



Project no. 027391

Project acronym: LT4eL

Project title: Language Technology for eLearning

Instrument Specific Targeted Research Project

Thematic Priority Information Society Technology

### **D3.1c Ontology validation on the basis of (multilingual) search – second cycle**

Due date of deliverable: 31-05-2008

Actual submission date: 08-07-2008

Start date of project: 1-12-2005

Duration: 30 Months

Organisation name of lead contractor for this deliverable: Bulgarian Academy of Sciences (IPP-BAS)

Revision [1]

| <b>Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)</b> |   |   |
|--|---|---|
| <b>Dissemination Level</b>   |   |   |
| <b>PU</b>  | Public  | x |
| <b>PP</b>  | Restricted to other programme participants (including the Commission Services)        |   |
| <b>RE</b>  | Restricted to a group specified by the consortium (including the Commission Services) |   |
| <b>CO</b>  | Confidential, only for members of the consortium (including the Commission Services)  |   |

# D3.1c

## Contents

- 1 Title
- 2 Summary
- 3 Introduction
- 4 Ontology Cleaning and Enrichment (wrt validation)
- 5 Formal evaluation of ontology semantic search vs. textual search (future work)
- 6 Mind Map
- 7 Motivation of Ontology as Interlingua
- 8 Tool for concept-LO tables
- 9 Conclusions
- 10 References
- 11 Scientific papers on the ontology, lexicons and annotation of learning objects

## 1 Title

### D3.1c Ontology validation on the basis of (multilingual) search – second cycle

## 2 Summary

During the second year of the project we developed an architecture for annotation of learning objects with concepts from the ontology. This architecture includes: *lexicons aligned to the ontology* and *concept annotation grammars*. The senses of the lexical items are expressed in the ontology. The annotation grammars encode the lexical items in the lexicons and the syntactic features of the corresponding language. On the basis of the defined architecture, during the second year, we have performed the following tasks: *Creation of Lexicons, which covered the whole ontology and all the languages (except for Maltese); Creation of Concept Annotation Grammars for all the languages; Semantic annotation of Learning objects with concepts; Formal evaluation of ontology-based semantic search vs. textual search; Crosslingual search and integration within ILIAS; Ontology revision.*

The main task for the last six months of the project was to support the evaluation of the new functionalities integrated in ILIAS.

## 3 Introduction

In this deliverable we describe the work done with respect to the above mentioned main task. It was guided by the requirements of the evaluation process. In order to achieve a better demonstration of the new functionalities and to have a better coverage of the domain we had to extend the ontology with new concepts and with relations. Thus, in this second cycle of ontology development and evaluation we repeated the same steps from the first cycle of evaluation, but this time on the new feedback from the experiments and the new usage scenarios. In general, we followed further the steps below:

- Enrichment of the Lexicons, which covered the whole ontology and all the languages;
- Enrichment of Concept Annotation Grammars for all the languages;
- Semantic re-annotation of Learning objects with concepts;
- Crosslingual search and integration within ILIAS.
- Ontology revision;

Additionally, a semi-manual disambiguation was performed over the semantically annotated files.

The resulting ontology contains 1002 domain concepts, 169 concepts from OntoWordNet and 105 concepts from DOLCE Ultralite. It also contains 107 object properties. Thus, the difference with the previous version is as follows: the additions concerned mostly the domain concepts and the relations (such as `part_of`, `used_for`).

Also, in order to support a better observation on the annotation of the learning object we have implemented a system for constructing a table which demonstrates the distribution of the concept annotation within the learning object. In this way we supported the construction of the evaluation experiments.

## **4 Ontology Cleaning and Enrichment (wrt validation)**

As a consequence of the evaluation of the ontology - the coverage check with respect to the LOs annotation, the feedback from the partners and the alignment to the upper ontology, we revised the ontology by focusing on the following tasks: better coverage of the domain and simplifying the upper part of the ontology.

### **■ Ontology covering of the domain**

The covering of the domain was extended on the basis of the annotation of the learning objects. All the partners inspected the semantic annotated LOs for missing annotation. The suggestions for new concepts were collected and added to the ontology. Also, new concepts were added on the basis of addition of missing siblings of already presented concepts. In this way more than 240 new domain concepts were added to the ontology.

In addition to these concepts suggestions for new extensions with concepts came from the preparation and testing of the user scenarios. In order for the scenarios to be supported by the necessary semantic information, we have added about 50 concepts.

