The Workshop Programme

09:15 - 09:30 Opening Remarks

09:30 - 10:30 Invited Talk – Hans Uszkoreit

10.30 - 11:00 Coffee break

Morning session Language Resources for teaching CL

11:00 - 11:30 "Language Resources in intensive Study Projects", Veit Reuer, Petra Ludewig

11:30 - 12:00 "Employing GermaNet in virtual CL courses", Claudia Kunze, Lothar Lemnitzer
12:00 - 12:30 Student Projects in Language and Speech Processing", Dan Cristea, Horia-Nicolai Teodorescu, Dan-Ioan Tufis

12:30 - 14: 00 Lunch

Afternoon session : Language Resources in e-learning

14:00 - 14:30 "Translation Memory Resources in multiple Languages to support E-Learning in multiple scenarios", Dragos Ciobanu, Karl-Heinz Freigang, Anthony Hartley, Uwe Reinke, Martin Thomas

14:30 - 15:00 "Terminological Grid and Free Text Repositories in computer-aided Teaching of foreign Language Terminology" Galia Angelova, Ognian Kalaydjiev, Albena Strupchanska, Svetla Boytcheva, Irena Vitanova

15:00 - 15:30 "Integrating Resources to realize a Self-contained Environment for Lexicon Learning", Sandro Pedrazzini, Alessandro Trivillini, Judith Knapp

15:30 - 16:00 "Dynamic teaching materials for ESSLI", Raffaella Bernardi, Ingo Dhan, Gilad Mishne, Michael Moortgat, Maarten de Rijke, Hans Uszkoreit, Willemijn Vermaat

16:00 - 16:30 Coffee break

16:30 - 17: 30 Summing –up
Contributions of the Workshop-Organisers

17:30 - 18:30 Panel Discussion

Workshop Organisers

Paola Monachesi, Utrecht University

Cristina Vertan, university of Hamburg

Walther von Hahn, University of Hamburg

Susanne Jekat, Zurich University of Applied Sciences Winterthur

Table of Contents

Author Index

Foreword

Language resources (LRs) are of crucial importance for research and development in language as well as for speech technology but and teaching purposes. E-learning added a new dimension to the usability of language resources and made them even interesting to the area outside of computational linguistics.

This workshop on "Language Resources: Integration and Development in e-Learning and in Teaching Computational Linguistics" focuses on the integration of LRs in the educational process and the cooperation among LRs and e-learning. Additionally, it discusses the use of LRs in the curriculum of computational linguistics.

The 8 papers included in these proceedings cover the following topics:

1. Case studies of the use of LRs in Linguistics and Computational Linguistics,
2. Additional skills acquired by the students when using or developing LRs (e.g. how to acquire standards),
3. Usage of LRs in the development of e-learning materials,
4. Adaptation of existent LRs for CALL environments,
5. Development of eContent localization resources.

And concern language resources for a broad variety of languages and types of applications. The articles are grouped in two parts.

The topic of the first part is "Language Resources in Teaching Computational Linguistics with three papers

Veit Reur and Petra Ludewig describe the use of LRs in two group projects for students at master level. In one project LRs are used for collocation extraction, in the other for the construction of a vocabulary trainer. The paper of Claudia Kunze und Lothar Lemnitzer focuses on the use of existing lexical resources, in particular GermaNet for case studies and explorative learning in virtual courses of Computational Linguistics and Language Engineering. Dan Cristea, Horia-Nicolai Teodorescu and Dan-Ioan Tufis report on LRs used for student projects both in language and speech technology. Cristina Vertan proposes a checklist of possible criteria when deciding whether in the instructional process LR's have to be created or reused

The second part of the proceedings consists of 4 papers and addresses relationships between LRs and e-learning.

Dragos Ciobanu, Karl-Heinz Freigang, Anthony Hartley, Uwe Reinke and Martin Thomas present a rationale for the development of a multilingual resource designed to support the training of translators in their use of translation memories. The following two papers focus on a special aspect of e-learning: computer aided language learning. In both papers the accent is on vocabulary learning. Galia Angelova, Albena Struchanska, Ognian Kalaydjev, Svetla Boytcheva and Irena Vitanova describe LRs used in a CALL-project for learning English financial terminology. The paper of Sandro Pedrazzini, Alexandro Trivilini and Judith Knapp shows how an existing LR can be adapted for e-learning purposes, i.e., language learning. The creation of an environment for dynamic teaching materials for ESSLI (European summer School on Logic, Language and Computation) is discussed in the paper of Rafaella Bernardi, I.Dahn, G. Mishne, M. Moortgat, M. de Rijke and H. Uzkoreit.

The organizers hope that this selection of papers will be of interest to a broad audience and readership, and will be a starting point for further discussion and cooperation.

Paola Monachesi
Cristina Vertan
Walther v. Hahn
Susanne Jekat

PART I

Language Resources in teaching Computational Linguistics

Language Resources in Intensive Study Projects

Veit Reuer, Petra Ludewig

Institute of Cognitive Science
University of Osnabrück
49069 Osnabrück, Germany
{vreuer, pludewig}@uos.de

Abstract

We describe two group projects carried out by Masters-level students at the University of Osnabrück. These projects gave students the opportunity to learn not only about the structure of language resources (LRs) but also to handle LRs and thus to gain professional knowledge of them. In the KoKs project (2001) a bilingual aligned German/English corpus was created and was used for contrastive collocation extraction. In the MAPA project students applied the GermaNet lexical resource to the construction of a vocabulary trainer for German second-language learners. In such projects, groups of approximately 8 students spend 12 months developing software applications. They invest more than 30% of their time on the project over this period and are awarded four times the number of credit-points (ECTS) as for standard courses. Such projects are an integral part of student training in Osnabrück, providing an environment in which the students can acquire the practical skills essential to successful collaborative work.

1. Introduction

In this paper we present two examples of so-called study projects which were originally developed for the course of Computational Linguistics and Artificial Intelligence (CL & AI Magister), and are now a compulsory learning unit in the new, international Master-Program Cognitive Science at the University of Osnabrück. In the KoKs project a bilingual German English corpus was developed by the students, and then in combination with a bilingual dictionary used for contrastive collocation extraction. In the MAPA project students used the lexical resource GermaNet (Lemnitzer and Kunze, 2002) in order to build a vocabulary trainer.

With respect to the use of LRs in study projects we would like to argue that this gives students the unique opportunity to not only learn *about* the structure and the use of language resources (LRs) but also to *handle* LRs in order to develop applications with them and thus to gain professional knowledge of LRs. The two projects presented here will demonstrate the range of possibilities for the use of language resources in teaching CL, which goes far beyond the factual knowledge about issues in CL and LRs or the simple, exemplary linguistic analysis of language data as it is done in traditional seminars.

2. Intensive Study Projects

From the perspective of the general course of studies, projects provide the ideal basis for fast, comprehensive and job-oriented education. On the one hand a group of approximately 8 students is required to develop software applications in the projects based on Computational Linguistics methods studied beforehand. On the other hand they should get to know the special techniques required to carry out a project successfully such as teamwork, presentational skills and careful project planning. As opposed to conventional courses students usually invest more than 30% of their weekly workload in a study project during a 12 month period, and they are expected to work through semester breaks. Also students earn four times the number of ECTS-points (European Credit Transfer System) than for standard seminars.

It must be mentioned that the topic of a project is only roughly determined by the lecturers who supervise a project in a team and play the role of the management level during a project. The development should be presented at least twice during the year to fellow students as well as members of the Institute and a final report needs to be written. The concrete issues are worked out by the students themselves which has the main advantage of creating a special feeling of responsibility and as a consequence more enthusiasm for the project among the members. As a consequence the students often present their results towards the end of the project at workshops or conferences that are relevant to the domain in question. Additionally each participating student has a chance to develop a personal preference for special topics during the project and he/she is explicitely advised in the study regulations to follow up on these topics for his/her final Master's thesis, an opportunity taken quite often by the students. Finally two more advantages need to be listed here.

1. Intensive study projects prepare a student for independent, self-determined work and life-long learning where the focus is no longer on the static acquisition of knowledge but on the *process* of acquiring knowledge in a rapidly changing environment.

2. The aspect of focussing on a certain topic for exactly one year forces a student to come to the end of his/her studies fast and additionally provides an ideal background for the preparation of a final thesis.

Therefore the goal of a study project is to crucially prepare the students for their future occupations in research and development positions. The fact that the results of the study projects have continuously been the basis for final theses and have been presented at major international conferences demonstrates the success of this teaching concept.

3. Corpus-based Search for Collocations - KoKs

The first project we describe here is called KoKs which stands for "Korpusbasierte Kollokationssuche" and lasted

from October 2000 to October 2001 with 6 participating CL&AI students advised by two university teachers. The primary goal of this study project was to exploit a bilingual parallel corpus in order to build up a list of German and English collocations with their respective translations (for a detailed description of the work see Erpenbeck et al., 2002). Since it was considerably more difficult than originally expected to have access to an aligned German English corpus without investing too much money a bigger part of the student workload had to be spent for corpus preparation tasks.

The KoKs system was integrated into an application for language learners, who use the program in order to retrieve translations for collocations, which are especially difficult to learn.

3.1. Collocations

During the last twenty years collocations, which can be regarded as a special kind of phraseologism (Burger, 1998), have attracted increasing attention, not only with respect to lexicography (Benson, 1985; Heid, 1998; Sinclair, 1987) and foreign language teaching (Granger, 1998; Lewis, 2000; Ludewig, 2001) but also within the context of natural language processing and theoretical linguistics (Erbach and Krenn, 1994; Lüdeling, 2002). Since collocations are a phenomenon of language use or language norm, and not of language as a rule-based system, it is not astonishing that collocations have built a core domain of corpus linguistics since the beginning of this field of investigation (Sinclair, 1987; Heid, 1998; Krenn, 2000). Thus the relation between collocations and language resources is two-fold: on the one hand collocations should constitute an integral part of lexical resources, on the other hand textual resources, corpora, play an inherently important role when trying to compile collocations.

Generally speaking collocations can be characterized as "frequent, recurrent, conventionalized building blocks of the lexicon [...] often not predictable; [...] non-collocational texts are not fluent, not elegant or just not the 'usual way' how one would express a given idea" (Heid, 1994, 228-9). Nevertheless the ideas of what a collocation is, are quite heterogeneous. Some scientists focus on the statistical aspect of collocations, defining collocations as word combinations of an above-average frequency (Smadja, 1993). Others emphasize their semantic peculiarities, indicating that one component of a collocation, the base, keeps its traditional meaning whereas the other one, the collocator, "has a special meaning that it cannot have in a free syntagmatic construction" (Breidt, 1993, 227).

The students in the KoKs project who had participated in a seminar on collocations beforehand decided to concentrate on the semantic peculiarities of collocations. Having the scenario of foreign language learning in mind the core assumption of the students was the idea, that the specific meaning taken by the collocator within a collocation is typically reflected in its collocation-internal translation, entailing a translation which is not word by word, but only "partially compositional". This can be understood as a translation of a complex expression where the translation of one constituent, the base, is given directly by one of the translations listed for this item in bilingual lexicons, whereas the translational counterpart of the other constituent, the collocator, is given by an expression which is not listed in the bilingual lexicons as a translation for this word as a context-less item. In this case it is likely that the expression under consideration is a collocation as defined above. In contrast, a compositional translation is interpreted as an indicator of compositional semantics, i.e. a free combination.

An example of such a "partially compositional translation" is *eine Rede halten - to give a talk*. The base is given by *Rede/talk*, and *halten/give* constitutes the collocator. Thus the approach to collocation extraction is rather contrastive than frequency driven. Of course there are collocations which cannot be identified this way because in some respect the two languages in question contain parallel collocations. An example is *vor Wut schäumen - to seeth with anger*. This kind of mistake can be characterized as a recall (vs. precision) problem. Inversely there may be phrases which are not translated word by word within a given context but which are nevertheless not collocations, but free combinations that might be translated word by word. In this case there is a precision problem.

In order to find "partially compositional translations" of phrases, the students of the KoKs project decided to do two things:

- detect phrasal correspondencies between English German translations within a bilingual corpus (subsentencial text alignment), and
- compare the translations of the identified phrase pairs with the corresponding single word translations given within bilingual dictionaries.

For this reason the students had to compile and preprocess German English bi-texts and dictionaries. The different steps undertaken in order to solve these tasks are described in section 3.2. and 3.3. and are displayed in figure 1.

3.2. Building up textual and lexical resources

3.2.1. Textual resources

At first, the students searched the internet for freely available aligned German English corpora but without success[1]. This experience was quite frustrating for them, as the scientific community seemed to raise the hope that there were a lot of corpora freely available or at least available at low cost. Following this disappointment the students decided to build their own corpus.

Now they had to look for electronic texts and their translations. Within this context a statistics-based program was written in order to automatically identify the language in which a given text is written, since some of the (EU) documents were not written in the language declared by their file name. Furthermore, the students had to establish criteria in order to decide which texts should be disregarded and which should be processed within the next steps. Here the quality of the translation as well as the text format played a decisive role. For example the students disregarded the

[1] Of course, the situation has changed since the beginning of 2001.

Bible texts located in the internet[2] because the German and the English version were not really translations of each other, but had different source texts written within different time periods. The Linux HOWTOs and FAQs[3] which originally raised the students' hopes were also rejected due to their structural differences. Thus, the students obtained and processed primarily two resources with rather short texts:

DE-News: [4] These are texts from German radio broadcasts translated by non-professional volunteers to English. Although the translation quality varies it is good in general. The texts are given in ASCII/HTML format. The students collected about 14.5 KB (approx. 2 mio. words[5]) given in about 2,200 files.

EU-publications: [6] These texts include press releases, news, political documents and contracts which are given in a HTML-like format. The students collected about 93.7 KB (approx. 11,5 mio. words) given in about 23,600 files.

Now due to the high number of files to deal with various scripts for file administration tasks were implemented for organizing directory structures, corpus browsing, standardizing file names, verifying file contents etc. Then the students normalized the selected texts in order to have a uniform format supporting the identification of headings, paragraphs and sentences. Furthermore, it should be tractable by a tagging and lemmatization tool. Some information irrelevant for tagging but supporting the alignment process was encapsulated via SGML-formatting so that the tagger could ignore it.

The students decided not to implement this annotation program themselves but opted for using the already existing and well-established tagging program DT-Tagger (Schmid, 1994) for reasons of economy and quality as well as for sentence boundary detection. During the phase of normalizing and tagging the students had to realize that the orthography of the German part had to be adapted to the lexical representation used by the tagger[7]. Moreover the students implemented a script for improving sentence identification which is very important for sentence alignment.

3.2.2. Lexical resources

Bilingual German English lexicon entries play a major role within the KoKs project. Firstly they are made use of within the sentence and phrase alignment process (s. section 3.3.). Secondly they are consulted in order to determine whether a given German English phrase pair is a candidate for a collocation. And thirdly the lexicon is the place where to store German and English collocations and their respective translations. Some of these collocations are inherited from the imported lexica and others are to be added to the lexicon incrementally on the basis of the corpus analysis.

The lexical data initially used within the KoKs project is based on different lexical resources, especially Ding[8] containing about 150,000 entries at that time, and Tyler and Chamber[9] comprising about 10,000 entries. For normalization purposes a dictionary entry parser was implemented in order to extract lexical items and their translations and to store them in a unified format.

For multi word expressions tagging was done in the same way as for textual resources. For the lexical resources alignment was skipped because it was already given by the respective translations. Altogether texts, lexical entries as well as identified phrases were stored in a highly structured database.

3.3. Corpus analysis

After this preparation work the bilingual corpus was aligned from paragraphs down to the level of phrases[10]. Especially the sentence alignment was a very important aspect of the KoKs project as it was one significant prerequisite for the following phrase alignment. After disappointing attempts to use the algorithm described in (Gale and Church, 1993) the students developed their own alignment tool containing a "Matrix Alignment Visualisation Tool", called Mavis. The alignment was mainly based on a lexical and length-based distance measure. In order to calculate the lexical distance the bilingual dictionaries are used to count the number of those word translations found in a sentence pair which are legitimated by the bilingual dictionaries. Furthermore trigram correspondences were calculated for the remaining open-class words.

Finally the detection of phrase correspondences had to be done. In this phase noun and verb phrases were identified again recurring to the results of the mentioned DT-Tagger enhanced by additional chunking rules. A statistical analysis of the corpora revealed which chains of parts-of-speech were frequent and from these chunking rules were derived. With the help of the rules every sentence was divided into multiple phrase candidates which were reduced to the major categories verb, noun and adjective, i.e. function words were deleted. The phrase alignment was done recurring once more to the translations of single words (and collocations) given in the lexical resources. As described in section 3.1. the method for identifying collocations was strongly depending on the definition of a collocation saying that a collocation contains a lexical element with a meaning in the context that it does not have as a single lexical element without the given context. Therefore the dictionary entries for single words were used in order to determine

[2] http://www.hti.umich.edu/index-all.html

[3] http://www.tldp.org/docs.html#howto

[4] http://www.isi.edu/~koehn/publications/de-news/

[5] The number of tokens was calculated using the Unix tool wc and are just rough approximations.

[6] http://europa.eu.int/rapid/start/welcome.htm

[7] On the 1st of August 1998 an orthographic reform had come into effect and was adhered to by the news agencies one year later.

[8] http://www.tu-chemnitz.de/dict

[9] http://www.june29.com/IDP/

[10] It was not always obvious, which of the collected texts of one language was the translation of which text of the other language. Within the DE-News the assignment of corresponding news was marked by preceding identifiers. Unfortunately sometimes two corresponding news items didn't have the same ID or some news items existed in one language only. To tackle the cases of permuted items the students applied the alignment tools developed in the project to (news) headlines.

whether the direct translations of the lexemes of a candidate phrase showed up in the corresponding, i.e. aligned phrase. Here the lexical distance measure sketched at the beginning of this section seemed to be a good measure of collocativity. As nevertheless the identification of phrases and their translational equivalents were quite uncertain, only those phrase pairs detected a given number of times were accepted to be collocations. It must be noted that this kind of statistics is completely different from that pursued by the advocators of a frequency based collocation definition.

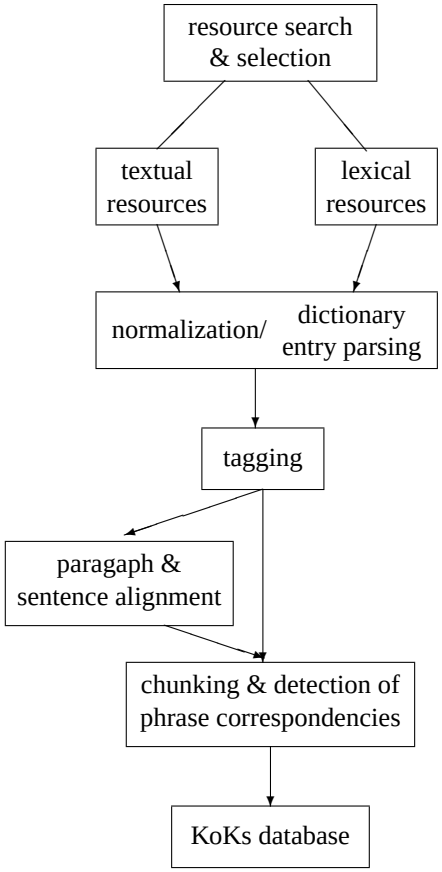


Figure 1: Core building blocks of the KoKs project

3.4. Pros and Cons

In the KoKs project the students on the one hand developed most of the computational tools for the analysis of the corpora by themselves and on the other hand worked with the resources from a linguistic perspective in order to identify and extract collocations.

Towards the end of the project the students had reached the situation from which they originally had planned to start their investigation. Of course this was to some degree frustrating. Perhaps in a classical seminar the teacher would have given the electronic German English texts directly, i.e. the DE-news and the EU publications. In this case, the time spent a) to look for affordable and possibly well-aligned German English texts as well as b) to search for texts and c) to select texts for future processing would have been saved. Furthermore, in a more guided course, the teacher certainly would have prevented the students from spending too much time for storing the corpus and the lexical data into a complex database. The design of the latter took a lot of time without really improving the collocation acquisition process originally the aim of the course. Perhaps within a more guided course the students could have spent more time on collocation extraction in a narrower sense.

However the students were highly motivated and, as a consequence, did a very good job given the opportunity to regard the system they developed under their own responsibility as their personal affair. This was recompensed by the opportunity to present the common results at an NLP workshop held at the international conference EUROCALL in Nijmegen in August 2001, and at the Workshop on Computational Approaches to Collocations held at Vienna in July 2002[11] (Kummer and Wagner, 2002).

Furthermore there is essential spin-off to be pointed out. The bilingual sentence-aligned corpus compiled within the KoKs project was integrated as an additional part into the dictionary-cum-corpus system LogoTax[12] which helps advanced students of German as a foreign language to compile and administrate a personal dictionary of German verb noun collocations. Additionally the KoKs corpus is available for future seminars and study projects to be held at the Master's Program Cognitive Science at the University of Osnabrück.

Finally there were two high-grade "Magister" theses written by student participants of the KoKs project that would have not been possible without the broad experience on corpus compilation gained by their authors within the study project. The first thesis treats sentence alignment of English-German parallel texts using linguistic knowledge (Tschorn, 2002), the second has "data driven machine translation" as its subject (Wagner, 2003). Thus advantage could be taken even of the additional load to build up a proper sentence-aligned German English corpus, which first seemed to be a long way round.

4. Mapping Architecture for People's Associations – MAPA

Within the MAPA study project 10 students from the Universities of Osnabrück, Tübingen and Bochum participated in a joint effort from October 2002 to October 2003 to develop a framework that allows the mapping of knowledge in a cognitively adequate way. The approach connects to existing techniques such as Mind Mapping, Concept Mapping and the like (Althaus et al., 2003).

4.1. Constructing Knowledge

A major assumption is that the internal mental representation of knowledge is network-like and that therefore its externalisation should be done accordingly (Spitzer, 2000). In the MAPA-system knowledge cues are usually represented in two ways, either as nodes or as relations between nodes, both of which are subsumed under the term 'entity'. An entity may contain any type of information, ranging

[11] http://www.ai.univie.ac.at/colloc02/workshop_ prog.html

[12] This system is available via internet: http://cato.cl-ki.uni-osnabrueck.de/~logotax/ and was developed at the Institute of Cognitive Science within a habilitation program

from simple text to images or any other data file (Bernedo and Elbers, 2003).

As a knowledge structuring tool the system is designed as a *personal* knowledge storage and retrieval device and therefore does not provide any predefined, i.e. typed elements. For this reason only one simple data structure 'entity' is needed, which may be parameterized by the user to be respresented as a link, a node or even a set of nodes. This reflects the goal of the students to build a system which limits the user with respect to possible knowledge items as little as possible. First and foremost the system should be used as a tool for knowledge *construction* in order to help the user to structure and organize his/her personal knowledge. The idea is to view the process of construction as the main task for gaining knowledge, i.e. learning (Novak, 1998).

4.2. GermaNet as a LR for Vocabulary Training

In opposition to the constructionist approach the students integrated data from GermaNet (Lemnitzer and Kunze, 2002) in an experiment in order to test navigational aspects for larger networks and to see whether a tool for second language vocabulary acquisition could be designed. GermaNet is based on the same principles as the original WordNet (Fellbaum, 1998), i.e. it is a lexical database for German with the lexems structured according to the more or less standard relation types from the field of lexical semantics, such as hyperonymy, hyponymy, antonymy etc. The network like structure of the GermaNet data was utilized to visualize the semantic relations of vocabulary items in some way similar to that of the 'Visual Thesaurus' (http://www.visualthesaurus.com). However only a restricted set of relations from the original GermaNet set was used in the transformation for the MAPA system: Synonymy, antonymy, hyponymy, hyperonymy, meronymy, holonymy and cause. Additionally two further relations were seen as necessary to be included. These relations are not used in GermaNet but are available in MAPA for a learner (and possibly a teacher) to add new material to the database.

- Association: This relation marks connections, which seem rather loose and are usually difficult to determine in linguistic terms as they are commonly derived from world knowledge.
- Lexical: The morphological relation of identical stems is marked with this type[13].

With the context, i.e. the lexical semantic relations being visualized, the memorization process should be more meaningful and enhanced than learning by simple word lists. Three program modules were developed for the vocabulary trainer:

1. The so-called 'Exploration Module' allows the user to explore, i.e. to surf the structure of the GermaNet structure and to possibly extract subnets. This mode might be used for the consultation of the database for a specific purpose or for getting the general idea of a certain branch of lexical knowledge.

[13]However this relation is included in the original WordNet set of relations.

