

# Improving the search for learning objects with keywords and ontologies

Lothar Lemnitzer<sup>1</sup> and Cristina Vertan<sup>2</sup> and Alex Killing<sup>3</sup> and Kiril Simov<sup>4</sup>  
and Diane Evans<sup>5</sup> and Dan Cristea<sup>6</sup> and Paola Monachesi<sup>7</sup>

<sup>1</sup> Seminar für Sprachwissenschaft, Universität Tübingen, Germany

<sup>2</sup> Natural Language Systems Divison, Institute of Informatics, University Hamburg,  
Germany

<sup>3</sup> Center for Security Studies, ETH Zürich, Switzerland

<sup>4</sup> LML, IPP, Bulgarian Academy of Sciences, Sofia, Bulgaria

<sup>5</sup> The Open University, Milton Keynes, UK

<sup>6</sup> University of Iasi, Romania

<sup>7</sup> Utrecht University, Uil-OTS, Utrecht, the Netherlands

**Abstract.** We report on an ongoing project which aims at improving the effectiveness of retrieval and accessibility of learning object within learning management systems and learning object repositories. The project *Language Technology for eLearning* approaches this task by providing Language Technology based functionalities and by integrating semantic knowledge through domain-specific ontologies. We will report about the development of a keyword extractor and a domain-specific ontology, the integration of these modules into the learning management system ILIAS and the validation of these tools which assesses their added value in the scenario of searching learning objects across different languages.

## 1 Introduction

Significant research has been carried out in the area of Language Technology and within the Semantic Web. Several initiatives have been launched in the last years both at the national and international level aiming at the development of resources and tools in the areas of NLP, Corpus Linguistics and Ontology development. However, their integration in enhancing eLearning systems has not been fully exploited yet.

The aim of the *Language Technology for eLearning project* (LT4eL, [www.lt4eL.eu](http://www.lt4eL.eu)) is to improve eLearning with language technologies and resources in order to provide new functionalities which will enhance the adaptability and the personalization of the learning process through the software which mediates it. In our project, we show how language resources and tools can be employed to facilitate tasks which are typically performed in an LMS such as the search for learning material in a (multilingual) domain, semi-automatic metadata development based on keywords and generating glossaries on the basis of definitions of key terms.

However, the main objective of the LT4eL project is to improve on the retrieval of the learning material and we tackle this problem from two different but related angles: from the content end and from the retrieval end.

On the content side, a steadily growing amount of content cannot be easily identified in the absence of systematic metadata annotation. Providing metadata is a tedious activity and the solution we offer is to provide a Language Technology based functionality, that is a key word extractor which allows for semi-automatic metadata annotation on the basis of a linguistic analysis of the learning material. While keyword extractors have been provided mainly for English [11], the innovative aspect of our project is that we provide this functionality for all the eight languages represented in our project, that is Bulgarian, Czech, Dutch, English, German, Polish, Portuguese and Romanian and that we embed such a tool within the eLearning context.

On the retrieval side, the standard retrieval systems, based on keyword matching, only consider the query terms. They do not really take into account the systematic relationships between the concepts denoted by the queries and other concepts that might be relevant for the user. In the LT4eL project, we use an ontology as an instrument to express and exploit such relationships, which should improve the search results and allow for more sophisticated ways to navigate through the learning objects. Furthermore, by linking the ontology to language specific lexica, multilingual retrieval will be possible. We believe that the potential of ontologies in the area of multilingual retrieval is not sufficiently exploited yet and with our project we intend to make a contribution in this direction.

In this paper, we focus on how retrieval of learning objects within a Learning Management System can be improved on the basis of semi-automatically generated metadata as well as a domain ontology. As basis for the extraction of the keywords and the development of the ontology, we use linguistically annotated learning material which has been converted into XML. This process is described in section 2. Our approach on the extraction of keywords is presented in section 3 while the ontology developed to support the search process is introduced in section 4. The developed functionalities (i.e. keyword extraction and ontology) are integrated in the ILIAS Learning Management System and the integration process is discussed in section 5. Validation is briefly addressed in section 6 while section 7 contains some concluding remarks on future work.

