Integrating Natural Language Processing into E-learning — A Case of Czech

Pavel Smrž
Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: smrz@fi.muni.cz

Abstract

The paper deals with the application of NLP technology in e-learning. We report our research on intelligent platforms for computer-mediated education. Some of the methods described in the paper have already taken part in the end-user applications that are in everyday use, others still wait for their implementation in the form of software products. The main message of the paper is that the language technology, even in the imperfect form of the current state of the art, can significantly enhance today's computer-mediated teaching and learning activities. It is true especially for languages different from English, where the adopted learning management systems often do not support even the basic functionality of a language-oriented search and retrieval of learning objects. As a case study, this paper demonstrates the application of the given ideas for e-learning materials in Czech.

1 Introduction

Contemporary projects aiming at launching learning management systems (LMS) often focus on the introduction of an existing software tool, rather than on an innovation by means of the modern information technologies. In effect, there is almost no original research directed to the complex integration of e-learning systems with the relevant IT such as assistive technologies (dialogue systems, speech recognition and synthesis ...), knowledge acquisition and knowledge management systems, etc. Among others, the current LMS do not integrate the emerging natural language processing (NLP) applications. The adopted learning management systems often do not support even the basic functionality of a language-oriented search and retrieval of learning objects.

Futhermore, the present-day LMS are not directly linked to the wealth of relevant information and knowledge sources. In the case of the higher education institutions, these sources could comprise standard libraries that provide at least an electronic catalogue of their sources nowadays, local digital libraries that are usually freely available for academics from particular institution, and the access to comprehensive electronic archives or digital libraries that are provided by many publishers and other organizations on a commercial basis. Companies often neglect valuable knowledge sources too. For example, they should consider the integration of their knowledge bases in the form of recorded questions and answers from the call centers.

The current e-learning systems do not exploit the potential of available high-level personalization techniques and adaptability of the form, the content and the access to the education. Most of them cannot play the role of a showcase for the modern teaching methods.

This paper surveys several areas, where NLP techniques and technologies can enhance educational systems and applications. Some of them exist in the form of prototypes only and have not been applied in an end-user system yet. Others find their place in software tools that have been implemented by our team. They will be briefly introduced in the paper.

The range of LMS used or tested at Masaryk University, Brno, Czech Republic (MU) is rather broad. The most important ones are ILIAS (http://www.ilias.uni-koeln.de/) and MOODLE (http://moodle.org). Both systems are developed and distributed under the term of the GNU General Public License and provide open platform appropriate for the integration on NLP solutions. The actual project at MU aims at unification of the used e-learning platforms and their integration with the administrative information server of the university (Pavlovic et al., 2003). Even though separate systems would be more modular, easily maintainable and extendable, we opt for the integrated solution that will benefit from the per-

manent technical support and personal assistence of the administrative information server team. We strongly believe that NLP techniques as a part of the e-learning system can help to open doors to those faculties and departments that have not discovered the world of computer-mediated education yet.

The paper discusses also the incorporation of language resources to support the learner during his/her interaction with an educational system and to provide personalized learning. We also tackle the use of NLP technologies and resources to support the automatic assessment of learners' answers, especially those which are in free text or restricted free text form. Such assessment is useful to learners for controlling their learning progress (self-regulation), to teachers for gathering information about learners and to systems for personalizing interaction. Concept mapping is a knowledge elicitation technique, which stimulates learners to articulate and synthesize their actual states of knowledge during the learning process. We propose the use of NLP in concept mapping systems in order to interactively support learners, who build concept maps and automate the process of the assessment of concept maps. The availability of wordnet-like semantic networks resulting from several projects such as EuroWordNet (Vossen, 1998), BalkaNet (BWN, 2004), RussNet (Azarova, 2004), or broad-coverage ontologies such as SUMO (SUMO, 2003) provide a reasonable starting point for such an effort.

The NLP applications in the area of e-learning can be divided according to various criteria. They can be specific for synchronous or asynchronous mode of the course. The main focus of the methods can be stressed to address e.g. enhancements of the teaching material accessibility or the adaptability of LMS. Also the complexity of the needed NLP techniques can make the distinctions, whether the methods are already available and prepared to integration into LMS or they need further development. The availability of language resources or language technology (lingware) for the particular language can make the difference too. A related issue can be the portability of a solution for other languages or other subject area, where subject-specific information cannot be obtained fully automatically. One can also divide the NLP applications in e-learning according to the NLP modules that are integrated, e.g. a language-specific morphological module or named-entity analyzer could play a crucial role. As the educational process has two faces — learning and teaching, the boarder-line can also be drawn between the tools focusing on the students' side and those intended for the course authors and teachers.