2. The 'Training Module' presents certain types of exercises to a user. The learning task itself consists of naming deleted nodes or relations in the network by considering the neighboring nodes which makes it more powerful than simple multiple choice tests, as the distractors are not randomly chosen but are closely related to the sought-after item.

3. Finally there is a 'Profile Module' which acts as a student modeller well-known from the field of Intelligent Tutoring Systems (ITS) (Greer, 1994).

The following figure 2 with a question-mark for the obviously missing 'Haus' can exemplify one type of exercise.

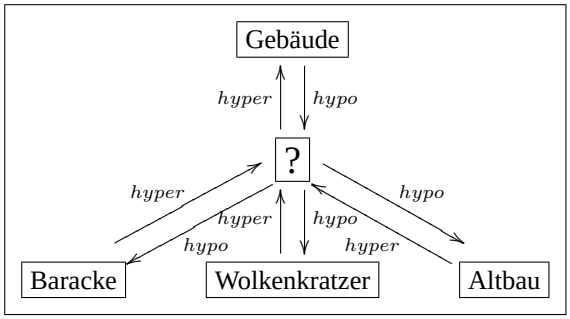


Figure 2: Example exercise (first 3 hyponyms for 'Haus' from GermaNet excluding compounds with '-haus')

Even though the usage of a lexical database such as GermaNet as a vocabulary trainer seems straightforward from a certain perspective, a number of difficulties were encountered that nevertheless lead to a deeper understanding of the technical handling of lexical databases and of the linguistic particularities of GermaNet. Two major technical challenges were encountered by the students in the preparation process of the GermaNet module in MAPA.

- As the students tried to transform the GermaNet data into the format used in the MAPA framework the number of nodes became increasingly difficult to handle. The reason for this was their first approach not to use a database but to store every 'entity' as a seperate file. Consequently the students had to learn to setup and maintain databases in order to handle to GermaNet data.

- One of the aims in the project was to use a XML-based framework for the data handling as is common in a lot of projects nowadays dealing with highly structured data sets. As the GermaNet data was encoded in a different XML-format than required in MAPA an intensive study phase of the foundations of XML was needed in order to make the data accessible.

With respect to the linguistic aspects of GermaNet a different set of challenges had to be mastered by the students. These were all related to presenting vocabulary items in a transparent and manageable way to a language learner.

1. One of the major problems while using GermaNet as the basis for language learning is the inclusion of

"none-lexem" nodes in the network. For example GermaNet contains the node ?*festes Nahrungsmittel* (solid food) as a hyperonym of *Grünzeug/s* (greens) and as a hyponym of *Nahrungsmittel* (food) in order to allow a more precise description and to avoid a hierarchy being too flat. Even though these "none-lexem" nodes are always marked with a question-mark, it is not always obvious whether one can simply delete the node and rearrange the connecting items in a simple manner. It is clear that exercises such as the one presented above cannot be used in circumstances where the deleted node is a multi word item. However in a program mode "exploration" multi word items might help a learner to structure his/her personal vocabulary knowledge in a more precise way as these nodes characterize the sub-nodes more accurately.

2. A second problem in the present scenario is the type of vocabulary contained in GermaNet. The sets of words to be added to GermaNet were selected in an almost random order by the developers so that some areas are covered more extensively than others. For language learning purposes however the ideal would be to have a GermaNet that does cover a basic stock of words with some extension to advanced vocabulary items. In our view no simple solution can be found for this problem. The students therefore explicitly included the possibility to develop one's own dictionary.

3. The inclusion of a learner dictionary pertains at least two aspects of the learning process. In the process of surfing the network the system can record the seen items and based on this produce a learner model for adaptation of the system. Additionally a learner dictionary can manage the learning steps by recording the tasks and items trained. The recordings can then be used for repeated but varied presentation of items until the vocabulary has been learned by the user.

4. Finally in the process of constructing networks for example tasks by themselves the students realized that more than the relations "predefined" by GermaNet were necessary for reasonable and effective exercises. This mainly stems from the paradigmatic view taken in GermaNet which is only one type in the array of lexical relations consisting also of e.g. syntagmatic (collocations) and phonologic (rhyme) relations both subsumed under "Association". Therefore two additional relations were introduced as mentioned before. However it was realized that also untyped relations should be allowed because on the one hand not all learners could be expected to be able to categorize every relation they wanted to enter into the system and on the other hand the high level goal of the MAPA project was to allow as much freedom as possible for creating knowledge networks.

In this project the focus was not on unstructured corpora but on a highly organized lexical database as the participants investigated the structure and the content of GermaNet in order to determine the feasibility of the project idea and to realize it. The students gained knowledge in the computational handling of structured language databases and the specific linguistic difficulties related to GermaNet. The results with respect to the vocabulary training module were presented at the GermaNet-Workshop 2003 at the University of Tübingen (Beck, 2003).

4.3. Project Related Experiences

In this subsection we would like report a few aspects of the project which are not directly related to the LR GermaNet but fall in the more general domain of project organization and management. The MAPA project was the first distributed study project started at the Institute of Cognitive Science in Osnabrück with different student/lecturer-ratios at the different locations: Osnabrück 7/3, Tübingen 2/1, Bochum 1/0. The single student in Bochum was also counseled by the teacher in Tübingen.

One of the main problems experienced was the "adequate communication problem" as the communication between the students is probably even more important in a distributed project. At the beginning of the project it became evident that a communication effort via email and a web-based WIKI-platform was not enough. Also the communication via a Chat-Tool tested in the project is cumbersome and requires a high degree of preparation, experience, discipline and a high level of tolerance. Finally telephone conference ermerged as a suitable communication device because participants can react more spontaneously and results can be reached quickly. Video-Conferencing was not an option as it was not available at all locations without too much technical preparation. The difficulties in communication emerged as a loss of coherent goals of the project for some time which could be reversed through a few face-to-face meetings of the whole group.

Finally a further difficulty in the project were the diverging levels of competence among the students with regard to formats, technical know-how and linguistic content of LRs. Some of the main efforts at the beginning of the project had to be invested in bringing the students towards an (almost) equal level of knowledge about LRs and their use.

4.4. Summary

We have presented the example of an intensive, distributed study project which included the extensive analysis of the LR GermaNet as the basis for a vocabulary trainer. With the perspective of using the LR for a clearly defined task at hand, a thorough analysis and idenfication of advantages and difficulties with this type of LR had to be carried out by the students.

The general concept of the project resulted from the students' wish to explore the topic of knowledge mapping and to develop a system for this type of mapping on the one hand and the requirement from us to include a module for computer-assisted language learning on the other hand.

It has to mentioned that no final theses have emerged from this project so far as the project was only finished in October 2003 and all of the students are still studying towards their degree.

5. Conclusion

As an alternative to classical seminars we have used LRs within the context of product oriented study projects, which allow the students to engage into the topic more deeply. Apart from the fact that this leads to greater commitment and responsibility by the participating students towards the success of the project, this also allows them to acquire extensive knowledge about corpus linguistics and relevant methods of CL. As opposed to a 'standard' view on LRs from a simple seminar where usually the access to LRs based on some predefined linguistic question is presented the students in projects are able to experience the creation and usage of corpora hands-on through all of the relevant stages.

Additionally a certain type of language awareness is raised in the presented type of handling LRs which in turn should lead to a more critical view on linguistic theories. As most theories make no difference between more and less frequent linguistic constructions it is important that the handling of LRs can help evaluate the theories with actually occurring constructions.

Finally study projects and the work with LRs allows students to gain more general research skills, which can not be taught in seminars and are an important step towards the learning goals of an university education. The fact that the results of the study projects have continuously been the basis for final theses and have been presented at major international conferences (e.g. EUROCALL '02, Colloc'02 Workshop on Collocations, EuroCogSci'03 and the GermaNet Workshop'03) demonstrates the success of this teaching concept.

However study projects also allow the participating students to make some of the more negative experiences such as disappointments about decisions and dead-ends in research. Projects based on LRs can then demonstrate the enormous effort necessary for developing suitable LRs and the limits of adaptivity of LRs with regard to specific tasks. These experiences usually cannot be made in a classical seminar where the exercises are developed with a predefined solution by the teacher.

All in all, the participating students have gained a fundamental insight into corpus linguistics which corresponds exactly to an expert's research work. They have gathered a varied and valuable work experience and acquired relevant techniques of which they can make use of within future working activities.

6. Acknowledgments

The MAPA project as part of the larger MiLCA-project was supported by a grant from the German Ministry for Education and Research No. 01-NM-167. All responsibilities lie with the authors.

The student participants in the KoKs project were Arno Erpenbeck, Britta Koch, Norman Kummer, Philip Reuter, Patrick Tschorn and Joachim Wagner with Helmar Gust and Petra Ludewig as teachers. The student participants in the MAPA project were Nadja Althaus, Kathrin Beck, Jasmine Bennöhr, Gordon Bernedo, Manuel Boeck, Michael Elbers, Felix Kugel, Stefan Scherbaum, Tobias Widdra and Jens Wissmann with Karin Krüger-Thielmann, Petra Ludewig, Veit Reuer and Claus Rollinger as teachers.

7. References

Althaus, N., K. Beck, J. Bennöhr, G. Bernedo, M. Boeck, M. Elbers, F. Kugel, S. Scherbaum, T. Widdra, and J. Wissmann, 2003. Abschlussbericht des Studentenprojekts MAPA im Masterprogramm Cognitive Science. Technical report, Institut für Kognitionswissenschaft, Universität Osnabrück.

Beck, K., 2003. Ein Vokabeltrainer auf der Grundlage von GermaNet und MAPA (Mapping Architecture for People's Associations). In *Proceedings des GermaNet-Workshops 2003*. Tübingen.

Benson, M., 1985. Collocations and idioms. In R. Ilson (ed.), *Dictionaries, Lexicography and Language Learning*. Oxford: Pergamon, pages 61–68.

Bernedo, G. and M. Elbers, 2003. MAPA – a platform for collaborative, cognitively adequate knowledge mapping. In *Proceedings of the EuroCogSci'03*. Osnabrück.

Breidt, E., 1993. Extraction of v-n-collocations from text-corpora: A feasibility study for german. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*. Columbo, Ohio.

Burger, H., 1998. *Phraseologie: eine Einführung am Beispiel des Deutschen*, volume 36 of *Grundlagen der Germanistik*. Berlin: Erich Schmidt.

Erbach, G. and B. Krenn, 1994. Idioms and support verb constructions in HPSG. In J. Nerbonne, K. Netter, and C. Pollard (eds.), *German Grammar in HPSG*. Stanford: CSLI, pages 365–396.

Erpenbeck, A., B. Koch, N. Kummer, P. Reuter, P. Tschorn, and J. Wagner, 2002. KoKs – Korpusbasierte Kollokationssuche. Technical report, Institut für Kognitionswissenschaft, Universität Osnabrück.

Fellbaum, C. (ed.), 1998. *WordNet: an electronic lexical database*. Cambridge, MA: MIT Press.

Gale, W. A. and K. W. Church, 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19:75–102.

Granger, S., 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (ed.), *Phraseology: Theory, analysis, and application*. Oxford: Clarendon Press, pages 145–160.

Greer, J. (ed.), 1994. *Student modelling: the key to individualized knowledge based instruction*. Berlin: Springer.

Heid, U., 1994. On ways words work together – research topics in lexical combinatorics. In W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg, and P. Vossen (eds.), *Euralex '94 – Proceedings of the VIth Euralex International Congress*. Amsterdam.

Heid, U., 1998. Towards a corpus-based dictionary of german noun-verb collocations. In *Euralex '98 – Proceedings of the VIII th Euralex International Congress*. Liège.

Krenn, B., 2000. Collocation mining: Exploiting corpora for collocation identification and representation. In *Proceedings of KONVENS-2000*. Ilmenau.

Kummer, N. and J. Wagner, 2002. Phrase processing for detecting collocations with KoKs. In *Proceedings of the*

Intl. Workshop on Computational Approaches to Collocations (Colloc'02). Vienna.

Lemnitzer, L. and C. Kunze, 2002. GermaNet - representation, visualization, application. In *Proceedings of LREC2002*. Las Palmas.

Lewis, Mi. (ed.), 2000. *Teaching Collocation: Further Developments in the Lexical Approach*. Language Teaching Publications (LTP). Hove.

Lüdeling, A., 2002. The productivity of collocations. In *Proceedings of the Intl. Workshop on Computational Approaches to Collocations (Colloc'02)*. Vienna.

Ludewig, P., 2001. Logotax: Bridging the gap between lexikon and text. In B. Daille and G. Williams (eds.), *Proceedings of the ATALA-Workshop on Collocations*. Paris: Association pour le Traitement Automatique de Langues.

Novak, J.D., 1998. *Learning, Creating and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations*. Mahwah, NJ: Erlbaum.

Schmid, H., 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the Intl. Conference on New Methods in Language Processing*. Manchester.

Sinclair, J., 1987. Collocation: A progress report. In R. Steele and T. Threadgold (eds.), *Language Topics: Essays in Honour of Michael Halliday*. Amsterdam: John Benjamins, pages 319–331.

Smadja, F., 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Spitzer, M., 2000. *Geist im Netz - Modelle für Lernen, Denken und Handeln*. Heidelberg: Spektrum-Verlag.

Tschorn, P., 2002. Automatically aligning English-German parallel texts at sentence level using linguistic knowledge. Magisterarbeit im Bereich CL&KI an der Universität Osnabrück.

Wagner, J., 2003. Datengesteuerte maschinelle Übersetzung mit flachen Analysestrukturen. Magisterarbeit im Bereich CL&KI an der Universität Osnabrück.

Employing GermaNet in Virtual Courses of Computational Linguistics

Claudia Kunze*, **Lothar Lemnitzer***

*Seminar für Sprachwissenschaft, Universität Tübingen
Wilhelmstr. 19, 72074 Tübingen, Germany
{kunze,lothar}@sfs.uni-tuebingen.de

Abstract

This paper focuses on the use of existing lexical resources as material for case studies and explorative learning in virtual courses of Computational Linguistics and Language Engineering. We will report on our experience with GermaNet, the German wordnet, as a key resource for eLearning which has been applied in three different virtual courses. Student projects which are centered around existing resources seem to be highly motivating. On the one hand, students gain hands-on experience with the acquisition, modelling, maintenance as well as with the application of lexical resources in NLP tasks; on the other hand, they can contribute to their further development and enhancement. With our paper, we aim at raising content-related as well as didactic issues which we consider relevant to the eLearning community in Computational Linguistics and neighboring disciplines.

1. Introduction

eLearning through virtual courses (Schulmeister, 2001) has become one important topic in teaching and curriculum design for Computational Linguistics and neighboring disciplines like Language Engineering, Cognitive Science, and General Linguistics (Lemnitzer and Schröder, 2003). Virtual courses offer students the opportunity to widen their scope of knowledge and skills beyond the curricular limits of their home universities. Groups of students from several distant places are able to work and study co-operatively, as well as in an interdisciplinary setting.

The design of virtual courses challenges the traditional way of teaching. Merely presenting material to students, like in a real classroom, cannot and should not serve as an adequate mode of teaching and learning in the virtual classroom. In particular, education on a higher level should include the application of acquired knowledge and practical skills to concrete tasks. We propose the integration of case-based scenarios which require the (inter)active participation of the students involved as one promising mode of successful teaching in the virtual classroom.

On-going activities in the field of computational lexicography, like the extension and restructuring of existing lexical resources, provide an ideal framework for defining various well-defined tasks, which students can perform, dealing with real data of a considerable size and complexity.

This article is structured as follows: In the next section, we shortly describe the virtual courses in which the GermaNet ressource, a lexical-semantic wordnet for German, has played a central role and give a brief overview of the main characteristics of the GermaNet database. Then we present the different student assignments which are tailored to the use of GermaNet. In section 4, the challenges of a virtual learning scenario are summarized. Our conclusion emphasizes how motivating an application-oriented perspective within eLearning can be.

2. GermaNet in three virtual CL courses

We will show that lexical-semantic wordnets like GermaNet (Kunze, 2001; Kunze and Naumann, 2004) are useful subjects of teaching and learning in the virtual classroom. GermaNet is currently employed as a key resource within two virtual courses in the framework of a national eLearning project, MiLCA[1]: *Computational Lexicography* (CoLex), held in Tübingen, and *NLP tools for Intelligent Computer Aided Language Learning* (I-CALL), held in Osnabrück[2]. The third course, *Applied Computational Linguistics* (ACLing), has been designed as a part of VirtuGrade[3] and taught at the University of Tübingen. CoLex and ACLing are virtual courses open to students from different universities in Germany and Switzerland[4].

The underlying didactic scenario is characterized by a combination of asynchronous and synchronous computer-mediated communication modes. The students meet regularly in a virtual classroom. Technically, this is a text-based chat tool which offers the facilities of a room for plenary sessions and rooms for group work. The students have access to shared workspaces for their respective groups and to extensive teaching material. For asynchronous communication, the ILIAS[5] Learning Management System we are using provides mailing lists as well as forums. The facilities for synchronous and asynchronous communication are used by the students for solving the problem sets in their groups during the course and by the teachers for supervising the homework assignment after the course.

CoLex thematizes the acquisition, modelling and maintenance of lexical resources from a computational perspective and the use of lexical resources in NLP applications.

[1]The acronym stands for 'media intensive learning modules for the practical training of computational linguists', cf. http://milca.sfs.uni-tuebingen.de/.

[2]The latter project is subject of another presentation on this workshop. We will therefore constrain our presentation to that part which concerns the use of GermaNet.

[3]This acronym stands for 'virtual courses for graduate students', cf. http://www.virtugrade.uni-tuebingen.de/.

[4]Students from Potsdam, Osnabrück, Bonn, Gießen, Saarbrücken, Heidelberg, Tübingen and Zürich participated in these courses.

[5]ILIAS, an open source Learning Management System, is being developed at the University of Köln, cf. http://www.virtus.uni-koeln.de/ILIAS/. This platform has been used within the MiLCA project.

In *ACLing*, various language engineering skills for the development of a complex access engine to lexical resources are introduced. For both courses, GermaNet figures as a prototype of a lexical database. In the first course, aspects of data acquisition and data modelling are focussed, with special emphasis on lexical semantics. In the other course, GermaNet constitutes one of two reference resources, the second is a bilingual dictionary.

In the *I-CALL* course, which centers around the development of a cognitively inspired platform for knowledge modelling, a vocabulary trainer has been developed as a reference implementation. This program makes use of the GermaNet data which had to be reformated for that purpose.

We know from our teaching experience that students enjoy working with GermaNet. Lexical-semantic wordnets seem to be appealing for the clarity and simplicity of their structures, the richness of their contents and the variety of natural language processing tasks in which they may play a role. GermaNet basically follows the guidelines and design principles of the Princeton WordNet (Fellbaum, 1998). In wordnets, the central unit of representation is the so-called *synset*, which comprises the set of synonyms which express a given concept, e.g. {*unhappy, sad*}. GermaNet covers the most frequent and important concepts of the German base vocabulary and the prominent semantic relations that hold among the words, like synonymy, antonymy, hyperonymy, meronymy, causation, etc. Partially integrated into EuroWordNet (Vossen, 1999), GermaNet also provides for cross-lingual links.

Wordnets have evolved popular background resources in various NLP applications that rely on word sense disambiguation like

- information retrieval and extraction;
- machine translation;
- language generation;
- text summarization;
- building of language tools and resources;
- semantic annotation of corpora, etc.

Wordnets are also applied as background resources for ontology building and ontology engineering, as they provide for taxonomic relations. Thus, wordnets are crucial in various processing tasks of Computational Linguistics and on the Semantic Web so that it makes sense to introduce them as teaching content for CL courses. Wordnets are interesting research topics, both from a theoretical and technical point of view. Some larger student projects that have been assigned focus on different lines of current wordnet research, exemplified on the basis of GermaNet data:

- modelling of linguistic contents,
- data structure and presentation,
- tools for accessing and visualizing wordnet structures,
- wordnets as lexical resources for NLP applications.

3. Student projects dealing with wordnets

This section serves to give an overview about the wordnet related student assignments in the context of our virtual courses. Certain aspects of the overall tasks have also been integrated in the explorative exercises of the on-line classes in order to motivate students for a larger assignment.

3.1. Modelling of linguistic contents

A substantial part of the student assignments deals with modelling linguistic entities in wordnets, as with regard to the semantic relations which hold among the concepts and lexical units, and the organization of senses and sense distinctions within the taxonomic hierarchies.

3.1.1. Analysis of the meronymy/ holonymy relation and its encoding in GermaNet

The German wordnet so far only encodes one unique pointer for covering all instances of the part-whole-relation (meronymy). In contrast to GermaNet, Princeton WordNet accounts for three different types of meronymy relations (part, member, substance), and EuroWordNet even realizes one generic meronymy pointer for underspecified instances and five subtypes (part, member, substance, made of, region). It could be very useful to implement a similar subdivision of pointers for GermaNet. Concepts being encoded as meronyms should therefore be checked under the following aspects: a) Is the application of three, or even five, meronymy pointers feasible for GermaNet? b) How should we handle pairs of concepts for which the meronymy relation is not symmetric? c) Will a subdivision into different meronymy pointers yield transitivity for these relations or are there still instances for which transitivity is blocked?

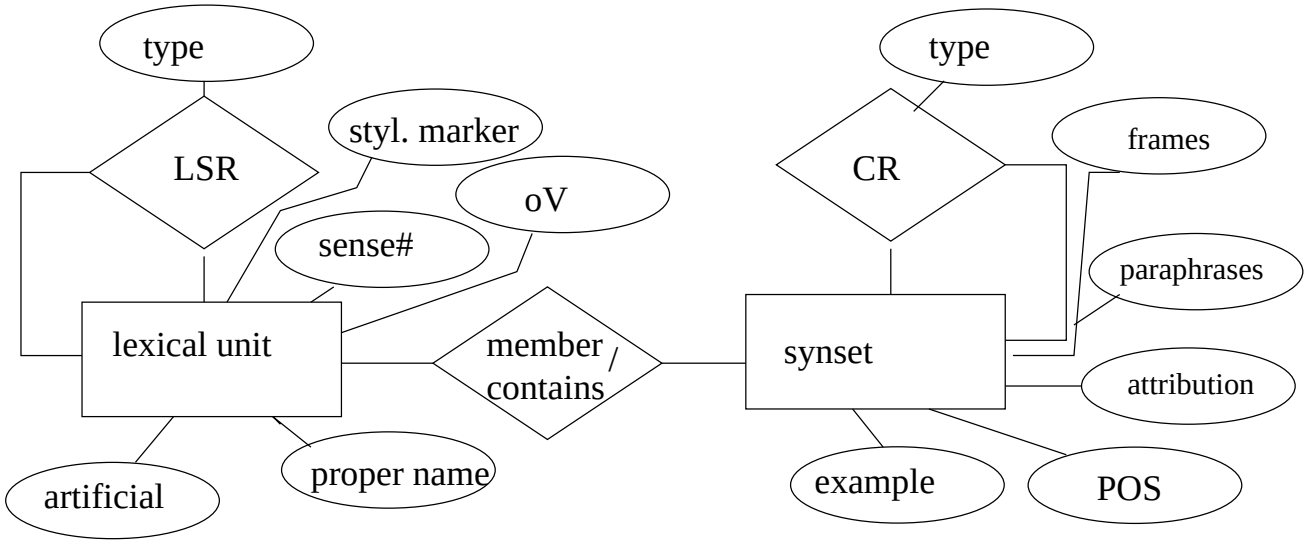
The investigation, which is currently carried out, considers the WordNet subdivision of meronymy and the classification proposed by Chaffin (1992), p. 274ff. This project aims at reviewing and refining the meronymy/ holonymy relation in GermaNet.

3.1.2. Analysis of the antonymy relation

Similarily to the case of meronymy, GermaNet implements a unique pointer for encoding the antonymy relation between lexical units. Different types of opposites, like 'man' vs. 'woman', 'busy' vs. 'lazy', 'warm' vs. 'cold' and 'arrive' vs. leave' are thus subsumed and uniformly treated under the label of antonymy. The student exercise consists in developing an adequate subclassification of antonymy, dividing the data in appropriate subgroups which should account for complementary opposites, scalar and gradable opposites respectively. Furthermore, a set of relevant features should be defined which captures opposites of, e.g., sexus or directionality, for nouns and verbs. The empirical analysis of the GermaNet antonyms should account for the categories being proposed in the descriptive approaches of Cruse (1986) and Agricola and Agricola (1987).