## 2 Preparing the data

The development and testing of the keyword extractor and ontology are based on domain specific corpora for the various languages. It was decided to collect corpora of learning objects of at least 200.000 running words per language. The topics of these learning objects are information technology for the non-expert, mainly introductory texts and tutorials for word processing, HTML etc., texts which convey basic academic skills, and texts about eLearning. Around one third of the corpora is truly parallel in the sense that we used translations of the same

basic text into the various languages. To this end, we chose the CALIMERA document (<http://www.calimera.org/>) because it is close to our domains.

The documents we collected with an opportunistic method vary considerably in size – from a few sentences to more than 50 pages. Wrt to the information extraction and search functionalities, the size of the documents has the following impacts:

- The statistical measures of the keyword extractor all rely, as we will show later on, on document frequency of a term. Therefore, the smaller the document and larger the number of documents in a collection, the more precise the statistics can capture the distributional behaviour of terms. So smaller documents are preferable from this point of view. On the other hand, the selection of keywords implies that the documents are not too small, though it is hard to determine where the limit is.
- With regard to searching, smaller documents are also preferable because the search can thus be more easily narrowed down the content which the user really needs, which might not always be the case with documents which are long and multi-thematic.

Therefore, the techniques we describe here are more well-suited for small to medium documents, which is realistic for the document type of a learning object, even if this is not reflected in all of our collections. We are aware of the fact that the individual corpora are rather small and cannot be considered to be representative for the text sort of instructive texts. But we assume that the corpora are large enough to build and test the extraction tools upon them. The evaluation results, about wick we report in section 3.6, prove this assumption.

The texts which we were able to acquire come also in different formats, namely PDF, DOC and HTML. Part of the work was to transform these texts into structurally and linguistically annotated documents. These documents serve as input to the information extraction tools and as resource for the ontology building. Some preprocessing was necessary to unify the different formats. We used third party tools<sup>8</sup>, some auxiliary scripts and modest manual intervention. As result, the text together with some basic structural and layout features is preserved in a project-specific format called BaseXML. This format serves as input to the individual linguistic annotation chains (LAC). In principle, the LAC for each language consists of a sentence and a word segmenter and a linguistic annotator which determines the base form (e.g. *derive* for the word *derives*) and the part of speech (e.g. VERB for the word *derives*), and its morphosyntactic features (e.g. 3RD PERSON SINGULAR PRESENT for the word *derives*). The latter is particularly important for the many morphologically rich languages in the

---

<sup>8</sup> Those were in particular conversion tools which convert the DOC format to text, tools, available under Linux, which convert PDF to text, and online tools on the ADOBE website, cf. [http://www.adobe.com/products/acrobat/access\\_onlinetools.html](http://www.adobe.com/products/acrobat/access_onlinetools.html).

project. For some languages, also noun phrases were detected and marked, as they play a central role as keyword and defined terms<sup>9</sup>.

Figure 1 presents an example of a fully annotated sentence from the German corpus, in the LT4el annotation format (LT4ELAna), which translates to *Write an e-mail!*

```
<par id="p63" name="p"><s id="s114">
  <tok base="schreiben" ctag="VVFIN" id="t1091"
    msd="pl,0,0,third,0,present,0,0" rend="ol,li">Schreiben</tok>
  <tok base="Sie" ctag="PPER" id="t1092"
    msd="pl,bot,nom,third,0,0,0,0" rend="ol,li">Sie</tok>
  <chunk category="NP" id="c247">
    <tok base="eine" ctag="ART" id="t1093"
      msd="sg,fem,acc,0,0,0,0,0" rend="ol,li">eine</tok>
    <tok base="E-Mail" ctag="NN" id="t1094"
      msd="sg,fem,bot,third,0,0,0,0" rend="ol,li,b">E-Mail</tok>
  </chunk>
  ...
</s>
...
</par>
```

**Fig. 1.** LT4ELAna example. Legend: par = paragraph; s = sentence, tok = token; base = lemma; ctag = part of speech; msd = morpho-syntactic description of the word, in the form of a feature vector; rend = layout information.