The last mentioned aspect has been taken into account in this paper. It is organized as follows: The next section discusses NLP techniques aimed at enhancements for the end-users of e-learning systems — students looking for an appropriate e-learning material or those who already enrolled in a course. The third section tackles the support of authors and providers of the e-learning facilities that can take the advantage of the language and text technology too. Of course, the boundary between those two cases is not strict at all, so there are NLP tools that can help both types of LMS users. The fourth section then covers supplementary information technologies such as multimedia and audio- or video- recording of courses that cannot be classified as NLP per se but are strongly related and, as our experience already shows, their integration should be at least coordinated with the employment language technology solutions. The paper concludes with future directions of our research.

2 NLP Support of LMS End-users

2.1 Basic Methods

Several standard NLP tools have been developed recently in Natural Language Processing Laboratory, Faculty of Informatics (NLPlab FI MU) that are beneficial for many areas. E-learning is no exception in this respect. The most important module is morphological analyzer AJKA (Sedlacek and Smrz, 2001) that serves as a base for various related modules.

The importance of the morphology for Czech can be demonstrated by the history of the web search engines in the Czech Republic. Czech is a representative of Slavonic languages characterized by abundance of inflective morphological processes. For example, nouns, adjectives, pronouns and numerals can be present in one of 7 grammatical cases in Czech (nominative, genitive, dative, accusative, vocative, locative and instrumental), three grammatical numbers (singular, plural and dual), and three genders (masculine, feminine and neuter), in which masculine exists in two forms - animate and inanimate. The most popular Czech web portals started with the search engine provided by Google, but

after a short time they replaced it by specialized language-aware systems developed usually in cooperation with Czech universities.

The story shows that even generally recognized implementations can be outrivaled, if they ignore language-specific features. It holds not only for web environment but also for the e-learning. Projects implementing LMS need to integrate strong language support and go beyond simple localization.

A lot of e-learning material has been produced in last years and the number grows as many new courses are prepared in an electronic form nowadays. One of the most important applications of AJKA is therefore the system that enables search in the available e-learning materials. As the document source language and the encoding are not always specified explicitly, a text technology module — language guesser based on the language samples in a particular encoding — has been also implemented and incorporated into the e-learning search engine.

Another language resource that proved to be very useful in our experiments is the query expansion module based on information from the Czech WordNet (Pala and Smrz, 2004). The crucial point here is the improvement of the user interface. Users are able to search for all word forms, synonyms, hypernyms and other semantically related words, exclude co-hyponyms, etc. They can also choose between orthographic, regional, style or register variants, words derived from the term searched, verbs related by the aspect, etc.

Of course, even a stemmer or a simple lemmatizer would suit the purpose of the search engine. However, the strength of the full morphological module becomes apparent when the incorporation of deeper analyses of texts is the issue. We currently work on a module that will be able to summarize the content of related documents. The immediate target of the tool is the summarization of messages in course discussion groups. Sometimes, the students are very active (the discussion is occasionally off-topic) and even a simple browsing through the discussion threads from previous runs of a course could present a tedious work. The automatic extraction of the most important points can significantly help newcomers.

Recently, we have added automatic classification of e-learning materials into given categories or at least clustering if there are no pre-defined classes. Both — learners and authors can take advantage of these options but the role of self-learning is stressed in the current version. The function that proved to be most useful for e-learning content is searching for documents similar to a given course or an additional learning material. Inspired by the biggest dictionary publishing servers, we will provide "a new document for this day" function and allow course authors to specify what should be the pool of searched documents. As there is currently an e-learning project covering all our university, we strongly believe that such functionality will draw a broad attention to the new methods of teaching and learning.

All the enhancements of the search for the additional e-learning material mentioned above are also applicable in the search for appropriate courses. The method based on the statistical tests known from the computational linguistics is provided which enables the automatic extraction of keywords that can be added to those specified by course authors. The available named-entity and multiword-expression analyzers for Czech have found their place in this task.