3.1.3. Applicability of regular polysemy in wordnets

Pustejovsky et al. criticize WordNet for ignoring existing regularities between word senses (Pustejovsky et al., 1997), like the systematic tree-wood alternation for 'oak', 'birch', etc. It is, however, still a controversial issue whether wordnets should implement regular sense rela-



CR=conceptual relation; LSR=lexical−semantic relation; oV=orthographic variant

Figure 1: Entity-Relationship graph of the GermaNet data model.

tions, and, if so, which should be the appropriate conceptual level for the application of such rules. The analysis concentrates on lexical (sub-)fields which are in the scope of a regular sense extension, e.g. instances of the type 'building-institution-staff' or instances of the type 'tree-wood-fruit'. It will be checked whether generic rules are feasible, and, if yes, on which level of abstraction they should apply and when a blocking of these rules would be necessary.

3.2. Data structure and presentation

XML and RDF provide suitable formats for making (wordnet) data interchangeable and accessible for NLP applications and on the Semantic Web. Thus, converting the original lexicographers' files into an XML representation, or integrating the GermaNet objects and relations into RDF, constitute suitable assignments.

3.2.1. Conversion of the lexicographers' files into an XML format

Neither the original GermaNet lexicographers' files nor the compiled database yield an ideal format for data exchange, presentation, and integration into NLP systems. XML is more convenient for these purposes. Based on the data model of GermaNet which was realized as an Entity-Relationship graph (cf. figure 1), several students have developed programs which convert the Lexicographers' files of GermaNet into an XML representation. The respective DTDs have been created jointly. The outcome of this project has been documented in some detail (Kunze and Lemnitzer, 2002a; Kunze and Lemnitzer, 2002b).

3.2.2. Integration of the GN objects and relations into the Resource Description Framework

Some work has already been carried out in view of integrating WordNet into the Resource Description Framework in the context of the Semantic Web initiative (Melnik and Decker, 2001), but the resulting files encompass only a part of its structure. Before starting to convert GermaNet accordingly, and even more exhaustively, we would like to figure out how well wordnet structures fit into the structures of full-fledged knowledge representation languages like DAML and OIL which are built on top of RDF. An examination of these languages with wordnet structures in mind should prove or disprove the usefulness of GermaNet objects and relations for the RDF and the KR languages.

3.2.3. GermaNet representation as Scalable Vector Graphics

SVG (Ferraiolo et al., 2003) might turn out to be a handy standard and tool for the visualization of wordnet objects and relations. A wordnet can be conceived as a large map where with a graphical representation of substructures at different levels of detail. The user can zoom in at a particular synset and inspect its context. The project should explore the feasibility of data conversion into the SVG format and the availibility of visualisation tools.

3.3. Tools for accessing and visualizing wordnet structures

Based upon the XML version of the GermaNet data, several tools for extracting taxonomic chains and/or conceptual neighborhoods of wordnet items as well as for visualizing concepts and partial structures of wordnets are built.

3.3.1. Development of tools for the extraction of the lexical and conceptual neighborhood of a lexical unit or a synset

The assignment in question addresses a user need for extracting data which are neighboring a particular synset or lexical unit. Currently there are two projects devoted to this task. One student employs a relational database for the intermediate representation of the data, the other accesses the data in their original format, using XSL Transformations to generate the output. Both results will be evaluated in terms

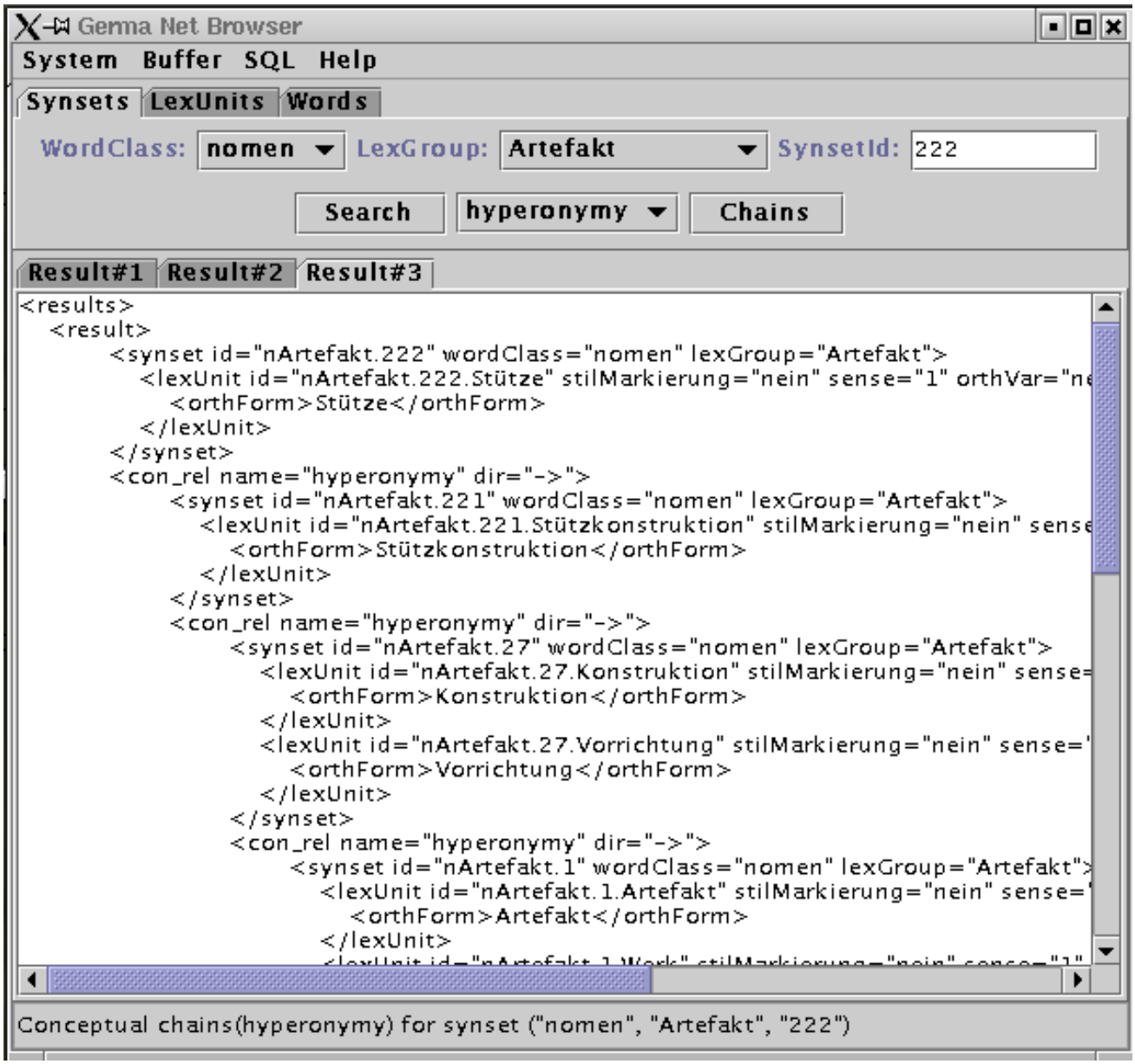


Figure 2: GUI of a GermaNet extraction tool.

of their speed and flexibility. The GUI of one of the tools is shown in figure 2.

3.3.2. Visualization of the wordnet

Within another student project, a visualisation tool which operates over the whole wordnet structure has been developed. The XML representation of the wordnet is used as a data base. The visualization of the data is very flexible but too slow for realistic user scenarios. Results of the project have originally been presented in Kunze and Lemnitzer (2002a). The outcome of the project has motivated our search for representation alternatives, e.g. Scalable Vector Graphics (see section 3.2.3.).

3.4. Wordnets in NLP applications

For numerous GermaNet users, this wordnet serves as a key resource in various NLP applications, like information extraction and retrieval, text summarization, etc. E.g., Steffen et al. (2003) use GermaNet for word sense disambiguation in the medical domain, Rösner and Kunze (2003) for domain specific document analysis. In this section, an application is described which integrates the GermaNet synsets and relations as a knowledge repository for a vocabulary trainer.

3.4.1. GermaNet as a lexical basis for a vocabulary trainer

A group of students in Osnabrück, Edinburgh and Tübingen has developed a network-like platform for collaborative work (MAPA = Mapping Architecture for People's Association). Within this framework, a prototypical programme for constructing networks of knowledge is being applied. Users should be enabled to link words and data of any kind such as images, videos, emails and documents, in order to represent their knowledge in a cognitively adequate way. Furthermore, MAPA allows for collaboration among users and integration of other users' networks in

the overall architecture. The development of a vocabulary trainer serves as reference application in this project. The GermaNet data are represented in a network-like fashion which helps the language learner to understand and learn German words in a wider semantic context. Several features of the application, e.g. a user profile which overviews the learning process, support the user's individual learning strategy. The outcome of the project was reported on the GLDV-Workshop "Applications of the German Wordnet in Theory and Practice" in October 2003 (Beck, 2003).

4. Challenges of the virtual learning scenario

In our opinion, virtual courses like the ones described above have two major advantages over traditional classroom teaching.

First, the virtual, computer-mediated environment enables access to various language resources, including lexicons and corpora. Thus, both hypotheses which guide the line of investigation and hypotheses which evolve from specific analyses can be tested against large bodies of data by the use of sophisticated techniques like pattern matching, concordancing, frequency counts and filtering.

Second, learning scenarios which make use of computer-mediated communication facilitate the access to contents which would otherwise not be available. Furthermore, and even more essential, special interest groups can be established with learners and experts from different sites and fields. Thus, interdisciplinary education and joint research are promoted.

For teachers who want to use existing language resources, the challenges are:

- to define projects of manageable size;
- to prepare the data such that students work only on those samples which are relevant for the task;
- to give an account of the state of the art as a context for the project and to provide for the relevant background literature.

5. Conclusion

With our presentation of various wordnet-related student projects, we have demonstrated the use of an existing large-scale lexical resource as an object of teaching and learning in virtual environments. The modelling, compilation, maintenance, evalution and use of language resources is a long-term activity which requires knowledge and skills in a wide variety of fields: Lexical Semantics, Theoretical Linguistics, Text Technology, Natural Language Processing etc. Well-tailored assignments which draw on these resources enable students to acquire a deeper knowledge in one or several of these fields. Assignments which cross the border of these fields are ideal cases for interdisciplinary groups.

The availability of lexical resources, e.g. wordnets, for several different languages offers students to choose a lexical resource of their mother tongue.

Last but not least, the perspective of contributing to an on-going research and development activity increases the intrinsic motivation of the students who are involved in these projects.

6. Acknowledgements

The MiLCA project has been funded by the German Federal Ministery of Education and Research (project ID: 01NM167). VirtuGrade has been funded by the government of Baden-Württemberg.

We would like to express our gratitude to the students who carried out the work in the course assignments, among them: Annika Böcher, Nicholas Dille, Alexander Grebenkow, Milen Kuylekow, Martin Müller, Sang-Ah Shim, Iris Vogel, Holger Wunsch. The authors of this paper are, however, fully responsible for its content.

7. References

Christiane Agricola and Erhard Agricola. 1987. *Wörter und Gegenwörter. Antonyme der deutschen Sprache*. Leipzig: VEB Bibliographisches Institut.

Kathrin Beck. 2003. Ein Vokabeltrainer auf der Grundlage von GermaNet und Mapa (Mapping Architecture for People's Association). In Claudia Kunze, Lothar Lemnitzer, and Andreas Wagner, editors, *Proceedings of the First GermaNet User Conference, Tübingen, 9-10 October 2003*, pages 46–55. GLDV.

Roger Chaffin. 1992. The concept of a semantic relation. In Adrienne Lehrer and Eva Feder Kittay, editors, *Frames, Fields, and Contrasts. New Essays in Semantic and Lexical Organization*, pages 253–288. Lawrence Erlbaum, Hillsdale.

D. Alan Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge/Mass.

John Ferraiolo, Jun Fujisawa, and Dean Jackson. 2003. Scalable Vector Graphics (SVG) 1.1 Specification. http://www.w3.org/TR/SVG11/.

Claudia Kunze and Lothar Lemnitzer. 2002a. GermaNet - representation, visualization, application. In *Proc. LREC 2002, Gran Canaria, May/June*, pages 1485–1491.

Claudia Kunze and Lothar Lemnitzer. 2002b. Standardizing Wordnets in a Web-compliant Format: The Case of GermaNet. In *Proc. LREC 2002 Workshop on Wordnet Structures and Standardization, a nd how these affect Wordnet Applications and Evaluation*, pages 24–29.

Claudia Kunze and Karin Naumann. 2004. Germa Net Homepage. URL: www.sfs.uni-tuebingen.de/lsd.

Claudia Kunze. 2001. Lexikalisch-semantische Wortnetze. In K.-U. et al., editor, *Computerlinguistik und Sprachtechnologie: eine Einfhrung*, pages 386–393. Spektrum Verlag, Heidelberg.

Lothar Lemnitzer and Bernhard Schröder, editors. 2003. *Computerlinguistik - neue Wege in der Lehre*. IKP, Bonn.

Sergey Melnik and Stefan Decker. 2001. Wordnet RDF Representation. http://www.semanticweb.org/library/.

James Pustejovsky, Bran Boguraev, Marc Verhagen, Paul Buitelaar, and Michael Johnston.

1997. Semantic Indexing and Typed Hyperlinking. In *Proceedings of the AAAI '97*. http://www.cs.brandeis.edu/ llc/publications/aaai97.ps.

Dietmar Rösner and Manuela Kunze. 2003. Issues in Exploiting GermaNet as a Resource in Real Applications. In Claudia Kunze, Lothar Lemnitzer, and Andreas Wagner, editors, *Proceedings of the First GermaNet User Conference, Tübingen, 9-10 October 2003*, pages 18–27. GLDV.

Rolf Schulmeister. 2001. *Virtuelle Universität – Virtuelles Lernen*. Oldenbourg, München.

Diana Steffen, Bogdan Sacaleanu, and Paul Buitelaar. 2003. Domain Specific Sense Disambiguation with Unsupervised Methods. In Claudia Kunze, Lothar Lemnitzer, and Andreas Wagner, editors, *Proceedings of the First GermaNet User Conference, Tübingen, 9-10 October 2003*, pages 79–86. GLDV.

Piek Vossen. 1999. *EuroWordNet: a multilingual database with lexical-semantic networks*. Kluwer Academic Publishers, Dordrecht.

Student Projects in Language and Speech Technology

Dan Cristea[♥♣] dcristea@infoiasi.ro
Horia-Nicolai Teodorescu[♦♣] hteodor@etc.tuiasi.ro
Dan-Ioan Tufiş[♠♥] tufis@racai.ro

[♥]University "Al. I. Cuza" of Iaşi, Faculty of Computer Science
[♦]Technical University of Iaşi, Faculty of Electronics and Telecommunications
[♣]Romanian Academy – the Iaşi Branch, Institute for Theoretical Computer Science
[♠]Romanian Academy, Research Institute for Artificial Intelligence

Abstract

The paper reports on term homework and projects in correlation with Natural Language Processing courses delivered at the Faculty of Computer Science of the "Al.I.Cuza" University of Iaşi (UAIC-FII) to both undergraduate students in Computer Science and graduate students enrolled in the Master Programme in Computational Linguistics, during the university year 2003-2004. All projects make heavy use of language resources, in either written or spoken form.

1. Introduction

At UAIC-FII, two categories of courses are taught in different areas of natural language and speech processing. At the undergraduate level, the terminal year students can take an elective course on natural language processing (which presents mainly theories and techniques for discourse level representation and processing of language), and at the master level, the students in Computational Linguistics, over the two years program, take a general introductory course on computational linguistics, a theoretical course on syntax, a course on corpus linguistics, one in lexical semantics, one on machine translation and one on speech processing. The laboratory activities of both undergraduate and graduate courses make heavy use of corpora and other linguistic resources. In these activities students usually work in teams to perform term projects and homework. The goal of all projects is three fold: a). to train students to exploit existing corpora through the program interfaces available on the web, by integrating function calls in their own applications; b). to train students to make use of annotations applied to texts, by first manually annotating a small corpus and then devising their own tools that exploit the annotation for NLP applications; c). to build small annotated corpora, including spoken language (speech) corpora, and appropriate exploitation software that remain in the Faculty for further NLP research.

2. Undergraduate projects

One category of undergraduate projects dealt with *Textual resources: acquisition and exploitation*. All projects were group projects, 4-5 students each.1

Lattice of XML annotation standards: the students working in this project had to propose a technique to transform an XML annotation standard into a node of a directed acyclic graph. Details on this project can be found in (Cristea and Butnariu, 2004). A node of the hierarchy records a set of tag names (XML element tags), their corresponding attributes, and possible semantic relations between attributes. Any node inherits all features of its parents in the hierarchy. A node (standard) A is said to subsume a node (standard) B in the hierarchy (therefore B is a descendent of A) if and only if:

- any tag-name of A is also in B;
- any attribute in the list of attributes of a tag-name in A is also in the list of attributes of the same tag-name of B;
- any semantic relation which holds in A also holds in B;
- either B has at least one tag-name which is not in A, and/or there is at least one tag-name in B such that at least one attribute in its list of attributes is not in the list of attributes of the homonymous tag-name in A, and/or there is at least one semantic relation which holds in B and which doesn't hold in A.

As such, a hierarchical relation between a node A and one descendent B describes B as an annotation standard which is more informative than A and/or defines more semantic constrains. On a lattice of annotation standards of this kind and a collection of documents annotated corresponding to these standards, all having the same hub document (original empty-annotation document), a set of operations could be defined. By proper definitions or by classifying a set of documents obeying different standards, a hierarchy can be build. Then, on such a hierarchy and a corresponding collection of documents, *merge* and *extract* operations can be defined. A *merge* operation combines two documents having identical hubs and corresponding to two distinct nodes of the hierarchy and produces the document containing the union of the annotation tags of the two original documents and the corresponding standard in the hierarchy. The resulted standard will then be placed in the hierarchy as a common daughter of the original document standards. An *extract* applies the reverse operation, extracting from a document, which corresponds to a certain node in the hierarchy, a document conforming to one of the node's ascendants in the hierarchy. *Concurrent-checks* could also be defined. Two annotations are called concurrent if they intend to represent the same linguistic phenomenon from different perspectives, therefore possibly resulting in different solutions. Viewed within the frame of the hierarchical graph representation, concurrent documents cannot be merged in a single file and its corresponding standard for the reason that the resultant annotation would contain crossing markings.

The project, realized in Java-DOM, intended to train student's abilities to model and realize complex

[1] The projects are posted (in Romanian) at the following URL: http://thor.info.uaic.ro/~dcristea/cursuri/LC/lcproiecte.htm.

applications dealing with annotation standards. More than a didactic benefit, the resulted tool proved to be already extremely useful for the corpora-driven research developed in our NLP laboratory.

Semantic structures for automatic translation. A large project aimed at training students on techniques of automatic translation. 10 groups of students, 4 members each, received the George Orwell's "1984" English-Romanian aligned parallel corpus, initially tagged in both languages to part-of-speech. Students had to recognize senses of words and to annotate these senses conforming to two aligned lexical thesaurus, the English WordNet (Fellbaum, 1998), as part of the EuroWordNet project (Vossen, 1997), and the Romanian WordNet (RoWN) (Tufis and Cristea, 2002), as part of the BalkaNet project (Tufis, Cristea and Stamou, 2004), and to build parallel semantic frames of translation-equivalent verbs. More precisely, their task was:

- to find all verb occurrences in English and to sort them in the descending order of their frequency;
- among the most frequent verbs, each group had to choose 10 English verbs and to select from the parallel corpus all the language-pair segments of occurrences;
- they had to annotate the occurrences of these verbs, in both English and Romanian, to senses (Inter Lingual Index codes), according to EuroWordNet and RoWN;
- then subcategorisation constituents of verbs had to be annotated: their syntactic role, the head word, and the sense of the head word – using also the EuroWordNet ILI codes;
- then, students had to select all occurrences in which a verb was considered to have the same sense and to generalize a semantic frame out of the set of constituents found around it. For a given constituent, say the direct object role, the generalization had to be the lowest concept in the wordnet hierarchy subsuming all senses of head words found on the role of direct object in the selected examples. If no generalization of this kind could be found, due to the fact that, for each part-of-speech, wordnet contains a collection of graphs, not just one, the union of the lowest computed role-concepts was computed;
- the final goal was to report a collection of English-Romanian frames around verbs that have given rise to parallel translations, which could be considered the kernel of a semantic transfer grammar.

Although the results were rather unequal, from spectacularly good to poor, overall the project was successful, since after finishing it most students had a very clear sense of the advantages of using annotated corpora in NLP, and they learned the technology to obtain annotated corpora and to exploit them. Moreover, the best rated projects have thrown the seeds for furthers master-level research.

Acquisition of a corpus of Romanian texts. All students had to find on the Web, collect and annotate conforming to the PAROLE schema (Villegas et al., 2000) (document header and paragraph level annotation) a corpus of Romanian texts summing one million words. One group was responsible to build an interface aiming to facilitate the uploading of individual corpora, the validation of the PAROLE headings and the filtering of uploads, in order to allow only PAROLE-conformant new texts. The uploading site is http://www3.infoiasi.ro/~toni/lingvistica/index.php. Although being perceived as tedious by most students, excepting by those who had to build the uploading and filtering interface, the theme raised the interest for a concentrated web search group activity that can result in the acquisition of a consistent linguistic resource, extremely useful for NLP research.

3. Postgraduate projects

The following series of projects were given to master students in Computational Linguistics, first and second year.

Assembling parts of a wordnet. The project dealt with the Romanian wordnet, part of the BalkaNet multilingual wordnet, using PWN as an inter-lingual index. After a detailed presentation of the Princeton WordNet (PWN) and of the multilingual architecture of the BalkaNet project, the students were trained in using the WNBuilder acquisition tool. WNBuilder (Tufis&Barbu, 2004) is a user-friendly interface integrating various language resources for Romanian and English (Explanatory Dictionary, Dictionary of Synonyms, Princeton WordNet, Romanian-English translation dictionary) and allowing for collaborative work in synsets definition and their linking to the counterpart synsets in PWN. Each student had to construct for every synset in distinct sets of English synsets extracted from PWN the corresponding Romanian synsets. Building a synset assumed identifying a synonymy list from the Romanian Dictionary of Synonyms (RDS), assigning sense numbers to each literal, based on the numbering of senses in the Explanatory Dictionary of Romanian (EDR), choosing from EDR the most adequate definition for the synset and finally establishing the interlingual relation with the starting point synset from PWN. There are various interlingual relations as defined in EuroWordNet (Vossen, 1997): EQ-SYN, EQ-HYPERONYM, EQ-HYPERONYM, EQ-MERO, EQ-HOLO. The students had to overcome (and explain their solutions) different difficulties arising from: different granularities of the underlying dictionaries, PWN and RDS&EDR, lexical gaps, missing senses in EDR, splitting conjunctive definitions, etc.

A special tool WNCorrect (Tufis&Barbu, 2004) was used to evaluate and correct each student's synsets. The evaluation tool provided detailed reports on the syntactic and semantic errors that allowed an objective assessment of each student's work.

Sense disambiguation. To further refine the students understanding of lexical semantics issues, in a second task they were engaged in a word-sense disambiguation exercise. Each student was given a set of English sentences containing several occurrences of different target words. The students had to semantically disambiguate all the targeted words by choosing the sense numbers from PWN. The context for sense disambiguation exercise was defined by the sentence containing the targeted word. Additionally, each student had the possibility to add comments whenever in doubts on the appropriate sense assignment. The comments indicated, among other thing, insufficient context and too fine-grained sense distinctions.

The set of English targeted words were extracted from the parallel corpus "1984" so that all their senses (at least two per part-of-speech) defined in PWN were also implemented (and interlingually aligned) in the RoWN. There resulted 211 words with 1832 occurrences. An extraction script generated, for each student a set of sentences containing occurrences of the targeted words. The extraction process ensured that the same sentence was in at least three student-sets. Therefore, in the end, each occurrence of a targeted word was sense disambiguated by at least three students. The same-targeted words were automatically disambiguated by the WSDtool system (Tufis et al. 2004a; Tufis, Ion and Ide, 2004b). WSDtool exploits the multilingual wordnets in BalkaNet (in this case the pair RoWN&PWN) and build on the TREQ-AL word-alignment program (Tufis, Barbu and Ion, 2003). The previous evaluation of the WSDtool performance on the same data has shown high accuracy (>85%) and as a result of that evaluation, we were able to construct a gold standard against which the students' assignments were evaluated.

The evaluation files contained detailed information for each occurrence of the targeted word:

- the name of the student that evaluated the occurrence and the sense he/she assigned;
- the comments the student had on the sense assignment in case;
- the sense in the gold standard;
- a majority-voting sense number as resulted from the students' sense assignments.