In the project, we provide a Document Type Definition (DTD) which defines the structural, the layout and the linguistic information of these documents. This DTD, called *Lt4ELAna*, is derived from the widely used XCESAna DTD for linguistic corpus annotation. This guarantees that our annotated corpora will be re-usable in other research projects.

On top of the linguistic annotation, the DTD allows for the markup of keywords and definitions. This has been done manually in the first project phase. At least 1000 keywords and 450 definitions have been identified and marked in the texts. These pieces of information are used for the training of the information extraction tools as well as for their testing and evaluation<sup>10</sup>.

---

<sup>9</sup> We provide more details of the conversion and annotation process in [9], with the German corpus as an example.

<sup>10</sup> The annotated corpora will be made available towards the end of the project, which is May 2008 – at least those documents for which the IPR issues can be cleared.

### 3 Keyword extraction

#### 3.1 Purpose of the tool

As has been said above, one of the aims of the LT4eL project is to improve the retrieval and accessibility of eLearning content through the identification of the learning material by means of descriptive metadata. Since it is not yet current practice for authors to provide keywords, but, on the other hand, effective retrieval of learning objects relies on them, we want to assist authors with the extraction of keyword candidates from their texts. The keyword extractor draws on qualitative and quantitative, in particular distributional, characteristics of good keywords.

#### 3.2 Measuring keywordiness

Good keywords are supposed to represent the topic(s) of a text. They therefore tend to appear more often in that text than could be expected if all words were distributed randomly over a corpus.

A well-established way to measure the distribution of terms over a collection documents is  $tf*idf$ , cf. equation 1.

$$tf * idf \quad \text{where} \quad IDF = \log_2 \frac{N}{df} \quad (1)$$

Another quite useful statistics used to model the expected distribution of words in texts is Poisson distribution or a mixture of Poisson distributions (cf. [4] and equation 2).

$$\pi(k; \theta) = \frac{e^{-\theta} \theta^k}{k!} \quad (2)$$

While the distribution of e.g. function words like *of*, *the*, *it* is close to the expected distribution under the Poisson distribution model, good keyword candidates deviate significantly from the expectation. The score of this deviation can be used as a statistics by which the lexical units are ranked ([5]). The deviation of the observed distribution of a word from the expected distribution under the Poisson model, i.e predicted IDF (cf. equation 3) is called Residual RIDF (short: RIDF).

$$-\log_2(1 - e^{-\theta}) \quad \text{where} \quad \theta = \frac{cf}{N} \quad (3)$$

During our experimenting with these metrics we recognized that RIDF does not take the term frequency in the analysed document into account. Since this is the most important factor in our statistics, we added it and arrived at a statistics which we call Adjusted Residual IDF (short: ADRIDF. cf. equation 4).

$$(IDF - PredictedIDF)\sqrt{tf} \quad (4)$$

The evaluation of the keyword extractor for all languages is described in the section 3.6.

### 3.3 Using linguistic information

The linguistically annotated text provides us with the base form, the part of speech and morphosyntactic features for each word. This information is used to remove words of those categories which are unlikely to be keywords. For most languages, only nouns, some verbs and words marked as unknown are taken into account as keyword candidates. These restrictions are defined in the so-called language models of the keyword extractor. These models can easily be adjusted to new domains or languages.

### 3.4 Multiword keywords

Lexical items which span more than one word, e.g. *learning management system*, *font selection menu*, play an important role as keywords and should therefore be treated as such by the extractor. The manually annotated keywords in our reference texts showed that while for languages like Dutch and German the single word keywords make for more than 90 % of all keywords, the share of multiword keywords is nearly two-thirds for Polish. We therefore put some effort to properly deal with these items. The implementation of the keyword extractor can be parameterized to take multi-word sequences up to a certain length into account, which is useful for e.g. Polish, or to ignore them, which might be good for Dutch and German.

