nom),
             (Cat N)(Gender N)(Num PL)(Case Gen),
             (Cat N)(Gender N)(Num PL)(Case Acc)
           häusern
             (Cat N)(Gender N)(Num PL)(Case Dat)
           haeuser
             (Cat N)(Gender N)(Num PL)(Case Nom)(Flach auml),
```

```
            (Cat N)(Gender N)(Num PL)(Case Gen)(Flach auml),
            (Cat N)(Gender N)(Num PL)(Case Acc)(Flach auml)
          haeusern
            (Cat N)(Gender N)(Num PL)(Case Dat)(Flach auml)
          haus
            (Cat N)(Gender N)(Num SG)(Case Nom),
            (Cat N)(Gender N)(Num SG)(Case Dat),
            (Cat N)(Gender N)(Num SG)(Case Acc),
          hause
            (Cat N)(Gender N)(Num SG)(Case Dat)
          hauses
            (Cat N)(Gender N)(Num SG)(Case Gen)
```

- The *WMTrans Word Formation Analyzer* analyzes the first level of word formation history for any legal lexeme. The result of an analysis query is a list of source lexemes from which the given lexeme derives.

```
query   -> kennenlernen
result  -> kennen
             (Cat V)(Aux haben)
           lernen
             (Cat V)(Aux haben)
```

- The *WMTrans Word Formation Generator* generates the first level of word formation history for any legal lexeme. The result of a generation query is a list of derivated lexemes created by derivation and word formation.

```
query   -> bosco
result  -> abbracciabosco
             (Cat N)(Gen M)
           boscaglia
             (Cat N)(Gen F)
           boscaiolo
             (Cat N)(Gen M)
           boschetto
             (Cat N)(Gen M)
           boschivo
             (Cat Adj)(Manner Qual)
           boscoso
             (Cat Adj)(Manner Qual)
```

One product line has been realized (currently) only for German: the so-called *Unknown Word Analyzer* products. Because of the high productivity of German word formation rules a lexicon cannot contain all possible entries, since many words are “ad hoc” combinations of existing words. This has the effect that a traditional lemmatizer or a spell checker recognizes only a part of the words in a German text. In order to increase the rate of recognized words the “unknown word” analyzer was developed. By using Word Manager word formation rules [160] this analyzer not only recognizes words contained in the lexicon (lexicalized words) but also “potential” words, i.e. correct German words produced through generative word formation. This kind of analysis has shown an improvement of the word recognition rates of up to 20%, measured for generic German text corpora.

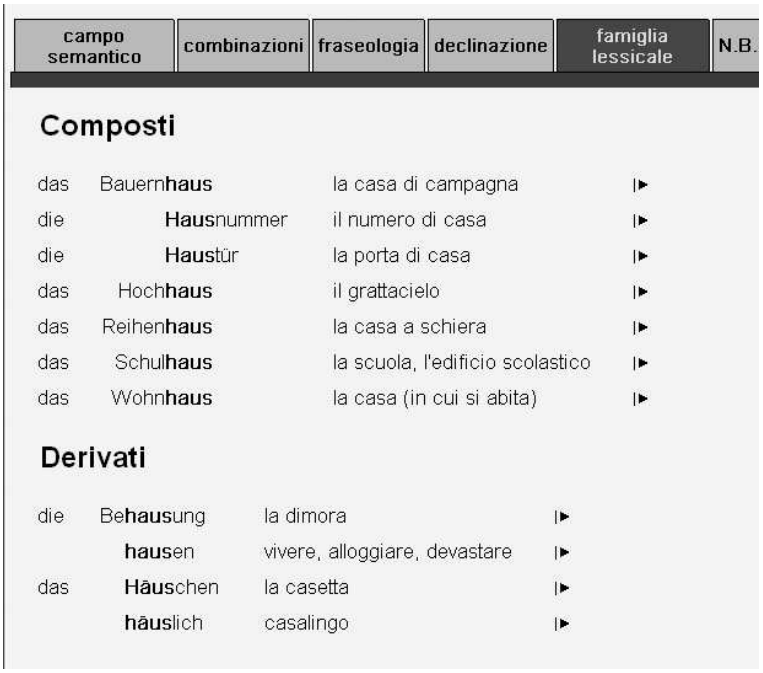


Figure 5.4: Describing word formation in ELDIT

Inflection Issues in ELDIT

For ELDIT we needed a WMTrans Lemmatizer to allow navigating from each word form to its corresponding citation form and any relevant explanation found in the dictionary. This navigation is carried out in search operations as well as when establishing the link structure.

In order to integrate explicit information on inflection we also needed the WMTrans Inflection Generator which is able to generate the whole inflection paradigm starting from a citation form (see Figure 4.5 and Figure 5.3).

In the future we will add a module to generate lexicon exercises. In order to be able to give meaningful feedback to the user, we will need a WMTrans Analyzer and an WMTrans Unknown Word Recognizer.

Word Formation Families in ELDIT

In order to extend the derivation families of ELDIT, we could make use of the WMTrans Word Formation Analyzer. Derivations suggested by a WMTrans Word Formation Analyzer are inspected by the linguists and eventually added to the existing list (see Figure 4.6 and Figure 5.4). In this way we come closer to a self-contained system, i.e. a system in which all educational information needed for the glossary is contained as head word or as element in a word formation list.

5.2.3 ELDIT and WordNet

In this section we describe our idea how WordNet could be integrated into the ELDIT system. Princeton WordNet [64] as the first in a range of wordnet constructions is an

online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each of which represents one underlying lexical concept. Different relations link the synonym sets.

The aim of the EuroWordNet[7] project was to develop a multi-lingual database with wordnets for several European languages which can be used to improve recall of queries via semantically linked variants in any of these languages. Wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian) have been created in the same way as the Princeton WordNet. However, each wordnet represents a unique language-internal system of lexicalizations. In addition, the wordnets are linked to an Inter-Lingual-Index. Via this index the languages are interconnected so that it is possible to go from the words in one language to similar words in one of the other languages.

Within the EuroWordNet project German[8] and Italian[9] wordnets have also been developed. We are currently using the Italian and also the English wordnets as reference tools for the development of the semantic fields. However, for several reasons it would be very interesting to include them into ELDIT:

- More semantic relationships between words could be shown to the learner.
- More detailed feedback could be given for quizzes, for example, when no semantic field has been elaborated.
- More links to the dictionary could be established, for example, if a word is not in the dictionary, the link could be set to a synonym of it.
- A wordnet could help to find more appropriate additional example sentences, especially in these cases, in which the current module does not work very well (example sentences for definitions and verb valency).

The integration of WordNet into ELDIT could be done by providing a java API to access the data or the WordNet data itself in XML format. This last format exists already for GermaNet [105]. We have already established first contacts with the authors of GermaNet and Italian EuroWordNet. However, a systematic co-operation is a plan for the future.

5.3 Interface for External Applications

All modules in ELDIT can be reused outside the system. This has successfully been demonstrated in another educational project carried out at the European Academy of Bolzano [153].

[7] http://www.illc.uva.nl/EuroWordNet/
[8] http://www.sfs.uni-tuebingen.de/lsd/
[9] http://wnit.ilc.cnr.it/

Integration of ELDIT and BISTRO

Our goal was to combine content modules and procedural modules of different projects for the creation of a new learning environment. Presupposition was the modular concept of both ELDIT[10] and BISTRO[11] (a terminological database developed at the European Academy Bozen/Bolzano). Each module of the two systems can be accessed independently.

GYMN@ZILLA

Figure 5.5: Gap-creation and annotation feature in GYMN@ZILLA

The result of our work was GYMN@ZILLA[12], a browser-like program exploiting Perl and CGI technologies, that allows freely browsing the Internet while dynamically annotating text with open learning resources, e.g. with a dictionary providing definitions, translations and explanations for unknown words. The browser integrates dictionaries such as ELDIT as content modules and the term extraction tool of BISTRO as procedural module. It has the following interesting aspects:

The links which are contained in a text can be followed with the processing mode maintained for the following texts: if a user chooses the option to annotate an Internet page with ELDIT annotations, this operation will be applied automatically to the following pages which are accessed through the browser unless specified differently.

A language recognition feature allows guessing the language of the accessed document, which makes proper processing possible.

A term extraction feature allows determining the words that should be annotated. This feature is used for the creation of the glossary and might additionally be useful if

[10] http://www.eurac.edu/eldit
[11] http://www.eurac.edu/bistro
[12] http://www.eurac.edu/gymnazilla

annotations in picture form or video clips (e.g. for showing the term in sign language) are added.

The terms inspected by the user are extracted and can be viewed separately. They form a special glossary which may be printed or read by text-to-speech programs. The term's translation or other information (grammar, morphology) is also collected in this glossary.

A quiz generation tool dynamically creates various types of completion tasks. With a technically very simple style variation the software substitutes the terms and dynamically creates a gap-filling quiz. The correct answer is linked.

By combining several options a text can be presented as a gap-filling quiz while other words are annotated with dictionary entries, etc. A screenshot of such a combined approach is shown in Figure 5.5.

Currently the languages German, Italian, English, French, Ladin (a Romanic minority language spoken in some valleys in South Tyrol and in other parts of the alps), Russian, Chinese and the American Sign Language are supported. It is very easy to extend the system to other languages by simply adding a corresponding dictionary and possibly some stemming rules.

5.4 Customization and Adaptation

The ELDIT system contains a large amount of information for each word and each text. Moreover this information can be reused for quizzes and questions in different ways. Collaboration activities can be carried out via the eTandem module. Learning with the adaptive tutor or with a human teacher is possible. Concerning information, different language levels as well as different interest domains are supported.

Evaluations of our system revealed that different users are interested in different aspects of the system and need different content [4]. These problems can be tackled by adapting content and presentation of the system to the individual user [69]. In ELDIT we distinguish adaptable and adaptive features. Adaptable features are personalization features that allow for the manual, a-priori customization of the system. These features will be described in section 5.4.1. Adaptive features are features that allow for adaptation of the system according to user observations. These features will be described in section 5.4.2.

5.4.1 Customization

When a learner registers the first time for ELDIT, the system requests some personal information: native language, proficiency in the target language and in the use of the Internet, the level of the exams in bilingualism the user is interested in (if applicable), professional background, personal interests, etc. These pieces of information are used to customize the system to the user's preferences and needs. Table 5.1 summarizes the customization features. They have already been implemented, but due to lack of data they are not enabled in the online version of the system yet.

Model The user can choose between a crosslingual and a monolingual version of ELDIT. In the crosslingual version the user interface (menu entries, labels on buttons,

etc.) and explanations such as linguistic differences between the two languages appear in the native language of the learner. In the monolingual version the user interface as well as the explanations appear in the language of the word entry itself and there are no translation equivalents given in the other language.

Domain ELDIT makes extensive use of so-called lexicographic examples to show various aspects of the language. These examples can be adapted to the user's professional background. Up to now only a few exemplary words have been elaborated with examples of different domains. They contain different lexicographic examples for users with a general, medical, or technical background. Another idea is to classify the existing examples of the large example pot in ELDIT electronically into corresponding groups.

Proficiency Depending on the language skills of the user (indicated directly or indirectly by the level of the exams in bilingualism) more or less detailed information on the words can be given. For example, to avoid information overload detailed differences between the word meanings or complicated idiomatic expressions can be hidden for beginners.

Goal ELDIT is based on the vocabulary and the texts which are used for the exams in bilingualism in South Tyrol. These resources are completed by a large number of additional information. Depending on the learning goal a student might be interested in this additional information or not. If a student wants to prepare as quickly as possible for the exams in bilingualism only the core information which is relevant for this exam is shown.

Certainly not all parts of the advertisement page are interesting for each user. Thus, when a German native speaker inspects the advertisement page, the advertisements of Italian native speakers will be shown and vice versa. Similarly, only the advertisements of the users who registered for the same exam level are shown. Users can also be joined by personal interests or similar learning preferences.

Help Another feature is the customization of the content of the help files. Depending on the familiarity with the use of computers ("novice" or "familiar") the online help can be more or less detailed. Moreover, for a novice the dictionary entries contain several direct links to the corresponding help sections.

Feature	Choice of user	Affected elements
model	monolingual, crosslingual	translations, labels, explanations
domain	general, medical, technical	lexicographic examples
proficiency	beginner, advanced	meanings, idiomatic expressions
goal	language, exam	all elements
help	novice, familiar	content of and links to help file
annotation	—	annotation content
text questions	—	answers' content

Table 5.1: Customizable features in ELDIT.

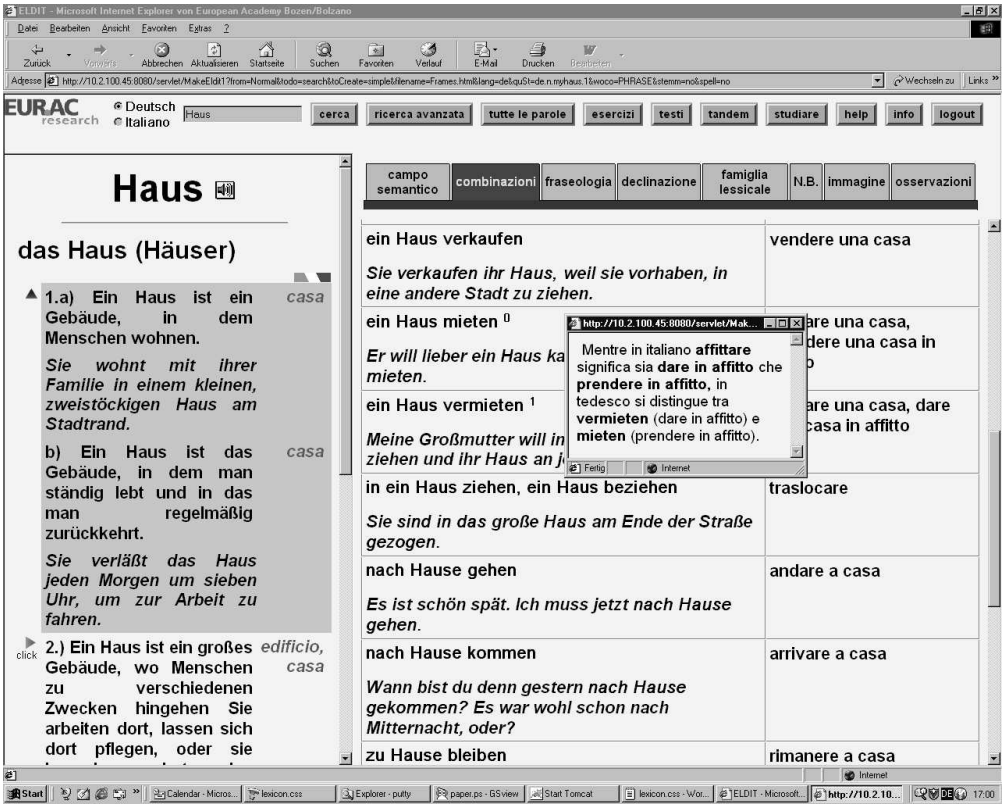


Figure 5.6: Default presentation of the German word ”Haus”.

Annotations and Answers Finally, the learner has the possibility of saving personal annotations for each word entry and the answer to the questions for each text.

An example of customization is given in Figure 5.7 which shows the screenshot for the German word “Haus” for a typical beginner. The information presented is different from the one in Figure 5.6 which shows the default presentation of the same word. In the customized version the description of the word meanings on the left hand side is less detailed and only the most important meanings are presented. Since the user prefers a monolingual dictionary, no translations are shown and the menu and the footnotes are in the same language as the dictionary entry. The learner’s professional background is medicine, hence the lexicographic examples contain an increased number of words from the medical domain.

5.4.2 Adaptation

We will now concentrate on the capabilities of the system to observe the user, to store this information in the user model, and to adapt the system accordingly. However, these features have not been fully implemented yet. Table 5.2 summarizes the adaptive features and the elements which would be affected.

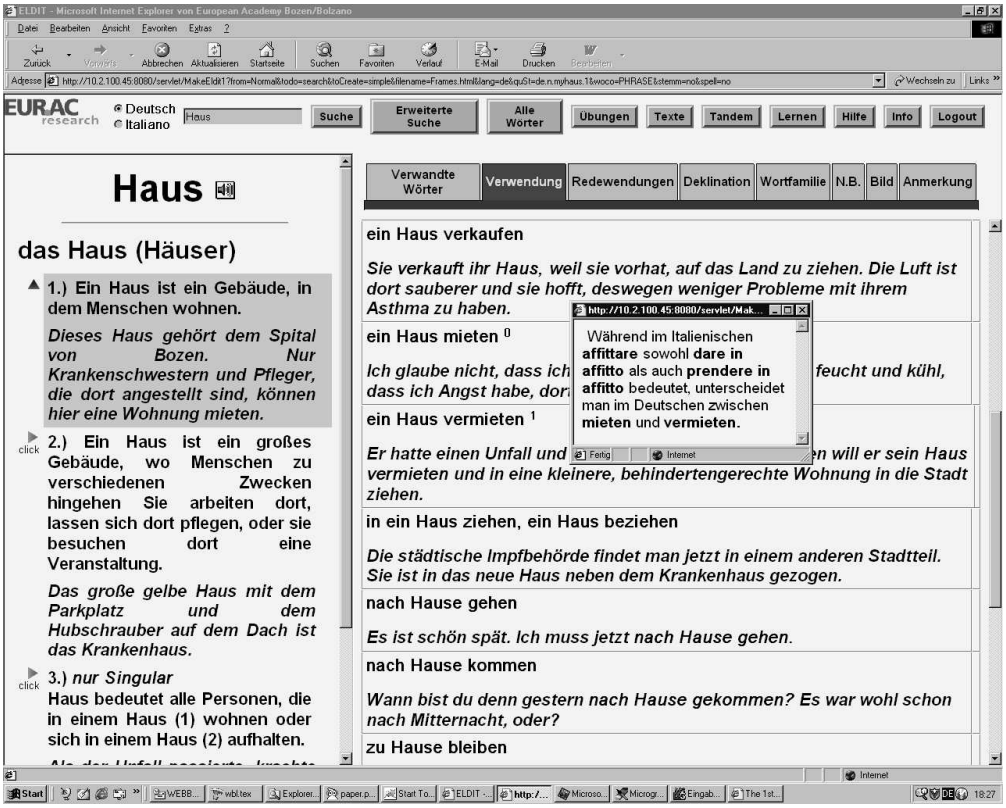


Figure 5.7: Customized screenshot for the German word ”Haus”.

Introductory Page When a user enters the ELDIT dictionary for the first time, a general introduction is provided which explains in detail how the system can be used. In the following sessions this introductory page is replaced by a link to the page.

Interface An evaluation of the system revealed that additional hints are required to improve the user interface, e.g. direct links to the specific help sections or labels which indicate how to interact with the system. These labels and links have been added to the system but will incrementally be removed when the user has read the corresponding help sections or has more experience with the system.

Content Activation If a user shows preference for a specific piece of information, the system can automatically activate the corresponding part. For example, if a user always listens to the pronunciation of a word the system will automatically play the sound file when a new word is accessed. Similarly, if the user always looks at the pictures they might be shown automatically when the user selects a specific word meaning.

Interactivity Due to the large number of quiz types, quizzes, feedback types, remediation types, etc. in the quiz module it might be necessary to adapt the material to

Feature	Observation	Affected elements
introductory page	first or subsequent login	introduction page
interface	experience with system	interface, labels, links to help
content activation	information preferences	info activated automatically
interactivity	background knowledge	quiz and quiz type selection
help	working history	reminding messages
tutor	learning history	next best items to be learned

Table 5.2: Adaptable features in ELDIT.

different levels of difficulty and the feedback and next best quizzes to different learning situations or user backgrounds:

- Novices will get the quizzes on an easier level, experts will get more difficult quizzes.
- Interest domains or known words which are saved in the user model can be considered when e.g. choosing an example sentence of which a gap-filling-quiz is created.
- The stage within one lesson can be considered, i.e. when the user starts working on a quiz, the pure quiz is shown, in further attempts the level of difficulty could decrease while explanations and feedback could become more detailed.
- Past and frequently made mistakes can be collected in the user model and system choices can be adapted to this knowledge.

Help Feature Since the user has the possibility of doing many different things with the system an adaptive help feature might be necessary. By considering the user's working history this help feature could suggest next best working steps to be carried out. Such working steps are mainly combined with the tandem module, since there the user is not only a learner but also the teacher of his or her tandem partner. Hence, advertisements have to be placed, incoming messages have to be read, the work of the partner has to be corrected, own learning has to be done, etc. An electronic help feature which keeps track of these tasks and reminds the user what to do next might be very helpful.

Adaptive Tutor Finally, working with the dictionary, practicing the presented vocabulary information, and applying the words on texts should be coordinated by an adaptive tutor that keeps track of the users learning history, gives suggestions about next best items to be learned, and indications about past performance.

Chapter 6

A New Approach to CALL Content Authoring

The authors of the system (linguists and language teachers) asked the software developers for the implementation of very innovative but complicated features. Approaching the implementation of these features in the usual way would have required the manual input of a large amount of data and a very detailed encoding of these data (see the DTD descriptions in appendix B). This extensive and detailed encoding was judged as not feasible by the authors. Moreover, due to the fact that the linguists did research in their conceptualization as well, a complete system specification was never available, but specification, development, evaluation, and adaptations happened alternating. Thus, the demands – and consequently also the data model and the software – underwent frequent changes. For this reason the system was required to be highly flexible, so that it could easily be extended or changed.

In section 6.1 we summarize the considerations that led to our approach to data and software management. In section 6.2 we give an overview of the method. In section 6.3 we will describe in detail the tools we have developed. For further details we refer to the general system description in chapter 4. The concrete implementation will be described in chapter 7.

6.1 Motivation

A modular approach to software development is essential for the realization of the previously described system. Furthermore, we propose an approach to CALL data management which according to the author's knowledge has never been applied before. Departing from semi-structured manually elaborated educational data, a second data version which is uniquely rich in information is generated electronically. In fact, the core consideration in our approach is *the encoding of the data in a sufficiently extensive and detailed way*. "Encoding the data in a sufficiently detailed way" means that the data is encoded *down to the level of single words and further*. "Encoding the data in a sufficiently extensive way" means that *as much information as possible is added to each individual element, word and word fragment*.

With data and software modules prepared in this way it is possible to fulfill the following demands:

- Sophisticated linguistic and didactic requirements regarding data presentation can be fulfilled more easily (e.g. “on an on-mouse-over-event the corresponding words of patterns and example sentences should be highlighted at the same time and in a specific color that indicates the phrase of a sentence”, or “each word used in the system should be linked to the corresponding dictionary entry”).

- Data and software reuse up to a high degree becomes easier (e.g. “text sentences should be used for reading, as example sentences illustrating language applications, as automatically correctable fill-in-the-blank exercises, etc.”, or “the spell checker should be used for the data generation process, to propose an alternative if a user tries to search a misspelled word, for the correction of the exercises, etc.”)

- The inclusion of external modules even if they work only on a very specific language feature is supported (e.g. “clicking on the prefix of a derivation should lead to a grammar section that provides an overview of word formation possibilities with this prefix”, or “clicking on a verb should open a window in which the entire conjugation of this verb is shown”).

Of course, these encoding requirements cannot be realized by humans, least of all if we are speaking about a real world system which includes thousands of documents and concepts. Hence electronic tools have to be developed to support the manual generation process of the educational content. These tools are not only used for content preparation but also for the system itself, either for analyzing answers submitted by the learner or for online authoring of new educational material by a teacher.

6.2 Overview

In the ELDIT project the data encoding and software development process occurs in several steps. The general approach has been visualized in Figure 6.1.

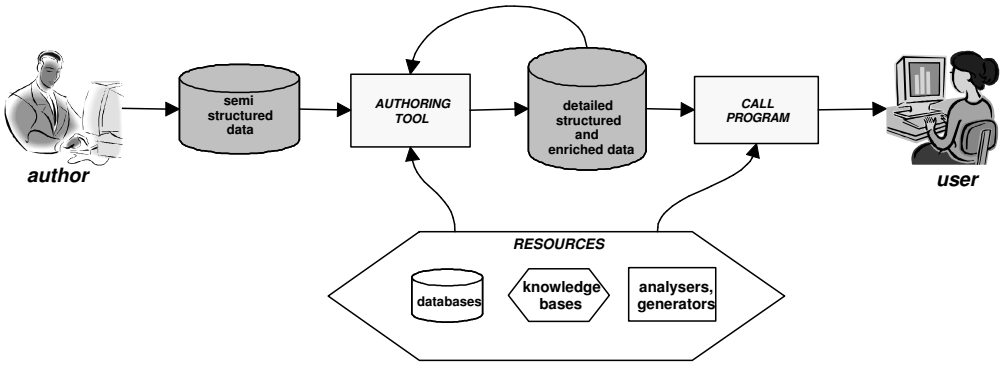


Figure 6.1: Overview of the approach to CALL development we propose.

Manual Data Elaboration The first step is to manually elaborate the language learning content in a possibly already semi-structured way and to submit and save it using a suitable editor.

Converting the Data Then an authoring tool converts the manually elaborated data into the final format.

Autonomous Learning Starting from these fully encoded data and by reusing the basic transfer modules intermediate and high level educational modules for autonomous learning can be created.

Guided Working Finally, guided working with the material can be taken into account. For this task Artificial Intelligence and collaboration features can be considered and corresponding modules can be added to the system.

6.3 Development of the ELDIT System

Our application of this approach has been visualized in Figure 6.2. Our linguists and language teachers create a semi-structured version of the data. An authoring tool converts these data in several steps into the extensive version needed by the system. Different resources such as the ELDIT data model, the WMTrans Modules, and several indices of the data are used both during the conversion process and during runtime.

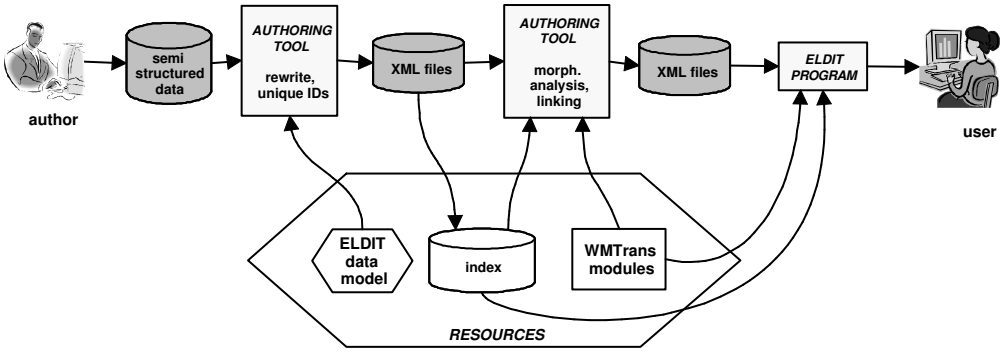


Figure 6.2: Overview of the approach to CALL development we applied for ELDIT

6.3.1 Manual Elaboration of Educational Data

Dictionary The linguists enter the data in a semi-structured way by using an XML-editor. For this task an XML file was made available to them in the form of an empty template. When creating a new word entry the linguists open this template, "copy-paste" XML-elements if they are needed more often than prepared in the template, fill the template with content, and save it with a new name.

Furthermore, a set of abbreviation characters was agreed upon. This set allows the linguists to give specific indications about a word in an easy and rapid way. For instance, when writing the sign "#" in front of some words such as in "Er #ging glücklich #nach #Hause" (he went home happily) the transfer program encodes these word with the attribute "emphasized". Similarly the characters "_" and "." are used to indicate prefixes, basis and suffixes of a derivation or compound word such as in "un.ver_besser_lich" (incorrigible). Figure 6.3 shows part of the ELDIT data (a derivation) at this stage.