Speech processing. A second category of master student projects exploited *speech data* in correlation with the course Speech Processing (Analysis, Recognition, and Synthesis). The goal of this project was multifold:

- to help student better understand speech signal processing, and more broadly, speech technology;
- to help student better understand the meaning of the prosody and its characteristics;
- to help students understand speech production;
- to help students understand inter-speakers and intra-speaker speech variability;
- to help students master the technical tools to analyze voice signals;
- to help students understand formantic and concatenative synthesis, their principles, their relative advantages and their disadvantages;
- to train students in acquiring small collections of speech data.

We believe that all these goals structure the knowledge needed to build sound speech corpora.

We have an experience of two years of teaching this course. It is important to say that the class has been, during both years, quite heterogeneous. Classes included about 50% students with a background in linguistics, about 25 % students who graduated computer science (informatics) and the rest of about 25% of the students with other background studies (including such varied studies as philosophy).

Figure 1 illustrates some results obtained by a student. A special emphasis has been put on hands on work by students in the laboratory and during their supervised work for a mini-project. The mini-project has been tailored to encompass a large section of the theory covered in the class. Namely, the project asked students to fulfill the following steps:

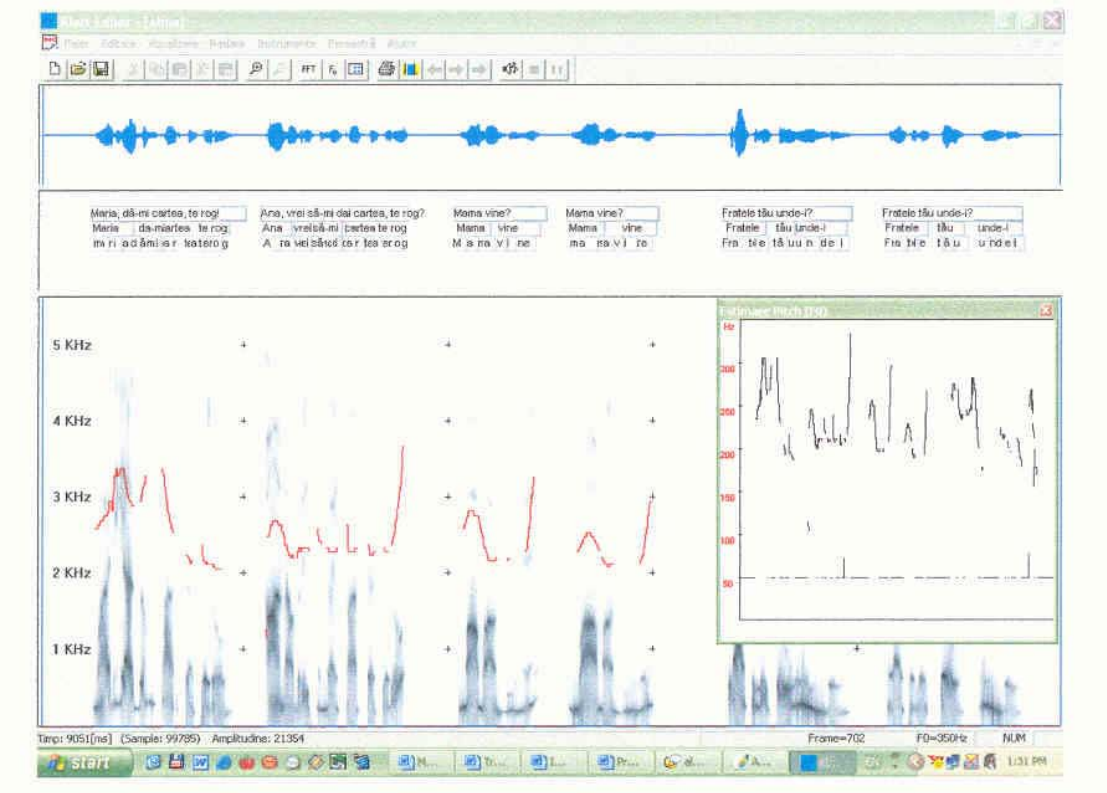


Figure 1: Example of sonogram of a phrase, annotated at the sentence, word and phoneme levels in a student project. The red curve represents the trajectory of the pitch and is used to determine intonation (prosody).

1. make several recordings with their own voice for spelled with various intonations, and under various experimental conditions;
2. perform basic analysis of the uttered vowels, words and phrases: Fourier spectra, sonograms, guessing the formant traces and determining the pitch evolution;
3. draw the "vowel triangle";
4. view the waveforms and determine their basic properties (amplitude, number of zero-crossings, periodicity, period);
5. study the stationarity/unstationarity aspects in the waveform;
6. study the influence the recording conditions have on the waveform and on the spectrograms, chiefly on the amplitude envelope, pitch and formants;
7. segment propositions, words, syllables, and phonemes;
8. annotate the spectrograms for the above;
9. perform a basic statistical analysis of the formant frequencies for vowels;
10. synthesize – using a Klatt voice synthesizer (Jitca et al., 2002) – the vowels and assess their quality.

Moreover, comments on the formants and the corresponding time series have been required for higher grades. The results of their activities had to be saved in speech files annotated with information that indicates words, syllables and phonemes boundaries.

4. Conclusions

The project on XML standards resulted in the development of a handily and very useful interface to exploit the richly XML-based annotated corpora. We intend to apply the methodology of hierarchical organisation of standards, developed as part of the project, and the resulted interface to further class projects in CL and to the development of our own resources, (mainly acquired by master students in CL).

Many students did not interpret the semantic structures project as an easy-to-do one. Most of the groups accomplished the sense annotations of verbs, but fewer realized also the sense annotation of constituents. Only two groups accomplished all tasks on the specification list, realizing the sense generalization programs and assembling the parallel frames.

Although initially we had no great expectations with respect to the outcomes of the corpora acquisition individual tasks, the initiative resulted in the acquisition of an opportunistic corpus of about 82 million Romanian words. This will be integrated into the resources for Romanian language as part of the activity of the Consortium for the Informatization of the Romanian Language (http://consilr.info.uaic.ro/~pic/).

At the time of this writing, we haven't finished yet to analyse the output of the project on Wordnet and sense disambiguation, but the results will be soon available on the Internet. The findings of this investigation will be discussed during the workshop.

The project in speech technology offered students the opportunity to grasp the knowledge and skills needed in building spoken language corpora. Most students have shown a great interest in the project, after they have completed about 25-40% of it. However, the starting work has been painful for some of them, because of their limited knowledge in one or several fields that the project involves (signal processing theory, phonetics, basic human physiology, or statistical analysis and pattern analysis). The best students were encouraged to continue this class project as research projects. One of them successfully contributed, as an assistant research team member, in a research based on a national grant. Others have had small contributions to the analysis of a phonologic thesaurus (created by the Romanian Academy) and have become familiar with the generation of linguistic atlases for the Romanian language. It should be emphasized that such studies are also continued at the doctoral level, and aim to deal with more intricate subjects, such as the nonlinear speech analysis (Rodriguez et al., 2000; Grigoraş, Teodorescu and Apopei, 1998) starting from deeply annotated speech corpora.

Overall, the projects raised the students' interest for corpus-based applications related to NLP. Many of the projects, perceived by students as rather complex, were intended also to challenge students for research activities. For instance the two teams that performed the best the semantic structures for translation, who have reached the expected final level, reported that the project overtook in complexity any of the projects they had to accomplish over the university years, but also helped them to understand more intimately the methodology of a real linguistic research activity. We believe that terminal year undergraduate students and master students, faced to activities of high degree of complexity, could be stimulated to join real research projects at master or doctoral level.

References

Cristea, D. and Butnariu C. (2004). Hierarchical XML representation for heavily annotated corpora, in *Proceedings of the LREC 2004 Workshop on XML-Based Richly Annotated Corpora*, Lisbon, Portugal.

Grigoras, Fl., Teodorescu, H.N., Apopei, V. (1998). Nonlinear Analysis and Synthesis of Speech, in *Studies in Informatics and Control*, vol. 7, no. 1, March.

Fellbaum, C. (1998) WordNet An Electronic Lexical Database. The MIT Press.

Jitca, D., Teodorescu, H.N., Apopei, V., Grigoras, Fl. (2002). Improved Speech Synthesis Using Fuzzy Methods, in *Int. J. Speech Technology* (Kluwer), September, Volume 5, Issue 3.

Rodriguez, W., Teodorescu, H.N., Grigoras, Fl., Kandel, A., Bunke, H. (2000). A Fuzzy Information Space Approach to Speech Signal Non-Linear Analysis, in *International Journal of Intelligent Systems* (Wiley), vol. 15, no. 4, April.

Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufiş, D., Koeva S., Totkov, G., Dutoit, D., Grigoriadou, M. (2002). BALKANET A Multilingual Semantic Network for the Balkan Languages, in *Proceedings of the International Wordnet Conference*, Mysore, India.

Tufiş, D., Cristea, D. (2002). Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet, in *Proceedings of the Workshop on Wordnet Structures and Standardization, and how these affect Wordnet Applications and Evaluation*, workshop

in conjunction with The Third International Conference on Language Resources and Evaluation, LREC-2002 28-31 May, Las Palmas, Spain.

Tufis, D., Barbu, A.M., Ion, R. (2003). A word-alignment system with limited language resources, in *Proceedings of the NAACL 2003 Workshop on Building and Using Parallel Texts; Romanian-English Shared Task*, Edmonton, Canada.

Tufis, D., Barbu, E. (2004). A Methodology and Associated Tools for Building Interlingual Wordnets, in *Proceedings of LREC2004*, Lisbon, Portugal.

Tufiş, D., Ion, R., Barbu, E., Barbu, V. (2004a). Cross-Lingual Validation of Multilingual Wordnets, in *Proceedings of Global Wordnet Conference*, Brno, Czech Republic.

Tufiş, D., Ion, R., Ide, N. (2004b). Word sense disambiguation as a wordnets validation method in Balkanet, in *Proceedings of LREC2004*, Lisbon, Portugal.

Tufis, D., Cristea, D., Stamou, S. (2004 forthcoming). BalkaNet: Aims, Methods, Results and Perspectives. A General Overview, to appear in *Romanian Journal on Science and Technology of Information*, Romanian Academy, Bucharest, Romania.

Villegas, M., Bel N., Lenci, A., Calzolari, N., Cataldo, G., Zampolli, A., Sadurni, T., Soler i Bou, J. (2000). Multilingual Linguistic Resources: From Monolingual Lexicons to Bilingual Interrelated Lexicons, in *Proceedings of the LREC 2000 2nd International Conference on Language Resources & Evaluation*, Athens, Greece.

Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval, in *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, March 5-7, Zurich, Switzerland.

Language Resources In Teaching CL - Between Reuse And Creation -

Cristina Vertan
University of Hamburg, Dept. of Natural Language Processing
cri@nats.informatik.uni-hamburg.de

Abstract:
Language Resources (LR) are a main component of any natural language processing system. Features like linguistic coverage, language coverage, adaptability can influence considerably the quality of the system in which they are embedded. Therefore their design and development is always an elaborated process requiring a lot of time and human resources. An alternative solution is the reuse of existing resources, under the constraints that they fulfill several criteria that we describe below. The present paper aims to offer a kind of checklist, whenever arises the dilemma "creation versus reuse" of linguistic resources. Special attention is paid to LR for "toy" systems, i.e. systems to be developed in the educational process.

1. Introduction

During the lat 10 years a large amount of Language Resources (mono and multilingual lexicons, annotated corpora, speech databases) were developed very often according to the particular requirements of the application, or the specific language(s) involved. Although the development process is extremely time consuming and requires specialized linguistic knowledge many of these resources are not reusable, or their transformation would require the same effort as the designing of a new one. In this context the development of standards appeared as an urgent priority to the scientific community. The problem is of particular importance in Europe due to the multitude of used official languages (and the number will increase in the near future). Several European projects as EAGLES (EAGLES, 1996), ISLE (Calzolari&al., 2001), TELRI (http://www.telri.bham.ac.uk/) had as goal the development of standards for language resources, and as consequence a new question was triggered: when to reuse and when to create new resources.

In this paper we will neither describe existing standards nor discuss their usability for adaptation of old language resources. However considerations about the latter point can be found in (Vertan & von Hahn, 2002).

The present paper concentrates on the use of language resources for teaching purposes and summarizes Pros and Cons to be considered whenever the question "Reuse or creation" arises. It is the result of several years of teaching experience of the author as well as of round discussions at recent events dealing with Computational Linguistics (CL) and Natural Language Processing (NLP) in the educational process.

2. The place of CL and NLP in curricula

Although CL and NLP are used very often as synonyms we will use them here with different meanings:

- Computational Linguistics (CL) is a discipline (like Computational Geometry or Computational Biology) which deals mainly with the study of the language with the help of computers. The computer applications are seen as helpful tools in proofing or exploring a linguistic theory.
- Natural Language Processing (NLP) aims on providing automatic tools (machine translation, text summarization, information extraction) partially with the help of linguistic theories. In the last years, NLP is more and more identified with HLT (Human Language Technology) as it borrowed methods from other fields (mainly mathematics and statistics)

There is no strict delimitation between CL and NLP, on one hand because linguistic phenomena's (e.g. co-occurrences) can be easily with statistical corpus-based approaches, on the other side because many of the CL-tools are used in NLP applications.

Although for research this delimitation is not of crucial importance, in the educational process it describes somehow the target group of students, and implicitly their skills i.e.:

- CL is taught mainly to students in linguistics. They have to understand the general principles of computer programs, and they have to be able to use existent tools and languages in order to model linguistic phenomena. But he will not be asked to design or modify internal processes in the tool (e.g. a student aiming to describe a HPSG grammar for language X with a certain tool will have first to understand how to manipulate the tool and then, how to encode his grammar in the tool's language, but he will not write or modify the processes interpreting this language).
- NLP is mainly a topic for Computer Science students. They are offered very often a linguistic theory or a statistical method, which they have to implement. Usually because of lack of time in acquiring the linguistic knowledge, the students prefer statistical or empirical methods.
- There is even a third group of students who have to be familiarized with CL and NLP : the translators which make more and more use of CAT (Computer aided Translation) Tools.

The most frequent situation is that these three groups of students are instructed separately, so they cannot communicate and exchange knowledge. The only exception are Master studies in CL or NLP where after some preparatory semesters both linguists and computer science students have a common knowledge basis. However, these are quite rare situations and cover a relative reduced number of students.

Both NLP and CL tools, even if they used are only for educational purpose make intensive use of LR's. "Reuse or create" is a frequent question for many instructors. It is to be mentioned that the workshop in which this paper is presented is the first one dealing use of LR's in CL. Until now the topic was only marginally addressed in the context of teaching Machine Translation. However these represent good "lessons to be learned" and will be summarized in the following section.

3. Case studies of using LR's in teaching Machine Translation

It is interesting to look at different approaches of teaching Machine Translation firstly because this is a typical subject to be taught (following different methods) to linguists, computer scientists and translators) and secondly because any MT-System is based on at least one type of LR (lexicons). Recent approach to MT make also use of parallel corpora, three-banks, etc.

Three recent workshops discussed different approaches of teaching Machine Translation (MT):

- Teaching Machine Translation, satellite Workshop at MT summit VIII, September 22, Santiago de Compostella, Spain
- 6th EAMT Workshop "Teaching Machine Translation, November 14-15, 2002, Manchester, UK
- Teaching Translation Technologies and tools, satellite Workshop at MT Summit IX, September 27, 2003, New Orleans, USA

Analyzing the presented papers we can summarize the followings:

1. students in Translation studies simply learn to use MT-Tools, and to asses how helpful they can be. (Forcada, 2003), (Gaspari, 2002). In (Yuste, 2002) a survey on trainees in Translation Studies in Switzerland mentions "Concerning electronic language resources (LR), only those respondents (about 3%) that have taken university modules in computational linguistics or corpus linguistics knew what corpora are. Yet, they conception seemed to limit to reference corpora (e.g. British National Corpus-BNC). There was no or little indication that they could-without extra training – get involved in the use, let alone the creation, of various types of corpora for language work".
2. there is no unified approach for computer science students. (Knight, 2003) and (Way, 2003) argue for teaching empirical MT-approaches (Statistical/example-based), where it is obvious that existent large corpora have to be reused. (von Hahn & Vertan , 2002) and (Vertan & von Hahn, 2003) describe an experiment of teaching rule-based MT and asking the students to produce their "toy" LR's. The main argument is that in this way the students learn additional subjects like XML-like standards, lexicon development, even if these toy LR's allow not very high performances.
3. There is very few material about teaching MT to computational linguists or simple linguists. It seems that MT is simply used to illustrate specific phenomena in different languages. In this sense the contributions in the current workshop will contribute with additional information.

4. "Create or reuse ". Checklist of possible criteria when designing a student project

Due to the variety of curricula in which CL or NLP is included as well as of the type of courses , it is impossible to offer standard receipts on when to reuse existent LR's or when to let the students to create them themselves. On the other hand the instructor should be prepaid to explain to the student's why one or other decision was take. This will help the students further, in their career when they will face partially the same problem. We present here a checklist of possible criteria taking this decision.

1. What is the main goal of the course? Should the students learn to develop LR's or LR's are simply used for another purpose?
2. If the student's will develop their only LR's how much will this affect the performance of the system to be tested? Example: a ad-hoc developed Lexicon will never cover the same vocabulary as one given by LDC (Linguistic Data Consortium)
3. If the students develop their own LR, and the performance of the system decrease, which is the gain; did they learn something additional (for example: do they learn in other courses XML, or is this an opportunity to acquire such knowledge, or what do they learn for aligning corpora)
4. How much time , from the whole duration of the project will be dedicated to development of LR's.? This is also a matter of how many students are involved in the project as well as the composition of the group (if there are mixed linguists and computer scientists the own development of LR's can contribute to a balanced distribution of work)
5. Is there any in-house or free-available LR for the involved language(s) and type of application? In case of a negative answer, the acquisition depends on several other criteria out of the aim of the paper here.
6. If there are already developed LR's available: how much time will cost the students to adapt them to their particular needs? Is there any tool for reading the resource? (This is quite easy with XML encoded Lr's for which parsers are already implemented)
7. How rich is the annotation of such resources? If it is too reach, and goes beyond of what the students need for the concrete project, it can happen that they do not become aware of the importance of some features; they simply use it because it was there, but they do not know what will happen if the respective feature was missing.

8. How exhaustive is the domain coverage. It is for example not a very big use to use a parallel corpora in medicine, for a system in a quite other domain.
9. Do the annotations correspond to the needs of the current project? For example from German-English lexicon, the German part will be only partially reusable in a German-Romanian application, simply because the Romanian morphology impose other differentiations.
10. How many resources exist already for the language(s) in discussion. Especially for small languages any contribution is welcome.

5. Conclusions

In this paper we presented an overview of existent methodologies in introducing LR's in student projects and we discussed their role in the educational process. The checklist in section 4 is far away of being exhaustive but it is a start point for future discussions in this direction. An possible idea will be the collection of adequate LR's for teaching purposes. In this way the instructor will have all the time an overview of the available material, and he can estimate easily which is the best approach for his course.

6. References:

Calzolari, N. & (2001) Lenci, A. & Zampolli, A. & Bel, N. & Villegas, M. & Thurmair, G., The ISLE in the Ocean – standards for Multilingual Lexicons (with an Eye to machine Translation), Proceedings of MT Summit VIII, Santiago de Compostella, 2001

EAGLES (1996) Input to the EAGLES architecture work: survey of MULTILEX", retrieved at http://www.ilc.pi.cnr.it/EAGLES96/lexatch/node4.html

Forcada, M.R. (2003) , A 45-hour computers in Translation course", in Proceedings of the MTSUMMIT IX Workshop on Teaching Translation Technologies and Tools, New Orleans, USA, September, 2003, pg. 11-16

Gaspari, F. (2002), Using free on-line services in MT-Teaching, Proceedings of 6th EAMT Workshop on teaching Machine Translation, Manchester, November 2002, pg 145-153

von .Hahn, W. & Vertan, C. (2002), Architectures of "toy" systems for teaching Machine Translation, Proceedings of 6th EAMT Workshop on teaching Machine Translation, Manchester, November 2002, pg. 69-78

Knight, K. (2003) "Teaching Statistical machine Translation", in Proceedings of the MT SUMMIT IX Workshop on Teaching Translation Technologies and Tools, New Orleans, USA, September, 2003, pg17-19

Vertan, C & v.on Hahn, W. (2002), Towards a generic architecture for lexicon management", Proceedings of the Workshop International Standards of Terminolology and language resources management, LREC'2002 Las Palmas de Gran Canaria, pag. 45-48

Vertan, C. & von Hahn, W. (2003), Specification and evaluation of Machine Translation Toy systems –Criteria for Laboratory Experiments, in Proceedings of the MT SUMMIT IX Workshop on Teaching Translation Technologies and Tools, New Orleans, USA, September, 2003, pg. 44-48

Way, A. (2003), Teaching and Assessing empirical approaches to Machine Translation, in Proceedings of the MT SUMMIT IX Workshop on Teaching Translation Technologies and Tools, New Orleans, USA, September, 2003, pg. 49-55

Yuste, E. (2002), MT and the Swiss language service providers: an analysis and training perspective, Proceedings of 6th EAMT Workshop on teaching Machine Translation, Manchester, November 2002, pg. 23-32

PART II

Language Resources in e-learning

Translation Memory Resources in Multiple Languages to Support E-Learning in Multiple Scenarios

***Dragos Ciobanu** ****Karl-Heinz Freigang** ***Anthony Hartley** ****Uwe Reinke** ***Martin Thomas**

*Centre for Translation Studies
University of Leeds
Leeds LS2 9JT, UK
{smldc, a.hartley, m.thomas} @leeds.ac.uk

**Angewandte Sprachwissenschaft
Universität des Saarlandes
D - 66123 Saarbrücken
kh.freigang@rz.uni-sb.de, u.reinke@rz.uni-saarland.de

Abstract

This paper offers a reflective account of progress to date on the eCoLoRe project, which stands for 'Creating Shareable and Renewable eContent Localisation Resources'. It is funded under the Leonardo da Vinci programme to deliver Translation Memory resources to support a variety of e-learning scenarios. The eCoLoRe is at http://ecolore.leeds.ac.uk.

Introduction

This project aims to remedy the 'severe skills shortage' identified in the EC-sponsored SPICE-PREP II report on eContent localisation (Nicholas & Lockwood. 2000). SPICE was a valuable, motivational complement to LETRAC (1999), which had already defined a model syllabus for integrating a range of technologies into translator training. What was lacking was suitable linguistic resources; eCoLoRe is providing these.

eContent localisation is the translation and cultural adaptation for local markets of digital information that is published on any Internet platform. This information covers business reports, marketing material, research literature, technical documentation, public service information and software applications. To be efficient, eContent localisation relies heavily on specialised computer tools requiring intensive training. One widely-used class of tools is Translation Memories (TMs), which store previous translations for re-use when translating similar texts (see, for example, LISA 2002).

SPICE highlights a severe skills shortage for technical translators, with European specialists being recruited into US organisations. As eContent localisation expands, this shortage poses a serious threat. A recent survey by industry analysts IDC estimates the eContent localisation market worldwide to grow to €11.7 billion by 2005. The fastest-growing sub-sector is centered on linguistic tools for translation and content management, projected at €325.4 million by 2005. The EU and US markets represent some 78% of this total, with an expected EU turnover of €5.5 billion in 2001-2004.

Accordingly, SPICE identifies the building of eContent localisation expertise as a high-impact priority necessitating long-term collaboration with industry. eCoLoRe represents a unique collaboration between industry stakeholders (SAP and the suppliers of the Déjà Vu X TM system), professional associations (ITI - Institute of Translation and Interpreting) and BDÜ - Bundesverband der Dolmetscher und Übersetzer) and university trainers (Saarland and Leeds).

eCoLoRe strategically targets trainers themselves, in academe, professional associations and industry, as well as university students and practising translators. Its products are being made available via the eCoLoRe website, and regular training events will be hosted by the professional associations and university partners.

What trainers lack at present is the resources to create and renew authentic materials for translator training suitable for use with the ICT tools employed by the leading professionals. They also lack a support network for sharing materials and examples of best practice with other trainers. eCoLoRe aims to remedy all three of these gaps, by providing: a repository of shareable materials, a forum for the exchange of ideas, and hands-on use of a representative computer-assisted translation tool. Despite themselves catering for a much larger audience, professional associations and individual universities lack the capacity to develop large-scale training materials for the benefit of their membership and students, especially across a wide range of languages. The eCoLoRe partnership can deliver such resources,

The deliverables of the eCoLoRe project can be divided into two categories: primary materials, which represent raw texts which can be used in localisation exercises. They belong to a wide range of text types and are presented in a multitude of file formats. The other category is that of secondary materials, guidelines covering various topics related to the selection of texts suitable for localisation, the design and production of training materials, as well as an overview of the localisation process and the associated tools.