### 3.5 Structural and layout information

Good keyword candidates tend to appear in certain salient regions of a text. These are the headings and the first paragraphs after the headings as well as an abstract or summary. Salient terms might also be highlighted or emphasised by the author, e.g. by using italics. We give an extra weight to terms which show this behaviour.

### 3.6 Evaluation

The quantitative evaluation of the keyword extractor comprised three parts: a) assessing the response time(s) of the tool, b) comparing the output of the tool to the human annotation, and c) an experiment in inter-annotator agreement.

The response times of the tool, i.e. the time it takes to analyse the document and return the keyword candidates, is good enough to use the tool in real time. Once the language model, i.e. information extracted from all analysed documents is loaded into memory, which is done only once, the time needed to extract keywords from one document ranges from 25ms up to 1.5 seconds, depending on the document size. That has been measured on a 1.5GHz Pentium machine with 512MB RAM, with a language model of around 400000 tokens.

In the second part of the evaluation, we compared the automatically extracted and ranked keywords (according to either of the three statistics mentioned above) with the manually marked keywords. For a document where  $n$

keywords have been marked manually (with  $n > 5$ ), we selected the  $n$  best keywords according the ranking of the keywords and recorded the overlap. From this evaluation across all languages, the three statistics, and different maximum lengths of keywords it could be observed that:

- Results varied significantly from slightly more than 40 % overlap at average for the German documents to more than 60 % overlap for the Czech documents;
- In general, tf\*idf and ADRIDF nearly produced the same results, with one outperforming the other on one language, and vice versa for another language, while RIDF performed worst for almost all settings;
- Results improved for all languages if multi-word keywords up to a length of 3 words were included. Using keywords of even higher length improved the results slightly for few languages (e.g. Bulgarian) and decreased results for most other languages. Therefore, including keywords up to a length of three words seems to be the best decision.

This part of the evaluation, however, relies completely on the quality of the manual keyword selection which seems to be good for some and less good for other languages. In order to control this experiment, an evaluation of inter-annotator agreement (IAA) on the keyword selection task has been performed. For each language, a group of at least 12 persons selected keywords from the same document, a document of modest size. We used kappa statistics to measure IAA, following the approach of Bruce and Wiebe ([3]), which seems to be appropriate for our type of data.

Table 1 presents the results of the inter-annotator agreement for each language:

Language	average human annotators	Keyword Extractor
Bulgarian	0.2008	0.0683
Dutch	0.2150	0.1373
English	0.1318	0.08964
German	0.2636	0.13208
Polish	0.1996	0.1651
Portuguese	0.1811	0.0893
Romanian	0.2102	0.215784

**Table 1.** Inter-annotator agreement for human annotators and for the Keyword Extractor compared to the human annotators

Results of these experiments reveal that the inter-annotator agreement for this task is low for all languages, indicating that the task of selecting keywords cannot be well defined. The average IAA for the annotators ranged between 0,1 and 0,4. Neither is there a significant difference between languages, nor between unexperienced and experienced annotators. The keyword extractor was at the

lower end of this scale for all languages except for Romanian, so there is space for improvement. The generally low IAA might have consequences for our search scenarios though. Documents which are assigned keywords of a wide variety might also be searched by such a wide variety of search terms.

## 4 Ontological support of searching

Current eLearning systems offer only full-text or keyword-based search facilities. We will outline in this section the steps we took in order to implement an ontology search facility and also describe planned extensions for crosslingual search. We will explain the methodology for ontology creation, the semantic annotation of learning objects with concepts from the ontology, and show how lexical items of various languages have been mapped onto the concepts of the ontology.

### 4.1 Ontology creation

The domain chosen for our ontology is computer science for non-computer scientists. Details about the choice of the domain and the related documents are given in section 2. Since we were not able to find any ontology which covers our domain in a satisfactory way we proceeded with the creation of our own.