The summarization module is applicable here too as not all authors provide enough metadata to facilitate searching courses or relevant parts of them. An automatically generated index as well as glossary or a small encyclopedia from a comprehensive course enables non-linear passing through the learning material, which becomes a norm.

Another technique that enables students to concentrate on the selected parts of a course is the matching of the course content to the student's knowledge of a given topic. The method applied to the matching is derived from the standard assessment procedure for language learners. As the task for the first assignment, students are asked to write an essay covering their current knowledge about the subject of the course. The content of the document is compared with the sections (lectures) of the given course and the parts that are not covered sufficiently are presented for further study.

The described approach can serve as the launching point for the LMS personalization. Some e-learning courses already contain special mechanisms to offer a path through the teaching material adapted to the needs of particular user. Such an enhancement can improve the acceptance of the content by users that often face learning materials that are too in-depth in some parts and too sketchy in others. Also,

the speed of presentation of the teaching material can vary according to user needs. Besides the above-mentioned assignment of essays, the simple statistical technique has been adopted that automatically evaluates the match of students' answers and determines knowledge levels of the students. The solution is now ready to be applied generally to all e-learning courses provided by MU. It can be generalized to allow semi-automatic generation of tests for each part of the subject matter that would "send" students (back) to the parts they should go through again.

The presented adaptive components, which adjust the system to particular users, are not the only ones that we thought about and that are covered in our e-learning environment. As FI MU hosts centre Teiresias, which is responsible for helping students with special needs from all departments of MU (`http://www.muni.cz/teiresias/`), we focused also on the adaptation of LMS for disabled. The activity of the center is not limited to the electronic form of learning but the context of the computer-mediated education fits its pursuance perfectly. We currently plan the first e-learning course that would be fully available both in Czech and in Brail for visually impaired people. The content different from a plain text still presents a problem as it is sometimes very difficult to find Brail equivalents for mathematical expressions or various symbols and it is always an extremely time-consuming work.

We should also mention the role of the speech synthesizer developed by our team that is available for the needs of visually impaired participants of the e-learning courses. The users are able to help other students providing the pronunciation of the words, which is not correctly derived from its orthographical representation by our automatic engine.

2.2 Question Answering and Exercise Generation

The automatic question answering (QA) is another task where the morphological as well as surface syntactic analyses (Smrz and Horak, 2000) play the crucial role. We gain from our experience with "Encyclopedia Expert" (Svoboda, 2002), which is able to answer free questions based on the information extracted from an available Czech encyclopedia. A set of semantic frames and corresponding syntactic structures is defined that enables analysis of the most frequent question types. The rest of queries are answered by the full-text search and the identification of the relevant part of the document containing the answer. Both cases are based on the question-type analysis module determining what information should be looked for.

The employment of the same strategy in the context of e-learning material is straightforward. The current system is able to apply the mechanisms described above to answer questions based on the content of a particular course. The information about the part of the course, which contains the answer, can be also returned. The evaluation of the QA module showed that the precision on queries that can be answered just on the content of a given course is pretty high. Even for information not covered by the pre-defined frames (answered by the full-text search) it was about 87 percent (errors are mainly due to the incorrect or incomplete analysis by the morphological module — unknown words). On the other hand, practical assessment of the module showed that the described functionality has only a limited use for the course participants. Many students would need and appreciate a search for answer in a much broader range of relevant e-learning materials. Therefore, we are going to provide such an option. Of course, the number of errors can increase in such a setting and it will be the responsibility of the course author to check the function of the QA module on the sources he or she identified and to improve the results by means of additional linguistic resources needed in the analysis phase.

The promising results of automatic QA led us to the idea to engage similar NLP methods the other way around and automatically generate questions and perhaps whole tests based on the content of particular e-learning courses. It is usually easy to extract "what is" questions or ask about a particular fact explicitly stated in the text. Sometimes, the structure of the documents itself helps to identify the important knowledge; sometimes, the abovementioned keyword extraction algorithm can be employed.

The tricky part of the process is therefore to create a procedure that would be reliable enough for the automatic checking of answers. Again, the basic QA module is satisfactory for the questions that expect factoid as the answer. However, it is much more difficult to automatically evaluate more elaborate answers.

Although we are not able to present final results for this part yet, the preliminary ones show that the application of the same method as for the comparison student essays with the content could provide at least a good approximation of the assessment that would be given by human teachers.