Description of the primary resources

A key aspect of eCoLoRe is that it builds on current best practice in multilinguality. The primary and secondary materials, and the website in which they are embedded, provide content in at least 14 languages, covering all the official languages of the EU, as well as the languages of some new accession states.

Several criteria were taken into account when designing and producing primary resources. The results of a survey of current trends and needs of translators conducted among ITI and BDÜ members (Höcker 2003) served to guide the production of materials that would answer the needs of professional translators. Figures 1-3 illustrate, respectively, the most frequently translated text types, commonly translated file formats, and the usage of TM technology.

The survey has enabled us to provide trainers and students with samples of the most frequently translated text types in the most popular file formats.

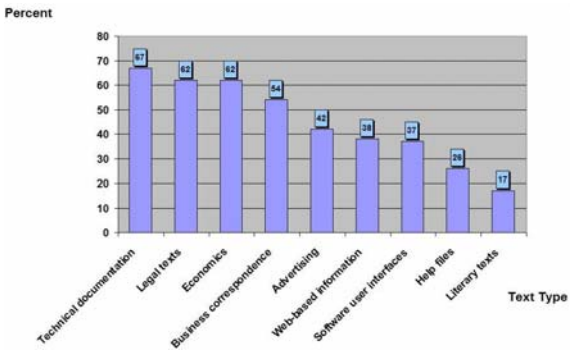


Figure 1: Most frequently translated text types

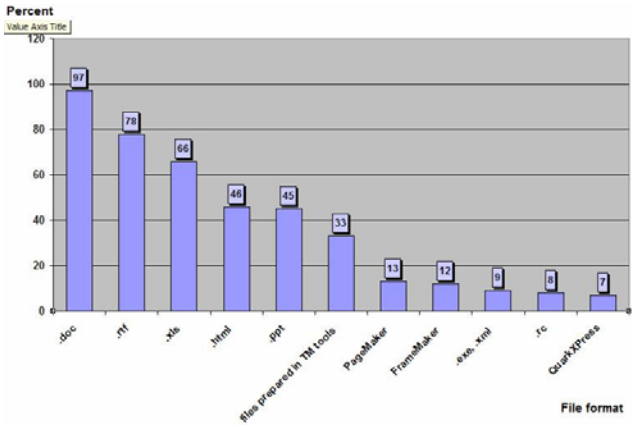


Figure 2: File formats of translation source materials

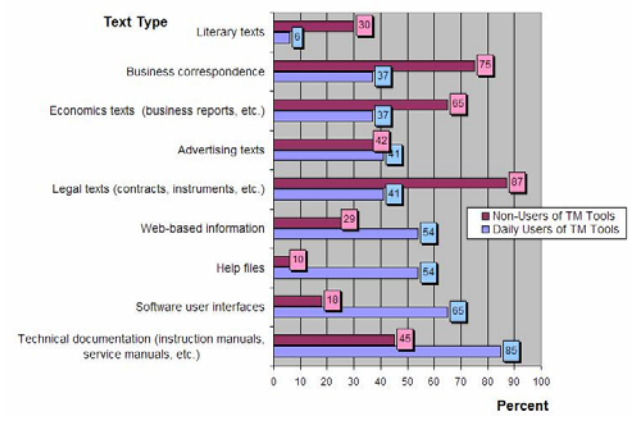


Figure 3: TM use to translate text types (data from the ITI and BDÜ survey)

We also investigated the issue of how suitable these text types are for localisation using TM tools. Special attention was paid to the amount of repetition that occurs both within the text to be translated and across related texts. We refer to *internal* or *intratextual* and *external* or *intertextual* repetition, respectively. Only in cases where this amount is significant is the use of TM tools worth the investment of time and money.

This leads us to introduce the two types of real-life localisation tasks translators are faced with. First of all, there is the 'from scratch' scenario, where there are no reference materials – such as previous translated versions of the current file – available. Secondly, there is the 'reference material' scenario, where the current file is an update of previously translated one(s), for which there may be translation memories available.

The 'from scratch scenario' is currently catered for by sets of .html files obtained with permission from the EU website, and which exist in between 13 and 24 languages, aligned on English.

The remaining primary files are being packaged into translation kits which mirror real-life reference material scenarios. One such translation kit in, for example, Romanian, will include:

– the source file in English;
– its translation into Romanian;
– the resulting translation memory in a non-proprietary TMX format;
– one or more updates of the source file; and
– a catalogue card.

The source file is joined by its translation into Romanian, which has been produced using the TM software Déjà Vu X, which ensures that the layouts of the source and target files are the same.

The resulting translation memory is provided in TMX format, which is a standard that allows users to import the translation units into the TM software of their choice in order to modify and use it in localisation tasks. The big difference in terms of accessibility between TMs in proprietary and TMX formats is obvious when attempting to open them in common text-editing applications such as Notepad or WordPad – see Figure 4.

The related files, essential for update scenarios, ensure that the use of TM applications is worthwhile. It is very hard to find authentic texts whose amount of intratextual repetition would justify the use of TM tools; so much more so when aiming to provide materials suitable for translator training, which need to be authentic, effective and relatively short. Therefore, although some of our primary texts have up to 25% intratextual repetition, the majority of our translation kits include updated versions of the source file, thus raising the degree of intertextual repetition to 68%.

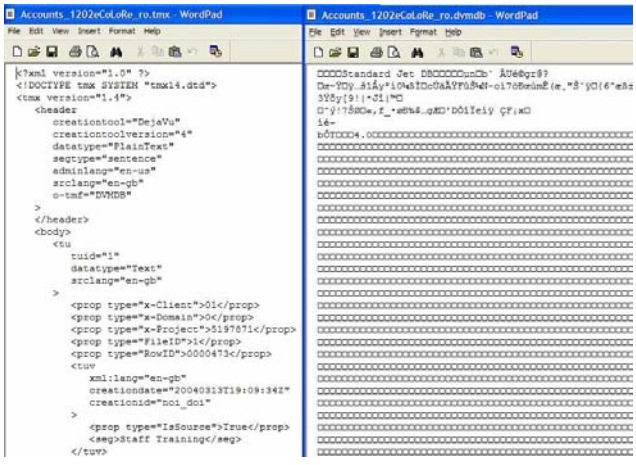


Figure 4: Viewing a translation memory in TMX format vs. one in the proprietary Déjà Vu X format

The catalogue card includes important information in the target language regarding the name and format of the source file, its word count and content, its availability in other languages, the text type and function, the existence of any related files, suggestions for use and relevant additional comments. Users can quickly see if that particular translation kit is suitable for their needs.
In fact, these catalogue cards, apart from being translated and thus featuring in the TMX files provided for the two languages in question – e.g. EN and RO – offer a first glimpse of what the secondary materials delivered by the eCoLoRe project aim to do. Their 'Suggested use' field is designed to assist trainers and students alike, indicating ways in which they be effectively exploited in localisation exercises.
The catalogue cards are not the only secondary materials translated into all the project languages. Most of the website content, including some secondary materials, such as the *Localisation Process and Tools Overview*, has been translated, and the TMs made available as additional primary materials.
The 'from scratch' translation scenarios which involved the project consortium working on the primary materials in order to provide effective translation kits in all the project languages have also been monitored using an online feedback form and the most relevant findings will be made public.

E-learning

E-learning is defined in broad terms. The UK Department for Education and Skills explains that "If someone is learning in a way that uses information and communications technologies (ICTs), they are e-learning". Such broad definitions embrace countless e-learning scenarios, including classroom-based teaching, distance learning, as well as supplementary uses of electronic media. Equally, users of e-learning resources range from structured (if sometimes virtual) classes to autonomous self-teachers. eCoLoRe attempts to cater for the full spectrum of scenarios and users.
Clearly, whoever is involved and whatever the nature of their involvement, e-learning requires materials. In order for these to be useful, a means of distribution must be provided. Moreover, structures enabling users to discuss experiences and contribute knowledge will add value to materials.

The Role of the Website

The eCoLoRe website (http://ecolore.leeds.ac.uk) is the principal means of distributing our materials. Accessibility and usability – in terms of multilinguality and interoperability, and for people with disabilities – are essential.
Clarity and consistency of presentation help visitors to remain oriented and aid navigation. A restricted colour palette and plenty of white space help to ensure clarity. The separation of form and content, inherent in our XML-based publishing framework, helps to maintain consistency of both.
In addition to maintaining consistency and clarity, our design creates lightweight pages, which translates into short download times. The conscious exclusion of animation and other multimedia content also means that visitors do not need to download plug-ins.
eCoLoRe aims to implement good practice in multilingual web design. In order to cater for the diversity of characters, Unicode is used throughout. Client identification is not sufficient to indicate language preference. A language selection menu appears on every page. If a page is not available in the requested language, alternatives are offered, based upon current availability.
Wherever possible, suitable characters are used in punctuation, e.g. quotation marks. The separation of style from content allows such punctuation to be generated automatically according to the needs of the current language. Similarly, lists, such as glossaries, can be sorted alphabetically when generating output, independently of the source-document order. This means that there need be no manual intervention to re-arrange translated content.
It is important that new languages can be added with a minimum of fuss. The site has extensible, modular architecture. Each language version of a given page of content is stored as a separate XML fragment. A wrapper file incorporates all language versions. On the one hand, new translations can easily be dropped in. On the other, we can work on different language versions independently, thus simplifying translation and maintenance.
Moreover, a hyperlink to a page always points to the wrapper file. An XSLT style sheet appends parameters to URLs in the output file and also uses them to extract and deliver content in the requested language. Thus language preference is isolated from markup. Language preference persists from page to page (until changed), without dependence on client software or setup.
In order that techniques used on our site might be reproduced, it is important that they conform to recognised standards, to which others have access, and that financial constraints play no greater part than necessary. The use of standard-compliant code and open source software to store and present the content, helps us realise our intentions.
Compliance with standards also means that our material should be readable by as many different applications on as many different operating systems on as many different devices as possible. We have taken time to make sure that the content of our pages is suitably rendered by a number of web browsers. We also invite visitors to offer feedback.
Our website infrastructure and secondary materials are encoded in DocBook XML. Files are transformed upon request into HTML4.01, but the XML source is also available. This approach helps to maintain compliance and interoperability. Moreover, it allows for the generation of other end-user formats, such as PDF.
While we recognise the value of separating style from structure, only features of the CSS1 (Cascading Style Sheets) standard that are widely supported have been used.
In addition to the techniques and design considerations above, we have paid special attention to the needs of disabled users. Indeed, aside from legal requirements, this approach is likely to improve the accessibility of our materials for everyone.
Firstly, we do not rely on the client to produce dynamic content - all generation is done on the server side.
The simple, tabular layout of our pages is handled gracefully by text-only web browsers, such as Lynx. This contrasts with the use of frames, which also inhibit search engine robots.

Additional links are provided to aid navigation in non-graphical user agents. For example, navigation links, which occupy the left-hand side of a page when laid out in tabular form, will come above the main content when linearized by non-graphical browsers. The inclusion of a "jump to content" link means that visitors do not have to read (or listen to) the same links every time they move from one page to the next. Similarly, alternatives are provided to the JavaScript which otherwise handles language selection. Lack of JavaScript support does not constrain functionality.

We use relative (and therefore scalable) font sizes, thus allowing users to read material in comfort. Graphics are labelled to describe their content for those who cannot view them. Moreover, the use of graphics is limited. Most illustrate certain tools or techniques. These are held in separate "examples" files, outside of the normal flow of the material.

Thus far, we have described how the website is optimised as a vehicle for storing and distributing resources to our users. This is not sufficient to go beyond knowledge transfer and facilitate knowledge exchange.

Once the website has been officially launched, we will set up a public email forum to allow our user community to share experiences.

Finally, the extensible architecture of the site allows for the easy incorporation of new material as it is given by external donors. It is hoped that this will provide a growing archive of primary materials from which e-learners and e-teachers can draw and to which they can contribute.

Description of secondary materials

It is obvious that using the primary resources described in the previous section requires an overall knowledge of the localisation process. On the one hand, it cannot be taken for granted that target groups, i.e., both self-learners and trainers, have already acquired this general knowledge; on the other hand, there is hardly any educational material available conveying this kind of information. The few existing text books mainly focus either on software localisation in particular (Esselink 2000, Schmitz & Wahle 2000) or on the description of tools rather than processes (Bowker 2002, Somers 2003). Moreover, to our knowledge, there is no similar educational material in European languages other than English and German. Furthermore, it is also obvious that a 30-month project like eCoLoRe can only offer a limited amount of original materials in as many relevant file formats as possible. Therefore, another major aim of the project is to enable trainers to create their own course materials based on raw material obtained from the eCoLoRe website and from elsewhere.

To achieve these aims, eCoLoRe will provide different kinds of secondary (or meta-) material along with the primary resources. These meta-materials are modular in design, allowing trainers and learners to adapt resources to suit their needs. The first set of documents, intended for self-learners and trainers alike, aims to provide a beginner's introduction to the tools and processes involved in the localisation of e-content (Figure 5).

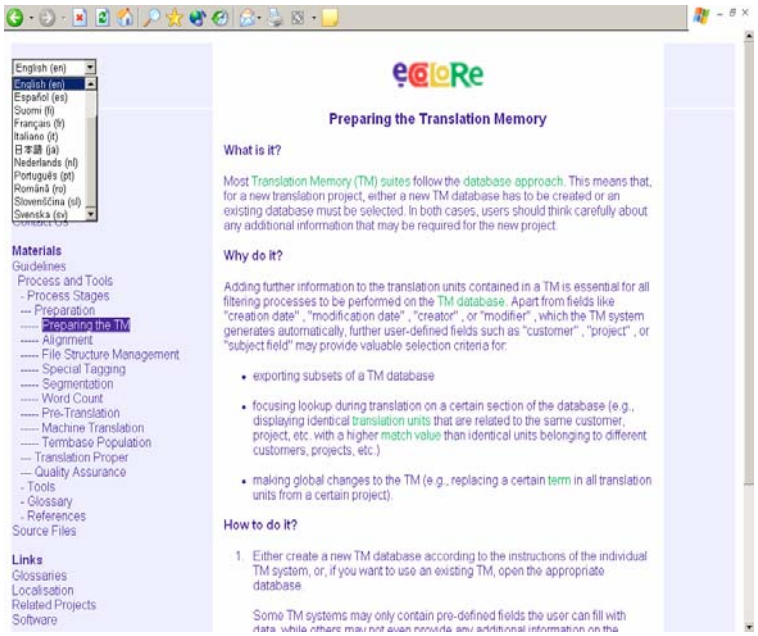


Figure 5: A section from the introductory guidelines on localisation processes and tools

As well as describing in generic terms the major steps in the process, a narrative thread is offered, so that the relationships between the various tools and procedures become clear. The set constitutes a background motivation and source of online reference for e-learners. The resources will be available in 18 languages. The XML source files will also be available in TMX as additional primary resources. The first draft of this set assumed knowledge that the intended users of the eCoLoRe materials may not all possess. At the ITI Annual Conference in September 2003 it became obvious in a panel discussion that the level of awareness of TM technology among professional translators is still surprisingly low. Therefore, materials immediately related to the tools and processes involved in e-content localisation were subsequently rewritten. Style and terminology were simplified and the material was compiled as a hypertext resource. This reflects the intention that the text should be accessible to a heterogeneous audience, i.e., self-learners as well as trainers.

Two other guidelines are intended for trainers rather than for students. They provide meta-information on the selection and production of further source texts and TMs. The guidelines on *Selecting Text Material for Education in eContent Localisation* give an overview of criteria trainers need to consider when selecting text material for educational scenarios in e-content localisation. The focus is on material for translation memory (TM) technology. Based on the results presented in (Höcker 2003) and (Cross et al. 2002), the document briefly discusses the question of domains and text types suited for TM tools and takes a look at the different aspects and degrees of 'textual relatedness' in reference material scenarios. Recurrences and similarities between texts play an important role when assessing the suitability of texts for reference material scenarios. Trainers should therefore be aware of the major intertextual parameters that determine

the degree of variation between related texts. Finally, the document also provides information on the more technical aspects of file formats and suggests some simple procedures for estimating the suitability of selected text material for use with TM tools.

The second document that is mainly intended for trainers is called *Guidelines for Using Reference Material to Create 'Related' Source Texts for Education in eContent Localisation*. The guidelines describe procedures and criteria for creating related source texts (updates, follow-ups) from existing material. They are designed to help trainers adapt source language documents to their special needs in cases where no genuine update material is available. Introducing real-world reference material scenarios into educational environments is fairly difficult, because TMs are specialised resources which are usually built up by companies, organisations and individual translators over the years, i.e., the material is not generally accessible to trainers. It is therefore quite likely that often all that is available are source language documents and their target language equivalents. In these cases, the follow-up texts to be used for translation can be created by systematically modifying the available material. In order to be able to systematically modify a source text for demonstrating the matching mechanisms of TM applications, it is essential to be aware of the difference between those mechanisms and the criteria relevant to similarity judgements of human translators. Drawing upon a chapter from (Reinke 2004) the guidelines therefore start with an outline of a translation-oriented typology of similarity for TM tools. Based on this typology further chapters offer principled guidance for artificially adapting source texts in order to engineer exact and fuzzy matches and so achieve the desired training effects by exercising the core functionality of a TM application.

Evaluation framework

Both primary and secondary materials in eCoLoRe are subject to evaluation and feedback by the project consortium and end users. Evaluation of primary materials at the current stage of the project is carried through by project participants (subcontractors), who are engaged in translating the sample texts which will be included in translation kits made available to the public on the eCoLoRe website. Once the first version of the primary materials has been made available on the website, end users will also be asked to evaluate the usability of the material for their respective purposes. Secondary materials are currently also being evaluated by all project partners (including subcontractors who are translating parts of the material into multiple languages). In their final version, these documents will be made available to the public on the eCoLoRe website, and end users will also be asked for feedback.

Evaluation of primary materials

The evaluation framework for primary materials is presented in a feedback form which, at the moment, is accessible for project partners and subcontractors on the project website. After the official launch of the website the first translation kits will be made available to the public and will then also be accompanied by a feedback form for end users. Evaluators are asked for information on the environment in which the primary material is currently being translated, and on the teaching scenarios in which this material might be used in future. The sample texts to be translated are evaluated on the basis of four criteria, which are considered relevant for e-content localisation using TM systems:

- degree of repetition within texts
- consistency of style
- consistency of terminology
- text length

During the evaluation process, the usability of the text material according to these criteria has to be rated in connection with different teaching/learning environments:

- courses for introduction to TM systems (with main emphasis on learning the major functionalities of the tools)
- general language translation classes
- LSP translation classes
- self-learning situation (learning major functionalities of the tools)
- vocational training courses for professional translators
- other scenarios

The degree of repetition within texts is regarded as a relevant criterion for the selection of texts to be translated in a scenario where no TM material coming from previous versions of the text or from similar texts is available. It uses a quantitative measure of repetition of sentences within the text.

The criterion consistency of style asks whether in the sample texts identical or similar content is expressed in the same or a similar way within the texts and/or between texts coming from the same source and covering the same subject. This criterion is in the first place intended to measure the number of possible fuzzy matches within and between texts, but might also be regarded as a criterion for evaluating the usability of technical texts in general, not only from the point of view of translation.

The usability of texts with respect to the term recognition feature of TM systems is evaluated by the criterion consistency of terminology. Besides measuring the usability of the texts for teaching and learning the terminology recognition feature, this criterion again also considers some general aspects of the quality of technical texts.

In different teaching scenarios the length of the translation texts might be an important criterion for assessing their usability. Evaluators are asked to rate the length of the texts as suitable, too short or too long for the respective scenarios.

Besides these characteristics of the text material, evaluation also concerns the material accompanying the translation kits. Users are asked to comment on the operability and usability of this accompanying material (catalogue cards, translation guidelines) for the task of translating the texts and for teaching purposes.

Evaluation of secondary materials

The feedback forms accompanying the secondary materials for evaluation by project partners and subcontractors are currently geared towards improving the final versions of the materials which will be made available on the website. When the materials are available to the public, they will be accompanied by feedback forms asking for their usability in the different teaching/learning

scenarios outlined above. There will be different evaluation forms for teachers, who want to use the secondary materials in various teaching scenarios, and students or other learners, who want to have an introduction to tools and methods in localisation of e-content.

Evaluation pilot tests

To obtain first reactions from potential end users, a first three-month pilot evaluation of primary materials is being carried out. For this purpose, two translation students from German universities not involved in eCoLoRe have been hired. The two students are not familiar with TM software, but are sufficiently computer-literate and computer-curious. They are typical representatives of the self-learner target group. However, their university background may also enable them to evaluate the didactic suitability of the material in various classroom scenarios.

For the pilot evaluation, 21 English-German 'translation kits' covering various translation scenarios for different file formats have been created from the eCoLoRe primary material. Apart from the translation kits, the two students have been provided with written instructions describing their task, as well as with background material on TM tools and the localisation process. Furthermore, they have been given individual introductions to the Déjà Vu X TM suite. In order to prevent the test from becoming an exercise in translation and to put the focus on evaluation, German translations of the English source texts to be worked on have been added to the translation kits, so that, if required, the actual translation task could be reduced to copying target language units from the German document into the editor of the TM tool.

Apart from their final assessment, the students are asked to deliver monthly interim reports. They have been briefed to distinguish as thoroughly as possible between problems related to the TM software and issues related to the nature and design of the eCoLoRe material.

Conclusions

We have presented the rationale for the development of a significant multilingual resource designed to support the training of translators in the use of translation memory applications. The primary resources consist of source texts representing a range of text types and in a variety of file formats, also delivered in the form of (aligned) translation memories in the generic TMX format. In many cases, they are accompanied by related texts displaying a high degree of repetition with respect to the TMs and thus enabling the enactment of real-world scenarios where updated or follow-up documents are translated. These materials are freely downloadable from the project website.

To further support e-learning activities, the website also offers, in 14 languages, illustrated textual materials explaining the principles and processes of e-content localisation. Other secondary materials aim to guide trainers in selecting and adapting suitable texts in order to create further resources. The materials in preparation and the whole development process are currently being evaluated.

The website will provide a forum for end users and solicit ongoing feedback on the usefulness of the materials. End users are invited to contribute descriptions of how they have incorporated these and other materials into e-learning scenarios, to serve as possible models for others.

References

Bowker, L. 2002. Computer-Aided Translation Technology: A Practical Introduction. Ottawa: University of Ottawa Press.

Cross, G.; McKenna, Sh., Smith, J. 2002. Survey of the UK Translation Market. Salford: Regional Language Network Northwest (in conjunction with the Institute of Translation and Interpreting (ITI)). URL: http://www.rln-northwest.com/shop/publications/survey%20results.pdf.

Esselink, B. 2000. A Practical Guide to Localization. Amsterdam, Philadelphia: John Benjamins.

Höcker, M. 2003. eCoLoRe Translation Memory Survey 2003. Berlin: Bundesverband der Dolmetscher und Übersetzer e.V. (BDÜ). URL: http://ecolore.leeds.ac.uk/downloads/2003.05_bdue_survey_analysis.doc

LETRAC. 1999. Language Engineering for Translator Curricula: Final Report. http://www.hltcentral.org/usr_docs/project-source/letrac/LetracFinal.doc.

LISA 2002: LISA 2002 Translation Memory Survey: Translation Memory and Translation Memory Standards. Geneva: The Localization Industry Standards Association (LISA). URL: http://www.lisa.org/products/survey/2003/tmsurvey.html

Nicholas, L. and R. Lockwood. 2000. eContent Localisation. Final Report. SPICE Prep II: Export potential and linguistic customisation of digital products and services. London, Cambridge: EPS Ltd. and Equipe Consortium Ltd.

Reinke, U. 2004. Translation Memories: Systeme - Konzepte - linguistische Optimierung. Frankfurt am Main: Peter Lang.

Schmitz, K.-D., Wahle K. 2000. Softwarelokalisierung. Stauffenburg: 2000.

Somers, H. (ed.) 2003. Computers and Translation: A translator's guide. Amsterdam, Philadelphia: John Benjamins.