During the creation of the ontology we made sure that the ontology covers (most of) the topics of our learning objects well and in great detail. We used the keywords which have been manually annotated, as well as the results of the keyword extractor which has been described above. A fine granularity of the concepts is required in order to ensure better text annotation. In addition to this, the ontology was aligned with an upper ontology in order to ensure consistency with respect to a general ontology development methodology.

The creation of the ontology can be summarized in the following steps: Processing of keywords and formalisation. On a later stage of ontology development new concepts (not related to the keywords) were added in order to improve the coverage of the domain.

**Processing of keywords** In order to ensure a relatively wide coverage with respect to the learning objects we based the construction of the ontology on the keywords which have been annotated in these documents. The ontology is therefore based on lexical items from all the languages of the project. English translation equivalents were provided for these lexical items. This reduced the complexity of mapping the concepts of the ontology to lexical entries of all languages. The processing itself was done in the following way. First of all, the keywords were classified into the conceptual space of the domain. Only those keywords which are connected to the subject of information technology for non-experts were selected.

Secondly, definitions for the concepts were collected. The WWW was searched for definitions of the selected keywords. The rationale behind this step was to

define in human readable way the concepts connected to the keywords. We collected a set of multiple definitions for most of the keywords because different definitions highlight different features of the related concept.

Thirdly, the terms were disambiguated and a canonical definition chosen for each meaning. These definitions are the human-readable meaning explications of the concepts which are included in the ontology. When necessary, two or more meaning explications were given for a concept. For instance, the terms *header* and *word* have more than one meaning in our domain. Other keywords show regular polysemy. For example, *MPEG* might signify an organization as well as a standard. Our rule of thumb was to prefer the more general meaning to the more specific ones. For example, we skipped the meanings indicating programming terms because we considered them to be too specific.

**Formalisation** In this step formal definitions of the extracted concepts and relations have been defined using OWL-DL (cf. [2]). OWL-DL was chosen due to the availability of reasoners for that language subset. The concepts were formalised in two separate steps. First, for each meaning, an appropriate class in the domain ontology was created. The result of this step is an initial formal version of the ontology. In order to ensure appropriate taxonomic relations between the concepts in the ontology and to facilitate the mapping to an upper ontology, each concept was mapped to synsets in the OntoWordNet version of Princeton WordNet ([6], [7]), which is a version of WordNet which is mapped to the DOLCE ontology. The mapping was performed via the two main relations *equality* and *hypernymy*. The first relation is between a class in the ontology and a synset in WordNet which (lexically) represents the same concept, while the second is a relation between a class in the ontology and a synset denoting a more general concept. Thus, the taxonomic part of the ontology was created. The connection of OntoWordNet to DOLCE allows an evaluation of the defined concepts with respect to meta-ontological properties as they are defined in the OntoClean approach (cf [8]).

## 4.2 Annotation of learning objects with concepts

From the perspective of the Learning Management System, the ontological annotation concerns only the metadata section of the learning objects. In the metadata, according to the Learning Object Metadata (LOM, [1]) standard, some ontological information can be stored and used later on to index the learning objects for retrieval. The annotation needs not be anchored to the content of the learning object. The annotator of the learning object can include in the annotation all concepts and relations she considers to be important for the classification of the learning object. In order to accurately link a learning object and/or its parts to the proper places in the conceptual space of the ontology, an inline annotation of the content of learning objects is an obligatory intermediate step in the annotation of the learning objects with ontological information. The inline annotation is done by regular grammar rules attached to each concept in the

ontology reflecting the realizations of the concept in texts of the corresponding languages. Additionally rules for disambiguation between several concepts are applied when a text realization is ambiguous between several concepts.

Within the project we performed both types of annotation, inline and through metadata. The metadata annotation is used during the retrieval of learning objects from the repository. The inline annotation will be used in two ways: (1) as a step to metadata annotation of the learning objects; and (2) as a mechanism to validate the coverage of the ontology. Additionally we have implemented a ontology-oriented search engine based on the full-text search engine Lucene<sup>11</sup>. It allows searches for documents, paragraphs or sentences that contains annotations of some concepts from the ontology. These searches provide a basis for detailed requests by the users in order to find the appropriate learning object for their needs.