```
<derivation>
  <pattern>
    <content>die Be_haus_ung</content>
  </pattern>
  <translation>
    <content>la dimora</content>
  </translation>
</derivation>
```

Figure 6.3: Semi-structured data encoded in XML

Text corpus When teachers use the online text authoring tool, they insert a new text in a semi-structured way via an HTML form. Title, text body, and questions are entered into different fields, further encodings, such as paragraphs and sentences are determined by the system through return-characters and punctuation marks. Figure 6.4 shows the corresponding interface.

6.3.2 Converting the Data

Dictionary After the linguists have saved a file they run the transfer program of the authoring tool which elaborates the file further and stores it in the final database. This transfer program consists of several modules (see Figure 6.5): the *ReWriter* determines paragraphs, sentences, words, and - using the abbreviation characters - word attitudes and word parts. All these tokens are then encoded with explicit XML-elements according to the ELDIT data model.

Since each word is accessed separately, a *spell check* can be performed. Mostly typing mistakes are discovered. Errors which occur on sentence level could be detected, too, if a Parser were included.

Next the *IDs-generator* generates a unique ID for each element. The ID consists of the path along which the element can be found in the XML tree and of a number which indicates the place of the element in a list of equal elements (see for instance the IDs of the elements <w> in Figure 6.6).

The *Lemmatizer* and *Morphological Analyzer* determine for each word its citation form and word class using the WMTrans modules. During this step ambiguities may occur, since the analysis is carried out at word level: for instance, the Italian word “posto” has two meanings, it could be the singular form of the noun “place” or the past participle of the verb “put”.

An analysis on sentence level is needed to eliminate these ambiguities. This can be done by a *POS-tagger*. We are currently using a tagger based on some very simple rules. If the tagger cannot eliminate the ambiguity all possibilities are considered in the following steps.

The next step in the process is *linking*. For this task cross references between the files have to be established. The files are not processed any further yet but an index is created of them which is consulted by the transfer program during the subsequent steps. By consulting the index links between each word and the corresponding dictionary entry are established.

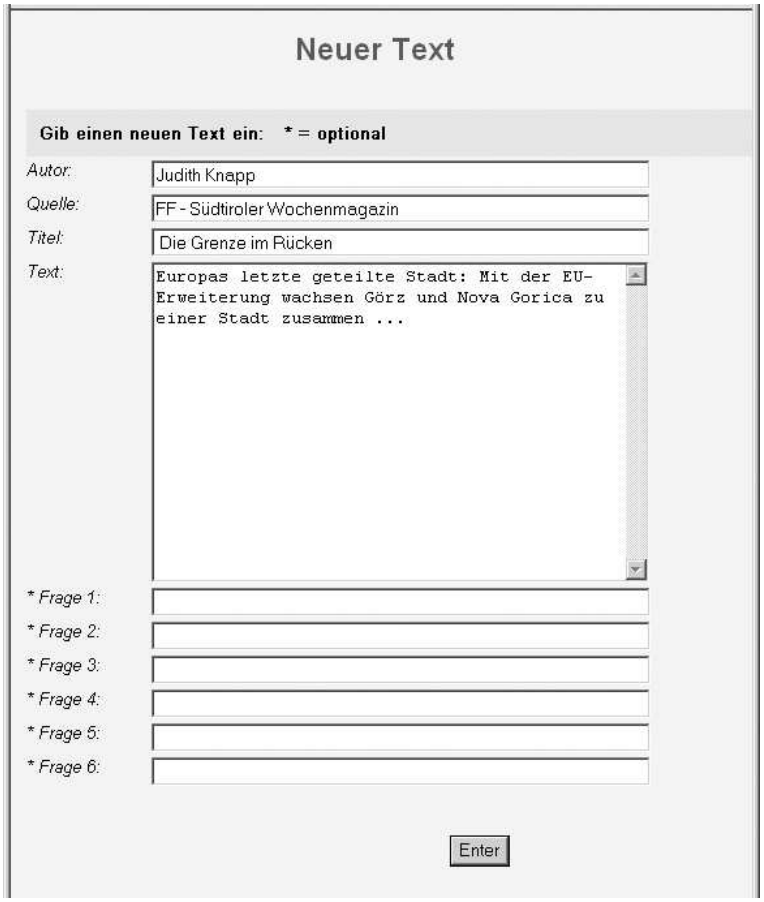


Figure 6.4: Authoring a new text in ELDIT

The inclusion of an *extended concordance tool* such as “Phrase Manager” [131] would make it possible to identify collocations in a sentence and to link the words of these collocations to the corresponding collocation in the dictionary. Similarly compound words not listed in the dictionary could be decomposed into simple lemmas by the WMTrans Word Formation Analyzer. Then links could be set to both lemmas. The choice could be shown in an intermediate window, as it is done in ambiguous cases.

Of course a variety of *other tools* and resources could be included in this transfer program. For instance, grammatical constructs could be determined by a parser, and links to a grammar section could be added. Links to a speech generation program would certainly be interesting, too. The more information is added to the data the more possibilities for an innovative presentation are given. Adaptation features could then be used in the system to adapt the information to different user interests and needs.

Figure 6.6 shows the same part of the ELDIT data as Figure 6.3, but at the fully encoded stage. Every XML-element got a unique ID (see for instance id=”de.n.haus.1.deriv2” in the element <derivation>). The words have been equipped with citation form and part-of-speech (see the attributes base=”Behausung” and ctag=”N” in the element <pattern> or the attributes base=”dimora” and ctag=”N” in the element <w> within <translation>). Wherever possible a reference to the corresponding dictionary entry is

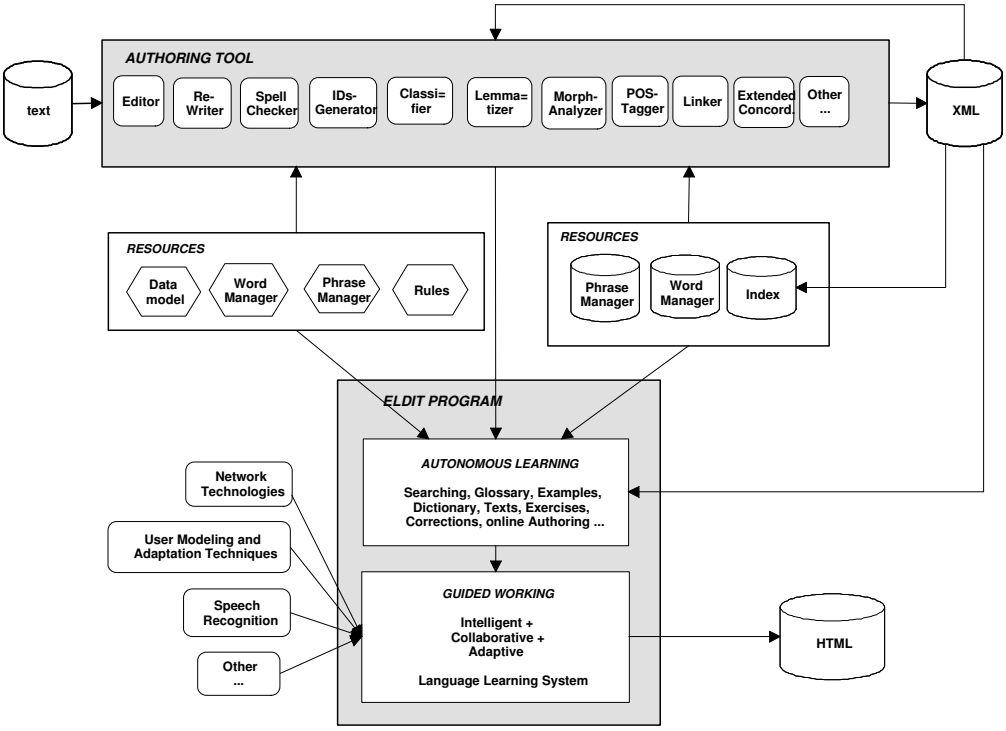


Figure 6.5: Development of ELDIT in detail

added (see for instance the attribute lexref=”it.n.dimora.1.sense2” in the element <w> of <translation>). Also references to explaining sections are added (see the attribute explref=”de.prae.h.be” which points to a section within which word derivation with the prefix ”Be” are explained).