Terminological Grid and Free Text Repositories in Computer-Aided Teaching of Foreign Language Terminology

Galia Angelova, Albena Strupchanska, Ognian Kalaydjiev, Svetla Boytcheva, Irena Vitanova

Bulgarian Academy of Sciences,
Institute for Parallel Processing
25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
{galia,albena,ogi}@lml.bas.bg

Sofia University "St. Kliment Ohridski"
Faculty of Mathematics and Informatics
and Faculty of Economics
svetla@fmi.uni-sofia.bg, itv@gmx.co.uk

Abstract

This paper describes the language resources used in the project Larflast[1] and their role in the Web-based prototype STyLE (Scientific Terminology Learning Environment) which supports adaptive learning of English financial terminology with a target group of adults, non-native speakers with intermediate level of English proficiency. Larflast attempts to improve the language learning process by intelligent integration of advanced natural language technologies (deep semantic analysis of free utterances and personalised information retrieval) into a single coherent system. The learning environment STyLE is implemented as a self-tuition workbench which offers a number of drills testing learner's comprehension of financial terminology and assessing his/her knowledge. User evaluation showed the positive and negative features of our approach in general and STyLE in particular. The conclusion is that language technologies have a long way to go, until they find the proper wrappings for integration of advanced applications and the necessary resources into useful systems.

1. Introduction

Computer-Aided Language Learning (CALL) is a hot area of research but no universal solutions are attained so far regarding the most desired features like learner-system communication in Natural Language (NL) and adequate processing of learner's language errors. The market applications of CALL systems are often perceived by learners and teachers as dumb and inflexible, which is demotivating for the learner and restricts the independent use of CALL systems considerably (Murphy & McTear, 1997). However, supporting free NL input requires integration of sophisticated techniques for semantic analysis, esp. parsing and checking the semantic correctness of the learners' answers. A number of prototypes try to support (almost free) NL input but "so few of these systems have passed the concept demonstration phase" (Holland & Kaplan & Sams, 1995). The early prototypes in the classical collection (Holland & Kaplan & Sams, 1995) contain mostly modules for checking students' competence in vocabulary, morphology, and correct syntax usage (parsers). The most sophisticated semantic analysis is embedded in BRIDGE/MILT which *matches* the learner's utterance (a lexical conceptual structure) against the prestored expected lexical conceptual structures in a dialog based on question-answering scenario (Dorr et all, 1995). The authors point out that the syntactic and semantic correctness of the student utterances have to be checked as well as the appropriateness of the answer at the given dialog point (therefore matching to expectations is a good solution). More recent systems like CASTLE in (RECALL 1997) and SLALOM (McCoy et all. 1996) still focus on spelling, morphological, and syntactic errors. Another example is CIRCSIM-Tutor (Glass, 2000) which expects quite short answers, permissively extracts whatever is needed and ignores the rest. Recent systems rely on (spoken) dialog, partial and/or incremental analysis, and combine rule-based and data-driven approaches (see e.g. (VanLehn et all 2002) and (Rose et all, 2002)) without much progress in checking the correctness and the appropriateness of the learners' utterances. To conclude, the present CALL solutions especially for semantic analysis are far from being perfect.

This paper presents the language resources in STyLE where, most generally, semantic analysis is systematically approached and personalised Information Retrieval (IR) is dynamically tuned to the content of the learner model. STyLE integrates formal semantic techniques for maintaining student input as free text. Up to our knowledge, STyLE is the only system that attempts proving the appropriateness of the learner utterance in real time, based on predefined minimal and maximal expected answer. Focusing on domain knowledge we invested much effort in the acquisition on the conceptual resources which were encoded as conceptual graphs (Sowa, 1984). STyLE is a coherent environment where the student accomplishes three basic tasks:

(i) reading teaching materials,
(ii) performing test drills and
(iii) discussing her own learner model with the system.

An initial user study (Vitanova, 1999) investigated how erroneous answers appear in terminology learning. Errors are usually caused by the following reasons:

- **Language errors** (spelling, morphology, syntax);
- **Question misunderstanding** which causes wrong answer;
- **Correct question understanding**, but **absent knowledge of the correct term**, which implies usage of paraphrases and generalisation instead of the expected answer;

[1] *Learning Foreign Language Scientific Terminology*, a Copernicus'98 Joint Research Project, funded by the European Commission in 1998-2001, with partners CBLU Leeds, UK; UMIST Manchester, UK; LIRM Montpellier, France; Academy of Sciences, Romania; Simferopol University, Ukrain; Sofia University and Virtech Ltd., Bulgaria.

- **Correct question understanding**, but **absent domain knowledge**, which implies specialisation, partially correct answers, incomplete answers and wrong answers.

This classification influenced considerably the design of the knowledge-based tutoring environment STyLE which assists non-native English speakers in English terminology learning. More details about STyLE components, functionality, architecture, and implementation are given in (Angelova et all, 2002).

The paper describes only the resources and technologies developed by the Bulgarian team in Larflast and is structured as follows. Sections 2, 3 and 4 discuss the three kinds of language resources in STyLE and give hints about their role in the learning process, their volume and the relevant technologies using the resources. Section 5 presents the evaluation results and the conclusion.

2. Terminology as a Conceptual Resource

The learning environment STyLE contains terminology organised as a conceptual hierarchy linked to the lexicon. We consider the distinction between the conceptual and the lexical resources as very important, since it imposes differences in the internal representation, the techniques providing the internal processing and the role of the two resources in the learning process. There are two important requirements imposed on the conceptual representation: firstly, it should be clear and intuitive enough to be shown to the learner with pedagogical purposes and should allow for simple graphical visualisation and secondly, it should be sofisticated enough to serve as an input to the natural language understanding component, providing the semantic analysis of the learner's answers. Acquiring the domain knowledge in this project was an effort-consuming manual activity which required proper goal-oriented combination of middle and upper models from well-known knowledge resources like CyC, WordNet, MikroKosmos, Sensus etc. We show that in a practically situated task-dependent paradigm, most ontological choices like granularity of concept types, choice of conceptual relations, engineering of the explicit and implicit hierarchy, etc. are influenced by the task requirements.

Ontological Choices for Acquisition of the Type Hierarchy

Looking for more universal principles and solutions, knowledge acquisition aims at the elaboration of a knowledge base fitting to the specific project goals. We consider the choices described below as task-dependent because there might be other ways to model the same domain. Acquiring the domain model, we try to answer questions like: *which concepts, relations and facts are important for the STyLE learner*? as well as *how should knowledge be encoded in order to better satisfy the specific project requirements?*

One of the reasons to support explicitly a type hierarchy is that some fragments of the domain knowledge are shown to the learner (visualised as domain facts) when student's misconceptions are detected. This means that the student observes almost directly the internal structure of the knowledge base. Because of this project-specific aspect, we partition the types in the ontology according to the features which seem to imply the most important characterstics and differentiation to be communicated to the learner (a foreigner who studies English financial terms). So we omit types that are considered insignificant to the student. Let us consider Fig. 1 which presents a fragment of the type hierarchy for SECURITY. Another possible classification for SECURITies can be built with respect to the issuing authority. But we consider the distinction BOND-STOCK as the central one to be taught to our learners and therefore ISSUING_AUTHORITY is connected to SECURITY as a feature of the concept.

We choose label-terms whenever possible. Most financial terms are noun phrases (NPs) containing more than one word. All concept types in Fig. 1 are real terms in financial dictionaries, which are to be considered in the terminology learning course (but there are also some labels, such as PRODUCT_OF_FINANCIAL_MARKET, that are not real-life terms). It might be misleading to arbitrarily synthesize "dummy labels" for providing a more ordered ontology, because the visualisation to the learner might give rise of wrong impressions and misconceptions about external collocations of financial terms. So, we prefer to synthesize somewhat explanatory dummy labels (phrases like ISSUED_BY_A_COMPANY instead of COMPANY_SECURITY). To summarise, in the hierarchy we place either label terms, or explanatory dummy labels.

Encoding Different Kinds of Partitions in One Hierarchy

There are many ways to partition a domain, at least because of the different goals and the numerous possible view-points that might exist. The compact hierarchy in Fig. 1 encodes several kinds of partitions in one lattice, by assigning one *isa_kind* clause per partition. We use a predicate *isa_kind*/4 (see examples in Fig. 1):

isa_kind(PartitionedType, [Subtype(s)],
[PartitionKind(s)], 'PartitionName').

The fourth argument of *isa_kind* is a text string to be shown when displaying the "legend" of the partition color to the learner. A visualisation fragment is given in Fig. 2 which uses the interface of the knowledge acquisition tool in Larflast (Dobrev&Toutanova, 2000) Focusing on a single concept, graphical representation of different classification perspectives with different colors is shown as a simple and natural way for system-learner communication.

In the LARFLAST project, we considered the ontological perspectives of **natural** and **role** partitions. Natural subconcepts are classified according to unchangeable features while roles are distinguished according to temporary features. For instance, in the world of finances, one DEALER can be a BULL and/or a BEAR for different clients, so the classification *DEALER is BULL and/or BEAR* is a role partition. We mix all partitions into one hierarchy, as shown in Fig.1, and distinguish them only by the corresponding *isa_kind* predicate. In a similar way, we mix the disjoint/joint partitions and the exhaustive/non-exhaustive partitions in the same hierarchy. The default partition in Fig. 1 is a joint and unexhaustive classification into natural types.

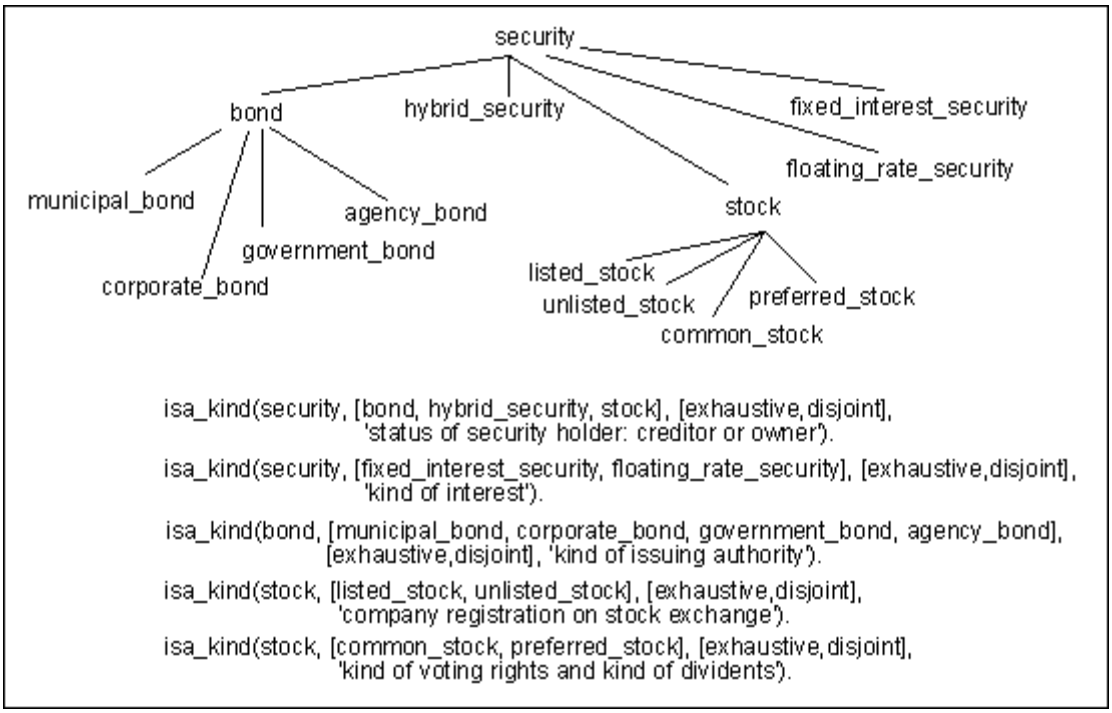


Figure 1. Ontology of terminological units and *isa_kind* perspectives.

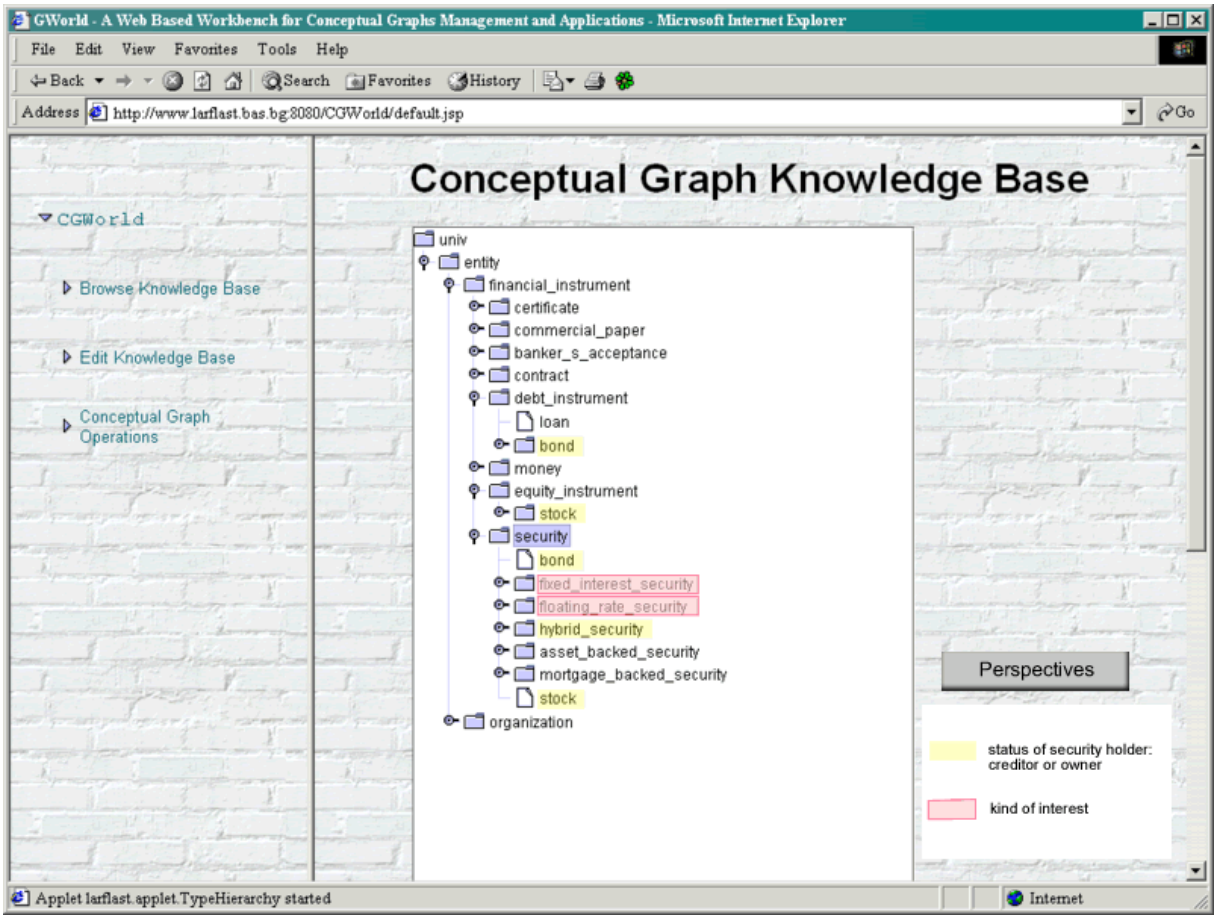


Figure 2. Visualisation of the ontology. Perspectives are marked by colors. Clicking on a type shows its partitions.

Our present knowledge acquisition experience shows that there are no simple choices in the extremely complex domain of finances. One needs mixed perspectives in the internal complicated hierarchy. However, the learner is shown a simplified and partial view to a small sub-hierarchy, which is relevant to the subject discussed at the current pedagogical situation.

Granularity of concepts and conceptual relations

STyLE integrates PARASITE as a drill-checking machine which means that the NL semantics is treated compositionally, word by word, with basic granularity of meanings as defined by word senses. Thus we need a formal technique for shifting the conceptual granularity, to assure that the domain semantics of "complex" objects is translated correctly to "one-word" meaning postulates. The shifting technique we use is implemented by ontological operations like type expansion and type contraction (Sowa, 1984). For instance, the type relation ISSUED_BY is defined as follows (the next proposition is called type definition):

relation ISSUED_BY(x,y) **is**
[ISSUE] -> (AGNT) -> [ISSUING_AUTHORITY: **y**]
-> (THEME) -> [SECURITY: **x**]

In this way we obtain facts with suitable "cascade" granularity: one encoding to be shown to the learner, for instance when visualise ISSUED_BY in the fact

[BOND]-> (ISSUED_BY)-> [COMPANY]

and another encoding with the corresponding word-by-word granularity, provided by type expansion (which corresponds to the granularity of the meaning postulates. The careful ontological elaboration at such a depth required much time-consuming efforts of knowledge engineers, language teaching experts and domain experts.

Role of the terminological grid

STyLE prototype uses about 300 English terms in financial markets, organised in a lattice built according to the different perspectives of the *is-a* relation (hyponyms, hyperonyms). The terminological hierarchy is the resource supporting the adaptive navigation through the pedagogical material, while the pedagogical agent plans learner's moves by suggestions for *performing drills* or *further readings* in case of wrong and incomplete answers and necessity to acquire more knowledge on a certain topic. *Terms* as words are the topics to be thought and they are searched in the repository of relevant texts from Internet, which become *suggested readings* if the filtering procedures decide so. *Terms* as conceptual labels participate as well in the Learner Model, where indications about the student performances to all drills are kept (after performing drills, indications like *know*, *not_know*, and *know_wrongly* are stored for every user and every term, for details see (Angelova et all., 2002)).

3. Resources encoding lexical semantics

STyLE integrates the system Parasite (developed in UMIST by Allan Ramsay, see (Ramsay&Seville, 2000)). In this section we discuss briefly the lexicon entries, which describe the terms as words, and the meaning postulates which define the lexical semantics of the terms (as we said, we distinguish between the conceptual and lexical semantics). Other linguistic resources like English grammar rules are integrated in Parasite as well, to provide the syntax and semantic analysis, but we do not consider them here.

The lexicon entries are defined in Prolog clauses and terms are entered there as nouns, verbs and adjectives to describe their morphological features. In fact, as STyLE is implemented in Prolog, the coincidences of the labels provide the links between the concepts in the type hierarchy and the words in the lexicon entries. We are interested here in the meaning postulates which are encoded manually and define the lexical semantics of the general lexica and the terms in the closed world of the project. Let us consider first several simple examples of meaning postulates for common words expressing facts that *apples are fruits* (but also have the role of *food*) and *all birds fly*:

```
lexicalMP(forall(X, apple(X) => (X is fruit) & soft(X))  ).

lexicalMP(criterial(lambda(X, apple(X)), lambda(Y,food(Y)))  ).

lexicalMP(forall(X :: {bird(X)}, fly(X))  ).
```

Further examples of meaning postulates show the way we define the semantics of financial terms, for instance

```
lexicalMP(forall(P1:: {bank(P1)}, (P1 is institutions))  ).

lexicalMP(
forall(X::{budget(X)},
    plan(X) & financial(X) &
    exists(Y::{summarize(Y)},
      exists(I::{income(I)}, theta(Y,$object,I) &
        exists(E::{expenditure(E)}, theta(Y,$object,E) &
            exists(T::{period(T)}, theta(Y, $over, T))))))   ).

lexicalMP(
forall(X::{capacity(X)},
    maximum(X) &
    exists(Y::{produce(Y)},
     exists(Z::{firm(Z)}, theta(Y,$agent,Z) &
       forall(U::{unit(U)}, theta(Y,$object,U) & count(U,X)))))  ).

lexicalMP(
forall(P1:: {company(P1)}, (P1 is institutions))  ).

lexicalMP(
forall(X::{expenditure(X)},
    money(X) &
    exists(Y::{spend(Y)}, theta(Y,$object,X)))  ).

lexicalMP(
forall(X :: {export(X)},
   (good(X) or service(X))
   & exists(Y :: {sell(Y)}, theta(Y,object,X)
       & exists(Z :: {theta(Y, agent,Z)},
           exists(ZC :: {country(ZC)}, location(Z, ZC)
             &  exists(T::{buyer(T)}, theta(Y,$to,T)
                 &  exists(TC::{country(TC)},
                     location(T, TC)
                     & not(ZC = TC)
                     & forall(D :: {to(X, D) & country(D)},
                       TC = D)))))))  ).
```

Please note that the meaning postulates impose a hidden hierarchy of lexical meanings which however is different from the conceptual hierarchy as the latter reflects all perspectives interesting for the learners and assigns labels to these perspectives, to provide multiple inheritance along the different lattice branches. No doubt the two lattices – the ontological and the lexical one - are similar at an abstract level, in the sense that they contain the same information, as the conceptual partition features can be alternatively encoded as attributes of the words in the lexical hierarchy. In our case, however, we included more knowledge in the ontology as we preferred to use the visualisation utilities elaborated especially for the project, and thus to show to the learner more information in graphical format.

STyLE contains about 150 logical expressions which are either distributed with the Parasite system or developed in Larflast. They describe the semantics of words expected in the utterances, which answer to especially designed drills where the student is allowed to write down free text. Every free text input is first processed by the system Parasite which checks the syntax and the semantic correctness of the learner's free text input. After a logical form is produced – which happens for linguistically correct utterances only – an additional prover called STyLE-Parasite checks whether the logical form of the answer is "between" the logical forms of the predefined minimal and maximal expected answers for the current drill (Angelova et all, 2002). The comprehensive diagnostics allows to recognise cases like answer generalisation, answer specialisation, paraphrases using the concept definition, incomplete answer, partically correct and wrong answer. This sophisticated tool makes STyLE a very powerful environment (from formal linguistic perspective), which goes very deeply into the semantic processing compared to other systems. Fig. 3 illustrates the diagnostics options.

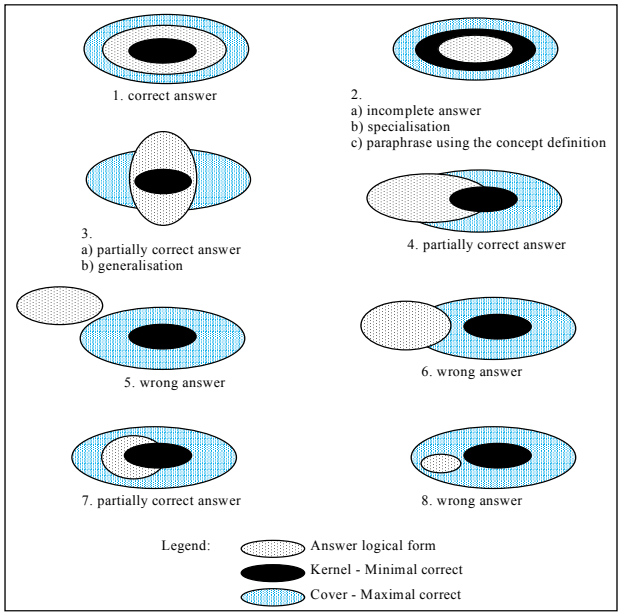


Figure 3. Diagnostics of conceptual errors

4. Collection of free texts

Trying to dynamically retrieve documents from widely-known financial sites, STyLE uses advanced filtering to determine the most relevant documents to be recommended as "suggested readings" in a particular learning situation. In this way STyLE can enlarge and continuously update its text archive. Agents - Web spiders search and deliver texts that correspond to some list of given keywords (the terminology covered by the prototype STyLE). These documents are stored on system's servers and periodically updated by newer documents with higher relevance scores. The filtering process is off-line performed by an original implementation of Latent Semantic Analysis (LSA) (Landauer et all, 1998; Deerwester et all, 1990)). It analyses all texts collected from the Web and generates a relevance measure for each text with respect to each of the terms in question. Only the documents whose proximity is higher than some threshold are kept and the others are discarded. Complex terms (consisting of more than one word) are placed as one term in the LSA matrix. Only the top most relevant documents for each of the domain terms are kept. An annotation table supports fast access to the STyLE archive, containing the key terms together with a list of their best corresponding relevant documents. Practically we work only with terms tested in exercises, because only these terms can appear as **unknown** or **known_wrongly** in the learning module and therefore only for them relevant readings are suggested.

STyLE text archive contains 800 most relevant readings, which are html-pages containing mostly text (we excluded tables and other information that signals prevailing technical content). These texts are offered as suggested readings but are also used for building dynamic concordances which show samples of terms usages to the learner. The latter samples may be displayed in cases of language errors to drills where the student makes linguistic mistakes. Choosing this option (*view samples*) is up to the student. The dynamism of the text collection ensures the appearance of new samples, which makes the browsing interesting at every run.