### 4.3 Mapping lexicons onto the ontology

Terminological lexicons represent the main interface between the user's query and the ontological search engine. The terminological lexicons were constructed on the basis of the formal definitions of the concepts within the ontology. In this approach of construction of the terminological lexicon we escaped from the hard task of mapping different lexicons in several languages as it was done in EuroWordNet Project [10]. The main problems with this approach of construction of terminological lexicons are that (1) for some concepts there is no lexicalized term in a given language, and (2) some important term in a given language has no appropriate concept in the ontology which to represent its meaning. In order to solve these problems we, first, allow the lexicons to contains also non-lexicalized phrases which have the meaning of the concepts without lexicalization in a given language. Even more, we encourage the lexicon builders to add more terms and phrases to the lexicons for a given concept in order to represent as many ways of expressing the concept in the language as possible. These different phrases or terms for a given concept are used as a basis for construction of the regular grammar rules for annotation of the concept in the text. Having them, we could capture in the text different wordings of the same meaning. In order to solve the second problem we modify the ontology in such a way that it contains all the concepts that are important for the domain.

We could summarize the connection between the ontology and the lexicons in the following way: the ontology represents the semantic knowledge in form of concepts and relations with appropriate axioms; and the lexicons represent the ways in which these concepts can be realized in texts in the corresponding languages. Of course the ways in which a concept could be represented in text are potentially infinite in number, thus, we could hope to represent in our lexicons only the most frequent and important terms and phrases.

Here is an example of an entry from the Dutch lexicon:

```
<entry id="id60">
```

<sup>11</sup> Apache Lucene is a full-featured text search engine: <http://lucene.apache.org/>

```

<owl:Class rdf:about="http://www.lt4e1.eu/CSnCS#BarWithButtons">
  <rdfs:subClassOf>
    <owl:Class rdf:about="http://www.lt4e1.eu/CSnCS#Window"/>
  </rdfs:subClassOf>
</owl:Class>
<def>A horizontal or vertical bar as a part of a window,
  that contains buttons, icons.</def>
<termg lang="nl">
  <term shead="1">werkbalk</term>
  <term>balk</term>
  <term type="nonlex">balk met knoppen</term>
  <term>menubalk</term>
</termg>
</entry>

```

Each entry of the lexicons contains three type of information: (1) information about the concept from the ontology which represent the meaning for the terms in the entry; (2) explanation of the concept meaning in English; and (3) a set of terms in a given language that have the meaning expressed by the concept. The concept part of the entry provides minimum information for formal definition of the concept. The English explanation of the concept meaning facilitates the human understanding. The set of terms stands for different wordings of the concept in the corresponding language. One of the terms is the representative for the term set. This representative term will be used where just one of the terms from the set is necessary to be used, for example as an item of a menu. In the example above we present the set of Dutch terms for the concept <http://www.lt4e1.eu/CSnCS#BarWithButtons>. One of the term is non-lexicalized - attribute `type` with value `nonlex`. The first term is representative for the term set and it is marked-up with attribute `shead` with value 1.

## 5 Integration into the ILIAS Learning Management System

### 5.1 Purpose and method of integration

The primary focus of the integration of the LT functionalities into ILIAS is to build a running prototype of a learning management system that provides extended functionalities supported by the use of the language technology tools. The basis for the integration process are use cases which have been defined for the keyword extractor and the ontology enhanced searching and browsing capabilities. The use cases have been the major input for the specification of a web service interface between the language technology tools and the learning management system. It is a major goal of the project to make the language technology based functionalities re-usable for other learning management systems. To make the integration of the tools as easy as possible, the interface of the tools will be well-documented and standards-based. The implementation of the interface as web services should ensure that these goals are met.