Thus, in contrast to the previous phases, in which we preferably relied on the prompts from the keyworded learning objects, in this third phase the feedback came from a real application and user-centred point of view. The main feedback concerned the granularity of the hierarchical structure within ontology. As a result, the structure became deeper and more detailed.

### **■ Restructuring of the upper part of the ontology**

In parallel to addition of new concepts we also did further simplification of the upper part of the ontology. This simplification was mainly done via compacting the unary branches in the hierarchy. These unary branches contained concepts necessary for the structuring of OntoWordNet, but unnecessary for our ontology.

Restructuring was done on the level of the domain ontology where some new concepts were introduced in order to keep the structure of the ontology balanced with respect the generalizations over the most specific concepts. This action aimed at making the ontology more comprehensible to the users.

## **5 Formal evaluation of ontology semantic search vs. textual search (future work)**

In the previous version of the deliverable we have reported on an experiment for comparing textual search with the ontology semantic search. The experiment was done by formulating the same query in terms of concepts from the ontology and terms from the lexicons. Then on the query formulated in terms of concepts from ontology query expansion operation was performed. In this case it included only addition of one superconcept and all subconcepts of the concepts from the query. For example, for the concept "program" the concept "software" was added and all the kinds of programs represented in the ontology were also added. In the case of the textual search a very limited query expansion was applied: the terms were lemmatised and some better known terms were added. Thus, in case of the terms for "program" the terms for "software" and "editor" were also added. The results were very attractive.

As it was pointed out by the reviewers, the comparison was not very fair, because there exist more

advanced textual search techniques. Thus, we performed a study on the different kinds of textual search [1]. Unfortunately, we did not find a free system which to support the most advanced textual search. We also had no resources within the project to implement such a system by ourselves. Therefore, our conclusion is that the experiments we did are very preliminary and the experiment has to be re-run with a better textual search, but this work is not possible to be done with the resources of the project.

Still the results from the experiments are encouraging to proceed with the research and application of the ontology semantic search.

## 6 Mind Map

We thank the reviewers for their suggestion we to use Mind Map as a visualization tool: "The consortium should consider advanced forms of search result visualisation such as WikiMindMap (see, e.g., [http://www.wikimindmap.org/viewmap.php?wiki=en.wikipedia.org&topic=natural\\_language\\_processing](http://www.wikimindmap.org/viewmap.php?wiki=en.wikipedia.org&topic=natural_language_processing))"

We considered this suggestion with respect to visualization not only of the search result, but also with respect to the other resources used in the system, such as ontologies, keyword prompts, etc. We concluded that Mind Map is very attractive for these purposes, but it was not feasible during the last period of the project, because the available systems for Mind Map can not be used directly. They need to be tuned to the representation of resources in LT4eL. This task required more work time than we had available on the project. Therefore we have decided to exploit Mind Map technology within the new European project LTfLL (<http://www.ltfll-project.org/>) in which we will develop further the technologies we have created within LT4eL.

The general idea in the new project is to define a common semantic framework which to include two types of elements - resources and tools. Resources include ontologies, lexicons, learning materials, communication notes, etc. Tools are resource specific and general tools. Resource specific tools are inference engine over ontology, lexicon management tools, learning material annotation tools, for example. General tools are XML search engine, different types of editors, etc. Here Mind Map will be used for visualization of the content of the resources (each resource will have a specific view as a mind map), visualization of the relations among the resources, organization of the personal learning space, etc.

## 7 Motivation of Ontology as Interlingua

Here we provide some answers to the following remark of the reviewers: "The decision to use the ontology as an interlingua for the cross-lingual tasks is not well justified; the text should clearly state why this approach has been followed."

There are at least three motivations for using ontologies as interlingua among several languages:

- **Existence of ontologies.** In many domains there exist ontologies which are created for other purposes and they could be reused in cross-lingual tasks.
- **Addition of a new language.** As interlingua, the ontology facilitates the addition of a new language. This addition would require only a mapping from the ontology to the lexicon of the language in question, and this would automatically ensure a mapping from the new language to all other languages, already included in the system (this was done for Maltese, for example).
- **Reasoning.** Ontology supports reasoning which provides better searching in documents annotated with ontology information. If the ontology is not used in the mapping between languages, then reasoning will not be available during the cross-lingual search. This could reduce the precision and recall of the search.