The weak point of the described automatic generator of exercises is its focus on the factography and impossibility to verify that students really understand the content, that they got the heart of the matter and are able to apply the obtained knowledge. Obviously, it is still far from today, when computers will be able to substitute human teachers in this respect. An interesting step, that at least simulates such a function and that is investigated by our current experiments, is the employment of standard ontologies for this purpose. The interest in ontologies increased with the recognition of their importance for the Semantic Web. The emerging information systems need the definition of common understanding for their application domains. Even though ontologies are meant as a means providing human knowledge to machines, they can be very useful for e-learning too. The formal specification of the concepts and relations between them takes usually advantage of the new XML-family standards, RDF (Beckett, 2003) and OWL (van Harmelen et al., 2003). The latter serves as a base for our recent research on the possibility to automatically generate exercise questions asking for relations between concepts given by an ontology. As the number of available ontologies for various subject domains will surely increase this direction of research has significant potential.

2.3 Language Learning

In our work we pay a special attention to NLP methods applied in the area of language learning and teaching. The role of empirical data in the form of corpora — large collections of written or spoken language in the digital form — is generally recognized in the computational linguistics. Corpora are crucial for language learning too. For example, the English courses taught at our faculty bear on BNC (the British National Corpus) and other available English corpora (e. g. Times Corpus). The standard form of the vocabulary test is automatically generated which lists concordances of a word that is deleted from all the presented occurrences and students have to fill the gap.

Corpora are beneficial not only for generation of queries. Students often use them as the primary source to learn about the usage of a word. What complicates the process is the presence of words students are not familiar with. Another direction of our research that is currently under development is the effort to sort the concordance list according to the estimated complexity of words in them. To be able to compute such a measure efficiently even for extensive concordance lists, the evaluation is based on heuristics that take into account frequencies of the words in the contexts.

New approach that will find its role in the modern computer-mediated language learning is the employment of the word sketch engine described recently in (Kilgarriff et al., 2004). Word sketches are brief automatic corpus-based summaries of a word's grammatical and collocational behavior. The sketch engine is a corpus tool developed by our team which takes as input a corpus of any language and a corresponding grammar patterns and generates word sketches for the words of that language as its outputs. The most helpful feature of the system from the language e-learning point of view are not word sketches per se but the ability to automatically compare sketches generated for different words. It is crucial especially for semantically similar words, e.g. for near synonyms or co-hyponyms. Students are able to compare collocates and grammatical relations of the words and the system can also automatically generate tests checking whether learners know the difference between semantically close words. The word sketch engine also generates a thesaurus that can be directly used in inspecting the knowledge of a particular semantic field.

Correct answers are usually required to enter next levels of a course. However, *errors* can play a significant role too. A special attention has been paid to the errors produced by students in project Czenglish. It is a join project of NLPlab FI MU and the Department of English at the Faculty of Arts. The e-learning course is based on popular book "English or Czenglish" which is intended for students of English at the advanced or the professional levels. Students produce translations of sentences presented by the system. If a translation does not match a stored one, the correct answers are displayed. At the same time, the actual answer is compared with the examples of listed incorrect ones. If a match is found the explanation of the error

is presented. Students may also indicate, that they still believe their translation is possible. The message is sent to the teacher that has to decide whether it should be added to the list of correct answers. Such a decision automatically affects the assessment of the student's test.

3 NLP Techniques for Authors and Teachers

3.1 Course Preparation

The quality of LMS raised enormously in last decades. Students are able to work with e-learning applications that are much more user-friendly than some years ago. However, simultaneously with the quality increased the complexity of the systems for content providers. To prepare a good course, teachers need to learn how to use the authoring system, how to determine the possible passes through the material etc. The authoring systems are usually linked to particular LMS. At least a basic understanding of the technical stuff and e-learning standards behind the particular LMS is often required. Our experience clearly shows that text and language technology can significantly help the e-learning especially in the phase of course preparation and is extremely beneficial for teachers.

The context analysis and the evaluation of the similarity between a new e-learning content and the existing courses can be employed to speed-up the process of course preparation. We currently prepare an expert system that will serve as an assistant for authors. The system compares the provided content with the stored courses and uses found similarities to propose standard links (to additional materials, on-line encyclopedias etc.) It also groups the possible actions to enable authors to perform many necessary actions at one-click. It is advantageous especially for providing metadata for the learning objects that are understood as a must for all the future e-learning content.