## 5.2 Integration Setup

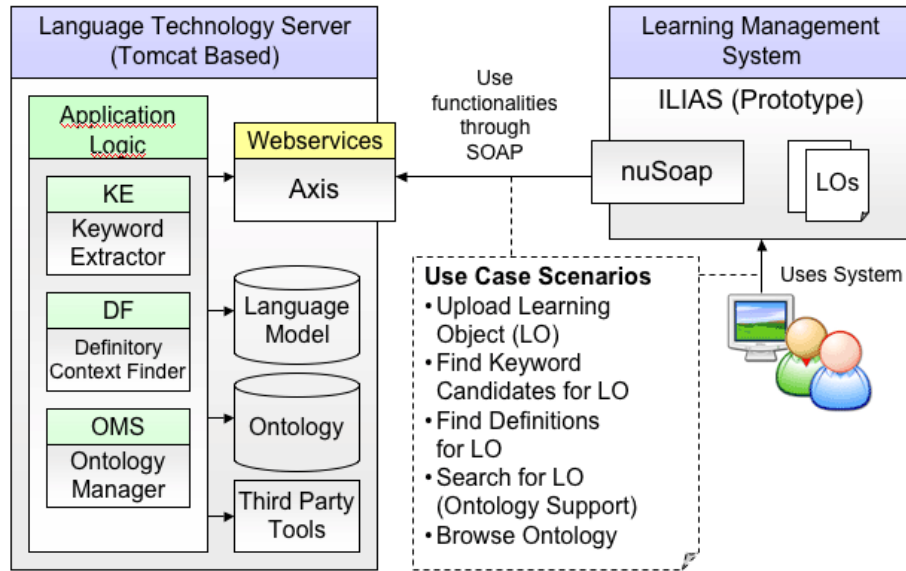


Fig. 2. Architecture of the Language Technology enhanced LMS

Figure 2 shows the major components of the integration setup. The language technology server on the left provides the keyword extractor, definitory context finder and ontology management system functionalities. The tools are developed using the Java programming language and are hosted on a Java web server. The functionalities can be accessed directly on the webserver for test purposes or they can be used by the learning management system through the web service interface.

The fact that multiple developers are working on different parts of the overall structure has led to the decision to setup a Subversion server as a central code repository. The project partners have also decided to make the results immediately available to the general public and to give everyone the opportunity to join and collaborate with the project. The source code is available under an open source licence and it is hosted on the SourceForge portal for open source projects at <https://sourceforge.net/projects/lt4e1/>. Figure 3 shows the first integration of the keyword extractor into the ILIAS learning management system. The function is embedded into the existing LOM metadata handling of ILIAS to enable a semi-automatic generation of keywords. In the future the definitory context finder operations will be used to provide the new functionality of semi-automatic glossary generation within the ILIAS authoring environment. The web service operations of the ontology management system will extend the



**Fig. 3.** User interface to the Keyword Extractor

functionalities of the ILIAS search function by enabling semantic and multilingual retrieval of learning objects<sup>12</sup>.

## 6 Validation of the tools and value added

The tools described in this paper are currently tested and validated with real users. For this purpose we designed validation scenarios. According to the different functionalities of the system the validation scenarios can be classified as either being monolingual or bilingual.

**monolingual scenario** From the teacher's perspective, the keyword extractor will be used to select and add keywords to a new document. The ontology can be used to find related keywords which might not have been appeared in the document. From the student's perspective, the ontology and keyword based searching will be used to elicit information from a set of learning objects, e.g. for answering a quiz.

<sup>12</sup> A demonstration of these functionalities will be given during the talk.

**bilingual scenario** Teachers as well as students will use the ontology to retrieve contents and terms in other languages than their own native language. Texts of different languages can be combined from multilingual learning packages.

These very general, high-level scenarios will be detailed further to get exact instructions for the test persons to follow. We will measure the added value of the language technology by comparing the outcome of the tasks with and without them as well as by evaluating the satisfaction of users with the new functionalities.