```
<derivation id="de.n.haus.1.deriv2">
  <pattern id="de.n.haus.1.deriv2.patt0" base="Behausung" ctag="N" lexref="">
    <article base="der" ctag="art" lexref="de.g.articles.1.item1">die</article>
    <praefix explref="de.prae.h.be">Be</praefix>
    <basis>haus</basis>
    <suffix explref="de.suff.h.ung">ung</suffix>
  </pattern>
  <translation id="de.n.haus.1.deriv2.trans0">
    <w id="de.n.haus.1.deriv2.trans0.w0" type="content"
       base="il" ctag="art" lexref="it.g.articles.1.item2">la</w>
    <w id="de.n.haus.1.deriv2.trans0.w1" type="content"
       base="dimora" ctag="N" lexref="it.n.dimora.1.sense2">dimora</w>
  </translation>
</derivation>
```

Figure 6.6: Fully structured data encoded in XML

Text Corpus For texts the same procedures as for the dictionary data are carried out, but some additional steps are included:

By analyzing the text according to its sentence syntax and by applying different computational linguistic algorithms the *Classifier* elaborates meta information: the

submitted text can be classified according to the languages German or Italian, according to the two levels of difficulty, and according to the twenty interest domains. Such an analyzer has been implemented and tested within the projects BISTRO[1] and MIRIS[2] carried out at the European Academy Bozen/Bolzano [154]. It will be adapted and used in a similar way for the ELDIT project.

If the author submits any questions the significant words (nouns, verbs, and adjectives) in the text are searched and sentences containing many of these words are marked as *hints* for the corresponding question.

Once the text has been processed at the server a framed HTML page is generated. On the left hand side the text is shown, on the right hand side the electronically determined information is shown for *disambiguation* and correction.

6.3.3 Autonomous Learning and Guided Working

Once the data is encoded in this way a large number of possibilities arise. In the ELDIT project the fully encoded data and the modules elaborated for the authoring tool were reused to create intermediate modules such as the glossary and the example feature (see Figure 6.5).

Finally high level educational modules for autonomous learning such as the dictionary and the text corpus were created. Other ones such as the quizzes will be created in the future. Together with external resources these modules have been combined to a language learning system for self-directed learning.

Last, modules for guided working with the material will be added to the system: an adaptive tutor that coordinates the material or guides a user individually through the learning content, a user model that allows considering the users' individuality and applying adaptation techniques, and network technologies that allow collaboration between language learners on a local or an international level.

[1] http://www.eurac.edu/bistro
[2] http://www.eurac.edu/miris

Chapter 7

Architecture, Data Model, and Implementation

We will now give some more details regarding implementation issues. In section 7.1 we describe the architecture of the ELDIT system. In section 7.2 we describe the data model, and finally in section 7.3 we describe the implementation of the ELDIT data model by using XML.

7.1 System Architecture

Figure 7.1 shows the architecture of the ELDIT system. The basic architecture is a client-server model exploiting the standard Internet and WWW protocols. The server is implemented using Java Servlet technology and runs on any Web-Server supporting the Java Servlet API. Java Servlet technology is a powerful and easy to use technology to dynamically create HTML pages. Many useful Java APIs are available for free and can easily be integrated into an existing system. On the client side the system can be accessed by any Web-browser. Currently newer versions of Internet Explorer, Netscape, Mozilla, and Opera are supported.

When a user wants to retrieve a particular dictionary entry, the client sends a request to the server. This request is processed in several steps. First, the Request Handler invokes the Searcher which is responsible to identify the requested dictionary entry. Then the corresponding WMTrans Lemmatizer retrieves the citation form of the requested words. This expression is looked up in the index by using the java-API Lucene[1]. If no results are obtained, the Searcher performs a spell check. If the search process is again negative, a “not found” message is sent back to the user.

If the requested word entry is found, the Request Handler invokes the Code Creator which is responsible for the compilation of the HTML answer page. We follow a clear separation between content, structure, and presentation: the Code Creator invokes the Data Provider which retrieves the entry and selects those pieces of information which are relevant for the user according to the data stored in the user model. The Code Creator generates an HTML-page which includes Java Script arrays containing the data itself as well as a link to an external Java Script and CSS file. When the browser inter-

[1] http://jakarta.apache.org/lucene/docs/index.html

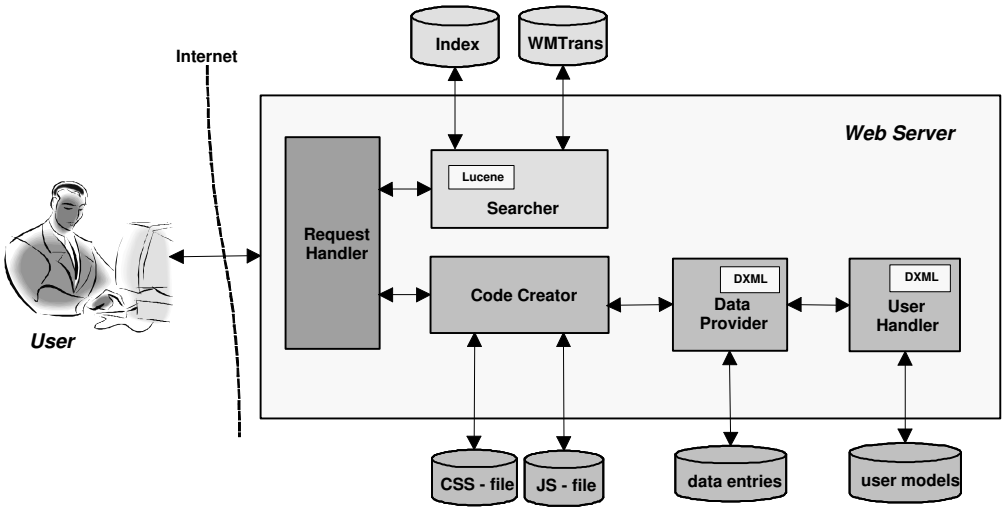


Figure 7.1: System Architecture of the ELDIT system.

prets the answer page, HTML elements such as tables, lists, paragraphs, and layers are generated dynamically by Java Script functions. The presentation of the information is realized by CSS.

This approach allows great flexibility. The use of Java Script did not create many problems (what browser differences concerns) since we have always tried to keep things as simple as possible, i.e. from the beginning we restricted the use of Java Script functions to the basic ones which are understood and interpreted in the same way by almost all graphic-based browsers. Unfortunately it will be necessary to eliminate this approach (possibly by the introduction of templates) if in the future a professional Web-designer will implement the final design of the system.

As a uniform data and knowledge representation formalism we use the XML language. The Data Provider and User Handler access the educational data and user models using DXML, a Java package for handling XML data. This approach will be explained in more detail in section 7.3.

7.2 Data Model

The main characteristics of the learning material and resources in the ELDIT system can be summarized as follows [73]:

- *Semi-structured data:* Semi-structured data are characterized by the lack of a regular structure and clear schemata. The manually elaborated ELDIT data contain only a rough structure and much information that is implicitly encoded, e.g. sentences are indicated by punctuation marks, prefixes and suffixes by special characters, etc.

- *Detailed level of granularity:* While in traditional learning material the typical level of granularity is that of a learning object which represents a concept in the

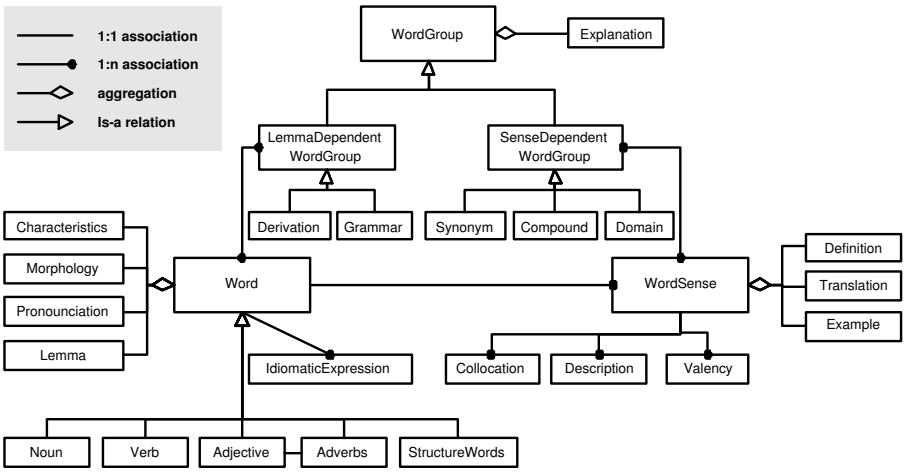


Figure 7.2: Simplified representation of the data model of the ELDIT dictionary.

specific domain [24, 89, 147, 156], our didactic demands require a more detailed level of data representation down to the level of words and parts of words.

- *Highly interlinked elements:* The ELDIT data are further characterized by a large number of links between the various pieces of information.

Data Model for the Dictionary

Figure 7.2 shows a simplified version of the data model of the ELDIT dictionary. The main entities in our domain of discourse are

- words,
- word senses, and
- word groups.

Each of the entities is composed of and related to several other entities. All entities are chunks of text which contain additional information as will be discussed later. There is a clear distinction between the mainly lexical information about a word (represented by the word entity) and the mainly semantic information about the different meanings of a word (represented by the word sense entity).

Words are classified into different categories: nouns, verbs, adjectives, adverbs, and structure words. Each word entity is composed of various pieces of lexical information which are independent from a particular meaning of the word. This includes the lemma of a word (e.g. “casa” (house)), morphological information such as the article and the plural form (e.g. “la casa, case” (the house, houses)), an adverb form in the case of adjectives (e.g. “ampiamente” (widely) for “ampio” (wide)), idiomatic expressions (e.g. “a casa mia” (with me at home)), and optionally some remarks about special characteristics of the word such as linguistic difficulties and pitfalls.

Each lexical word form can have several meanings which we call *word senses*. A word sense entity is composed of a short definition in the same language as the word

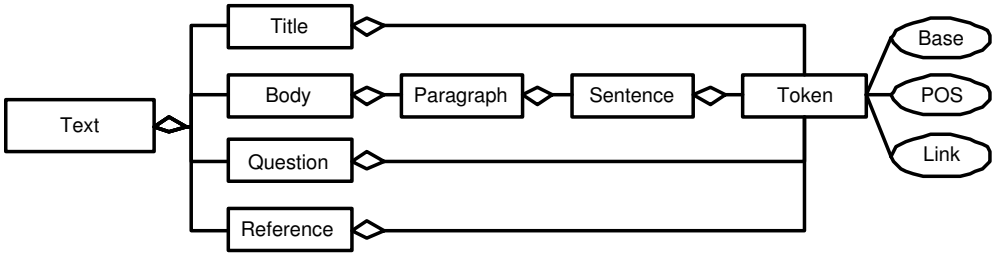


Figure 7.3: Data model of texts.

itself, of one or more example sentences, and of one or more translation equivalents in the other language. A word sense is further associated with additional information which helps to get a more comprehensive understanding of the word meaning as well as to use the word in the correct way. The additional information includes collocations, a description element which is a list of adjectives that typically occur with a noun, verb valency which describes the use of a verb together with subjects and objects, a prototypical picture, etc.

The third type of primary entity are groups of related words which we call *word groups*. In ELDIT these relations are not only named but also annotated, e.g. with explanatory information. We distinguish between word-dependent groups, which are independent of a particular meaning, and sense-dependent groups.

A typical example of a word-dependent group are derivations. A derivation entity is composed of all words which are derived from the same base word. Another type of word-dependent group is formed by grammatically related words, for instance prepositions, conjunctions, etc.

Sense-dependent word groups are composed of words depending on specific word meanings. Figure 7.2 shows a few of them. Synonyms are groups of different words with the same meaning. The group of the compound words is composed of all words which are build of the same base word. The domain groups are composed of words belonging to different interest domains, for instance sports, music, traveling, etc.

Data Model for the Text Corpus

Figure 7.3 shows the data model of the texts in ELDIT. Each text consists of a title, a text body, a reference element, and several questions to be answered by the learner. The body is further divided into paragraphs and sentences. All sentence-like text pieces consist of quotes (e.g. spoken language) and words.

The Need for More Details

Figure 7.2 and 7.3 show the data models at a rather coarse level. To be able to provide the requested support, we need more details.

Sub-Entities: Many of the main entities in Figure 7.2 and Figure 7.3 are composed of several sub-entities. For instance, idiomatic expressions and collocations are described by a pattern, by examples, and by one or more translations. Verb valency is described by a pattern, by comments, and by examples. Derivations and compound words are described by a pattern showing the formation rule, by comments and by one

or more translations. Synonyms, hyperonyms, antonyms, etc. are described by their names, relations, and differences, etc.

Word-Level Annotation: Most of the sub-entities consist of one or several short text sentences or some words. In order to reuse this information for different purposes and to evidence particular aspects of a language we have to annotate these text pieces at the word level and even below. The derivations and compound words are split up into prefixes, basis, and suffixes.

The Link Structure: A particular characteristic of the ELDIT learning material is that all pieces of information are highly interlinked. In particular, all words in definitions, example sentences, texts, etc. are linked to the corresponding dictionary entry to provide a fast access to meaning explanatory information. Pre- and suffixes are linked to word formation and inflection rules. Linguistic terms are explained. The explanations are treated as the rest of the educational data, hence unknown words are linked with the corresponding dictionary entry, too, etc.

7.3 Implementation

7.3.1 XML

XML is a meta-markup language for defining semantic tags that break a document into parts. Such tag-sets are defined via a document type definition (DTD). An XML tag-set describes the actual content of a document since the information is encoded within the names of the tags or their attributes. Thus, XML documents can be shared among and reused for various applications. Further advantages are the clear separation of content, structure and presentation, the ease of use, its clearness, and the free availability of many XML tools and utilities [84].

There are several possibilities of handling XML-documents. XML documents can be directly formatted using Cascading Style Sheets. However, in this way neither structural information can be added nor can a document be split. Page-to-page translations can be realized using the XSLT language. XSLT allows specifying a set of rules to translate an XML document into another format including HTML. XT[2] and Coccon[3] are two well known XSLT translators. SAX[4] is an event-driven interface to access XML documents which can also be done only partly. This has the advantage that large documents can be parsed and new data structures can be constructed. Tree-based APIs like DOM[5], JDOM[6], or DXML[7] allow reading and writing XML documents as if they were regular Java objects. This approach is the most flexible one, but puts great constraints on system resources if the XML document is very large.

Since our XML documents are not very large (5 to 20 KB), and since on each request we need almost all information in a document, we decided to use the most flexible approach, namely a tree-based API. Up to now we have used DXML to handle XML files in the ELDIT system and, because of the clear error messages, we also

[2] http://www.blnz.com/xt/index.html
[3] http://xml.apache.org/cocoon/
[4] http://www.saxproject.org/
[5] http://www.w3.org/DOM/
[6] http://www.jdom.org/
[7] http://www.objectspace.com/products/prodDXML.asp

used it to validate XML files in the authoring tool. The advantage of DXML is that unlike with DOM or JDOM no type conversion is necessary which results in a clean programming code. For example, the following code retrieves the <*article*> element shown in Figure 6.6:

```
String article = word.getDerivationAt(2).getPatternAt(0).getArticle();
```

The disadvantage of DXML is that no general applications but only applications for one specific DTD can be programmed. Since ELDIT is now growing very rapidly, we prefer substituting DXML with JDOM.

7.3.2 Document Type Definitions

We defined DTDs for words, texts, and different word groups (the corresponding DTDs are listed and documented in the appendices B.1 - B.5). For our purposes a rather detailed representation of information is required. For example, the DTD for a word entry contains 39 elements, and the DTD for a semantic field contains 45 elements.

In all DTDs the element <*w*> is the lowest level element. Almost all other higher level elements (definitions, translations, examples, footnotes, explanations, grammatical indications, etc.) are composed of this element. The corresponding part of the DTD is:

```
<!ELEMENT w (#PCDATA)>
<!ATTLIST w type CDATA #IMPLIED style CDATA #IMPLIED
            base CDATA #IMPLIED ctag CDATA #IMPLIED
            lexref CDATA #IMPLIED collref CDATA #IMPLIED
            nbref IDREF #IMPLIED>
```

The attribute *type* gives meta information, it can have values such as “content” or “remark”. The attribute *style* may have impacts on the font style, it can have values such as “plain” or “emphasized”. The attribute *base* shows the citation form of a word, the attribute *ctag* the corresponding word class or part-of-speech. Since each element can have several links, and since we generate all links dynamically and different for different situations and different users, we encode only references to other XML elements and no *xlink* links into the document: the attribute *lexref* contains a reference to an element of another word entry, the attribute *collref* points to a multiple word expression, i.e. a collocation or idiomatic expression. The attribute *nbref* contains a reference to a footnote in the same XML document.

7.3.3 User Model and Help Files

While word entries, word groups and texts contain the core information, ELDIT also stores other kinds of information in XML files. For instance, the user model stores all data collected about a user in an XML document. Different versions of the help information are also stored in several XML files. Currently, only one version is available for each language. Future work includes the adaptation of the help file to the users’ knowledge about the use of computers in general and the use of ELDIT in specific.

7.3.4 Advantages of XML

XML has several appealing features as a data representation and communication language, especially for Web applications. One of the greatest advantages is the strict *separation of logic, content, structure, and presentation*. Rules regarding presentation are defined in external files, hence changes are easy to implement and do not affect the data. This has several implications: (i) linguists and computer scientists were able to work quite independently from each other, (ii) data can be reused for several applications, and (iii) data can be presented in different ways to different users which supports the inclusion of adaptable and adaptive features.

Another important aspect is the *flexibility* of XML. XML supports the modeling of tree based data at a level of arbitrary detail. Our dictionary entries are good examples of rather complex tree based data sets. Thanks to the sophisticated DTD we can easily access very small pieces of the information and emphasize it in the presentation. Indexing each element makes it possible to provide a structured full-text search which means that the search operation can be limited to specific parts of the word entries. Again, the implementation of adaptive features is supported by this property: adaptation techniques, such as “hiding information”, “conditional text pieces”, or “stretch text” can easily be implemented with XML files.

Although it is very flexible, XML is *simple*, *human readable*, and *easy to use*. These advantages over other data representation languages are very important in the prototyping phase when the requirements often change. Moreover, we experienced that the communication with the linguists has been facilitated because the XML data model was easy to understand, since data are represented in a natural way leaving coherent information together. This is a quite important aspect since knowledge engineering is known as a difficult task, where the knowledge engineer has to mediate between domain experts and the formal representation of the domain in a computer.

Another advantage of XML is the *open standard* and *the free availability* of an increasing number of tools to process XML data. System developers can test different tools and then choose those which are most suitable to implement a running system at low costs. This not only helps to reduce the development costs, but to a certain extent also allows adapting the system to individual needs. For example, each ELDIT author in our team can use a favorite XML editor to enter the data.

Chapter 8

Evaluation

We will now outline the advantages our approach to content authoring offers. Its flexibility for the implementation of didactic demands has been confirmed each time we programmed a new module for ELDIT. Moreover, this approach made it possible to additionally implement many interesting features the linguists and language teachers had even not considered. In fact, all modules and features described from section 4.2 to 5.4 were initiated by the author of this thesis.

The first part of the present chapter deals with some evaluations in which we examined the didactic validity of some of these additional features (see section 8.1). The second part deals with some questions of system use and discusses the users' opinions of the ELDIT dictionary [4] (see section 8.2). Whenever possible we ask people what they think about our ideas for the future by showing them simple prototypes. The encouraging feedback is stated in section 8.3.

8.1 Didactic Evaluations of ELDIT Features

In this section we examine the didactic validity of some of the features the ELDIT system provides.

8.1.1 Glossary

Concerning the reuse of educational content by linking all words used in the system to explanatory information, we analyzed the following questions:

- How many words are linked to explanatory information?
- How many of these links are unambiguous?
- Are the ambiguous links useful for the learner from a didactic point of view?

The evaluation was carried out on texts that had been authored with the prototype of the online authoring tool. However, the results can be generalized, since the language used in the dictionary, i.e. in definitions, explanations, lexicographic examples, etc. is similar to the one used in the text corpus: intermediate level of difficulty and slightly reformulated, i.e. made appropriate for language learning.

For each language six texts on intermediate level of difficulty were examined. We counted the number of words for which the following holds: (i) base form and word class were found, (ii) base form and word class were unambiguous, (iii) a link to a dictionary entry was found, and (iv) the obtained link was unambiguous. Table 8.1 summarizes the results.

	Base form and word class		Links	One
	found	unambiguous	found	meaning
German	93	92	89	60
Italian	92	89	88	58

Table 8.1: Number of links found for the glossary - the numbers are percentages.

More than 90% of the words were annotated with base form and word class (93% for German and 92% for Italian). In most cases base form and word class were also unambiguous (92% for German and 89% for Italian). If base form and word class were found for a word, almost all of these words could be linked with some explanatory information in the dictionary (89% for German and 88% for Italian). Approximately 60% of the words (58% for Italian) were linked to a single source of explanatory information. In these cases either the target dictionary entry is described by only one meaning or the word is a compound, a derivation, or an adverb. In the remaining 30% of the linked cases (29% for German and 30% for Italian) the learner is faced with ambiguous links.

The results were then shown to the linguists and teachers in our team, and they were asked about their opinion of the usefulness of the feature, even if the links were not unambiguous in some cases. We can distinguish the following cases:

- An unambiguous base form and word class could not be determined. In these cases the system shows the links to all corresponding dictionary entries. The teachers in our team appreciated this solution and stated that a learner could even take advantage of coming across all forms of an inflected word.

- Several dictionary entries exist for the same word such as in the case of homonymy. In these cases the word is linked to all dictionary entries. The linguists suggested adding a note of the different meanings or to show the first definition as a quick indication.

- Several meanings of a word exist for one word such as in the case of polysemy. Currently, it is not possible to distinguish automatically between different word meanings and to find the correct one. Hence, all meanings are listed and the learners have to find the correct one by themselves. According to the teachers this might be too difficult for the learner, since some words have more than ten different meanings. The possibility of manually authoring such links and assigning words to a special meaning was considered to be necessary.

A rudimentary prototype of such an authoring tool has already been implemented, its final elaboration, however, is future work. Since no wrong information is shown, in

the mean time we have enabled this feature in the production version of our system for all previously mentioned cases.

8.1.2 Example Feature

For the reuse of illustrative content we analyzed the following two questions:

- How many additional examples were found?
- How many of the additional examples were useful/valid from a didactic point of view?

We systematically inspected and analyzed the generation of additional examples for some words. The results for the word *Haus* are summarized in Table 8.2. The first column indicates the pattern or word for which additional examples were retrieved. The second column shows the number of additional examples that were retrieved automatically. The third column shows the number of additional examples that were considered valid by our linguists. "Validity" means that the example applies the pattern in a correct way and is hence useful for the learner.

	Number of retrieved examples	
Pattern	Total	Valid
Derivations and compounds		
die Behausung	0	0
hausen	1	1
das Häuschen	13	13
häuslich	0	0
das Bauernhaus	2	2
die Hausnummer	2	2
die Haustür	17	17
das Hochhaus	7	7
das Reihenhaus	1	1
das Schulhaus	2	2
das Wohnhaus	3	3
Typical adjectives		
groß	4	4
klein	5	5
hellhörig	0	0
modern	0	0
neu	13	13
alt	5	5
baufällig	0	0
Collocations		
ein Haus bauen, errichten	18	15
ein Haus renovieren	2	2
ein Haus besitzen, haben	2	1
ein Haus kaufen	4	3
ein Haus verkaufen	2	2

ein Haus mieten	1	1
ein Haus vermieten	1	1
ein Haus einrichten	0	0
nach Hause gehen	31	31
nach Hause kommen	55	55
zu Hause bleiben	29	29
Haus an Haus wohnen	1	0
zu Hause sein	17	17
Idiomatic expressions		
aus dem Haus sein	0	0
noch zu Hause wohnen	0	0
jemandem ins Haus platzen	0	0
ein offenes Haus	0	0
von Haus aus	5	0
frei Haus	1	1
Definitions		
definition 1	22	21
definition 2	33	32
definition 3	0	0
definition 4	0	0
definition 5	0	0

Table 8.2: Additional examples found for the elements of the dictionary entry for the German word *Haus*

The first observation is that for some patterns and words no examples were found at all. There are at least two reasons for this result. The first reason is the fact that we use rather strict rules and patterns to reduce the number of invalid results as much as possible. With some additional tuning or more sophisticated NLP tools the number of retrieved examples can certainly be increased. The second reason was the incompleteness of the dictionary at the time of evaluation. Up to that time only about 2/3 of the 3,000 dictionary entries have been elaborated and the text corpus was not integrated yet.

Finding valid examples for derivations and compound words worked well, since only one word has to be searched and different meanings are not considered in the presentation. Hence in our evaluation we reached a precision of 100%.

Finding examples for typical adjectives was also quite easy and worked with a high precision, since we apply the restrictive rule that the adjective has to occur immediately before the noun for German words and immediately before or after the noun for Italian words. The advantage of this strict rule is that the retrieved examples are valid. The price to pay is the fact that only attributively used adjectives (e.g. a beautiful house), but no predicatively used adjectives (e.g. the house is beautiful) are found.

For collocations the first difficulties arise with invalid results. Although many examples were considered valid, there were some example sentences which did not describe the pattern. For example, for the collocation *ein Haus kaufen* our system retrieves four examples, where only three of them are valid. One of these sentences contains *ein Grundstück für ein Haus kaufen* (to buy a site for a house) which is obvi-

ously not a valid example for the collocation under consideration.

For the idiomatic expressions the feature does not work very well. Hardly any examples were found and most of those ones found were considered invalid. The reason for this is the fact that the examples in ELDIT are kept simple, since they should help the learners to reach a fuller understanding. Hence idiomatic expressions are hardly used for them.

The last section of Table 8.2 shows the number of examples found for the different word meanings. Finding such examples requires some kind of semantic reasoning. Our pragmatic approach to use the manually elaborated example sentences to find additional examples works only partially. Many (mostly valid) examples were found for the first meanings which are the most frequently used ones, but only very few examples were found for the less frequently used meanings.

After this evaluation 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 will be disabled in the production version until we have appropriate tools to improve the results.

8.1.3 Text Corpus

We carried out several evaluations of the electronic encoding features we used to prepare the text corpus.

POS-tagging

Our text files were POS-tagged at the *Institute for Computational Linguistics* of the University of Stuttgart by using the TreeTagger [144]. We did not systematically inspect the POS-tag of each word, but discovered some errors when linking the words to the corresponding dictionary entry by systematically collecting and listing unlinked words according to their frequency. In this way mistakes with lemmatizing and some mistakes in the POS-tags were discovered. The number of mistakes discovered amounts to 2,5% of all of the words for German and to 2,3% for Italian. In the following we list some error classes:

- The tagger has been trained with a corpus consisting of texts spelled according to the old German spelling rules valid until 1998. Our texts, however, are written according to the new spelling rules. Hence all words which changed with the new spelling rules (e.g. “mussten”, “muss”, “musste”, “isst”, “lässt”, “wusste”, “hinterlässt”, “dass”) were tagged wrongly or were not recognized at all by the tagger.

- The German words “weitere”, “weiterer”, “weiteres” have two meanings: (1) “further” and (2) the comparative of “wide”. In the first case the lemma is “weiter-”, in the second case the lemma is “weit”. The tagger did not recognize this difference, all words were tagged with “weit”. The mistakes were corrected manually.

- The Italian verb “sono” (they are) was always tagged with “sonare” (=suonare, make music) instead of with “essere” (to be). The verb “sia” (he may be) was al-

ways recognized as a conjunction and tagged with “sia” instead of with “essere” (to be).

- Many conjugated forms of “avere” were tagged with “riavere” (to have something back) instead of with “avere” (to have). Many conjugated forms of “andare” were tagged with “riandare” (to go back) instead of with “andare”.

- Abbreviated forms of Italian words (such as “bel”, “vuol”, “pur”, “fin”) were tagged as nouns and with the original form as lemma.

- Some Italian words which exist both as nouns and as past participles (such as the word “successo” (the success, it happened)) were tagged with the wrong word class.

Frequency Value

We evaluated the frequency number of the texts by looking at the content of some texts with a high importance value and some with a low importance value to understand whether the number really gives indication about the generality of a text. When looking at the words used in the corresponding texts, one gets the impression that special vocabulary occurs more often in the texts with a low frequency number. However, not all texts hold this impression.

The three texts with the highest frequency values are about the Antarctic, about pets in the 20th century, and about a school (a circus school) in which an entire family and only this family is educated. The last two texts may be considered general and important, since they contain many words about children, families, and school live.

The three texts with the lowest frequency values are about what people can learn from a cobra which lives in the jungle, about shoplifting, and about nutrition. Since they do not speak about daily live events, they can be considered special texts. Moreover, they contain special vocabulary such as the things that are stolen most often (perfume or leather goods) and the names of food articles (noodles, cereals, sweets, etc).

8.2 Evaluations of ELDIT Use

For the evaluation of ELDIT use we examined various data sources to gather information about how ELDIT is used and about how much it is accepted by its users:

- In ELDIT a log file analysis program[1] generates online statistics. This program parses the log files of our Web Server[2] and generates statistical tables and visualizations. The visualizations are included in the ELDIT system.

- ELDIT includes a user model in which for each user statistical and other information about the way ELDIT is used is recorded (e.g. number of logins, number

[1] http://www.mrunix.net/webalizer/
[2] http://jakarta.apache.org/tomcat/

of words looked up, number of clicks per visit). Using JCharts[3] we programmed a software which evaluates the user models and generates statistics in tabular and graphic form. 78 user models were examined. Each of them contained at least three logins.

- Furthermore, the software for online questionnaires was created. This was done similarly as described in [95]. A questionnaire on a certain topic can be submitted in XML. The software creates online interactive questions which are filled out and saved by the user. Finally the software evaluates the answers similarly as for the user models.

- We conducted an evaluation in which we asked the participants to explore the system, to carry out some simple tasks, and to fill in a questionnaire. The questionnaire consisted of 38 questions about dictionary use in general and on the use of the ELDIT dictionary specifically. 40 people have tested the system and conducted the questionnaire.

- At first new features are usually implemented as rudimentary prototypes and shown to users and teachers to gather feedback about acceptance and usefulness.

8.2.1 Server Log Files

In the log-files of the Web-Server all accesses to the ELDIT system are recorded. We evaluated these records using the Webalizer software and now show the results online.

In February 2002, when we went online with the dictionary, the average daily access rate started at about 5 visits a day. It increased to about 50 visits a day up to May 2003. The graphic in Figure 8.1 shows the daily use of the system in May 2003.

From 19th May onwards a strong increase can be noticed. This was the day when we went online with the text corpus. Up to that day only the dictionary was enabled. We sent out a newsletter to inform about this new feature. Daily access rates rose from about 50 to about 80 visits a day. Apparently our users are interested in this additional module, since access numbers remained high during the following weeks as well. These are quite satisfying numbers, since up to then we hardly advertised our product.

8.2.2 User Models

A second source of information is provided by the user models which store relevant information about the interaction with the system for each user. Each user must explicitly permit the recording process.

Unfortunately, more than 75% of the user models examined contain only one login. It is not possible to find out whether these users visited ELDIT only once or whether they regularly use ELDIT but register each time they log in. For the evaluation described in this section we only used models which fulfilled the following three criteria:

- permission to the data recording process was given

[3] http://jcharts.sourceforge.net/

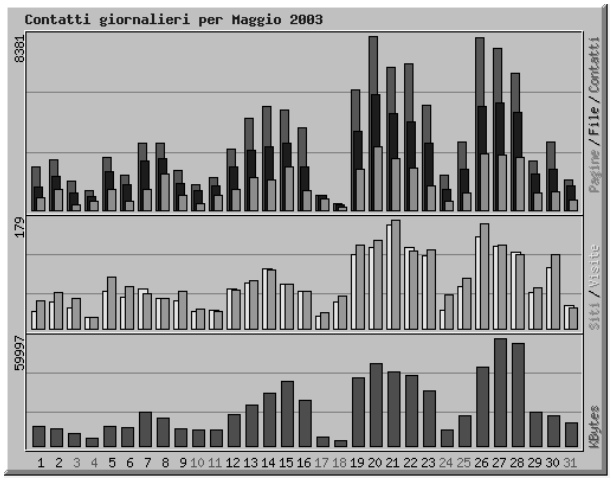


Figure 8.1: A screenshot of the online statistics in ELDIT

- a minimum of three logins were recorded in the user model
- the user was not a collaborator of the European Academy Bozen/Bolzano

There were 78 user models which fulfilled these demands.

The average login time was between 1 and 7 minutes. Only 3 of the 78 users used ELDIT for less than 1 minute on average. There were also some users who regularly used ELDIT for more than 7 minutes, namely for an average of up to 40 minutes per login.

The average number of accessed word entries per login is between 2 and 8 words per login. Only 8 out of 78 users on average checked only one word per login, some users on average even looked up 20 to 40 words per login.

Concerning the average number of clicks per login most of the users carry out between 3 and 20 clicks in each session. This means that many users do not only enter and exit the system, but consult more of the information each time.

The average number of clicks per word can be calculated by dividing the sum of all clicks carried out by the number of all words accessed. This number is 1.5 clicks per word which is lower than we have expected.

8.2.3 Questionnaires

In spring 2001 we took part in an exhibition about language learning material which was organized by the local government[4]. For this occasion we composed a questionnaire which consisted of 38 questions on dictionary use in general and on the use of the ELDIT dictionary in specific. 38 people tested the system and filled out the questionnaire, 51% of them were male and 49% were female. The participants were between 10 and 60 years old (more or less equally distributed).

[4]http://www.fieralingue.it/cgi-bin/WebObjects/Fiera.woa/wa/Main

General Outcomes

As we expected the younger participants were more familiar with electronic dictionaries and the Internet than the older ones. Many participants stated that they had no difficulties with dictionary use (51%), in another question, however, many stated that they found it necessary to be taught how to correctly use a dictionary (64% very important, 25% important). Colors are used in ELDIT to emphasize linguistic information and many users liked this approach. However, about 33% of the users were in favor of a more solid interface. The font size was okay for 74% of the participants, 26% suggested a bigger font, no participant wanted a smaller font size.

Description of Word Meanings

In ELDIT word meanings are explained by a definition, a lexicographic example and a translation. For verbs a short pattern of verb use is added as well. We wanted to know in which order users read this information and whether they use all sources of information (definitions, translations, examples, etc) or only some or one of them (e.g. the translations). It turned out that about two thirds of the participants read the definition first (namely 24). For many, however, the translation was the first thing to read (11 read the translation first, 11 read it second). This result shows that both definitions and translations are consulted for word understanding which encouraged us in our crosslingual approach. However, such results should be explicitly tested, since sometimes the subjective impression of a test person is not very reliable.

Correct Use of the Tabs

We wanted to know whether the users were able to understand and use the tab metaphor. Therefore we asked them to carry out some small tasks and to answer some questions on them afterwards. Some of the questions were: ”Where do you look up specific characteristics of a word?” etc. Most users turned out to cope well with this part of the interface. Depending on the tab tested between 70% and 97% of the users gave the correct answers.

Evaluation of Verb Valency

We asked the participants whether they found the use of colors for verb valency (i.e. to show the construction of a sentence) useful. 95% of the participants liked our approach and found the use of colors helpful, 5% had no opinion of this topic and nobody stated not to find it useful.

We also checked whether the users understood the meaning of the colors. This question had to be answered with their own words. The following answers were given: 56% of the participants understood that the colors indicated the construction of a sentence. 22% said the colors were used to make it easier to find the pattern elements applied in an example sentence. 11% stated that the colors showed the correspondence between pattern and example sentences. 5% assumed that the colors were used to indicate linguistic characteristics. All these answers are correct and show that the users understood the purpose of the colors. Only 6% of the participants gave a poorer answer, namely ”to attract attention”, 3% of the participants gave other answers.

Correct Use of the Search Features

In the default search mode all search features are applied automatically, i.e. if an expression is not found, a spell and inflection check are carried out immediately. We also implemented an ”extended search” interface by which the users themselves can determine which feature should be applied. The restriction of the search process to single elements as well is accessible only through this interface. This search feature was developed for experts, i.e. for people who are used to searching the Internet, for people with linguistic knowledge, and for us programmers and lexicographers to be able to rapidly examine our data. Nevertheless, we wanted to test its intuitivity on all the participants. It was necessary to show the users how to access the interface, but then they generally succeeded in using at least some of the features: 83% of the users were able to use the inflection feature, 85% were able to use the spell check feature, but only 16% were able to understand how to restrict the search process to single elements.

General Acceptance

In the end we asked the participants to express their general opinion and to explain in their own words what they liked and what they explicitly disliked. They could also suggest improvements. The participants were allowed to give more than one answer. The number of users who gave a specific answer is indicated in brackets.

The users liked the following features of the system: the large amount of information (7), the combination of definitions and translations (6), speed (6), simplicity (5), the large amount of examples (5), use of the colors (5), everything or almost everything (4), the electronic medium (4), concentration on meaning and use of a word (3), the translations (3), spell check feature (1).

When we asked them what they did not like, the following answers were given: the extended search feature is too complicated (9), we dislike nothing (5), we dislike selecting a meaning before accessing the tabs (2), we miss images (1), the dictionary in general is too difficult to use (1).

Finally we asked the users what they would change in the system. Some very interesting answers were given. Many recommended providing the interface in the native language of the user (13), some of the users liked the concept of the dictionary and suggested to implement it for other languages as well (2), one user recommended to add audio files (1).

8.2.4 Implications

We now want to summarize the implications this evaluation had on our work.

The daily access rates were quite satisfying, since up to the time of evaluation we hardly advertised our product.

The number of clicks per word, however, is quite low, namely 1.5 clicks per word. This may be interpreted in two ways: (1) the users quickly find the desired information so that further clicks are not necessary. Because of the efficient search feature which takes the user directly both to single words and to multi-word-expressions this assumption might be quite realistic. (2) Our users used the ELDIT dictionary mainly as a reference tool. ELDIT, however, is meant to be a learners’ dictionary, i.e. a dictionary for language learning. Users should explore the information given for one word

and possibly remember at least parts of it. This problem encouraged us in our plans to enlarge the dictionary with quizzes, texts and an intelligent tutor which should help to take fully advantage of all the information collected.

When exploring the system and carrying out the tasks many users expressed their desire to see further example sentences. Especially when learners had a specific context or sentence in mind, they wanted to see an example sentence which was rather close to their specific needs. "I would need the example just a bit different" was said more than ones. This was when we started to think about content reuse and about the realization of the "more example" feature.

Some users suggested adding multimedia content. Images and sound files are planned for ELDIT and we have already done some studies of which kind of images are useful for learners. A small database of material has already been collected. However, the final elaboration and inclusion of this material is future work.

Not only in our questionnaire but also orally some users told us that they found the interface (or program) too difficult. Thus, we developed a help page. This page is shown to each user when registering and can be accessed via a help link at any time. However, in our questionnaire we also asked the participants whether they usually cope well with the help function of a program. Only about half of them stated to have no problems at all, 19% stated that they had problems using a help function and 32% stated that they had never used it! Hence the use of the many features ELDIT offers will always cause problems, since it is difficult to teach these features to users who are not willing to use the help functions included in the system. A first step to reduce these problems was to add a small label for each link in the system on which the function of this link is explained. In the future we will include video clips as well and offer face-to-face introduction courses which should familiarize interested users with the most important features of the system.

An interesting outcome was the recommendation to provide the interface in the native language of the users (13 participants). As a first step we try to provide a better default language for the interfaces: the interface of the German words is given in Italian and vice versa, since we expect the Italian users to study the German words and vice versa. A better solution, namely providing the interface always in the users' native language, is planned for the future, when adaptation components will be enabled.

8.3 Feedback for Future Ideas

We have implemented demonstrations of the modules of ELDIT that have not been realized yet to be able to explain our ideas and to possibly acquire funds for the realization of them. These demonstrations are either very rudimentary prototypes or some static HTML pages. Whenever it is possible (on conferences, presentations, education and training sessions, expositions, etc.) we ask people to express their opinions and suggestions. Usually the reactions are positive and encouraging. Some frequent suggestions and considerations are reported below.

Quizzes We showed the possibility of electronically generating quizzes to some language teachers. We asked them whether in their opinion such quizzes are useful for language learning and whether they would use this module in their lessons. Some

teachers were impressed by the variety of quizzes that can be created. Other ones appreciated quizzes such as cross word puzzles since they are considered motivating for the students.

Some teachers, however, told us about the following experience they already made when working with traditional hypermedia systems: learners might give a correct answer but not one of the answers the systems expects (e.g. in cases of prefabricated answers). In such cases the answers are usually marked as wrong which is either irritating or de-motivating for the learner.

We hope that this experience might be less a problem in ELDIT. The system includes semantic fields and synonyms for many words which we can consider in the correction process. The inclusion of WordNet would further augment the identification of correct cases. Thus, the number of incorrectly graded answers might be significantly lower than in traditional hypermedia systems.

Authoring Tool The possibilities of online authoring of new texts was also shown to learners and teachers. Especially teachers liked this feature a lot. Some of them reported that they had seen similar authoring tools before, but that most of these tools were only able to link words which occur in the citation form. The teachers considered the underlying dictionary a big help, since traditional system usually only give translations of the words, but do not provide the large amount of information as the ELDIT dictionary does.

Tandem In spring 2002 we presented our idea of the eTandem module to some local government authorities. The reaction was very enthusiastic and funds were promised almost immediately. We started implementing the module before the contract was signed. Unfortunately some months later the promise was repudiated, since the project was considered to be too expensive by the local government offices.

Feedback about our ideas was also given by some face-to-face tandem experts of a local language school[5]. Our approach was considered as very helpful. Future work might include a systematic realization and evaluation of this component in collaboration with this language school.

Tutor The ideas about adaptive tutoring has been presented very often as final vision of our project. Especially language teachers think that “contextualized, adaptive vocabulary acquisition” is a useful method to overcome speed *and* superficiality problems in vocabulary acquisition.

8.4 Written Feedback

Sometimes we are receiving e-mails with encouraging feedback. These people have just seen the user version of the system and do not necessarily know a lot about ongoing work and future ideas:

- *OK, spent some time browsing. I must admit: this is becoming a nice ED indeed with many nice and rather new features. Congrats!*

[5] http://www.alphabeta.it

Good luck with the continuation,
Maurice.
(Gilles-Maurice de Schryver [51], Ghent University, Belgium)

- *hallo eldit-team*
  ich bin deutschlehrer an der supsi lugano (wirtschaftsabteilung).
  habe mich zu eldit durchgeclickt, mich eingelogt und mir das ganze angeschaut. sieht ziemlich interessant aus! werde mich sicher weiter umsehen und mich einarbeiten! möchte euch unbedingt ein grosses kompliment machen für die grossartige arbeit. freue mich schon jetzt auf die weiterentwicklung des projekts.
  herzlichst
  martin saurer, daf-lehrer supsi manno

- *carissime! vi voglio fare I miei complimenti; il sito è utilissimo; io lo utilizzerò per preparare l´esame di bilinguismo; ho provato a navigare un pò ed é bellissimo*
  grazie
  ciao Roberta
  (Roberta Medda, European Academy Bozen/Bolzano)

- *liebe eldit-/erinnen,*
  ich finde das projekt höchst spannend und würde gerne etwas mehr hintergrundinformationen erhalten...
  (Lukas Wertenschlag [120], CLAC AG, Schweiz)

- *Sehr geehrte Damen und Herren,*
  mich freut es im Internet endlich ein gutes Wörterbuch für Deutsch - Italienisch gefunden zu haben. Obwohl sich das von mir gesuchte Wort noch nicht im derzeitigen Stand Ihres Wörterbuches befindet, gefällt mir die Art des Aufbaues sehr gut. Vor allem gefällt mir, das zur einfachen Übersetzung auch Redewendungen und Anwendungsbeispiele angeboten werden.
  Fragen...
  Meine Glückwünsche zur bisherigen Arbeit e buon lavoro
  Cordiali Saluti,
  Friedrich Forthuber
  (Friedrich Forthuber, B&R Automation)

- *Hallo miteinander,*
  unser Regionalverein hat im Rahmen des Projektes "Via Claudia Augusta" sehr viel mit italienischsprachigen Partnern zu tun und ich habe Eure website schon zwei oder dreimal in Anspruch genommen, wenn mir Vokabeln gefehlt haben. Gratulation, die Seite ist sehr gut gemacht!
  Mit freundlichen Grüssen,

MIAR
mittelfristige Initiative für eine angepasste Regionalentwicklung für den Bezirk Landeck
Assistentin Evi Jörg

Chapter 9

Discussion

In this chapter we discuss our work. First we compare our system and the technologies used to related work (see section 9.1). Then we discuss some possible generalizations both of the system itself and of our approach to content authoring (see section 9.2).

9.1 Related Work

The development of the ELDIT system has been inspired by a lot of other work. In the following paragraphs we refer to related work by emphasizing the differences between our system and other systems and by explaining how we have tried to improve the outcomes of our work.

Dictionary Learners' dictionaries are developed with the objective to support the language learner in text decoding and text production [83, 168, 114, 94, 150]. Dictionaries in paper form, however, show the problems mentioned in section 3.2.1, and often their electronic versions are just a one-to-one transfer to the electronic medium. The improvements we achieved in ELDIT by exploring multimedia and other technologies are summarized in section 3 and explained throughout this thesis.

Improvements based on hypermedia technologies can also be found in the electronic learners' dictionaries KirrKirr [97], Alexia [38], or the Plumb Design Visual Thesaurus [135] which all try to show a graphical representation of the so-called mental map. These dictionaries provided the basic ideas for the semantic fields in ELDIT.

The main difference between our dictionary and lexical databases such as the terminology database BISTRO[1] (European Academy Bozen/Bolzano), the dictionary LEO[2] (Technical University of Munich), or WordNet projects [64], lies in the different objectives of the systems. Lexical databases are usually intended as a reference tool and try to describe knowledge as completely as possible. The ELDIT dictionary is a learners dictionary, especially designed for language learners. As a consequence the information stored partially differs from the information stored in reference tools. The main difference, however, lies in the elaboration and presentation of the material. The ELDIT dictionary includes only the so-called basic vocabulary for each language.

[1] http://www.eurac.edu/bistro

[2] http://dict.leo.org/

For each word entry only the most important usage patterns are given. The presentation of this information, however, has been carefully designed for language learners. Relations between words and between the two languages are explained, and a large amount of well-designed illustrative material, such as examples, pictures and sound files is provided. Many lexical databases provide valuable input for the creation of the ELDIT learning material, but cannot be used directly for didactic purposes.

Text Corpus The approach to link every word in the text corpus to the corresponding dictionary entry which allows checking unknown words easily has been implemented in several reading systems [44, 161, 126, 134, 148]. In ELDIT additionally much support is provided for the learners when answering the questions about a text. Moreover, while departing from raw text material, all these features were created or added electronically. The texts were reused from a former project, manual work was only necessary for correcting ambiguities or processing errors.

Another characteristic that distinguishes our system from other reading systems is the online authoring tool which allows adding new material. This material can be prepared and integrated in real time in the same way as the original ELDIT material was elaborated. Most of the authoring tools that have been created up to now support teachers only in simply uploading and manually editing material [14, 29, 75]. There are a few highly sophisticated systems which support the development of mathematical content [10, 140], and some which allow the creation of adaptive hyperbooks for a general subject [23, 60, 122]. But these tools do not support particular language-related problems. Even in some intelligent language tutoring systems all correct answers have to be indicated by the authors [164].

Last, we want to compare the ELDIT reading module with a system that only recently has been developed at our institution as well. GYMN@ZILLA is a program that browses the Intra- or Internet while dynamically annotating pages with dictionary information [153]. Already in section 5.3 we mentioned that the ELDIT dictionary has been included in GYMN@ZILLA. Working with GYMN@ZILLA is much more general and flexible than working with ELDIT, reading works for many languages and annotation for arbitrary dictionaries. However, users sometimes browse special or difficult texts. In these cases erroneous annotations may occur. Moreover, the amount of links set by GYMN@ZILLA depends on the amount of entries in the corresponding dictionary. Due to these facts, fewer links than in ELDIT are established and they are more erroneous than in ELDIT.

Quizzes Quizzes in paper form of the type true-false, multiple choice, matching, or gap-filling are described in [120], systems which exploit multimedia capabilities to provide such environments are described in [78, 98]. The use of gap-filling quizzes to practice grammar rules or entire sentence construction is described in [19, 87, 116, 146, 152, 172].

ELDIT differs from these systems in that a large number of quizzes can be generate automatically of the existing educational data without manual authoring. The use of parameters augments the number of different quizzes that can be generated. In this way a large amount of practicing material for different topics becomes available. Moreover, since knowledge about inflection, word formation, synonyms, spelling, etc.

is included in the ELDIT dictionary, the system can provide much more than simple "wrong/correct" feedback, but corrections can be provided and new quizzes can be created for remediation. Additionally, just as all the information in our system, every word used in the quizzes is linked to the corresponding dictionary entry and can be inspected by a simple mouse click. Finally, thanks to the user model included it is possible to adapt the quiz material to different levels of difficulty. Remedial material and feedback can be adapted to users' performance, and guidance according to different learning situations or users' backgrounds can be provided.

Tandem Tandem learning via e-mail or chat is a language learning method which has been successfully applied in several NCALL projects [25, 34, 47, 102, 109, 113, 177]. Our system provides the advantage that the messaging modules are integrated into the learning platform and hence can profit from computational linguistics knowledge and adaptation features.

Tutor Several systems try to combine incidental and intentional vocabulary acquisition. In CAVOCA [78] vocabulary is taught by guiding the learner through three stages of the mental acquisition process. In PET2000 [43] a learner corpus was designed, and students use concordance tools to create their own dictionaries of words to be learned. Similar to these two systems, ELDIT combines incidental and intentional vocabulary acquisition, but includes an additional step where words are applied by practicing them on a text.

For the teaching process adaptive content presentation and adaptive annotation techniques are explored. Adaptive educational hypermedia systems have mainly been developed for teaching natural sciences and computer science [23, 30, 60, 89, 125, 174]. Only very few AEH systems exist for computer-assisted language learning [148, 172]. ELDIT exploits common adaptation techniques by slightly modifying them for language learning. Similarly to the AHA system [23], an introduction page is provided only when the user logs in for the first time. This page is replaced by a link in the following sessions. Adaptive link annotation is used in a similar way as it was done in the KBS-Hyperbook [125], Isis Tutor [30], ELM-ART [174], and InterBook [60]. However, while these systems look into the future, i.e. indicate whether an item is or is not ready yet to be learned, ELDIT gives feedback about the past, i.e. how well an item has been studied and whether it should be repeated. Similarly to ELM-ART [174] and InterBook [60], a "next" link is provided in ELDIT. However, since words (or texts) are independent from another, our system does not recommend ready-to-be-learned items, but suggests groups of important and interesting items to be studied. The number of items in a group depends on the users' learning speed. A word is considered to be the more important the more frequently it occurs in the texts. A text is considered the more important the more frequent words it contains. An item is considered to be interesting if its domain matches the users' interests. These considerations are a support for adaptively structuring the huge amount of quite unrelated words to be learned and texts to be practiced [93]. Finally, unlike all the systems mentioned previously ELDIT will provide possibilities to systematically repeat formerly acquired concepts which is a feature especially important in vocabulary acquisition.

Extensions ELDIT includes some special features that affect the entire system and make it innovative and outstanding from other educational systems. It includes a powerful search engine especially designed for language learners, an extensive link structure so that each word used in the system is linked to the corresponding dictionary entry, and a large number of lexicographic examples for almost each piece of information. The modules can be combined and used outside of the system as well. The inclusion of externally developed products such as Word Manager has also been successfully realized.

Corpus Annotation and DDL Using large text corpora for language learning has a long tradition. Data-driven learning (DDL) is an approach that provides learners the opportunity to research language materials, to discover language rules by themselves, and to learn through problem-solving activities. The corpora have usually been annotated with NLP like tools, and frequency analyses and concordancing tasks were performed to determine what lexical and grammatical concepts to teach, to find example sentences, and to create personalized dictionaries [11, 12, 43, 106, 148]. Also parallel corpora are used, mostly for teaching translation [171]. Another approach consists of collecting learner corpora, for instance of written essays, and to analyze them for particular phenomena of second language acquisition (SLA) or foreign language teaching (FLT) research purposes [77]. In teaching language for special purposes (LSP) techniques such as term extraction and the creation of a glossary are also popular [22]. Traditionally the focus is on a bottom-up approach. Learning starts with a word and goes on to word combinations, relations, and context.

ELDIT differs from these approaches in that the underlying material has been carefully designed by linguists and language teachers for the purpose of language learning. Only afterwards the material has been annotated. Beyond the bottom-up approach (learning word groups), ELDIT offers also a top-down approach (practicing texts). Moreover, it is possible to combine the two approaches.

Data Model Recently much research has been carried out to develop systems, data models, and standards for Web-based learning in general, e.g. WebCT, Hyperwave, KBS-Hyperbook, InterBook, SCORM, LOM, etc. Several authors developed data models which are especially designed to represent and share teaching material over the Web. In [89] a data model to support constructivist learning in the KBS-Hyperbook system is described, in [156] a meta-modeling approach to adaptive hypermedia-based electronic teachware is described. These approaches focus on document structures and navigation services. The main difference of these systems 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 model which breaks the learning material down to the level of single words and further.

Encoding in XML Recently the advantages of XML and related standards have been increasingly exploited for language learning. The KirrKirr system is an application that allows users to explore a Warlpiri (a Central Australian language) dictionary [97]. The data are represented in XML. XSL is used for enhanced customization, and the

XQL[3] language is used to query the dictionary entries. CoCoaJ is a system for writing in Japanese [129] which allows students and teachers to exchange documents over the Internet. The system includes a writing error analysis model by which typical morphological errors can be detected. For the annotation of documents with remarks and comments the eXtensible Communicative Correction Mark-up language (XCCML), which is based on XML, was developed. The authoring tool WHURLE [24] and the multi-agent learning system IDLE [147] are two learning tools which use a similar approach to data management as we do in ELDIT. Educational content is provided in chunks of XML data. They are automatically linked according to an indicated lesson plan represented in a dependency graph. The last two systems also include a user model and adapt content presentation to each individual learner. In ELDIT, however, we have encoded the data in much more detail. The fine-grained annotation allows a rich link structure as well as reusing data and providing a better adaptation to the individual learner.

9.2 Generalizations

There are several interesting possibilities to generalize the results of our work. It would be possible to implement the ELDIT system for other language pairs. Different contrastive aspects as well as different combination possibilities of e.g. the dictionaries would arise. The system could also be implemented for a further language. Monolingual word entries would have to be developed and in the default version of the dictionary a second translation could appear below the first one. A text corpus would also have to be developed. The users could then indicate their native language as well as the target language they want to learn. Hiding techniques could be used to adapt the system to the users, e.g. by showing only the relevant translations.

Such extensions would be an interesting task for linguists and language teachers. For computer scientists, however, it would be interesting to apply the approach we developed for content authoring in a more general way. This idea will be explained in the following.

Adaptive educational hypermedia systems (AEH) are often used to support explorative learning and constructivist teaching. They aim at encouraging individual learning and deal with the problem of "not ready yet to be learned" material by hiding it or by showing it with a warning annotation. In this approach, however, the learning material has to be structured very well and in a special way: the author has to define a multi-dimensional dependency graph and to give explicit indications about all learning dependences between the concepts. Preparing a content space in this way, however, is a large amount of work.

A second problem arises when evaluations of AEH systems are studied. Especially weaker learners tend to continuously click on the "next" button when working with the system [174, 93]. Hence it is questionable if in these cases adaptation techniques are really the best way to support explorative learning.

Our approach to content authoring may be interesting not only for a language learning system. It could also be useful for general educational or information retrieval systems, eventually as a complement to other techniques such as user modeling, tutoring

[3] http://www.ibiblio.org/xql/

and adaptation. Such systems usually include many technical terms in their descriptive material. However, in all systems where some kind of exploration is asked it is very likely that a user has not learned yet or forgotten the technical terms used in the explaining text sections. Hence it would be very interesting for such a system to encode the data down to the level of single words, to include a dictionary, and to link all special terms to the corresponding dictionary entry. In this way free browsing and exploring is supported, since all unknown terms can be checked by one simple mouse click.

Content authoring would not require much extra work, since the corresponding encoding and linking tools can be provided by a computer program as it has been done in the ELDIT system. Depending on the system specifications and users' needs other tools which allow for an innovative data presentation could be included.

Chapter 10

Conclusion

The result of our work is a new approach to CALL content authoring and a fully-fledged language learning system which is sophisticated and easily extendable. As resumed in section 10.1 our approach to data management has proved its power and flexibility in a number of cases. In section 10.2 we summarize the innovative features of the ELDIT vocabulary acquisition system. In section 10.3 we summarize the efforts we made to create not only a research prototype but a real world system. Section 10.4 is dedicated to describing the current situation of the project and to outlining future plans.

10.1 A New Approach to CALL Content Authoring

Inserting the data in the proposed semi-structured way was appreciated and judged as feasible by the linguists in our team. Rewriting the data electronically allowed enriching the data by additional information which yielded a powerful database for the computer scientists.

The effectiveness of our approach to data management, its flexibility regarding the implementation of didactic demands, and its robustness regarding system changes has been confirmed for each new module we implemented for ELDIT:

- The detailed data model allowed fulfilling sophisticated linguistic and didactic requirements regarding presentation and interaction.
- Rewriting and enriching the data made it possible to create such a large amount of data that the system can be used as a real world system.
- Due to the flexibility of the transfer program frequent changes to the demands were no more an insuperable problem.
- The detailed data model allowed reusing data and software up to a very high degree.
- Again due to the fine grained data model the inclusion of external modules that worked only on specific language features was possible.

10.2 An Innovative Language Learning System

ELDIT provides many innovative features, both at a didactic and at a technological level. In the following we list just the most important ones.

Dictionary The dictionary is a completely new type of dictionary, namely a crosslingual dictionary. It is not a copy from paper to screen but fully exploits the electronic medium. It has been created according to the most modern psycholinguistic and didactic demands. The implementation of some of the modules, for instance verb valency and inflection, are unique.

Texts The text corpus is outstanding in that learners are strongly supported during the reading process. About 90% of the words will have a link to the corresponding dictionary entry. An authoring tool allows teachers to submit their own texts and to prepare them in real time in the same way as the system texts have been prepared. Text sentences can be reused as example sentences and texts can not only be read but also practiced as a gap-filling exercise.

Quizzes It is possible to electronically create a large number of quizzes of the existing data set. The quizzes are parameterized and vary according to different dimensions. High level corrections, detailed feedback, as well as meaningful remediation can be provided. The large number of quizzes and possibilities to present them can be handled by adaptation techniques.

Tandem This module can help to establish international friendships and might contribute to facilitating the complicated sociolinguistic situation in South Tyrol. Adaptation techniques help to establish learning partnerships and bring together users who are well suited to each other thanks to similar personal interests and learning preferences.

Tutor Up to now only a few adaptive educational hypermedia language learning systems have been developed. In ELDIT an adaptive tutor provides guidance through the learning material by exploiting traditional adaptation techniques and combining them with computational linguistic technologies. The tutor teaches vocabulary according to a new method we call “contextualized, adaptive vocabulary acquisition” which should help to eliminate some commonly known problems in vocabulary acquisition.

Extensions Both educational and illustrative content are reused throughout the system. Additional examples can be requested for each piece of information. Every word used in the system was electronically linked to the corresponding dictionary entry. These features can be combined with each other and reused for external applications as well. The system includes a user model. Hence content presentation and navigation can be adapted to the individual user.

External Products An advanced search feature was developed especially for language learners. This feature is used by the users as well as by the system to retrieve relevant and didactically meaningful information. The inclusion of Word Manager

into ELDIT has been successfully realized. The inclusion of GermaNet would easily be possible. Other products can be included if they provide a platform independent interface, e.g. in Java or in XML.

10.3 A Real World System

The ELDIT system is not only a research prototype but a real world system that is suitable for learning and working. The data that have been developed up to now amount to more than 8,000 word entries, 2,000 word groups, and 800 texts. The system is easily accessible via the Internet, a stable client-server environment is provided, and usability features are included. The system is continuously available in several versions, namely for production (user version), for demonstrations (demo version), for data development (beta version), and for software development (alpha version).

Since the system contains the educational material of the exams in bilingualism it provides specific support for exam preparation. According to our opinion many of the problems mentioned in section 1.1 can also be addressed. Thanks to network technologies geographical and social separations of the two ethnic groups can be overcome. The unwillingness to study the other language might be reduced, since computer assistance has proved to be a successful motivator for language learning.

Furthermore the system is designed in a general way so that it can be used in other language learning situations as well. The increasing economic co-operation with the alpine regions that border South Tyrol requires the knowledge of the neighbor's language. Therefore the ELDIT system might be especially useful for the Austrian regions North and East Tyrol, the Italian region of Trento, and the Swiss region of Graubünden. In fact, the online statistics show that the system is used not only locally but in the whole of Europe and beyond.

10.4 Current Situation and Future Work

ELDIT is still ongoing work. The large interdisciplinary project has been going on since September 1999. Currently 3 German and 3 Italian linguists and 2 computer scientists (the author of this thesis included) work on the system.

The core part of the dictionary and its data elaboration process have already been concluded. The modules for content reuse, the search engine, the inclusion of Word Manager, as well as the interface for external applications have been concluded as well. Some additional modules for the dictionary (semantic fields and word formation) are scope of an ongoing Interreg project[1] and will be finished in September 2004.

The text corpus is scope of an ongoing project funded by the local government of South Tyrol[2], it has been implemented but will be refined with additional modules (generation of hints, recording of the answers in the user model, exemplary answers, etc.).

For the exercises and the online authoring tool prototypes have been implemented and tested, concrete didactic concepts and extensive evaluations are planned for the future.

[1] http://europa.eu.int/comm/regional_policy/

[2] http://www.provinz.bz.it

For the tandem module a small prototype has been programmed, the final realization is planned to be carried out in co-operation with a local language school that has extensive experience in face-to-face tandem learning[3].

Up to now user modeling has been implemented only in a restricted version and it has been used mainly to record statistical data and to carry out evaluations. Refinements would be necessary to implement the adaptation possibilities mentioned and the adaptive tutor.

The adaptive tutor is the core module of the system. Its realization depends on all the other modules. Hence it has been conceptualized but not implemented yet.

[3] http://www.alphabeta.it

Appendix A

Summary of Awards

We have received the following awards for the system or for scientific publications of it:

- In December 2002 we have been awarded the *Anerkennungspreis 2002 für Forschungs- und Studienprojekte der Eduard-Wallnöfer-Stiftung*[1] for the ELDIT vocabulary acquisition system.
- In June 2003 we received a *best paper award* at the World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA 2003) for our approach to a learner friendly description of verb valency in the ELDIT dictionary.
- In June 2002 we have received a *best session paper award* at the Sixth World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002) for our approach to adaptive tutoring in a language learning system.
- In June 2003 our work on providing the inflection paradigm for each ELDIT entry has been *nominated* for a *best poster award* at the Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA 2003).

[1] http://www.industrie-tirol.at/Initiativen/wallnoefer.html

Appendix B

The DTDs

B.1 DTD of Words