A course with a dense web of hyperlinks can turn to be a nightmare for maintainers in a dynamic environment. To keep track of all the links and relations between the e-learning content, a new system called DEB has been designed and implemented by our team (Smrz and Povolny, 2003). It is a client-server application that enables efficient storage and retrieval of XML documents. Also complex structural and content queries can be defined. For example, it is the main platform to host the above-mentioned Czenglish project. It guarantees that the whole content will stay consistent when linked information is changed. The consistency checks are defined as XSLT sheets (Clark, 1999) and the system reports any violation of the predefined constraints.

The implemented mechanisms of XSLT transformations are powerful enough to be integrated in another important task of the current LMS — the support of open e-learning standards. All new learning objects developed at our faculty should be developed in such a way to facilitate sharing and interchange. As the field of LMS is rather hot and many commercial as well as public-domain systems emerged recently, the newly produced learning object should also reflect the possibility of a switch between platforms without an information loss. The currently developed authoring module defines XSLT transformations of the primary document (DocBook-like (Walsh and Muellner, 1999)) format into LOM/SCORM compliant forms. It could help to put the platform independent standards into wide use.

3.2 Course Run

The support to teachers during the run of courses is at least equally important as in the phase of the content preparation. It is true especially if a course has already run several times and the feed-back from the previous runs should be reflected in the actual one. The current systems usually ignore the possibility to reflect the experience from the past years. The students' answers, either correct or incorrect, are often thrown away or used just for elementary statistical profiles. One of the most valuable resources is frequently neglected.

The results of our recent research reveal that the detailed analysis of the students' outputs offers an extremely useful material directly applicable in the optimization of the e-learning content. For example, special tools have been developed for tagging and categorizing the grammatical and stylistic errors in student essays (Pala et al., 2003). Teachers mark the errors in the electronic documents. Students have to correct them in the new version but also to record the original (incorrect) form and the error type. The files in the source format (usually MS Word or LaTeX) are transformed into an appropriate XML format. A special form of a learner corpus has been developed that serves as a base for focusing the teaching in a special course. As a side effect, the created re-

source is used to develop the Czech grammar checker (Smrz and Horak, 1999) which will be the first one based on the large empirical data.

The attempts to make the assessment of student knowledge as objective as possible and to reduce the teacher's work led to the spread of multi-choice tests in last decades. It is a well-known fact that this form of testing has many disadvantages too and that the focus on them can easily produce test-experts rather than people understanding the subject. The current technology offers various means to implement intelligent tests and to escape the trap of multi-choice tests. Our current research concentrates on the integration of NLP techniques into the evaluation module of LMS. The experience shows that it is relatively easy to provide such functionality for short answers in the form of phrases described by simple grammatical patterns. The results of the first limited experiment are very promising as the method reduced the number of answers that needed to be processed manually significantly (only 31 % of the cases remained for the manual check). However, there are also many open questions concerning scaling-up the method of the answer patterns. The longer the answers are, the less matches between the pre-defined applications are found. A more general solution of the analyzing mechanism in the form of a general grammar is obviously needed. However, the available grammar for Czech is developed for a robust analyzer so the ambiguity of the analysis tends to be rather high. Also it is much more difficult for authors that do not specialize in the computational linguistics to define the acceptable forms of answers. The assessment of open questions is definitively one of the actual topics of our research.

Although the current state-of-the-art in NLP does not allow full computerization of the assessment of general tests, there are areas, where the technology can already take over the role traditionally falling upon the teacher. We have recently implemented the automatic evaluation of programs in the Java programming course at FI MU and also the phrase-structure grammars developed by students of the "Introduction to the computational linguistics". Students are obliged to provide their assignments in the designated form. Pre-defined tests are applied to verify the correctness of the provided solutions. We are to provide a similar function for a grammatical pre-checking of students' essays. Even though only a limited number of error types can be identified automatically, the method can still reduce a significant portion of the tedious teachers' work.

3.3 Additional Functions

The previous section discussed how the standard NLP methods can help students to search in the e-learning data. The same technique can be integrated into the LMS authoring subsystem. Especially, the tedious linking of the additional course sources can be facilitated by NLP. We have designed and implemented a system which automatically finds relevant articles and papers available for the students of our faculty or the university (digital libraries provided by ACM and Springer-Verlag), freely accessible on the web (www.arxiv.gov, ResearchIndex) or that are at readers' disposal at the faculty libraries. The experience shows that the personalization of the application interface is perhaps even more important for authors than for students. Too complex environment can scare off and discourage some authors, others call for more functions and more possibilities to determine the behavior of the search engine.