## 7 Further work and perspectives

The main objective of the LT4eL project is the Integration of Language Technology resources and tools in eLearning which should enhance the search and retrieval of (multilingual) learning material. In order to reach this objectives, we have:

- created a corpus of 200.000 words (1000 pages) of learning objects for all languages of the consortium;
- normalized and converted the corpus in XML;
- annotated the corpus with PoS for all languages;
- developed a keyword extractor based on different statistical measures;
- carried out a preliminary quantitative evaluation of the keywords extracted;
- developed an ontology of about 1000 concepts in an upper ontology and domain ontology in Computer Science for non experts);
- linked the ontology to OntoWordNet;
- semantically annotated learning objects in various languages on the basis of the ontology;
- developed lexica for all the languages to be mapped to the ontology.

One of the major challenges for the future is to make the Linguistic Annotation Chain available for each new document which is submitted. Currently the tools and search can only be applied to our corpora. Intellectual property rights on the annotation tools have to be solved for some languages before we can offer this service.

In parallel with the validation of the monolingual ontological search we are currently working on multilingual search. The main assumption is that users are able to read documents in languages other than their native language (usually at least in English), but traditional search engines will not find them. In the first phase of the project domain-related lexicons were created in all eight languages represented in the project. The lexicons were mapped on the ontology and they provide the interface between the users's query and the search engine. Further work consists in implementing the multilingual engine and validating the multilingual scenarios, as described in the previous section. The extension from monolingual search to multilingual search raises additional issues like:

- Ranking of documents over the various languages, as the user may be less interested in receiving one separate list of documents for each language. Another option is to display the complete list according to the same ranking criteria, and for each document indicate its language. In this way, the user can compare the relevance of two documents even if they are in a different language. A further refinement could be to include the language as a ranking criterion by giving a bonus which differs per language.
- Introducing parameters like possible languages of search query, retrieval languages, etc. to the search functionality.
- The inclusion of other ontologies is also still an issue. We are currently investigating the possibility of introducing relations corresponding to some pedagogical criteria.

In short, there is much room for improvement in the keyword and ontology driven annotation and search once the basic resources and tools are in place. Currently, we can provide these resources and employ the new functions. Other researchers and developers in the field of technology-enhanced learning are invited to join these efforts and to profit from our achievements.

## References

1. Standard for Learning Object Metadata, Final Draft, IEEE, 2002 – P1484.12.1
2. OWL. Web Ontology Language (Overview). <http://www.w3.org/TR/owl-features/>
3. R. Bruce and J. Wiebe. 1999. *Recognizing subjectivity: A case study of manual tagging*. In: Natural Language Engineering. Vol. 5, No 2, pp 187–205.
4. K. Church and W. Gale. 1995. *Poisson mixtures*. In: Natural Language Engineering. Vol. 1, No 2, pp 163–190.
5. K. Church, W. Kenneth and W. Gale. 1995. *Inverse Document Frequency (IDF): A Measure of Deviations from Poisson*. In: Proc. of Third Workshop on Very Large Corpora.
6. Christiane Fellbaum. 1998. Editor. WORDNET: an electronic lexical database. MIT Press.
7. Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The OntoWordNet Project: extension and axiomatisation of conceptual relations in WordNet. International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2003), Catania, Italy.
8. Nicola Guarino and Christopher Welty. 2002. "Evaluating Ontological Decisions with OntoClean." *Communications of the ACM*, 45(2): 61-65.
9. Eelco Mossel and Lothar Lemnitzer and Cristina Vertan. Language Technology for eLearning – A Multilingual Approach from the German Perspective. Proc. GLDV-2007 Spring Conference. Tübingen, April 2007, pp. 125–134
10. Vossen Piek. 1998. Editor EuroWordNet General Document. Version 3, Final, July 19, 1999, <http://www.hum.uva.nl/ewn>
11. F. Sclano and P. Velardi, TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities. To appear in Proc. of the 3rd International Conference on Interoperability for Enterprise Software and Applications I-ESA 2007, Funchal, Madeira Island, Portugal, March 28-30th, 2007.