5. Evaluation and Conclusion

Technically, from a learner's perspective, STyLE is a set of Web-pages containing exercises and readings. The embedded applications and components like Parasite, STyLE-parasite, generation of xml-pages, LSA, web-spiders run behind the scene or off-line. STyLE was tested by *(i)* two groups of university students in finance with intermediate knowledge of English, *(ii)* their university lecturers in English, and *(iii)* a group of students in English philology. STyLE was evaluated as a CALL-tool for self-tuition and other autonomous class-room activities, i.e as an integral part of a course in "English for Special Purposes". The learners could test their knowledge through the specially designed exercises, compare their answers with the correct ones using the generated feedback (immediate, concrete and time-saving, it comes in summary form which is crucial in order to accomplish the use of STyLE autonomously) and extract additional information from the suggested readings and concordancers.

Users liked the feedback after performing drills, immediately after they prompted erroneous answers to exercises where this term appears. All of them evaluated positively the visualisation of the hierarchy as well as the surrounding context of texts and terms usages organised in a concordancer which is dynamically built and centered on the terms discussed at the particular learning situation. The teachers were very pleased to have concordancers with contiguously updated term usages; they would gladly see such a language resource integrated in a further authoring tool, because searching suitable texts in Internet is a difficult and time-consuming task. Indirectly, these positive reactions show that the idea to keep separately the conceptual representation is a fruitful one, as it allows for easy visualisation as well as for a terminology-centred design of the dialog, the navigation choices, the suggestion of further moves and so on.

We concentrated especially on the evaluation of the free NL input, which attempted to provide complete NL diagnostics and is the most serious in CALL at present (up to our knowledge). Unfortunately the learners were not very enthusiastic regarding these modules, as they permit relatively restricted simply input and do not go beyond the human capacity of the teacher. The learners were not impressed that for instance the sentence "*primary market operates with newly issued securities and provides new investments*" is correct since it is between the minimal answer "*primary market operates with newly issued securities*" and the maximal answer "*primary market operates with newly issued securities and provides new investments and its goal is to raise capitals*". The main disappointment of learners and teachers is that STyLE cannot answer *why*, i.e. Parasite and STyLE-Parasite provide extremely comprehensive diagnostic about the error type but not about the error reason. Fortunately, all users liked the fact that there were numerous examples of terms usages in real texts whenever morphological or syntax errors were encountered in the free NL input. So we conclude with certain pessimism concerning the appropriateness of formal semantic approaches in CALL today and much optimism that data-driven corpus techniques, if properly applied, fit quite well to the adaptive CALL. A possible improvement of the current paradigm for formal analysis is to switch to partial semantic analysis, which – at the level of the interface - will give more flexibility to the students to enter phrases instead of full sentences. What is still desirable regarding the filtering module is to restrict the genre of the suggested readings since the current texts are freely collected from the Internet and some of them should be used as teaching materials (LSA cannot recognise the text educational appropriateness since it considers the terms occurrences only; other supervised IR techniques like text categorisation might improve the filtering if they are properly integrated).

The conclusion is that teachers as well as learners like CALL systems that are easy to integrate in the typical educational tasks, i.e. the area of language learning has well-established traditions and the experimental software is well-accepted only if it is really useful and facilitates the learning process. Our feeling is that all attempts to integrate language technologies in CALL should be closely related to testing the laboratory software with real students. At the same time cooperation with teachers is an obligatory condition as the necessary pedagogical background is often missing in the research environments where normally the NLP applications and language resources appear. Language technologies have a long way to go, until they find the proper wrappings for integration of advanced applications and the necessary resources into useful CALL systems.

References

Murphy M. and M. McTear (1997). Learner Modelling for Intelligent CALL. In: Proc. User Modelling 97, Springer.

Holland, V., M. Kaplan, J. and M. Sams (eds.) (1995). Intelligent Language Tutors: Theory Shaping Technology. Lawrence Erlbaum Ass., UK.

Dorr, B., Hendler, J., Blanksteen, S., and Migdaloff, B. (1995) On Beyond Syntax: Use of Lexical Conceptual Structure for Intelligent Tutoring. In (Holland & Kaplan & Sams, 1995), pp. 289-309.

RECALL (1997), a Telematics Language Engineering project, http://iserve1.infj.ulst.ac.uk/~recall.

McCoy, Pennington, Suri (1996). English error correction: A syntactic user model based on principled mal-rule scoring. Proc. User Modelling Conf.-96, pp. 59-66.

Glass, M. (2000) Processing Language Input in the CIRCSIM-Tutor Intelligent System. AAAI 2000 Fall Symp. on Building Dialogue Systems for Tutorial Applications. http://www.csam.iit.edu/~circsim

VanLehn, K., Jordan, P., Rose, C., Bhembe, D., Bottner, M., Gaydos, A., Makatchev, M., Pappuswamy, U., Ringenberg, M., Roque, A., Siler, S. and R. Srivastava. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. Proc. Intelligent Tutoring Systems Conf. 2002, LNCS vol. 2363, Springer, pp. 158-167.

Rose, C.P., Bhembe, D., Roque, A., Siler, S., Srivastava, R. and K. van Lehn. A hybrid language understanding approache for robust selectioon of tutoring goals. In Proc. of the 6th Inter. Conference on Intelligent Tutoring Systems, Spain, June 2002.

Sowa J. (1984) Conceptual Structures: Information Processing in Mind an Machine. Addison-Wesley, Reading, MA.

Vitanova I. (1999) Learning Foreign Language Terminology: the User Perspective. Larflast report 8.1, Sofia.

Angelova G., Boytcheva, S., Kalaydjiev, O., Trausan-Matu, S., Nakov, P. and A. Strupchanska (2002). Adaptivity in Web-Based CALL. Proc. ECAI-2002, IOS Press, pp.445-449.

Dobrev, P. and K. Toutanova. (2000) CGWorld - a Web-based workbench for conceptual graphs management and applications. Proc. ICCS-2002, Shaker Verlag, Aachen 2000, ISSN 0945-0807, pp. 243-257.

Ramsay, A. and H. Seville. (2000) What did he mean by that? Proc. Int. Conf. AIMSA-2000, Lecture Notes in Artificial Intelligence Vol. 1904, Springer, pp. 199-209.

Landauer T., Foltz P., Laham D. (1998) Introduction to Latent Semantic Analysis. Discourse Processes, vol. 25, pp. 259-284.

Deerwester, S., Dumais, S., Furnas, Landauer, Harshman R. (1990) Indexing by Latent Semantic Analysis. J. Am. Society of Information Sciences, vol. 41, 391-447.

Integrating Resources to Realize a Self-Contained Environment for Lexicon Learning

Sandro Pedrazzini, Alessandro Trivilini, Judith Knapp
SUPSI, University of Applied Sciences of Southern Switzerland, Manno, Switzerland
EURAC, European Academy of Bolzano, Bolzano, Italy
sandro.pedrazzini@supsi.ch, alessandro.trivilini@supsi.ch, judith.knapp@eurac.edu

Abstract

This paper describes the steps and the criteria used to set up a new bilingual environment for lexical learning, using available dictionary and morphological information.
A bilingual available dictionary (German-Italian) has been taken as starting point, integrating, step by step, morphological information such as inflection paradigms, derivation families, segmentation, surface variants, etc.
An important issue is that the added information has not been specified from scratch, rather it has been reused from modules derived from other projects. We will explain how this has been achieved, the criteria used to choose the right components, and how the latter will be used to automatically generate exercises on lexicon knowledge.

1. Introduction

When facing with the task of setting up a new dedicated environment for lexicon learning, the first problem you encounter is how to make best use of the many resources available in the field.
We started from an existing version of a bilingual electronic dictionary (Eldit, German-Italian), which represented a successful project carried on at EURAC, the European Academy of Bolzano, Italy, and wanted to extend it with different sorts of morphological information (inflection, wordformation, spelling, etc.) and with a tool for generating exercises.

Eldit is a pedagogical online dictionary module realized as a union of two learners' dictionaries, for German and Italian. Each dictionary contains approximately 3000 word entries as lemma. Each word entry is represented by a huge amount of information, which on one hand helps the learner to comprehend the correct meaning of a word and on the other hand helps to use the word correctly (Abel & Weber, 2000). The two dictionaries have been combined by linking included translation equivalents, which serve as entry points to the corresponding part of the other dictionary. All information has been carefully selected and prepared by linguists and language teachers according to modern psycholinguistic criteria.

Fig. 1: Eldit view of the information regarding the Italian word "pane"

A dictionary entry is presented to the user in two frames. The left-hand frame shows the lemma of the word and a list of different word meanings, each of which is described by a definition, an example sentence, and an optional translation equivalent.
The right-hand frame is organized in several tabs and shows additional information such as word relations, collocations, idiomatic expressions, a set of linguistic difficulties, a picture, the possibility to save personal annotations, verb valency, etc.
The new information on inflection and word formation has also been integrated as tab on the right-hand frame.
The linguistic difficulties can also be accessed directly at the place where they occur. They are indicated by a kind of footnote number and shown in a small window.
The system has been implemented during the last three years. For the implementation Java Servlet technology has been used in combination with XML (Gamper & Knapp, 2002).
We will first describe the system and then explain how we could extend it using available resources.

2. Available Information

We will 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.

2.1. Lemma and Morphology

The lemma (called "base form" or "citation form" as well) is the core description of a word, e.g. the Italian word "pane" on the left-hand side in fig. 1. In cases where there is both a male and a female version, such as for the German "Einwohner/Einwohnerin" (inhabitant), both forms are stated. Below the lemma some basic 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 itself is a compound word (see "Ausverkauf" (sale) and "aeroporto" (airport)). More morphological information will be added with the external resources.

2.2. Meaning Description

Each lemma may have several meanings, for instance "a *house* to live" is not the same as "the royal *house* of Scotland". We follow a new approach in dictionary design which we call "crosslingual": each word meaning is explained by a definition (a typical element of monolingual dictionaries) and one or more translation equivalents (a typical element of bilingual dictionaries) which are linked to the corresponding lemma in the other dictionary module (this generalizes the existing concept of semibilingual dictionaries and leads to the term "crosslingual dictionary") [Abel and Weber, 2002]. In this way by clicking e.g. on the German word "Brot" in fig. 1 the user can access directly the corresponding dictionary entry for "Brot" and all the information collected for this word.
A lexicographic example illustrates the meaning in context. The lexicographic examples consist of several short sentences. They have been created manually according to specific criteria in order to illustrate the word. Frequency analyses on text corpora have been carried out to inspire the creation process. 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 grey triangle on the left-hand side of each definition allows the activation of one particular meaning. This gives access to the information provided in the right-hand frame, which is often different for different word meanings.

2.3. Semantic Fields

Semantically related words, so-called "semantic fields" or "word fields", are shown within the first tab on the right-hand side, namely the tab "Verwandte Wörter/campo semantico". They are illustrated in two-dimensional graphs in ELDIT. We are currently elaborating about 250 word fields for each language (see fig. 2). 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 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 (Abel et al., 2004).

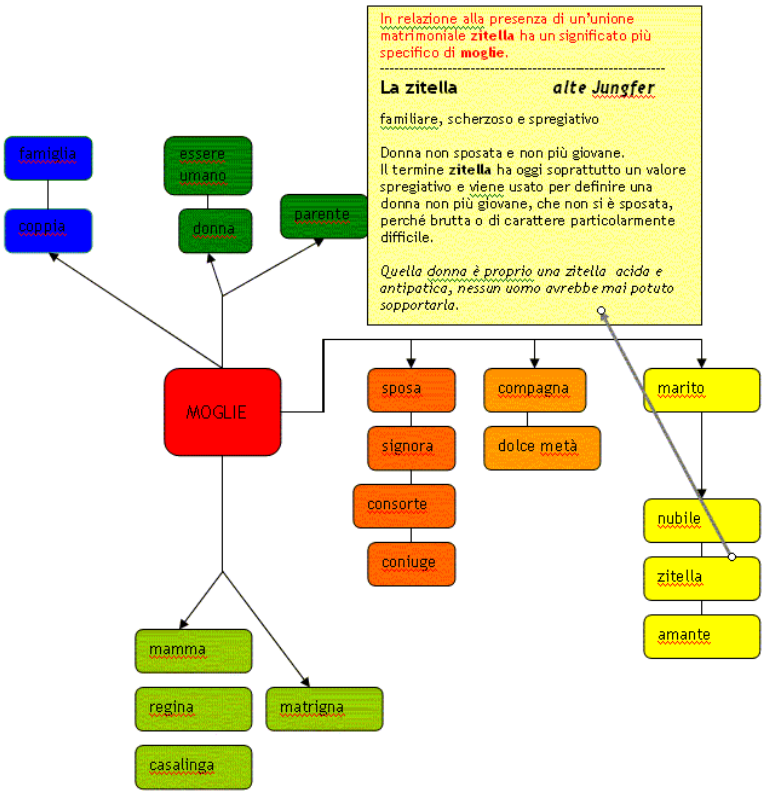


Fig. 2: The semantic field of the Italian word “moglie” (woman)

2.4. Collocations and Idiomatic Expressions

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 (Abel and Weber, 2000):

Within the tab “Verwendung/combinazioni” word combinations are listed for the learner. Each word combination or collocation is treated as a separate item supplied with a translation and an example sentence (see fig.3 1). Also information about the stylistic level of an expression (colloquial, literary, etc.) may be included. Furthermore, a number of adjectives which are typically used together with the lemma under consideration are listed together with respective translations and a few examples.

Within the tab “Redewendungen/fraseologia” different 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.

2.5. Verb Valency

For verbs we also show information on verb valency within the tab “Verwendung/costruzioni” (usage). This information is very important for the learners, since it indicates how a sentence is constructed in a grammatically correct way. Since space is limited in a paper based dictionary, verb valency is usually described in a very dense and complicated form by using abbreviations and special codes (for instance “01a v 1b C” in [Bianco, 1996], a special verb valency dictionary or “tr K jd fragt jdn [nach etw dat]” in [Hecht and Schmollinger, 1999], a monolingual learner’s dictionary). This module is an outstanding example how hypermedia features and semiotic didactics are used in ELDIT to describe complex information in an innovative and comprehensible way to the learner.

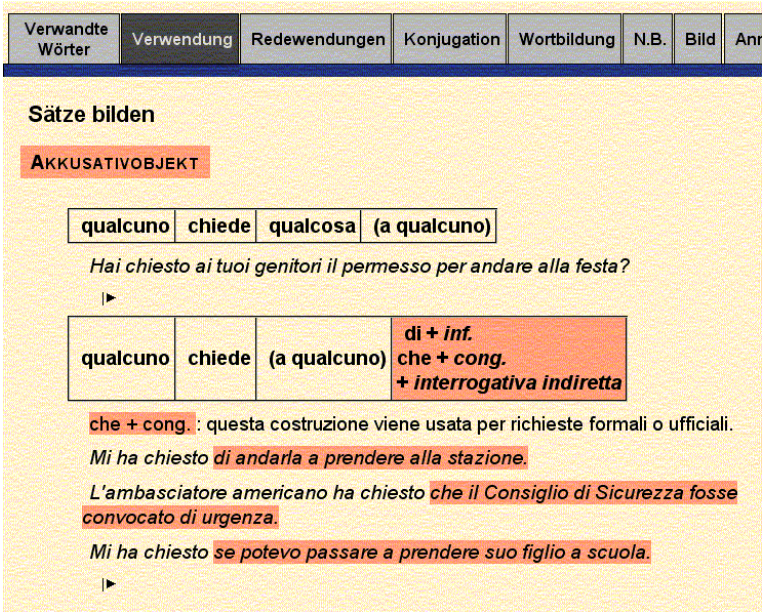


Fig. 3: Description of verb valency for the Italian word “chiedere” (to ask)

The core part of verb valency is represented in a simple way by a table (see fig. 3). 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. Below the pattern one or more lexicographic examples are listed which show the use of the verb in an authentic context. 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 by 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). (Abel, 2002)

2.6. Footnotes

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.
Note the small footnote next to the translation of the first collocation in Figure 1. 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, etc. The 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.

2.7. Pictures

Sometimes it is difficult to explain ambiguities between words in an easy way. Thus we decided to add pictures for concrete nouns in ELDIT. They are shown in the tab “Bild/imagine”, a further aspect related to a word. The pictures are selected according to psycholinguistic criteria [Kroll and Tokowicz, 2001; Weidemann, 1994]. Psycholinguistics suggests the use of prototypical representations of an object, 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 [Vettori, 2003]. Movements, certain actions, or scenarios will be visualized by animated graphics.

2.8. Sound Files

It is very important for a language learner to learn the correct pronunciation of a word. In an electronic

dictionary complex codes such as 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 a simple click the loudspeaker button next to the lemma

3. New Information

Here is a brief description of the kind of new information we needed and its impact on the functionality of the system.

3.1. Inflection

Inflection is the declension of nouns or adjectives and the conjugation of verbs. 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).

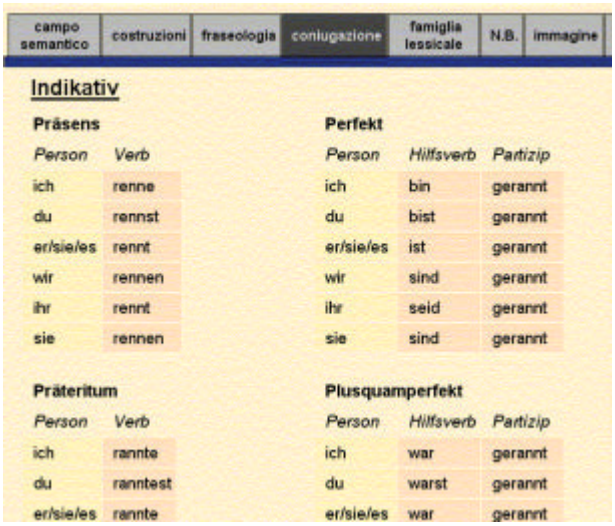


Fig 4: The conjugation of the German word "rennen" (to run) in ELDIT

The word forms are organized in tables according to modes (indicative, subjunctive), tenses (present, perfect, etc), persons (first, second and third), cases (nominative, genitive, etc.) and number (singular, plural) (Knapp et al., 2004).

3.2. 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" (word formation) by emphasizing prefixes, basis and suffixes of each word. In this way the user can see the underlying construction rule. Next to the words, small triangles indicate the possibility to search the ELDIT example database for lexicographic examples containing the word in question (Abel et al., 2004b).

Fig. 5: Derivations and Compounds

3.3. Integration

The first aim of the integration with morphological information was to be able to dynamically show all paradigms of the entries, all surface variants (ev. with information on the new German spelling reform) and build lexical families.

The modules we decided to integrate allowed us to do even more: dynamically create links from all words used for the explanations to the existing entries and derivations, automatically generate exercises on lexicon knowledge, with useful feedback.

In order to judge the quality of the available morphological modules and to choose the right ones to combine with Eldit, we had to specify some criteria.

4. Criteria of Choice

Here is the list of the main criteria used in order to choose the morphological enhancing modules.

- Data quality
  The first criterion is of course the suitability of the data offered. We knew exactly which kind of information we needed to our purposes, so the first choice was based on our needs on morphological data, for German and Italian.
- Data quantity
  Not only morphological information on the chosen entries for the pedagogical dictionary had to be available, but also on all dictionary corpus, which means all words encountered in all entry explanations. This was the main item behind the idea of self-contained environment.
- Performance
  In our project we wanted to emphasize the content interactivity, which we found as lacking in other e-learning projects in the lexicon field (Pedrazzini & Knapp, 2003). So one of the main criteria was to be able to dynamically generate entry paradigms, even filtered with some criteria, through a tool that could

fully guarantee the interactivity. A good performance in the inflection analysis, was also an important criterion in order to generate links on the fly between words contained in the dictionary, entry words as well as words within the explanatory texts.

- Integration
  The ideal solution had to be a black-box solution that could satisfy our needs, without forcing the dictionary developers to be concerned with the internal implementation. The runtime solution had of course to be compatible with the starting infrastructure, which is Java/Servlet based.
- Technology
  Even though the idea was to integrate the modules as a black-box solution, as explained above, we wanted the guarantee that the used technology was the state of the art in the field.
- Extensibility and support
  Support is important, in form of selling company or of Web community. Also important was the opportunity to have someone to refer to in case of needs of extensibility, for example in case we needed information for new entries not contained in the first release.
- Consistency between inflection and word formation data
  This is the first optional criterion. Different modules from different sources could show some inconsistencies among their coded data. This criterion would suggest to get both modules from the same source.
- Current and future modules available
  We basically needed information on inflection and word formation for German and Italian. Our criteria had to concentrate on these modules. But the availability of further modules could suggest new ideas to the pedagogical dictionary. So the fact that a supplier had further modules to offer, had a positive value on the evaluation.
- External references
  You can trust yourself and your ability to judge a product. If someone before you already did it, even to different purposes, this helps too. This is not a mandatory criterion, but we chose to consider it anyway, as added information.

5. Evaluation and Choice

We chose to integrate the Canoo WMTrans transducers, (analyzer and generator, http://www.canoo.com/WMTrans) for the following reasons.

The data for the Canoo transducer elements are based on a well known system, Word Manager (WM), used for the specification, test and maintenance of morphological dictionary databases.

Eldit could profit from the WM components in several ways. Hence a co-operation was started in which the two systems were joined.

Word Manager's lexical resources include morphological and orthographic knowledge for three languages, namely English, German and Italian (ten Hacken and Domenig, 1996; Domenig 1992). In WM 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.

In order to deploy WM 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 (Pree, 1999; Gamma et al., 1995). 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 developed5: WMTrans Word Recognizers, Lemmatizers, Analyzers and Generators for inflection and word formation. Each product type is available for different vocabulary sizes and processing volumes.

5.1. Data Quality and Quantity

A WM database describes the morphological system of a language, including inflectional, word formation and spelling rules. To deploy WM dictionary databases within complex applications, lexicon entries are compiled into finite state transducers (WMTrans).

Data quality is assured by the linguists who are working with it, and the tools used to continuously test the consistency.

Currently WM contains over 250'000 lexeme entries, for each of which the whole paradigm can be generated. This number is far bigger than what we needed.

The Italian database is smaller, about 60'000 lexemes, but we could specify a list of entries that we needed for our purposes. These entries were added to the database and new transducers were delivered.

Moreover an important issue for Italian was the ability to analyze clitics, i.e. single graphic words, like "dimmelo", "prendetevelo", etc., formed by different elements.

5.2. Performance

The finite-state implementation of the Canoo analyzers and generators was able to assure a good performance. Depending on the machine, the lemmatizer is able to analyze between 10'000 and 12'000 entries/second, which is more than enough for our purposes. We needed the analyzer to generate links on the fly between words contained in the dictionary and in the explanation texts.

5.3. Integration

One important issue was the integration. Our system is Java based. We imagined an external module as a black-box with a simple Java API. That was it. The API is composed of a method for the initialization and a method for the query. The integration was straightforward.

5.4. Extensibility and support

Here we wanted to assure the opportunity to specify some new data to add to the transducers. This was directly possible, but we were warned about the fact that the transducers themselves are not able to perform consistency check. The other opportunity was to send missing entries to Canoo, which linguists would then insert them into WM, performing checks and then generate an updated transducer. This is what we did, above all for some Italian entries.

5.5. Technology

The finite-state tools are considered state of the art technology for morphological analysis. Their implementation has been made in Java, which was our technology. So this part of the evaluation was straightforward.

5.6. Current and future modules available

Because we also needed a word formation analyzer and this was available in a similar way as the inflection module (same data, in fact they are generated from the same source, Word Manager, and same API), this turned out to be a very important criterion.

In fact we did not want to add two completely different (from the technology and data point of view) external modules, which, by the way, could have had some inconsistencies between them.

Moreover at that time Canoo already announced the availability of a new suite of interesting products, i.e. the "unknown word" analyzers, able to analyze unknown words. These products are interesting for languages like German, which are very productive at lexical level.

5.7. External references

Among others the Canoo analyzers, in particular the German ones, have been licensed by Google to improve their indexation. This is not of course alone a criterion of choice, but it is an added value in the overall evaluation.

6. The Search Engine

The search engine we implemented tries to address problems language learners commonly encounter: it makes it possible to find single words and multi word expressions (collocations and idiomatic expressions), conjugated or declined 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.

There are different kinds of search. Some of them could be allowed using the integrated Canoo modules.

6.1. Searching Single and Multiple Words

Single words and multiple words can easily be found in ELDIT. The user simply types the word or part of the expression into a search field at top of the screen, indicates the language, presses the enter key, and gets the result.

6.2. Wildcards

Sometimes a language learner might not know the correct spelling of a word. 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".

6.3. 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.