Many e-learning courses have been recently created all over the world; many of them are available for the general public via Internet now. Most of the resources are written in English. It is sometimes impossible to take directly the foreign version, at least a part of the content needs to be translated. The NLP technology known as translation memory finds its place here. The motivation of the application of the translation memories is the same as in the common localization of software. If the content of the e-course will be modified and a new version becomes available, the "memory" of the tool will help to translate parts that remain the same or changed in a limited extent only. Our experience of the localization based on DejaVu software (Dej, 2003) is rather positive as it enabled to "upgrade" our English-Czech translation in just a week.

The last piece of software that will be mentioned here is the experimental plagiarism identifier. Very simple methods comparing word n-grams showed to be efficient and precise enough to identify plagiarism in cases where both the original and the derived document are in one language (Czech). However, teachers often find essay that are word-for-word translations of an original text. A reliable automatic identification

of such cases is difficult as the n-gram methods could not provide reasonable precision due to the difference of the syntactical structures between Czech and English. This kind of checking forms another direction of our future research.

4 Web services and Multimedia Support

4.1 Service-Oriented Architecture for E-learning

A typical LMS is constructed as a monolithic piece of software or a set of tools fully integrated into another system such as administrative information system. The first really open systems able to communicate with the environment via platform-independent channels emerged in the area of LMS only recently. The long-term goal of our research is to provide NLP solutions for LMS via open technologies in the heterogeneous systems. The form of web services seems to be appropriate for the task.

Generally speaking, our view is based on the idea of the service oriented architecture. This emerging approach assumes that software applications implement only functions specific for their tasks, while more general functions are implemented as services available on the net.

If a method can be provided as a web service, it immediately evokes the idea of "outsourcing". A standard task in the preparation of learning courses is the search for additional material, its indexing and linking to the primary course texts. As has been presented above, NLP can offer supplementary functions, such as automatic question answering. However, as has been also shown, the preparation of necessary resources used in the analysis needs detailed knowledge about the way a particular NLP techniques work. One reason for the outsourcing lies therefore in the effort to enable users to focus on the content of e-learning courses only and to provide other services externally (free of charge or on a commercial basis).

To motivate the other reason, let us share the experience of the processing large corpora. The above-mentioned word sketch engine needs several hours to compute the necessary statistics on a 100,000-word corpus (BNC). The implementation takes advantage of the available powerful PC workstations. To process a 1-billion corpus in the same time, we would need a supercomputer to perform the task. However, such a computer would be idle almost all the time in our laboratory. It is much easier and cheaper to "hire" the computing power needed for the processing.

Many modern NLP methods are based on large language corpora (hundreds thousands of words) and lexical databases (e.g. Princeton WordNet (Miller et al., 1990) contains more than 200,000 lexical units and their relations). High-speed networks and Grid systems are very important in this context as they enable to transfer the processing of resource-demanding tasks such as the creation of indices for large collections of data to the available powerful systems. Such systems can moreover benefit from the analysis of various corpora. Such tasks can be fully outsourced without an allocation of own capacity — in terms of efficiency as well as language resources.

The described approach is one of the most actual directions of the integration of NLP and the Semantic Web. It is studied at FI MU especially as a part of the recently proposed project combining e-learning and the Grid technologies. We propose to integrate and evaluate services based on the new specification WSRF (Web Service Resource Framework) that defines a unifying view at the web and Grid services.

4.2 E-learning and Multimedia Support

The modern e-learning courses are often supplemented by a multimedia data. It can range of a simple recorded audio from the lecture to the full video recording and its streaming with integrated switching between the shots of the lecturer and his/her presentation. A recent topic we are working on deals with the software tools that can facilitate the preparation of such an e-learning material. One of the most important current issues is the research on the collaboration between LMS and the development platforms for multimedia applications (Authorware (Aut, 2004) in our case).

Processing and indexing the multimedia content is another large issue. We currently prepare a new methodology to link the presentation with the lecture recording. We also performed the first experiments aiming at the evaluation of automatic linking of the presentation content to the recording by means of automatic speech recognition. The obtained unsatisfactory results influence the changes in the recording setting with the aim to improve the acoustic quality of the recording. We would like to define a universal methodology applicable also outside the university.