```
<!--

word: the ELDIT dictionary contains about 4,000 word entries for each
language. A word can be a noun, a verb, an adjective or a structure
word.

The attributes are as follows:
language: the language of the word entry
class: the word class in abbreviated form (N,V,A,S)
index: 1, 2, 3, ... in the case of homonymy
domain: a number between 1 and 20 that indicates the interest domain
frequency: frequency of the word in the text corpus

-->

<!ELEMENT word (noun|verb|adjective|structural)>
<!ATTLIST word  language CDATA #IMPLIED class CDATA #IMPLIED
                index CDATA #IMPLIED domain CDATA #IMPLIED
                frequency CDATA #IMPLIED>



<!--

Nouns, verbs, and adjectives consist of the following elements:

lemma: the lemma, i.e. the base or citation form
morphology: article and plural form
comment: a comment to describe particularities, e.g. if a word
         is a compound word
adverb: for adjectives the corresponding adverb is given
insertElement: the ID of an external lemma dependent word group,
               for instance a group of derivations
sense: different word meanings
reference: a reference to another word or word group
idiomExpr: an idiomatic expression in which this word occurs
footnote: linguistic difficulty accessible by a footnote number
```

```
-->

<!ELEMENT noun (lemma?,morphology?,comment*,insertElement*,sense*,
                reference*,idiomExpr*,footnote*)>
<!ATTLIST noun id ID #IMPLIED>
<!ELEMENT verb (lemma?,morphology?,comment*,insertElement*,sense*,
                reference*,idiomExpr*,footnote)>
<!ATTLIST verb id ID #IMPLIED>
<!ELEMENT adjective (lemma?,morphology?,comment*,adverb?,
                     insertElement*,sense*,reference*,idiomExpr*,
                     footnote)>
<!ATTLIST adjective id ID #IMPLIED>



<!--

Single structure words are given only as provisory entries. They
are describe in more detail within the themes word groups.

-->

<!ELEMENT structural (lemma?,morphology?)>
<!ATTLIST structural id ID #IMPLIED>



<!--

Adverbs are described in the Italian adjectives, they consist of a
form ending in ‘‘-mente’’. Different senses are described.

-->

    <!ELEMENT adverb (form?, adverbSubsense*)>
    <!ATTLIST adverb id ID #IMPLIED>
        <!ELEMENT form (w*)>
        <!ATTLIST form id ID #IMPLIED>



<!--

Each sense is divided into subsenses. External sense dependent word
groups can be inserted by an ID given in insertElement. Collocations
and describing adjectives are listed.

-->

    <!ELEMENT sense (nounSubsense*,verbSubsense*,adjectiveSubsense*,
                     insertElement*,collocation*,description*)>
    <!ATTLIST sense id ID #IMPLIED>
```

```
<!--

Each word sense is described by a definition, translations and
lexicographic examples. Verbs are furthermore described by several
verb valency models.

-->

     <!ELEMENT adverbSubsense (definition?,translation*,example*)>
     <!ATTLIST adverbSubsense id ID #IMPLIED >
     <!ELEMENT nounSubsense (definition?,translation*,example*)>
     <!ATTLIST nounSubsense id ID #IMPLIED >
     <!ELEMENT verbSubsense (definition?,model*,translation*)>
     <!ATTLIST verbSubsense id ID #IMPLIED>
     <!ELEMENT adjectiveSubsense (definition?,translation*,example*)>
     <!ATTLIST adjectiveSubsense id ID #IMPLIED >



<!--

For the presentation of verb valency corresponding parts in the
pattern, comments and example sentences have to be indicated. Hence
the data is divided into cells and subcells. Correspondence is
indicated by an index given in the attribute ‘‘partofs’’.

-->

     <!ELEMENT model (realisation*)>
     <!ATTLIST model id ID #IMPLIED>
        <!ELEMENT realisation (modpattern*,modcomment*,modexample*)>
        <!ATTLIST realisation id ID #IMPLIED>
        <!ELEMENT modpattern (cell*)>
        <!ATTLIST modpattern id ID #IMPLIED>
        <!ELEMENT modcomment (cell*)>
        <!ATTLIST modcomment id ID #IMPLIED>
        <!ELEMENT modexample (cell*)>
        <!ATTLIST modexample class CDATA #IMPLIED id ID #IMPLIED>
           <!ELEMENT cell (subcell*)>
           <!ATTLIST cell partofs CDATA #IMPLIED id ID #IMPLIED>
              <!ELEMENT subcell (subsubcell*)>
              <!ATTLIST subcell id ID #IMPLIED>
                 <!ELEMENT subsubcell (w*)>
                 <!ATTLIST subsubcell id ID #IMPLIED>



<!--

Collocations and idiomatic expressions are described by a pattern, a
comment, translations, and examples. For idiomatic expressions also an
explanation can be added.