6.4. Stemming

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 the verb "ging" (went). In ELDIT the Canoo WMTrans Lemmatizer allows finding the citation form of a word. Searching for the verb "ging" yields the citation form "gehen" (to go). Also very complicated forms can be decoded, for instance Italian contractions such as "daglielo" (give it to him) lead to the infinitive "dare".

6.5. 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, grammatical hints, etc.). The search operations are restricted to single fields and the results are provided for each field separately.

6.6. 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.

All search operations are carried out automatically. 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.

7. Generating self-correcting Exercises

The integrated modules will be used to generate self-correcting exercises.

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 will be 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 such as the Canoo WMTrans Analyzer/Generator or the Spell Checker of our search engine highly sophisticated correction possibilities and individual feedback can be provided.

Matching quizzes are quizzes where the user has to match two representations of a word, for instance, the lemma with a picture, or the definition with a translation. *Direct questions* such as “What is the opposite of...?” can be generated from the relations between words stored in the semantic fields.

Morphology and syntax 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.

The interaction process between user and system varies according to different quiz types: *Gap-filling* quizzes are text pieces in which some words have been replaced by an edit field. *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.

8. Conclusions

We have presented a project in the lexicon domain, where the interactive experience represents the main goal. The project can be integrated within e-learning courses and uses external engines to achieve interaction at morphological level.

Although the system was already rather complete and extensive in what concerns the lexicon, we decided not to implement extensions that were available from other providers.

Instead we chose to start a collaboration and to reuse morphological functionality from modules derived from other projects. We performed an evaluation and we integrated the chosen modules in our system.

The result was very positive. We achieved the needed functionality in our system very quickly and we could offer some important extensions, which allow a more interactive user experience in our system.

The system is freely available via Web at http://www.eurac.edu/eldit.

9. References

Abel,A., Weber, V. (2000):. *ELDIT, prototype of an innovative dictionary.* In Proceedings of the 9th EURALEX International Congress on Lexicography (EURALEX'00), pages 807-818, Stuttgart, Germany.

Abel A., Gamper J., Knapp J., Weber V. (2003), Formative Evaluation of the Web-based Learners Dictionary ELDIT, In *Proceedings of World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA 2003)*, AACE, 2003.

Bianco M. T. (1996): *Valenzlexikon Deutsch-Italienisch*. Groos Verlag, Heidelberg.

Domenig M. & ten Hacken P. (1992): *Word Manager: A System for Morphological Dictionaries*, Olms, Hildesheim. ISBN 3-487-9677-3

Gamma, E., et.al..(1995): *Design Patterns: Elements of Reusable Software Design*, Addison Wesley.

Gamper, J., & Knapp, J. (2002): *XML for an Electronic Learners' Dictionary*. In Proceedings of the IADIS International WWW/Internet 2002 Conference, pages 427-434, IADIS Press, 2002.

Ten Hacken, P. & Domenig, M. (1996): *Reusable Dictionaries for NLP: The Word Manager Approach*, Lexicology 2:232-255.

Hecht, D. & Schmollinger, A. (1999): Pons Basiswörterbuch Deutsch als Fremdsprache: das einsprachige Lernerwörterbuch zum neuentwickelten Zertifikat Deutsch. Stuttgart (Klett International).

Knapp J., Pedrazzini S., ten Hacken P. (2004): Supporting Language Learners by Intelligent and Efficient Use of Technology, Paper to be presented at ALLC/ACH 2004.

Kroll J.F..,Tokowicz N. (2001): The development of conceptual representation for words in a second language. In *One Mind, Two Languages. Bilingual Language Processing*. Blackwell Publishers, Malden–Oxford, 2001.

Pree, W. (1999): *Hot-Spot-Driven Development*, in: Fayad, Schmidt and Johnson: Building Application Frameworks, Chapter 16, John Wiley & Sons.

Vettori (2003): La visualizzazione di informazioni in lessicografia. Technical Report, Europäische Akademie Bozen, 2001.

Weidemann B. (1994): *Wissenserwerb mit Bildern*. Verlag Hans Huber.

Dynamic Teaching Materials for ESSLLI

R. Bernardi[1] **I. Dahn**[2] **G. Mishne**[3] **M. Moortgat**[4] **M. de Rijke**[3] **H. Uszkoreit**[5]

[1] Free Unversity Bolzano Bozen
[2] University Koblenz-Landau
[3] Informatics Institute, University of Amsterdam
[4] Utrecht University
[5] Universität des Saarlandes

Abstract

In the context of the European Network of Excellence in Computational Logic (CoLogNet, `http://www.cologne t.org/`), the European Association for Logic, Language and Computation (FoLLI, `http://www.folli.org`) has started a project on E-Learning in Computational Logic and the development of Dynamic Teaching Materials for its annual European Summer Schools (ESSLLIs). The project has a double aim: (i) to enhance the (re)usability of existing ESSLLI teaching materials by creating a richly structured repository; and (ii) to develop dynamic teaching materials for the upcoming ESSLLIs, integrating textual presentation, exercises, and computational tools (theorem provers, parsers) into a user-centered "living book". This paper presents the background of the project, gives some brief information about ESSLLI and describes the two subtasks in which the project is divided.

1. Background

While the term E-learning has only recently entered our vocabulary, it has already picked up many readings: such as on-line repository of teaching materials, learning by means of electronic tools, virtual courses or long distance learning. However, as we will explain below, we believe that in all its meanings, it has acquired great importance in the educational endeavor.

For a start, the increase of undergraduate students and lecturers' mobility within Europe due to the ERASMUS exchange programmes, and recently to the development of European Masters and Double Degree Progammes will lead to the establishment of standards for educational programmes which can be mutually recognized between several universities. It should also lead to well documented descriptions of the state of the art in educational methods in a broad variety of scientific disciplines so as to keep lecturers updated on the last scientific and didactic developments in their field. Thus, sharing teaching materials on line is an important stepping stone towards the establishment of such standards.

In addition, electronic tools have already proved to be an effective teaching support. First of all — as shown by Barwise and Etchemendy (1996); Cox et al. (1995) for teaching logic — when students can see what they are reasoning about, and when proofs are presented as graphs, they achieve far greater understanding of the subject than otherwise. Moreover, students have different learning rhythms. Using a computational assistant enables them to follow their own pace, checking their mistakes and working out further solutions as they need. When mistakes are pointed out by a machine, instead of a teacher, students are more motivated to understand the problems (Hoover and Rudnicki, 1996). For these reasons, we believe that electronic tools are an important support in teaching. This holds for Language Resources in general, and more particular for Computational Linguistics Tools (such as Parsers, Corpora and Ontologies).

Furthermore, e-tools are becoming part of the basic tool kits of researchers in applied fields, like Question Answering, Natural Language Interfaces to Databases, Semantic Web, etc. The modularity of the systems and the complexity of the tasks addressed require different areas of expertise which may be hard to be gathered within a single research group. Having ESSLLI's learning resources on-line will help speed up the research in these fields and enhance collaborations which go beyond traditional faculty borders and physical distance.

2. ESSLLIs

The European Summer School, organized by FoLLI, is the key European educational event for interdisciplinary exchanges in the fields of Logic, Language and Information for students, researchers and industrials. It has been organized yearly since 1989, contributing in this way to strengthen the community, facilitate sharing of common interests, form young researchers and provide a framework for the contact between the different fields. It offers around 40 courses per year at different levels (foundational, introductory and advanced) besides workshops, where outstanding results are presented.

The final output of the project will be a rich repository with integrated teaching materials provided by leading researchers in the area, and equipped with navigational tools and content search facilities which will be an endurable learning environment, easily accessible for both students and teachers. Below, we briefly describe the two subtasks of the project.

3. The ESSLLI Archive

The first subtask of the project is to create an infrastructure for on-line exploitation of the vast collection of existing ESSLLI teaching materials.

This project is supported by CoLogNet, Network of Excellence in Computational Logic, Contract No. IST-2001-33123. Maarten de Rijke was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 365-20-005, 220-80-001, 612.069.006, 612.000.106, 612.000.207, and 612.066.302.

Requirements

One of the first issues to be addressed in disclosing the ESSLLI teaching material is dealing with conversion into a standardized, easily accessible format. The ESSLLI CD archive (ESSLLI CD Archive, 2004) is a mirror of the CDs that accompanied the summer schools of the years 1997–2003. This archive includes various information about the courses given in the summer school of that year; typically, this data includes textual meta-data such as author name, description, prerequisites etc, and some teaching aids which are usually slides, handouts or articles in PDF, Postscript, Powerpoint, or DVI format. In some cases, the courses also contain pointers to additional external resources for the subject.

On the CDs, the course content is not provided in a standardized way (i.e., it differs from year to year in design and content); it is not searchable, and is only browsable to a limited extent

Within the task framework, web access should be granted to the ESSLLI archive, with the following features:

- **Search.** The data should be fully searchable. The search facilities must include full-text search, as well as searching by year, author, title and subject. Combinations of the search are also permitted, e.g., searching for a string of text in an article whose author is specified. The search mechanism will also include standard features of similar search facilities with which the target users are accustomed, such as searching for a quoted phrase, boolean searches, highlighting query terms in the results and displaying relevant snippets of the results.
- **Browse.** The data should be hierarchically browsable in several hierarchy trees, according to categories such as: YEAR, SECTION, and LEVEL.
- **Added Value.** Additional data which is currently not in the ESSLLI archive and may assist the user should be provided; this data includes cross references between related information, links to relevant external resources, and so forth.
- **Look and Feel.** The data should be presented in a clear, intuitive, and uniform way, making use of current web technologies. The user will also be able to access the original, non-modified ESSLLI CDs.

Implementation

The conversion of the current archive includes a number of stages. First, the "raw" data (the meta-information about the courses and the teaching material itself) is converted to a standard format; we chose to represent the data as XML documents, the de-facto standard for information representation. Every course and document from the ESSLLI CD archive is converted to an XML document using both publicly available tools for extraction of data from various formats, and specific tools developed for mining the archive for data.

Next, the data is enriched with external information, such as pointers to home pages of course authors, cross references between the courses, and so on. Once it is enriched, the uniform data is indexed and stored for fast retrieval; this allows for various queries such as *display all introductory courses from year 2000* or *search for "X" in the contents of all documents from courses in 2003.* For this stage, we use open source tools which are part of the Apache (Apache Software Foundation, 2004).

Finally, the repository is integrated into a Web Server which enables dynamic content, i.e. creation of web pages "on the fly" from XML data. For this, we use the Cocoon servlet technology, also part of Apache and based on the Java programming language. This enables maximum flexibility and security in the generation of web-accessible pages from the XML repository.

Current Status and Future Plans

As a pilot, we have started the ESSLLI archive creation for ESSLLI 2003. A large part of the teaching materials from this particular year has been converted into the standard XML format, indexed and stored; a Web Server has been set up for the repository, and currently allows browsing and searching it as defined in the requirements section. A methodology has been defined for extending the archive with additional years, so no changes to the server-side technology are required (only generation of additional content in XML, conforming to some rules). The pilot is accessible online (ESSLLI Web Site, 2004).

Our plans include overcoming technical difficulties which prevent some of the data from being indexed, as well as expanding and enriching the access methods to the information, and, of course, indexing the rest of the ESSLLI years.

4. Dynamic teaching materials

A specific advantage of events like the ESSLLIs is that they go beyond a single course and cover a topic by a series of courses that are closely related. As a consequence there arises a natural need to reuse parts of one such course for the preparation of or within another course. One of the objectives of the second subtask is to support this reuse by supporting the automated creation of specific content collections on the fly. Another advantage of ESSLLI is that the experts are available to support the students in getting hands on experience with actual research tools. This is interleaved with the teaching objectives in the second subtask by making these tools accessible from within dynamically generated documents.

To give the reader a feeling for what we are aiming for, we first sketch an existing setup, which already provides some basic levels of interactivity. Then we discuss how we can use Slicing Books Technology to leverage dynamic teaching materials into a user-centered 'living book'.

4.1. Hypertext functionality of the hyperref package

Sebastian Rahtz' hyperref package (available from (Comprehensive TEX Archive Network, 2004) and see (Goossens and Rahtz, 1999, 35–66)) is standardly used for turning the inherent document structure and cross-referencing information of LATEX documents into active hyperlinks. In addition to providing a basic navigational structure, the package has extended levels of functionality, al-

```
\hyperbaseurl{...} % URL for the CGI-executables

\newcommand{\hyperfrag}{...} % grammar fragment URL

\newcommand{\parsescript}[3]{\href{netgrail?% call the parser engine
  url=\hyperfrag&% load fragment from URL
  struct=yes&% structural rule output
  sem=no&% meaning assembly shown
  lexsem=yes&% substitute lexical semantics, 'no' for proof terms
  unary=inactive&% ignore semantics for Diamond/Box operations
  mode=nd&% e.g. natural deduction format
  goal=#1&% goal formula
  test=#2}% your test phrase for the script call ('+'separator)
  {#3}}% your test phrase for typesetting
```

Figure 1: Calling a cgi-script with the hyperref package

lowing the user to specify hypertext links to external documents and URLs, including linking through the Common Gateway Interface (CGI).

In the Computational Linguistics programme at Utrecht University, this extended functionality is used in dynamic teaching materials familiarizing students with a number of computational grammar formalisms. The client-server interaction takes the following form (see (Moortgat et al., 2002) for a full description):

- Students present a linguistic analysis in the form of a grammar fragment. Formalisms currently supported are Stabler-style Minimalist Grammars, and type-logical grammars. A fragment, in these frameworks, consists of a set of lexical type declarations, structural options, and a sample of test phrases.
- Fragments are submitted to the server, where they are turned into dynamic PDF documents. The test sample is hyperlinked to CGI scripts interfacing with general parsers/theorem provers for the formalisms under consideration (The GRAIL type-logical theorem prover (Moot, 1996), CKY deductive parser for Minimalist Grammars (Stabler, 2001)).
- The student (or teacher) can produce derivations 'on-demand' for the test sample in a number of available formats. The server returns these derivations as PDF documents, which can then further be integrated, commented, etc.

The reader is invited to try out the setup at the portal site `http://grail.let.uu.nl`, or to inspect the two possible permutations for the sentence 'Naoko ate Hiromi's sushi' below (from a syntax take-home test). Clicking the sentences fetches their derivation from the server, and displays them in the format specified by the user — Natural Deduction style, in this case. Figure (1) gives the essential code.

1. Naoko ga Hiromi no osusi o taberu.
2. Hiromi no osusi o Naoko ga taberu.

Using these tools with a PDF-enabled web browser, one obtains a seamless client-server interaction. Still, the described hyperref-based architecture has certain limitations: automatic navigational features are restricted to the LaTeX-internal crossreferences, additional linking has to be provided by hand; similarly, the integration of server output with the documents from which the parser engines are called, requires manual post-editing. In the next section, we show how these limitations can be overcome with the aid of SIT (Slicing Information Technology, 2004).

4.2. Living book

The core idea of the SIT approach is to semi-automatically break up LaTeX source code into semantic units, thus providing mark-up that goes beyond the logical document structure. The semantic text-units become 'recombinant' components of a dynamically unfolding document, customized to fit individual readers' needs. Depending on his/her background knowledge and preferences, the user can integrate textual presentation of the material thaught with his/her own crossreferences, explanations, exercises and solutions, possibly obtained with the use of integrated external parsers, theorem provers and the like.

To implement the SIT approach in the preparation of ESSLLI course materials, we distinguish the following steps.

- The authors of selected courses prepare their teaching materials according to ESSLLI style guidelines, and submit them to the SIT Splitter for initial slicing. This phase provides mark-up at the level of general logical-mathematical knowledge.The SIT Splitter decomposes these documents into re-usable learning objects at an agreed granularity. For a learning object to be reusable it is not necessary that it makes sense on its own. Rather it is essential that it makes sense in a context which can be precisely described so that it can be automatically reconstructed whenever this specific learning object is to be reused. Experience shows that in mathematics related documents like those to be handled for this project an average granularity of 5 slices per page is necessary to achieve maximum possibilities of reuse. Note that providing a uniform LaTeXstyle for ESSLLI authors is an important help for a reliable automated slicing.
- The sliced manuscripts are returned to the authors,

who deepen the re-engineering transformation on the basis of their domain-specific knowledge. This phase concentrates on further assigning key phrases, defining extra semantic relations between document slices, and adding components relying on server interaction, such as described above.

- The results of the document re-engineering process are made available as content packages according to the open IMS Content Packaging Specification (see (IMS Global Learning Consortium, Inc)). This specification is supported by many e-learning platforms, opening up the possibility for later reusing specifically built ESSLLI documents in these environments. Added meta-information is encoded in XML in accordance with the open Trail-Solution Metadata Specification and Thesaurus Specification (see (TRIAL Solution, 2004)).

The tool for the dynamic generation of personalized documents, the SIT Reader, utilizes declarative descriptions of the intended structure of documents to be delivered for specific usage scenarios. Deep inside the tool there is an automated theorem prover, called sl-engine, which combines these descriptions with the learning object metadata and with information about the knowledge of the user in order to infer what should be proposed to the user for reading. Another application of the sl-engine is to provide internal inferences in order to obtain a more complete user model. We mention that sl-engine is in part based on methods which are taught to students at ESSLLI.

In order to use a sliced book, the student selects parts she is especially interested in and asks the server to complete the selection automatically for a specific purpose, for example by adding necessary prerequisites or exercises or material from related courses but omitting material that has been inferred to be known.

In a second attempt interactivity is added to the sliced book, turning it into what we call the Living Book. The approach taken here uses interactive pdf documents with embedded JavaScript which make up a connection from the dynamically generated teaching material to some tools running on possibly remote servers.

As an application a student may enter some formula into an input field in the pdf document and will on request receive a newly generated pdf document where the server has added some evaluation of this formula. Another application is the random generation of exercises which take the knowledge of the student into account.

Authors use a simple generic interface in order to bind interactive systems to their teaching materials. Input forms are described in the LaTeX source documents by using the possibilities of the aforementioned hyperref package. At the places where the reply from the server should go into the document, the author inserts a LaTeX command `\tsdynamic{<script>}{<template>}`. `<script>` denotes the name of a program (which can be written in any programming language) that is called by the SIT Reader with parameters describing the user input and how to access the SIT Reader user model. In addition `<script>` may also store data in a protected area of the SIT Reader server in order to correlate different requests. `<script>` is supposed to generate a fragment of LaTeX source code which replaces the `\tsdynamic` command. The `<template>` parameter can provide prepared LaTeX source code which is then filled up by the `<script>`. Finally the SIT Reader generates the pdf document for the learner, using the content generated by `<script>`.

A variant of this approach is to launch an interactive system in a separate window from within the personalized pdf teaching material and only integrate the final result of the students work into a new version of the material, but not following each interaction.

The reader is invited to investigate these possibilities at (Furbach, 2004). We mention that this installation allows as additional features editing of personal annotations (which can be typed or hand written with a tablet pc) and the generation of different views for print, PC screen, Palm Pilot or Pocket PCs. These are basically applications of the same technology for adding interactive systems which has been described above.

The current status of this component of the project is that for the upcoming ESSLLI (ESSLLI 2004. 16th European Summer School in Logic, Language and Information), a coherent set of courses in the logic and language field have been selected to serve as a pilot for the application of the Living Book approach. The pilot is aimed at providing a set of procedures and tools that can then be used for the preparation of future course materials, and that can be made available as author instructions (ESSLLI Repository, 2004).

5. Conclusion and Outlook

The implementation of the outlined project results in a flexible learning environment with a functionality clearly extending beyond the CoLogNet project period in which some of the described methods and tools have been applied to teaching material in Computational Logics. Together with related initiatives (such as Milca (MiLCA, Medienintensive Lehrmodule in der Computerlinguistik-Ausbildung) and LoLaLi (LoLaLi. Logic and Language Links, 2004), it can provide the starting point for the set-up of an encompassing web-based resource center of life-long education in the various disciplines represented at the transdisciplinary ESSLLI summer schools such as Computational Linguistics, Formal Semantics, Computational Logics and Artificial Intelligence.

The segmentation of teaching materials into 'recombinant' units is a novel contribution to the transformation of todays educational instruments into an e-learning setup. In principle the method can be applied to written teaching material such as text books and course scripts in any academic field. However, the method favors clearly structured texts with recognizable units and sub-units such as statements, arguments, conjectures, theorems, proofs, examples, derivations. It may not have been accidental that the origins and first testing grounds for the methodology have been in Computational Logics.

However, the method has far-reaching consequences. For centuries, printed books have constituted the preferred representation for the storage and transfer of human knowledge. Today distributed knowledge sources, hypertexts and

powerful systems for computer aided instruction demand more flexible representations of the same knowledge utilizing a richer explicit structure. While this change is taking place, more books are written. The transformation of existing and new textual knowledge into a representation supporting novel forms of learning and knowledge management is an important challenge for all scientific disciplines.

The linking of conceptual units with external resources such as other information sources, computational systems and visualization tools also has applications that extend far beyond the utilization sketched in this paper. One of these applications is of special interest to the field of language recources and technology evaluation. A revolutionary development for the study of language and for the evaluation of language technology has been the annotation of data with linguistically motivated interpretations. It started with simple part-of-speech annotation, progressed with treebanks and has recently led to some semantic interpretation of data such as in prop-banks and frame-banks. On the other hand, scientific publications link back to the data that were exploited to obtain and verify the linguistic and technological results. The gradual emergence of a solid empirical methodology linking data and their scientific interpretation is changing both theoretical and computational linguistics.

A next logical step in the preparation of teaching material in the ESSLLI disciplines will be the exploitation of the hyper-referencing mechanisms for the linking of statements on language, methods and processing tools with data that exemplify the observations, illustrate the effects of the methods and evaluate the processing components. In this way remote language resources can be integrated into the teaching material. This linking will be especially relevant for courses on empirical methods. But it can also be utilized for courses on formal linguistic analysis or computational methods for the processing of human language.

We expect that the developed methods will be successfully tested in future ESSLLI summer schools. In this way a larger number of students and teachers will contribute to their improvement and hopefully also become interested in their application within other contexts. We hope for a dissemination into regular academic programs, e-learning initiatives, and the European Master Programmes.

6. References

Apache Software Foundation, 2004. `http://www.apache.org/`.

Barwise, J. and J. Etchemendy, 1996. Visual information and valid reasoning. In G. Allwein and J. Barwise (eds.), *Logical Reasoning with Diagrams*. Oxford University Press.

Comprehensive TEX Archive Network, 2004. `http://www.ctan.org/`.

Cox, R., K. Stenning, and J. Oberlander, 1995. Contrasting the cognitive effects of graphical and sentential logic teaching: reasoning, representation and individual differences. *Language and Cognitive Processes*, 10(3/4):333–354.

ESSLLI 2004. 16th European Summer School in Logic, Language and Information, 2004. `http://esslli2004.loria.fr/`.

ESSLLI CD Archive, 2004. `http://www.folli.org/CD/`.

ESSLLI Repository, 2004. `http://www.folli.org/ESSLLI/`.

ESSLLI Web Site, 2004. `http://www.esslli.org/`.

Furbach, U., 2004. Logic for computer scientists. Electronic course. `http://slice.uni-koblenz.de/~sitlog/`.

Goossens, M. and S. Rahtz, 1999. *The LaTeX Web Companion. Integrating TEX, HTML, and XML*. Reading, MA: Addison Wesley.

Hoover, H. and P. Rudnicki, 1996. Teaching freshman logic with mizar-mse. In *Workshop on Teaching Logic and Reasoning in an Illogical World*.

IMS Global Learning Consortium, Inc, 2004. `http://www.imsproject.org`.

LoLaLi. Logic and Language Links, 2004. `http://lolali.net/`.

MiLCA, Medienintensive Lehrmodule in der Computerlinguistik-Ausbildung, 2004. `http://milca.sfs.uni-tuebingen.de/`.

Moortgat, M., R. Moot, and R. T. Oehrle, 2002. Teaching tools for logic-based grammar development. *Eutupon*, 9:1–14. Available at `http://obelix.ee.duth.gr/eft/`.

Moot, R., 1996. *Proof Nets and Labeling for Categorial Grammar Logics*. Master's thesis, Utrecht University. Code available at `http://www.labri.fr/Perso/~moot/grail.html`.

Slicing Information Technology, 2004. `http://www.slicing-infotech.de`.

Stabler, E., 2001. Minimalist grammars and recognition. In C. Rohrer, A. Rossdeutscher, and H. Kamp (eds.), *Linguistic form and its computation*. Stanford: CSLI Publications. Online access at `http://131.211.190.177/mgcky/`.

TRIAL Solution, 2004. `http://www.trial-solution.de`.