The next priority of our research is the shared teaching. A pilot course with shared parts of selected lectures transmitted between two distant sites (Prague and Brno) will run in the next semester. The experience with the teleconferencing will surely be a helpful for this task, especially the verified camera tracking mechanisms. FI MU is also active in the field of streaming the lecture recordings via the high-speed networks.

5 Conclusions

The broader context of all the research described in the paper is given by the common e-learning strategy built by several leading universities in the Czech Republic. The initiative includes the support for choosing LMS platform, study of the standards in the area, publishing good practices for e-learning materials and the new forms of education, etc. The incorporation of NLP techniques to extend the functionality of the current LMS becomes one of the primary action lines in the program.

The experience gained by some faculties at MU as well as of other cooperating institutions demonstrates that preparation and providing of the e-learning courses can generate significant profit. It holds not only for commercial providers but also for the universities (distant learning courses). Moreover, the idea of computer-mediated education is in accord with the long-term aim to internationalize the universities in the Czech Republic. The described NLP techniques and other modern approaches help to open the institutions to students from all over the world.

Acknowledgements

This work was supported by Ministry of Education of the Czech Republic Research Intents MSM6383917201 and CEZ:J07/98:143300003, Grant Agency of the Czech Republic Grant GACR 405/03/0913 and by EU project BalkaNet IST-2000-29388.

References

2004. Macromedia Authorware 7, `http://www.macromedia.com/software/authorware/`.

Irina V. Azarova. 2004. RussNet — wordnet for Russian. `http://www.phil.pu.ru/depts/12/RN/`.

2004. Balkanet project website, `http://www.ceid.upatras.gr/Balkanet/`.

Dave Beckett. 2003. Rdf/xml syntax specification. `http://www.w3.org/TR/2003/WD-rdf-syntax-grammar-20030123/`.

James Clark. 1999. XSL Transformations (XSLT) Version 1.0. (`http://www.w3.org/TR/xslt`).

2003. DejaVu translation memory and productivity system, `http://www.atril.com/`.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *Proceedings of Euralex 2004*. (to be published).

George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Five papers on wordnet. Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton University.

Karel Pala and Pavel Smrz. 2004. Building Czech wordnet. *Romanian Journal of Information Science and Technology, Special Issue on BalkaNet*. (to be published).

Karel Pala, Pavel Rychly, and Pavel Smrz. 2003. Text corpus with errors. In *Proceedings of TSD 2003*, Berlin. Springer-Verlag. Lecture Notes in Artificial Intelligence.

Jan Pavlovic, Tomas Pitner, Pavel Smrz, and Jiri Verner. 2003. Customization of ILIAS and its integration with the university information system. In *ILIAS 2003*, Cologne, Germany.

Radek Sedlacek and Pavel Smrz. 2001. A new Czech morphological analyser ajka. In *Proceedings of the TSD 2001*, pages 100–107, Czech Republic.

Pavel Smrz and Ales Horak. 1999. Implementation of efficient and portable parser for Czech. In *Proceedings of TSD'99*, pages 105–108, Berlin. Springer-Verlag. Lecture Notes in Artificial Intelligence 1692.

Pavel Smrz and Ales Horak. 2000. Large scale parsing of Czech. In *Proceedings of Efficiency in Large-Scale Parsing Systems Workshop, COLING 2000*, pages 43–50, Saarbrucken: Universitaet des Saarlandes.

Pavel Smrz and Martin Povolny. 2003. DEB - Dictionary Editing and Browsing. In *Proceedings of the EACL03 Workshop on Language Technology and the Semantic Web: The 3rd Workshop on NLP and XML (NLPXML-2003)*, pages 49–55, Budapest, Hungary.

2003. SUMO — Suggested Upper Merged Ontology, `http://ontology.teknowledge.com/`.

Zdenek Svoboda. 2002. Znalec encyklopedie (encyclopedia expert). Master's thesis, Faculty of Informatics, Masaryk University, Brno.

Frank van Harmelen, Jim Hendler Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. 2003. OWL web ontology language reference. `http://www.w3.org/TR/owl-ref/`.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

Norman Walsh and Leonard Muellner. 1999. *DocBook: The Definitive Guide*. O'Reilly.