-->

        <!ELEMENT collocation (pattern?,comment?,
                              translation*,example*)>
```

```
        <!ATTLIST collocation id ID #IMPLIED>
        <!ELEMENT idiomExpr (pattern?,comment?,explanation*,
                             translation*,example*)>
        <!ATTLIST idiomExpr id ID #IMPLIED>



<!--

Describing adjectives are given for each noun. They consist of a kind
of title (statement), the describing adjective itself (quality), its
translation, and some lexicographic examples.

-->

        <!ELEMENT description (statement*,characteristic*,example*)>
        <!ATTLIST description id ID #IMPLIED>
            <!ELEMENT statement (w*)>
            <!ATTLIST statement id ID #IMPLIED>
            <!ELEMENT characteristic (quality?,translation*)>
            <!ATTLIST characteristic id ID #IMPLIED >
                <!ELEMENT quality (w*)>
                <!ATTLIST quality id ID #IMPLIED >



<!--

For all elements each word is encoded separately to realize the link
structure of the system.

-->

<!ELEMENT reference (w*)>
<!ATTLIST reference id ID #IMPLIED>
<!ELEMENT footnote (w*)>
<!ATTLIST footnote id ID #IMPLIED lang CDATA #IMPLIED>
<!ELEMENT morphology (w*,gramNotice*)>
<!ATTLIST morphology id ID #IMPLIED>
<!ELEMENT definition (w*)>
<!ATTLIST definition id ID #IMPLIED>
<!ELEMENT translation (w*)>
<!ATTLIST translation id ID #IMPLIED>
<!ELEMENT example (w*)>
<!ATTLIST example class CDATA #IMPLIED id ID #IMPLIED>
<!ELEMENT pattern (w*)>
<!ATTLIST pattern id ID #IMPLIED>
<!ELEMENT explanation (w*)>
<!ATTLIST explanation id ID #IMPLIED lang CDATA #IMPLIED>
<!ELEMENT comment (w*)>
<!ATTLIST comment id ID #IMPLIED>
<!ELEMENT gramNotice (w*)>
<!ATTLIST gramNotice id ID #IMPLIED>
<!ELEMENT insertElement (#PCDATA)>
<!ATTLIST insertElement id ID #IMPLIED target CDATA #IMPLIED>
<!ELEMENT lemma (#PCDATA)>
<!ATTLIST lemma id ID #IMPLIED >
```

```
<!--

At the lowest level for each word the following information is encoded:

type: indicates whether this word belongs to the content or
      whether it is a remark
style: this is an indication for presentation and has currently the
       values ‘‘normal’’ or ‘‘emphasized’’
nbref: reference-ID to a footnote
base: the citation form of a word
ctag: the word class (N,V,A,...) of a word
lexref: the ID to the corresponding meaning description in ELDIT
collref: the ID to the corresponding collocation description in ELDIT

-->

<!ELEMENT w (#PCDATA)>
<!ATTLIST w id ID #IMPLIED
        type CDATA #IMPLIED
        style CDATA #IMPLIED
        nbref CDATA #IMPLIED
        base CDATA #IMPLIED
        ctag CDATA #IMPLIED
        lexref CDATA #IMPLIED
        collref CDATA #IMPLIED>
```

B.2 DTD of Texts

```
<!--

text: the ELDIT system contains about 400 texts for each language. A
texts consists of meta information, a title, the text body, a
bibliographical element (reference) and 6 questions.

-->

<!ELEMENT text (meta?,title?,body,reference?,question*)>
<!ATTLIST text id ID #IMPLIED>



<!--

Meta information includes the author of a text (the name of a user),
the source (‘‘system’’ or ‘‘user’’), the ID of the user that has
authored this text, the language of the text, the level of difficulty,
the interest domain the text belongs to, a sequential number for
unique identification, and the text frequency number.
```

```
-->

  <!ELEMENT meta (author?, source?, userid?, lang?, level?,
                  domain*, index?, freq?)>
  <!ATTLIST meta id ID #IMPLIED>
        <!ELEMENT author (#PCDATA)>
        <!ATTLIST author id ID #IMPLIED>
        <!ELEMENT source (#PCDATA)>
        <!ATTLIST source id ID #IMPLIED>
        <!ELEMENT userid (#PCDATA)>
        <!ATTLIST userid id ID #IMPLIED>
        <!ELEMENT lang (#PCDATA)>
        <!ATTLIST lang id ID #IMPLIED>
        <!ELEMENT level (#PCDATA)>
        <!ATTLIST level id ID #IMPLIED>
        <!ELEMENT domain (#PCDATA)>
        <!ATTLIST domain id ID #IMPLIED>
        <!ELEMENT index (#PCDATA)>
        <!ATTLIST index id ID #IMPLIED>
        <!ELEMENT freq (#PCDATA)>
        <!ATTLIST freq id ID #IMPLIED>



<!--

All sentence-like elements such as title, question, etc. consist of
words (w), a piece of spoken language (q), or a translation of a
difficult expression in brackets. The translations stem from the paper
based material for the exams in bilingualism and are kept so that the
presentation of the online texts is equivalent to the presentation of
paper based ones.

-->

  <!ELEMENT title (w|translation|q)*>
  <!ATTLIST title id ID #IMPLIED>



<!--

The text body consists of paragraphs which consist of sentences.
Sentences consist of pieces of spoken language (quotes), translations
or single words.

-->

  <!ELEMENT body (p*)>
  <!ATTLIST body id ID #IMPLIED>
    <!ELEMENT p (s*)>
    <!ATTLIST p id ID #IMPLIED>
      <!ELEMENT s (w|translation|q)*>
      <!ATTLIST s id ID #IMPLIED>
```

```
<!--

The bibliographical indication is divided into name which is the
source of the text, and note which might be a comment such as
‘‘abbreviated and reformulated version’’.

-->

  <!ELEMENT reference (name?,note?)>
  <!ATTLIST reference id ID #IMPLIED>
    <!ELEMENT name (w|translation|q)*>
    <!ATTLIST name id ID #IMPLIED>
    <!ELEMENT note (w|translation|q)*>
    <!ATTLIST note id ID #IMPLIED>



<!--

Questions are sentence-like text pieces which consist of quotes,
translations or words (see element title). The attribute sref contains
reference-IDs to some text sentences that could be helpful for the
answer.

-->

  <!ELEMENT question (w|translation|q)*>
  <!ATTLIST question id ID #IMPLIED
                    sref CDATA #IMPLIED>



<!--

Translations can again contain quotes, and quotes can contain
translations. All elements consist of words (w).

-->

        <!ELEMENT translation (w|q)*>
        <!ATTLIST translation id ID #IMPLIED lang CDATA #IMPLIED>

        <!ELEMENT q (w|translation)*>
        <!ATTLIST q id ID #IMPLIED broken CDATA #IMPLIED
                    next IDREF #IMPLIED prev IDREF #IMPLIED>



<!--

The lowest level element w is the same as in the dictionary entries
and has been explained above.

-->

         <!ELEMENT w (#PCDATA)>
         <!ATTLIST w id ID #IMPLIED
```

```
                        type CDATA #IMPLIED
                        style CDATA #IMPLIED
                        base CDATA #IMPLIED
                        ctag CDATA #IMPLIED
                        lexref CDATA #IMPLIED
                        collref CDATA #IMPLIED
                        nbref CDATA #IMPLIED>
```

B.3 DTD of Semantic Fields

```
<!--

semantic field: the ELDIT dictionary contains about 250 meaning
dependent semantic fields for each language. A semantic field shows
words related to a word entry of the dictionary.

The attributes are as follows:
insertplace: an indication where the field belongs to
             (i.e. the ID of a specific word meaning)
instance: each field might consist of several instances
nbs: linguistic difficulties are explained in footnotes

-->

<!ELEMENT field (insertplace*,instance*,nbs*)>
<!ATTLIST field id ID #IMPLIED>

    <!ELEMENT insertplace (#PCDATA)>
    <!ATTLIST insertplace id ID #IMPLIED>



<!--

A field is divided into three levels: the upper level shows words
which are more general than the word entry, the lower level shows
words which are more special than the word entry, and the main level
shows relations at the same level. Words that form one relation are
grouped and ordered within these groups.

-->

    <!ELEMENT instance (upperlevel?, mainlevel?, lowerlevel?)>
    <!ATTLIST instance id ID #IMPLIED>

      <!ELEMENT upperlevel (hyperonymgroup*,holonymgroup*)>
      <!ATTLIST upperlevel id ID #IMPLIED>

      <!ELEMENT mainlevel (lemma?,relationgroup*)>
      <!ATTLIST mainlevel id ID #IMPLIED>

        <!ELEMENT relationgroup (refobject?,synonymgroup*,
```

```
                                quasisynonymgroup*,antonymgroup*,
                                entailmentgroup*,causationgroup*)>
        <!ATTLIST relationgroup id ID #IMPLIED>

      <!ELEMENT lowerlevel (f_hyponymgroup*,a_hyponymgroup*,
                            meronymgroup*,troponymgroup*,
                            particleverbgroup*)>
      <!ATTLIST lowerlevel id ID #IMPLIED>



<!--

Each group has a group name and consists of several field words.
Each field word is described either by a reference to the
corresponding dictionary entry if it exists already in the
dictionary (elditref) or by the following information:

lemma: the citation form
morphology: article and plural form
definition: the definition of a word meaning
translation: several translations
example: lexicographic examples
comment: comments and explanations
connection: this element describes the difference between the field
            word and the main lemma.
importance: within one group the field words can be sorted
occurrence: indicates in which word class this element occurs
relkind: indicated the kind of relation, e.g. is-a, or part-of
face: indicates an opposite relation, e.g. holonym is an opposite
      relation to meronym

-->



<!--

Hyperonyms form an is-a relation to the word entry, for instance
‘‘Mensch’’ (men) is a ‘‘Lebewesen’’ (living being).

-->

      <!ELEMENT hyperonymgroup (groupname?, hyperonym*)>
      <!ATTLIST hyperonymgroup id ID #IMPLIED>

          <!ELEMENT hyperonym (elditref?,fi_lemma?,morphology*,
                               translation*,comment*,definition*,
                               example*,connection*)>
          <!ATTLIST hyperonym id ID #IMPLIED
                              importance CDATA #IMPLIED
                              occurrence CDATA #FIXED "NA"
                              relkind CDATA #FIXED "is_a"
                              face CDATA #FIXED "hyponym">
```

```
<!--

Holonyms form a ‘‘part-of’’ relation to the word entry, for instance
‘‘Arm’’ (arm) is a part of the ‘‘Körper’’ (body).

-->

          <!ELEMENT holonymgroup (groupname?, holonym*)>
          <!ATTLIST holonymgroup id ID #IMPLIED>

              <!ELEMENT holonym (elditref?,fi_lemma?,morphology*,
                                 translation*,comment*,definition*,
                                 example*,connection*)>
              <!ATTLIST holonym id ID #IMPLIED
                                importance CDATA #IMPLIED
                                occurrence CDATA #FIXED "N"
                                relkind CDATA #FIXED "part_of"
                                face CDATA #FIXED "meronym">



<!--

Synonyms are word with (more or less) the same meaning as the word
entry, e.g. ‘‘Dame’’ (lady) is a synonym of ‘‘Frau’’ (woman).

-->

          <!ELEMENT synonymgroup (groupname?, synonym*)>
          <!ATTLIST synonymgroup id ID #IMPLIED>

              <!ELEMENT synonym (elditref?,fi_lemma?,morphology*,
                                 translation*,comment*,definition*,
                                 example*,connection*)>
              <!ATTLIST synonym id ID #IMPLIED
                                importance CDATA #IMPLIED
                                occurrence CDATA #FIXED "NVA"
                                relkind CDATA #FIXED "is_equivalent">


<!--

Quasi synonyms are words with a meaning similar to the word entry,
e.g. ‘‘Tussi’’ (a female) is a quasi synonym of ‘‘Frau’’ (woman).

-->

          <!ELEMENT quasisynonymgroup (groupname?, quasisynonym*)>
          <!ATTLIST quasisynonymgroup id ID #IMPLIED>

              <!ELEMENT quasisynonym (elditref?,fi_lemma?,morphology*,
                                      translation*,comment*,definition*,
                                      example*,connection*)>
              <!ATTLIST quasisynonym id ID #IMPLIED
                                     importance CDATA #IMPLIED
                                     occurrence CDATA #FIXED "NVA"
                                     relkind CDATA #FIXED "resembles">
```

```
<!--

Antonyms are a kin of opposite of the word entry, e.g. ‘‘kalt’’ (cold)
is the opposite of ‘‘warm’’ (warm), ‘‘Tag’’ (day) is the opposite of
‘‘Nacht’’ (night).

-->

          <!ELEMENT antonymgroup (groupname?, antonym*)>
          <!ATTLIST antonymgroup id ID #IMPLIED>

              <!ELEMENT antonym (elditref?,fi_lemma?,morphology*,
                                 translation*,comment*,definition*,
                                 example*,connection*)>
              <!ATTLIST antonym id ID #IMPLIED
                              importance CDATA #IMPLIED
                              occurrence CDATA #FIXED "NVA"
                              relkind CDATA #FIXED "contrasts_with">



<!--

Entailments describe an inclusion, e.g. ‘‘essen’’ (to eat) includes
the fact that the person is also doing ‘‘kauen’’ (to chew).

-->

          <!ELEMENT entailmentgroup (groupname?, entailment*)>
          <!ATTLIST entailmentgroup id ID #IMPLIED>

              <!ELEMENT entailment (elditref?,fi_lemma?,morphology*,
                                    translation*,comment*,definition*,
                                    example*,connection*)>
              <!ATTLIST entailment id ID #IMPLIED
                                   importance CDATA #IMPLIED
                                   occurrence CDATA #FIXED "V"
                                   relkind CDATA #FIXED "entails">



<!--

Causation describes an implication, e.g. ‘‘töten’’ (to kill)
causes ‘‘sterben’’ (to die).

-->

          <!ELEMENT causationgroup (groupname?, causation*)>
          <!ATTLIST causationgroup id ID #IMPLIED>

              <!ELEMENT causation (elditref?,fi_lemma?,morphology*,
                                   translation*,comment*,definition*,
                                   example*,connection*)>
```

```
            <!ATTLIST causation id ID #IMPLIED
                               importance CDATA #IMPLIED
                               occurrence CDATA #FIXED "V"
                               relkind CDATA #FIXED "causes">


<!--

F_hyponyms are words that indicate a specific function,
e.g. ‘‘Hausfrau’’ (housewife) is a specific ‘‘Frau’’ (woman), namely a
woman that works at home.

-->

          <!ELEMENT f_hyponymgroup (groupname?, f_hyponym*)>
          <!ATTLIST f_hyponymgroup id ID #IMPLIED>

              <!ELEMENT f_hyponym (elditref?,fi_lemma?,morphology*,
                                   translation*,comment*,definition*,
                                   example*,connection*)>
              <!ATTLIST f_hyponym id ID #IMPLIED
                               importance CDATA #IMPLIED
                               occurrence CDATA #FIXED "N"
                               relkind CDATA #FIXED "inv_is_a"
                               face CDATA #FIXED "hyperonym"
                               respective CDATA #FIXED "particleverb">


<!--

A_hyponyms are words that indicate a specific mode, e.g. ‘‘Villa’’
(villa) is a ‘‘Haus’’ (house) in a specific mode, namely a more
elegant house.

-->

          <!ELEMENT a_hyponymgroup (groupname?, a_hyponym*)>
          <!ATTLIST a_hyponymgroup id ID #IMPLIED>

              <!ELEMENT a_hyponym (elditref?,fi_lemma?,morphology*,
                                   translation*,comment*,definition*,
                                   example*,connection*)>
              <!ATTLIST a_hyponym id ID #IMPLIED
                               importance CDATA #IMPLIED
                               occurrence CDATA #FIXED "N"
                               relkind CDATA #FIXED "inv_is_a"
                               face CDATA #FIXED "hyperonym"
                               respective CDATA #FIXED "troponym">


<!--

Meronyms form a part of relation, e.g. ‘‘Fenster’’ (window) is a part
of a ‘‘Haus’’ (house).
```

```
-->

          <!ELEMENT meronymgroup (groupname?, meronym*)>
          <!ATTLIST meronymgroup id ID #IMPLIED>

              <!ELEMENT meronym (elditref?,fi_lemma?,morphology*,
                                 translation*,comment*,definition*,
                                 example*,connection*)>
              <!ATTLIST meronym id ID #IMPLIED
                                importance CDATA #IMPLIED
                                occurrence CDATA #FIXED "N"
                                relkind CDATA #FIXED "inv_part_of"
                                face CDATA #FIXED "holonym">



<!--

Troponyms show a specific kind to do something, e.g. ‘‘wandern’’ (to
walk) is a specific way of ‘‘gehen’’ (to go).

-->

          <!ELEMENT troponymgroup (groupname?, troponym*)>
          <!ATTLIST troponymgroup id ID #IMPLIED>

              <!ELEMENT troponym (elditref?,fi_lemma?,morphology*,
                                  translation*,comment*,definition*,
                                  example*,connection*)>
              <!ATTLIST troponym id ID #IMPLIED
                                 importance CDATA #IMPLIED
                                 occurrence CDATA #FIXED "V"
                                 relkind CDATA #FIXED "inv_is_a"
                                 respective CDATA #FIXED "a_hyponym">



<!--

Particle verbs are verbs with prefixes, e.g. ‘‘weglaufen’’ (to run
away) is a particle verb of ‘‘laufen’’ (to run).

-->

          <!ELEMENT particleverbgroup (groupname?, particleverb*)>
          <!ATTLIST particleverbgroup id ID #IMPLIED>

          <!ELEMENT particleverb (elditref?,fi_lemma?,morphology*,
                                  translation*,comment*,definition*,
                                  example*,connection*)>
          <!ATTLIST particleverb id ID #IMPLIED
                                 importance CDATA #IMPLIED
                                 occurrence CDATA #FIXED "V"
                                 relkind CDATA #FIXED "inv_is_a"
                                 respective CDATA #FIXED "f_hyponym">
```

```
<!--

Each element is coded down to the level of single words and can
include footnotes which indicate linguistic difficulties. Also in
the footnotes each word is encoded separately.

-->

                <!ELEMENT elditref (#PCDATA)>
                <!ATTLIST elditref id ID #IMPLIED>

                <!ELEMENT groupname (w*,nbs?)>
                <!ATTLIST groupname id ID #IMPLIED>

                <!ELEMENT lemma (w*,nbs?)>
                <!ATTLIST lemma id ID #IMPLIED>

                <!ELEMENT refobject (w*,nbs?)>
                <!ATTLIST refobject id ID #IMPLIED>

                <!ELEMENT fi_lemma (w*,nbs?)>
                <!ATTLIST fi_lemma id ID #IMPLIED>

                <!ELEMENT morphology (w*,nbs?)>
                <!ATTLIST morphology id ID #IMPLIED>

                <!ELEMENT translation (w*,nbs?)>
                <!ATTLIST translation id ID #IMPLIED>

                <!ELEMENT comment (w*,nbs?)>
                <!ATTLIST comment id ID #IMPLIED>

                <!ELEMENT definition (w*,nbs?)>
                <!ATTLIST definition id ID #IMPLIED>

                <!ELEMENT example (w*,nbs?)>
                <!ATTLIST example id ID #IMPLIED>

                <!ELEMENT connection (w*,nbs?)>
                <!ATTLIST connection id ID #IMPLIED>

                    <!ELEMENT nbs (nb*)>
                    <!ATTLIST nbs id ID #IMPLIED>
                        <!ELEMENT nb (w*)>
                        <!ATTLIST nb id ID #IMPLIED lang CDATA #IMPLIED>


<!--

For the lowest level element w we refer to the documentation in the DTD for
dictionary entries.

-->
```

```
                                <!ELEMENT w (#PCDATA)>
                                <!ATTLIST w id ID #IMPLIED
                                        type CDATA #IMPLIED
                                        style CDATA #IMPLIED
                                        nbref CDATA #IMPLIED
                                        base CDATA #IMPLIED
                                        ctag CDATA #IMPLIED
                                        lexref CDATA #IMPLIED
                                        collref CDATA #IMPLIED>
```

B.4 DTD of Word Formation

```
<!--

word family: the ELDIT dictionary shows about 3,000 meaning dependent
groups of compound words and about 1,000 lemma dependent groups of
derivations for each language. Compound words and derivations together
form the word family for a word entry and are shown to each dictionary
entry.

-->

<!ELEMENT family (derivation*,nbs*)>
<!ATTLIST family id ID #IMPLIED>



<!--

A derivation consists of the word itself (pattern), some comments,
and some translations.

The attributes have the following meanings:

insertplace: if this attribute has the value ‘‘yes’’ the derivation
             is an entry of the dictionary and the whole group should
             be shown within this word entry
padre: each family has one father element, the most simple element
       where the others derive from
dictfreq: indicates how often a word is occurring in the dictionary
textfreq: indicates how often the word occurs in the texts
degree: degree of derivation in Word Manager and in ELDIT respectively
parentderiv: the parent derivation the word is derived from

-->

    <!ELEMENT derivation (pattern*,comment*,translation*)>
    <!ATTLIST derivation id ID #IMPLIED insertplace (yes|no) "no"
                                       padre (yes|no) "no"
                                       dictfreq CDATA #IMPLIED
                                       textfreq CDATA #IMPLIED
```

```
                                        wmDegree CDATA #IMPLIED
                                        wmParentDeriv CDATA #IMPLIED
                                        elditDegree CDATA #IMPLIED
                                        elditParentDeriv CDATA #IMPLIED>


<!--

In order to show the word formation process to the learner a
derivation is indicated with article, prefixes, the basis, suffixes,
and possible some more words such as a reflexive form ‘‘sich’’.

The attributes are as follows:

splitting: indicates whether the splitting has been created manually by
           the linguists or whether it is a proposal of Word Manager
base: the citation form of a derivation
ctag: the word form (N,V,A) of a derivation
lexref: the ID to the corresponding word entry in the ELDIT dictionary if
        it exists

-->

            <!ELEMENT pattern (article*,praefix*,basis?,suffix*,w*)>
            <!ATTLIST pattern id ID #IMPLIED
                      splitting (eldit|wm) "eldit"
                      base CDATA #IMPLIED
                      ctag CDATA #IMPLIED
                      lexref CDATA #IMPLIED>



<!--

Comments, translations and linguistic difficulties consists of
several words.

-->

            <!ELEMENT comment (w*)>
            <!ATTLIST comment id ID #IMPLIED>

            <!ELEMENT translation (w*)>
            <!ATTLIST translation id ID #IMPLIED source CDATA #IMPLIED>

            <!ELEMENT nbs (nb*)>
            <!ATTLIST nbs id ID #IMPLIED nbid CDATA #IMPLIED>
                    <!ELEMENT nb (w*)>
                    <!ATTLIST nb id ID #IMPLIED lang CDATA #IMPLIED>



<!--

Article, praefix, basis, suffix, and w are given in plain
text˜(PCDATA).  Article and w have some more attributes (described
```

```
already in the DTD for dictionary entries above).

-->

                        <!ELEMENT article (#PCDATA)>
                        <!ATTLIST article explref CDATA #IMPLIED
                                  base CDATA #IMPLIED
                                  ctag CDATA #IMPLIED
                                  lexref CDATA #IMPLIED>
                        <!ELEMENT praefix (#PCDATA)>
                        <!ATTLIST praefix explref CDATA #IMPLIED>
                        <!ELEMENT basis (#PCDATA)>
                        <!ATTLIST basis explref CDATA #IMPLIED>
                        <!ELEMENT suffix (#PCDATA)>
                        <!ATTLIST suffix explref CDATA #IMPLIED>
                        <!ELEMENT w (#PCDATA)>
                        <!ATTLIST w id ID #IMPLIED
                                type CDATA #IMPLIED
                                style CDATA #IMPLIED
                                nbref CDATA #IMPLIED
                                base CDATA #IMPLIED
                                ctag CDATA #IMPLIED
                                lexref CDATA #IMPLIED
                                collref CDATA #IMPLIED>
```

B.5 DTD of Themes

```
<!--

theme: the ELDIT dictionary shows about 50 lemma dependent
groups of thematically related words (such as the days of the week)
and structure words (prepositions, conjunctions, etc).

title: the title of the group
subtitle: several subtitles of sub groups
item: the words described in this group
explanation: grammatical explanations (mostly for structure words)
example: lexicographic examples
nbs: linguistic difficulties accessible by a footnote
corresponding: groups are presented in contrastive tables, hence each
               group has a corresponding group in the other language.
template: presentation is carried out by a template engine, the
          corresponding template name is indicated in ‘‘template’’.
subitemstyle: indicates whether sub items should be presented in a
              horizontal or vertical order
examplestyle: indicates whether the examples should be shown next to the
              items or as a separate group.

-->

<!ELEMENT theme (title, subtitle?, template*,
```

```
                (explanation|corresponding)*,
                item*, example*, nbs*)>
<!ATTLIST theme id ID #IMPLIED
          examplestyle (top|bottom|right|auto) "auto"
          subitemstyle (horizontal|vertical|auto) "auto">



<!--

In order to emphasize the contrastive aspect, example sentences
and other elements have translations.

-->

        <!ELEMENT title (w*, translation?, nbs?)>
        <!ATTLIST title id ID #IMPLIED>

        <!ELEMENT subtitle (w*, translation?, nbs?)>
        <!ATTLIST subtitle id ID #IMPLIED>

        <!ELEMENT template (#PCDATA)>
        <!ATTLIST template id ID #IMPLIED>

        <!ELEMENT explanation (#PCDATA)>
        <!ATTLIST explanation id ID #IMPLIED
                              type CDATA #IMPLIED
                              lang CDATA #IMPLIED>

        <!ELEMENT corresponding (#PCDATA)>
        <!ATTLIST corresponding id ID #IMPLIED
                                language CDATA #IMPLIED>



<!--

Items can be shown in different ways, e.g. with a picture, such
as the time, and illustrated by translations and examples.

-->

        <!ELEMENT item (w*, prepref*, imgref*, description*,
                        template*, subitem*, translation?,
                        example*, nbs?)>
        <!ATTLIST item id ID #IMPLIED
                  titleline (true|false) "false">

                <!ELEMENT prepref (IDREF)>
                <!ATTLIST prepref id ID #IMPLIED>

                <!ELEMENT imgref (#PCDATA)>
                <!ATTLIST imgref id ID #IMPLIED>

                <!ELEMENT description (w*, translation?, nbs?)>
                <!ATTLIST description id ID #IMPLIED>
```

```
                <!ELEMENT subitem (w*, ref*, url*, template*,
                                   (example|translation)*, nbs?)>
                <!ATTLIST subitem id ID #IMPLIED>


        <!ELEMENT example (w*, nbs?, translation*)>
        <!ATTLIST example id ID #IMPLIED
                          localid CDATA #IMPLIED
                          needsexample CDATA #IMPLIED>

                <!ELEMENT translation (w*,nbs?)>
                <!ATTLIST translation id ID #IMPLIED>

                        <!ELEMENT nbs (nb*)>
                        <!ATTLIST nbs id ID #IMPLIED>
                                <!ELEMENT nb (w*)>
                                <!ATTLIST nb id ID #IMPLIED
                                             lang CDATA #IMPLIED>


<!--

The element w has been explained in the DTD for dictionary
entries above.

-->

                <!ELEMENT w (#PCDATA)>
                <!ATTLIST w id ID #IMPLIED
                            type CDATA #IMPLIED
                            style CDATA #IMPLIED
                            nbref CDATA #IMPLIED
                            base CDATA #IMPLIED
                            ctag CDATA #IMPLIED
                            lexref CDATA #IMPLIED
                            collref CDATA #IMPLIED>
```

Bibliography

[1] Flor Aarts. Syntactic information in OALD5, LDOCE3, COBUILD2 and CIDE. In *The Perfect Learners' Dictionary*, Lexicographica, page 16. Max Niemeyer Verlag, Tübingen, 1999.

[2] Andrea Abel. Ein neuer Ansatz der Valenzbeschreibung in einem elektronischen Lern(er)wörterbuch Deutsch-Italienisch (ELDIT). In *International Annual for Lexicography*, number 18 in Lexicographica, pages 147–167. Max Niemayer Verlag, Tübingen, 2002.

[3] Andrea Abel. *Alte und neue Problematiken der Lernerlexikographie in Theorie und Praxis*. PhD thesis, Universität Innsbruck, 2003.

[4] Andrea Abel, Johann Gamper, Judith Knapp, and Vanessa Weber. Evaluation of the Web-based learners' dictionary ELDIT. In *Proceedings of World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA 2003)*, pages 1210–1217, 2003.

[5] Andrea Abel and Vanessa Weber. ELDIT, prototype of an innovative dictionary. In *Proceedings of the 9th EURALEX International Congress on Lexicography (EURALEX'00)*, Stuttgard, Germany, 2000.

[6] Andrea Abel and Vanessa Weber. ELDIT - Electronic learner's dictionary of German and Italian: Semibilingual, bilingualised or a very new type? In *Proccedings of the Eleventh International Symposium on Lexicography*, Lexicographica, Tübingen, 2002. Max Niemayer Verlag.

[7] Gregory Aist. Speech recognition in computer-assisted language learning. In *CALL - Media, Design, & Applications*, pages 165–181. Swets & Zeitlinger, The Netherlands, 1999.

[8] Jean Aitchison. *Words in the Mind: An Introduction to the Mental Lexicon*. Blackwell Publishers Ltd, Oxford, UK, 2nd edition, 1994.

[9] Don Anderson. Machine translation as a tool in second language learning. *CALICO Journal*, 13(1):68–97, 1995.

[10] Ivon Arroyo, Agustin Schapira, and Beverly Woolf. Authoring and sharing word problems with AWE. In *Proceedings of the 10th International Conference on Artificial Intelligence in Education (AIED 2001)*, 2001.

[11] Guy Aston. Learning with corpora: An overview. In *Learning with corpora*, pages 7–45. Cluebb, Bologna, 2001.

[12] Guy Aston. The learner as corpus designer. In *Teaching and learning by doing corpus analysis*, pages 9–25. Amsterdam: Rodopi, 2002.

[13] Beryl T. Atkins and Francsi E. Knowles. Interim report on the EURALEX/AILA research project into dictionary use. In *Proceedings of BudaLEX'88*, pages 381–392, 1990.

[14] Peter Baumgartner, Hartmut Häfele, and Kornelia Maier-Häfele. *E-Learning Praxishandbuch - Auswahl von Lernplatformen*. StudienVerlag, 2002.

[15] Bernd Benedixen. Freiheitsgrade für den Nutzer eines Computer-Sprachlehrprogramms. In *Translationsdidaktik - Grundfragen der Übersetzungswissenschaft*, pages 352–360. Gunter Narr Verlag, Tübingen, 1997.

[16] Jared Bernstein, Amir Najmi, and Farzad Ehsani. Subarashii: Encounters in Japanese spoken language education. *CALICO Journal*, 16(3):361–384, 1999.

[17] Maria Teresa Bianco. *Valenzlexikon Deutsch-Italienisch*. Groos Verlag, Heidelberg, 1996.

[18] Peter Blumenthal and Giovanni Rovere. *PONS-Wörterbuch der italienischen Verben*. Klett Verlag, Stuttgart, 1998.

[19] Edwin Bos and Joke van de Plassche. A knowledge-based, English verb-form tutor. *International Journal of Artificial Intelligence in Education*, 5(1), 1994.

[20] Chris Bowerman. An overview of the state of the art in computer-assisted language learning with reference to LICE: An intelligent tutoring system to assist writing in German. In *Computers and Teaching in the Humanities*, 1990.

[21] Chris Bowerman. An intelligent language learning system for writing in German. In *Proceedings of the East-West Conference on Computer Technologies in Education (EW-ED'94)*, 1994.

[22] Lynne Bowker and Jennifer Pearson. *Working with Specialized Language - A practical guide to using corpora*. Routledge, London, 2002.

[23] Paul De Bra, Ad Aerts, David Smits, and Natalia Stash. Aha! version 2.0, more adaptation flexibility for authors. In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education (E-LEARN 2002)*, pages 240–246, 2002.

[24] Tim Brailsford, Craig Steward, Mohamed Ramzy Zakaria, and Adam Moore. Autonavigation, links and narrative in an adaptive Web-based integrated learning environment. In *Proceedings of the 11th International World Wide Web Conference (WWW2002)*, 2002. http://whurle.sourceforge.net/.

[25] Helmut Brammerts. International tandem network. http://www.slf.ruhr-uni-bochum.de/servers.html, 1994.

[26] Peter Brusilovsky. Adaptive educational systems on the World-Wide-Web: A review of available technologies. In *Proceedings of the Workshop on Intelligent Tutoring Systems on the Web at ITS'98*, 1998.

[27] Peter Brusilovsky. Methods and techniques of adaptive hypermedia. In *Adaptive Hypertext and Hypermedia*, pages 1–43. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[28] Peter Brusilovsky and John Eklund. A study of user model based link annotation in educational hypermedia. *Journal of Universal Computer Science*, 4(4):429–448, 1998.

[29] Peter Brusilovsky and Philip Miller. Course Delivery Systems for the Virtual University. In *Access to Knowledge: New Information Technologies and the Emergence of the Virtual University*, pages 167–206. Elsevier Science and International Association of Universities, 2001.

[30] Peter Brusilovsky and Leonid Pesin. ISIS-Tutor: an intelligent learning environment for CDS/ISIS users. In *Proceedings of the International Workshop on Complex Learning in Computer Environments (CLCE-94)*, 1994.

[31] Gabriella Brussino, Bernadette Luciano, and Cathy Gunn. Integrated CALL design: Crescendo in Italia, a language teaching package for intermediate Italian learners. *Computer Assisted Language Learning*, 12(5):409–426, December 1999. http://www.arts.auckland.ac.nz/mtsu/LOP.htm.

[32] Keith Cameron, editor. *CALL: Media, Design, & Applications*. Swets & Zeitlinger, The Netherlands, 1999. http://llt.msu.edu/vol4num2/review1/default.html.

[33] Canoo Engineering AG, Basel, Switzerland. *Getting a Grip on E-Content with WMTrans*. http://www.canoo.com/wmtrans/home/white_paper.pdf.

[34] Stephen Carey. The use of WebCT for a highly interactive virtual graduate seminar. *Computer Assisted Language Learning*, 12(4):371–380, October 1999.

[35] Stefania Cavagnoli and Francesca Nardin. Second language acquisition in South Tyrol: Difficulties, motivations, expectations. *Multilingua*, 18(1):17–45, 1999.

[36] Stefania Cavagnoli and Elisabeth Ramoser. *Zweisprachigkeit - Bilinguismo. Deutsche/Italienische Texte - Testi tedeschi/italiani - Schwierigkeitsgrad A/B.* Arcadia Edition, 2nd edition, 2000.

[37] Thierry Chanier. Special Issue Introduction. *International Journal of Artificial Intelligence in Education*, 1994.

[38] Thierry Chanier and Thierry Selva. The ALEXIA system. The use of visual representations to enhance vocabulary learning. *Computer Assisted Language Learning*, 11(5), 1998.

[39] Carol Chapelle. Call in the year 2000: Still in search of research paradigms? *Language Learning & Technology*, 1(1):19–43, July 1997.

[40] Carol A. Chapelle. Multimedia CALL: Lessons to be learned from research on instructed SLA. *Language Learning & Technology*, 2(1):22–34, July 1998.

[41] Carol A. Chapelle. *Computer Applications in Second Language Acquisition*. Cambridge University Press, 2001.

[42] Dorothy M. Chun. L2 reading on the Web: Strategies for accessing information in hypermedia. *Computer Assisted Language Learning*, 14(5):367–403, 2001.

[43] Tom Cobb. Breath and depth of lexical acquisition with hands-on concordancing. *Computer Assisted Language Learning*, 12(4):345–360, October 1999.

[44] Luigi Colazzo and Marco Costantino. Multi-user hypertextual didactic glossaries. *International Journal of Artificial Intelligence in Education*, 9(1–2), 1998.

[45] Roger E. Cooley. Vocabulary acquisition software: User preferences and tutorial guidance. In *Proceedings of the Workshop on Computer Assisted Language Learning held in conjunction with AIED'01*, 2001.

[46] Claire Cramsch and Roger W. Andersen. Teaching text and context through multimedia. *Language Learning & Technology*, 2(2):31–42, January 1999. http://www.humnet.ucla.edu/humnet/al/clrl/crphome.html.

[47] Vincent Crespi, Eric Loken, Josh Millet, and Lesleigh Cushing. Number2.com. http://www.number2.com/, 1995.

[48] Jonathan Dalby and Diane Kewley-Port. Explicit pronunciation training using automatic speech recognition technology. *CALICO Journal*, 16(3):425–445, 1999.

[49] Suyada Dansuwan, Kikuko Nishina, Kanji Akahori, and Yasutaka Shimizu. Development and evaluation of a Thai learning system on the Web using natural language processing. *CALICO Journal*, 19(1), 2001.

[50] Tullio de Mauro and Angela Cattaneo. *DIB - Dizionario di base della lingua italiana*. Paravia Editore, Torino, 1996.

[51] Gilles-Maurice de Schryver. Lexicographers' dreams in the electronic-dictionary age. *International Journal of Lexicography*, 16(2):143–199, 2003.

[52] William H. DeSmedt. Herr Kommissar: An ICALL conversation simulator for intermediate German. In *Intelligent Language Tutors – Theory Shaping Technology*, pages 153–174. Lawrence Erlbaum Associates, 1995.

[53] Giacomo Devoto and Gian Carlo Oli. *Il dizionario della lingua italiana*. Casa Editrice Felice Le Monnier, Firenze, 2000.

[54] Marc Domenig and Pius ten Hacken. *Word Manager: A System for Morphological Dictionaries*, volume 1 of *Informatik und Sprache*. Olms Verlag, Hildesheim, 1992.

[55] Sarah A. Douglas. LingWorlds: An intelligent object-oriented environment for second language tutoring. In *Intelligent Language Tutors – Theory Shaping Technology*, pages 201–220. Lawrence Erlbaum Associates, 1995.

[56] Günther Drosdowski. *Duden Deutsches Universalwörterbuch*. Dudenverlag, Mannheim, Leipzig, Wien, Zürich, 1996.

[57] Charles Egert. Language learning across campuses. *Computer Assisted Language Learning*, 13(3):271–280, July 2000.

[58] Kurt Egger and Karin Heller. *Kontaktlinguistik*, chapter Italienisch-Deutsch, pages 1350–1357. de Gruyter Berlin, 1997.

[59] John Eklund and Peter Brusilovsky. The value of adaptivity in hypermedia learning environments: A short review of empirical evidence. In *Proceedings of the Second Workshop on Adaptive Hypertext and Hypermedia held in conjunction with Hypertext´98*, 1998.

[60] John Eklund, Peter Brusilovsky, and Elmar Schwarz. Adaptive textbooks on the World Wide Web. In *Proceedings of the 3th Australian World Wide Web Conference (AusWeb'97)*, Southern Cross University, Australia, July 1997.

[61] John Eklund, Peter Brusilovsky, and Elmar Schwarz. A study of adaptive link annotation in educational hypermedia. In *Proceedings of World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA 1998)*, 1998.

[62] Maxime Eskenazi. Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning & Technology*, 2(2):62–76, January 1999.

[63] Michael Feiler. South Tyrol - model for the resolution of minority conflicts? *Review of international affairs*, 48, 1997.

[64] Christiane Fellbaum, editor. *WordNet — An Electronic Lexical Database*. MIT Press, 1998. http://www.cogsci.princeton.edu/˜wn/.

[65] Carl H. Frederiksen, Janet Donin, and Michel Décary. A discourse processing approach to computer-assisted language learning. In *Intelligent Language Tutors – Theory Shaping Technology*, pages 99–120. Lawrence Erlbaum Associates, 1995.

[66] G DATA Software. WebSpeech 4. http://www.webspeech.de/index2.php, 2003.

[67] Johann Gamper and Judith Knapp. Towards an adaptive learners' dictionary. In *Proceedings of International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2000)*, Lecture Notes in Computer Science, pages 311–314. Springer Verlag, 2000.

[68] Johann Gamper and Judith Knapp. A review of CALL systems in foreign language instruction. In *Proceedings of the 10th International Conference on Artificial Intelligence in Education (AIED 2001)*, pages 377–388. IOS Press Amsterdam, 2001.

[69] Johann Gamper and Judith Knapp. Adaptation in a vocabulary acquisition system. *KI - Zeitschrift Künstliche Intelligenz*, 3(2):27–30, 2002.

[70] Johann Gamper and Judith Knapp. A review of intelligent CALL systems. *Computer Assisted Language Learning*, 15(4):329–342, oct 2002.

[71] Johann Gamper and Judith Knapp. Tutoring in a language learning system. In *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002)*, volume 2, pages 465–470, 2002.

[72] Johann Gamper and Judith Knapp. A Web-based language learning system. In *Proceedings of the 1st International Conference on Web-based Learning (ICWL2002)*, Lecture Notes in Computer Science, pages 106–118. Springer Verlag, 2002.

[73] Johann Gamper and Judith Knapp. 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, May 2003.

[74] Robert Godwin-Jones. Language learning and the World Wide Web. In *Proceedings of the 2th International World Wide Web Conference (WWW2)*, 1994.

[75] Robert Godwin-Jones. Language testing tools and technologies. *Language Learning & Technology*, 5(2):8–12, 2001.

[76] Robin Goodfellow. Evaluating performance, approach and outcome in the design of CALL. In *CALL: Media, Design & Applications*, pages 109–140. Swets & Zeitlinger, The Netherlands, 1999.

[77] Sylviane Granger. A bird's-eye view of learner corpus reseach. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. John Benjamins Publishing Company, 2002.

[78] Peter J.M. Groot. Computer assisted second language vocabulary acquisition. *Language Learning & Technology*, 4(1):60–81, May 2000.

[79] Henry Hamburger. Tutorial tools for language learning by two-medium dialogue. In *Intelligent Language Tutors – Theory Shaping Technology*, pages 183–200. Lawrence Erlbaum Associates, 1995.

[80] Henry Hamburger, Michael Schoelles, and Florence Reeder. More intelligent CALL. In *CALL - Media, Design & Applications*. Swets & Zeitlinger, The Netherlands, 1999.

[81] Paddy Harben. An exercise in applying pedagogical principles to multimedia CALL materials design. *ReCALL*, 11(3):25–33, nov 1999.

[82] William G. Harless, Marcia A. Zier, and Robert C. Duncan. Virtual dialogues with native speakers: The evaluation of an interactive multimedia method. *CALICO Journal*, 16(3), 1999.

[83] Andrew Harley. Cambridge dictionaries online. In *Proceedings of the 9th EURALEX International Congress on Lexicography (EURALEX'00)*, 2000. http://dictionary.cambridge.org.

[84] Elliotte Rusty Harold. *XML Bible*. IDG Books Worldwide, Inc., 1999.

[85] Michael Harrington. Complex: A tool for the development of L2 vocabulary knowledge. *International Journal of Artificial Intelligence in Education*, 5(4), 1994.

[86] Dörthe Hecht. *PONS Basiswörterbuch Deutsch als Fremdsprache*. Klett International, Stuttgart, 1999.

[87] Trude Heift and Devlan Nicholson. Theoretical and practical considerations for Web-based intelligent language tutoring systems. In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems (ITS'2000)*, 2000.

[88] Gerhard Helbig and Wolfgang Schenkel. *Wörterbuch zur Valenz und Distribution deutscher Verben*. Max Niemeyer Verlag, Tübingen, 1991.

[89] Nicola Henze and Wolfgang Nejdl. Adaptivity in the KBS hyperbook system. In *Proceedings of the 2nd Workshop on Adaptive Systems and User Modeling on the WWW*, 1999.

[90] Melissa Holland. Tutors that listen. *CALICO Journal*, 16(3):243–250, 1999.

[91] Melissa Holland, Jonathan Kaplan, and Mark Sabol. Preliminary tests of language learning in a speech-interactive graphics microworld. *CALICO Journal*, 16(3), 1999.

[92] V. Melissa Holland, Jonathan D. Kaplan, and Michelle R. Sams, editors. *Intelligent Language Tutors – Theory Shaping Technology*. Lawrence Erlbaum Associates, 1995. http://acl.ldc.upenn.edu/J/J96/J96-4007.pdf.

[93] Kristina Höök and Martin Svensson. Evaluating adaptive navigation support. In *Social Navigation of Information Space*, pages 238–251. Springer Verlag, 1999.

[94] Albert S. Hornby, editor. *Oxford Advanced Learners' Dictionary*. Oxford University Press, 2002. http://www.oup.com/elt/oald/.

[95] Fabrice Issac and Oliver Hû. A Description Formalism for Complex Questionnaires. In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems (ITS'2000)*, Lecture Notes in Computer Science. Springer Verlag, 2000.

[96] Reiko Ito and Charles Hannon. The Effect of Online Quizzes on learning Japanese. *CALICO Journal*, 19(3):551–561, 2002.

[97] Kevin Jansz, Christopher Manning, and Nitin Indurkha. Kirrkirr: Interactive visualisation and multimedia from a structured warlpiri dictionary. In *Proceedings of the 5th Australian World Wide Web Conference (AusWeb'99)*, 1999. http://www-nlp.stanford.edu/kirrkirr/#Demos.

[98] Christopher Jones. Contextualise & personalise: Key strategies for vocabulary acquisition. *ReCALL*, 11(3):34–40, November 1999.

[99] Judith Knapp and Pius ten Hacken and Sandro Pedrazzini. ELDIT and Word Manager - a powerful partnership. In *Proceedings of World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA 2003)*, 2003.

[100] Bernd Kielhöfer. Psycholinguistische Grundlagen der Wortschatzarbeit. *Babylonia*, 1996.

[101] Keiki Kitade. L2 learners' discourse and SLA theories in CMC: Collaborative interaction in internet chat. *Computer Assisted Language Learning*, 13(2):143–166, April 2000.

[102] Tom Koet. ICT and language skills: An integrated course. *ReCALL*, 11(1):65–71, May 1999.

[103] Judith F. Kroll and Natasha Tokowicz. The development of conceptual representation for words in a second language. In *One Mind, Two Languages. Bilingual Language Processing*. Blackwell Publishers, Malden–Oxford, 2001.

[104] Anja Krüger and Simon Hamilton. RECALL: individual language tutoring through intelligent error diagnosis. *ReCALL*, 9(2):51–58, 1997. Project homepage at http://www.infj.ulst.ac.uk/˜recall.

[105] Claudia Kunze and Lothar Lemnitzer. GermaNet - representation, visualization, application. In *Proceedings of the Second International Conference on Language Resources & Evaluation*, pages 1485–1491, 2002.

[106] Marie-Noëlle Lamy and Robin Goodfellow. "Reflective conversation" in the virtual language classroom. *Language Learning & Technology*, 2(2):43–61, 1999.

[107] Alexander Langer. Chancen und Hindernisse für Zweisprachigkeit in Südtirol. In *Linguistic Problems and European Unity*. Franco Angeli Editore, 1982.

[108] Batia Laufer. Electronic dictionaries and incidental vocabulary acquisition: Does technology make a difference? In *Proceedings of the 9th EURALEX International Congress on Lexicography (EURALEX'00)*, pages 849–854, 2000.

[109] Amy S.C. Leh. Computer-mediated communication and foreign language learning via electronic mail. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2), October 1999.

[110] Ruddy Lelouche. Dealing with pragmatic and implicit information in an ICALL system: The PILÉFACE example. *International Journal of Artificial Intelligence in Education*, 5(4), 1994.

[111] Michael Levy. *Computer-Assisted Language Learning: Context and Conceptualization*. Clarendon Press, 1997. http://llt.msu.edu/vol2num1/review/levy.html.

[112] Mike Levy. Theory and design in a multimedia CALL project in cross-cultural pragmatics. *Computer Assisted Language Learning*, 12(1):29–57, February 1999.

[113] David Little and Ema Ushioda. Designing, implementing and evaluationg a project in tandem language learning via e-mail. *ReCALL*, 10(1), 1998.

[114] Longman.com. Longman dictionary of contemporary english. Longman, 2003. Web version at http://www.longmanwebdict.com/.

[115] Melissa Magliana. The autonomous province of South Tyrol: A model of self-governance? Arbeitshefte der Europäischen Akademie Bozen, 2000.

[116] Michael Mayo and Antonia Mitrovic. Optimising ITS behavior with bayesian networks and decision theory. *International Journal of Artificial Intelligence in Education*, 12(3), 2001.

[117] David McArthur, Matthew Lewis, and Miriam Bishay. The Role of Artificial Intelligence in Eductaion: Current Progress and Future Prospects. Technical report, RAND - Research and development, 1993. Retrieved on 19-07-2003 from http://www.rand.org/education/mcarthur/Papers/roleab.html.

[118] Wolfgang Menzel, Daniel Herron, Rachel Morton, Dario Pezzotta, Patrizia Bonaventura, and Peter Howarth. Interactive pronunciation training. *ReCALL*, 13(1):67–78, 2001. Project homepage with demo at http://nats-www.informatik.uni-hamburg.de/~isle.

[119] Jack Mostow and Gregory Aist. Giving help and praise in a reading tutor with imperfect listening - because automated speech recognition means never being able to say you're certain. *CALICO Journal*, 16(3):407–424, 1999. Project homepage at http://www.cs.cmu.edu/~listen.

[120] Martin Müller and Lukas Wertenschlag. Wortschatz-lernen ganzheitlich: effektiv und effizient. *Babylonia*, 2:25–31, 1996.

[121] Janet H. Murray. Lessons learned from the athena language learning project. In *Intelligent Language Tutors – Theory Shaping Technology*, pages 243–256. Lawrence Erlbaum Associates, 1995. Systems purchasable at http://web.mit.edu/jhmurray/www/Project_Sampler.html.

[122] Tom Murray, Tina Shen, Janette Piemonte, and Chris Condit. Adaptivity in the MetaLinks Hyper-book Authoring Framework. In *Proceedings of the Workshop on Adaptive and Intelligent Web-Based Education Systems at ITS'2000*, 2000.

[123] Marie J. Myers. Voice recognition software and a hand-held translation machine for second-language learning. *Computer Assisted Language Learning*, 13(1), 2000.

[124] Noriko Nagata. An effective application of natural language processing in second language instruction. *CALICO Journal*, 13(1):47–67, 1993.

[125] Wolfgang Nejdl and Martin Wolpers. KBS hyperbook – a data-driven information system on the Web. In *Proceedings of the 8th International World Wide Web Conference (WWW8)*, Toronto, May 1999.

[126] John Nerbonne, Duco Dokter, and Petra Smit. Morphological processing and computer-assisted language learning. *Computer Assisted Language Learning*, 11(5), 1998.

[127] Hilary Nesi. A user's guide to electronic dictionaries for language learners. *International Journal of Lexicography*, 12(1), March 1999.

[128] Stefania Nuccorini. Monitoring dictionary use. In *Proceedings of the 5th EURALEX International Congress on Lexicography (EURALEX'92)*, pages 89–102, 1992.

[129] Hiroaki Ogata, Yoshiaki Hada, and Yoneo Yano. Computer supported online correction for collaborative writing. In *Proceedings of International Conference on Information Society in the 21st Century (IS 2000)*,, pages 576–583, Aizu-Wakamatsu City, Fukushima, Japan, 2000.

[130] Open Learning Technology Corporation Limited. *Learning with Software - Pedagogies and Practice*, 1996. Retrieved on 30-07-2003 from http://www.educationau.edu.au/archives/CP/default.htm.

[131] Sandro Pedrazzini. *Phrase Manager: A System for Phrasal and Idiomatic Dictionaries*, volume 3 of *Informatik und Sprache*. Olms Verlag, Hildesheim, 1994.

[132] Sandro Pedrazzini. The finite-state automata's design patterns. In *Proceedings of Third International Workshop on Implementing Automata (IWA'98)*, 1998. Jacaranda Framework available at http://b.die.supsi.ch/˜pedrazz/jacaranda.

[133] Sandro Pedrazzini and Judith Knapp. From e-learning to complete software development projects: Canoo.net and ELDIT. In *Proceedings of 5th International Conference on New Educational Environments (ICNEE 2003)*, Lucern, 2003.

[134] Jan L. Plass. Design and evaluation of the user interface of foreign language multimedia software: A cognitive approach. *Language Learning & Technology*, 2(1):35–45, July 1998. http://www.gss.ucsb.edu/faculty/dmchun/cyberbuch/.

[135] Inc Plumb Design. Plumb Design Visual Thesaurus. http://www.visualthesaurus.com/online/index.html, 2003.

[136] Charlotte Price, Gordon McCalla, and Andrea Bunt. L2tutor: A mixed-initiative dialogue system for improving fluency. *Computer Assisted Language Learning*, 12(2):83–112, April 1999.

[137] Isabelle De Ridder. Are we conditioned to follow links? Highlights in CALL materials and their impacts on the reading process. *Computer Assisted Language Learning*, 13(2):183–195, April 2000.

[138] Claudia Maria Riehl. *Mehrsprachigkeit im Alpenraum*, chapter Schriftsprachliche Kompetenz und Zweisprachigkeit: der Fall Südtirol. Aarau, 1998.

[139] Claudia Maria Riehl. *Minderheiten- und Regionalsprachen in Europa*, chapter Deutsch in Südtirol. Opladen, 2000.

[140] Steven Ritter, John Anderson, and Michael Cytrynowicz. Authoring content in the PAT algebra tutor. *Journal of Interactive Media in Education*, 1998.

[141] Paul Robberecht. Critical review of TransIt-Tiger. Retrieved on 17-07-03 from http://calico.org/CALICO_Review/review/transtiger.htm.

[142] Bernd Rüschoff. Data-driven learning (ddl): the idea. Technical report, European Center for Modern Languages, 2004. http://www.ecml.at/projects/voll/menu_top.htm.

[143] Francesco Sabatini and Vittorio Coletti. *DISC - Dizionario Italiano Sabatini Coletti*. Giunti Gruppo Editoriale, Firenze, 1997.

[144] Helmut Schmid. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*, 1995.

[145] Anny Schweigkofler. Conflicting language realities: A case study of Bolzano/Italy. Contribution on the 4th European Conference on Immersion programs, September 1998.

[146] Sue Sentance. A rule network for English article usage within an intelligent language tutoring system. *Computer Assisted Language Learning*, 10(2), 1997.

[147] Yi Shang, Hongchi Shi, and Su-Shing Chen. An Intelligent Distributed Environment for Active Learning. In *Proceedings of the 10th International World Wide Web Conference (WWW10)*, 2001.

[148] Chi-Chiang Shei. FollowYou! an automatic language lesson generation system. *Computer Assisted Language Learning*, 14(2), 2001.

[149] Chi-Chiang Shei and Helen Pain. Learning a foreign language through machine translation: Focusing on sentence stems and collocations. In *Proceedings of the Workshop on Computer Assisted Language Learning held in conjunction with AIED-01*, 2001.

[150] John Sinclair, Gwyneth Fox, and Stephen Bullon, editors. *Collins Cobuild English Dictionary*. Athelstan Publishers, 1999. http://titania.cobuild.collins.co.uk/.

[151] Marcello Sofritti. Bilinguismo e patentino in Alto Adige. In *Heimat - identità regionali nel processo storico*, pages 343–351. Casa editrice Donzelli, 2000.

[152] Jesús Soria. Expert CALL: data-based versus knowledge-based interaction and feedback. *ReCALL*, 9(2):43–50, 1997.

[153] Oliver Streiter, Judith Knapp, and Leonhard Voltmer. GYMN@ZILLA: a browser-like repository for open learning resources. In *Proceedings of World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA 2003)*, pages 1371–1379, 2003. http://www.eurac.edu/gymnazilla.

[154] Oliver Streiter and Leonhard Voltmer. Document classification for corpus-based legal terminology. In *Proceedings of The Eighth International Conference of the International Academy of Linguistic Law*, 2002.

[155] Oliver Streiter, Daniel Zielinski, Isabella Ties, and Leonhard Voltmer. Term extraction for Ladin: An example-based approach. In *Proceedings of 10th Annual Conference on Natural Language Processing (TALN 2003)*, 2003.

[156] Christian Süß, Burkhard Freitag, and Peter Brössler. Meta-modeling for Web-based teachware managment. In *Advances in Conceptual modeling - Workshop on the World-Wide Web and Conceptual Modeling (ER'99)*, Lecture Notes in Computer Science, Berlin, 1999. Springer Verlag.

[157] Karen Swan. The Effectiveness of Online Learning: A Review of the Literature. In *Proceedings of World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA 2003)*, pages 2225–2232, 2003.

[158] Merryanna L. Swartz and Masoud Yazdani, editors. *Intelligent Tutoring Systems for Foreign Language Learning*, chapter Introduction. Springer Verlag, 1992.

[159] Pasi Tapanainen and Timo Järvinen. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 64–71, 1997.

[160] Pius ten Hacken and Marc Domenig. Reusable Dictionaries for NLP: The Word Manager Approach. *Lexicology*, 2:232–255, 1996.

[161] Pius ten Hacken and Cornelia Tschichold. Word Manager and CALL: Structured access to the lexicon as a tool for enriching learners' vocabulary. *ReCALL*, 13(1), 2001.

[162] Naoyuki Tokuda. New developments in intelligent CALL systems in a rapidly internationalized information age. *Computer Assisted Language Learning*, 15(4):319–327, 2002.

[163] Naoyuki Tokuda and Liang Chen. An online tutoring system for language translation. *IEEE Multimedia*, 8(3):46–55, 2001. Preliminary version at http://azalea.sunflare.co.jp.

[164] Janine Toole and Trude Heift. The Tutor Assistant: An authoring tool for an intelligent language tutoring system. *Computer Assisted Language Learning*, 15(4):373–386, 2002.

[165] María Dolores La Torre. A web-based resource to improve translation skills. *ReCALL*, 11(3):41–49, 1999. Resource accessible at http://www.hum.port.ac.uk/slas/babel/INDEX˜1.HTM.

[166] Anna Trifonova and Marco Ronchetti. Where is mobile learning going? In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education (E-LEARN 2003)*, pages 1794–1801, 2003.

[167] Mark Urban-Lurain. Intelligent tutoring systems: An historic review in the context of the development of Artificial Intelligence and Educational Psychology. Technical report, CSE 101 - Computing Concepts and Competencies, 1996. Retrieved on 19-07-2003 from http://www.cse.msu.edu/rgroups/cse101/ITS/its.htm.

[168] Dieter v. Götz, Günther Haensch, and Hans Wellmann. *Langenscheidts Großwörterbuch, Deutsch als Fremdsprache*. Langenscheidt, 1998.

[169] Anne Vandeventer. Creating a grammar checker for CALL by constraint relaxation: A feasibility study. *ReCALL*, 13(1):110–120, 2001. Project homepage at http://www.latl.unige.ch/freetext/index.html.

[170] Rita M. Vick, Martha E. Crosby, and David E. Ashworth. Japanese and american students meet on the web: Collaborative language learning through everyday dialogue with peers. *Computer Assisted Language Learning*, 13(3), 2000.

[171] Spela Vintar and Tomaz Erjavec. In two minds: How to teach translation students to learn from parallel corpora. In *Proceedings of the 5th Teaching and Language Corpora Conference*, 2000.

[172] Maria Virvou and Victoria Tsiriga. Web passive voice tutor: An intelligent computer assisted language learning system over the WWW. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT 2001)*. IEEE Computer Society Press, 2001.

[173] Richard C. Waters. The audio interactive tutor. Technical Report MERL-TR-94-04, Mitsubishi Electric Research Laboratories Cambridge Research Center, April 1994.

[174] Gerhard Weber and Markus Specht. User modeling and adaptive navigation support in WWW-based tutoring systems. In *Proceedings of the 6th International Conference on User Modeling (UM97)*, pages 289–300. Springer Verlag, 1997.

[175] Bernd Weidenmann, editor. *Wissenserwerb mit Bildern*. Verlag Hans Huber, 1994.

[176] Angelika Wöllstein-Leisten. *Deutsche Satzstruktur: Grundlagen der syntaktischen Analyse*. Stauffenburg-Verlag, Tübingen, 1997.

[177] Jane Woodin. E-mail tandem learning and the communicative curriculum. *ReCALL*, 9(1), 1997.

[178] Jie Chi Yang and Kanji Akahori. A discourse structure analysis of technical japanese texts and its implementation on the WWW. *Computer Assisted Language Learning*, 13(2), 2000.

[179] Ekkehard Zöfgen. *Lernerwörterbücher in Theorie und Praxis*, chapter Wortschatz-Lernen mit dem Wörterbuch? Lexicographica. Max Niemeyer Verlag, Tübingen, 1994.

Index

Curriculum Vitae

Judith Knapp was born on 21st March, 1972 in Brunico (Italy). From 1991 to 1998, she studied Mathematics at the Graz University of Technology.

In 1994/95, she spent one year as an ERASMUS-student at Queen Mary and Westfield College in London, GB. In 1997 she was selected to take part in the First European Intensive Course "Complex Analysis and Generalizations" in Aveiro, Portugal.

In 1997/98, she worked on her diploma thesis in numerical and analytic methods in nonlinear partial differential equations. She received her master's degree in Mathematics in November 1998.

In 1999, she undertook a nine month period of trainee-ship at the company Ceramica Anços in Coimbra, Portugal.

Since September 1999, she has been a scientific collaborator at the European Academy of Bolzano.

In 2002, her team was awarded the award "Anerkennungspreis 2002 für Forschungs- und Studienprojekte der Eduard-Wallnöfer-Stiftung" for the project ELDIT.

In October 2004 she received her PhD from the Information Systems Institute - Knowledge Based Systems of the University of Hannover

Currently, she is not only working on ELDIT but is also involved in some other educational projects (GYMN@ZILLA and LOGOS GAIAS), which are all together strongly focussed on technology-based language learning.