## 8 Tool for concept-LO tables

We implemented a tool that creates tables for manual inspection of the concept-LO relations. It makes use of the search functionality. It consists of the following classes:

- eu.lt4el.search.IndexLOsForTables.java
- eu.lt4el.search.MakeConceptDocTables.java

For each language, the tool creates three different flat representations (table in text format) of the concept-document relations. For the LT4eL domain concepts, the name spaces are removed, e.g. #HTMLPage instead of <http://www.lt4el.eu/CSnCS#HTMLPage>). Concepts from the upper ontology keep their name spaces.

### **First result file: Concept to Document**

("which docs are covered by a certain concept?", sorted by concept, then by document title)

Example lines:

```
#Abbreviation <tab> 3x <tab> swp2ontGuidelines_for_Writing_a_Scientific_Paper.xml
#Abbreviation <tab> 1x <tab> swp2ontHow_ICT_Can_Create_New__Open_Learning_Environments.xml
```

### **Second result file: Document to Concept**

("which concepts are contained in a certain doc?", sorted by document title, then by #occurrences)

Example lines:

```
swp2ontHow_ICT_Can_Create_New__Open_Learning_Environments.xml <tab> 4x <tab> #WebSite
swp2ontHow_ICT_Can_Create_New__Open_Learning_Environments.xml <tab> 3x <tab> #Acronym
swp2ontHow_ICT_Can_Create_New__Open_Learning_Environments.xml <tab> 3x <tab>
#AlphabeticCharacter
```

### **Third result file: Document <tab> Concept <tab> All super concepts of it**

(super concepts can be used to retrieve the document too, because of query expansion with sub concepts)

Example line:

```
swp2onteng09_Power_Point_Presentations.xml <tab> 51x <tab> #View <tab> supers:[
http://www.loa-cnr.it/ontologies/IOLite.owl#DigitalResource
http://www.loa-cnr.it/ontologies/DUL.owl#InformationRealization
http://www.loa-cnr.it/ontologies/DUL.owl#Entity ]
```

## **9 Conclusions**

This deliverable describes the final version of the LT4eL ontology within the project in the domain of the computer science for non-computer scientists. The ontology was constructed in first place to support the annotation of the learning objects collected by all the partners. However, it was also used for ontology browsing and for semantic search. For that reason, we extended it with respect to even distribution of the concepts in the hierarchy. The ontology was aligned to an upper ontology - DOLCE Ultralite. The user can navigate over the ontology via the lexicon in some specific language. Thus, the user can never see the actual encoding of the ontology which is used only for inference.

## **10 References**

[1] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. 2008. An Introduction to Information Retrieval. Cambridge University Press Cambridge, England. Preliminary draft (c) 2008 Cambridge UP.

## **11 Scientific papers on the ontology, lexicons and annotation of learning objects**

The following papers have been published and presented on conferences. The full papers are in the appendix. In the following, the bibliographical data are listed.

- Paola Monachesi, Kiril Simov, Eelco Mossel, Petya Osenova and Lothar Lemnitzer. What ontologies can do for eLearning. Presented at: IMCL 2008 (<http://www.imcl-conference.org/>).

# What ontologies can do for eLearning

P. Monachesi<sup>1</sup>, K. Simov,<sup>2</sup> E. Mossel,<sup>3</sup> P. Osenova,<sup>4</sup> L. Lemnitzer,<sup>5</sup>

<sup>1</sup> Utrecht University, Utrecht, The Netherlands

<sup>2,4</sup> IPP Bulgarian Academy of Science, Sofia, Bulgaria

<sup>3</sup> University of Hamburg, Hamburg, Germany

<sup>5</sup> University of Tübingen, Tübingen, Germany

**Abstract—** In this paper, we discuss the role that ontologies, a key element in the Semantic Web vision, can have within eLearning and how they can help improve the learning process. We take the LT4eL project as test case, since in this project ontologies play a crucial role in enhancing the management, distribution and retrieval of the learning material within a Learning Management System (LMS). We also sketch a potential use of ontologies in facilitating social learning.

**Index Terms—** ontologies, technology enhanced learning, information extraction, information retrieval.

## I. INTRODUCTION

The World Wide Web contains an enormous quantity of information but this information is designed for human consumption. The Semantic Web vision, however, is working towards an automatic computer-based processing of information on the Web, bringing thus a change to this situation [1]. More specifically, in order to make use of the information on the Web, we need to be able to interpret the large collection of available facts in the context of knowledge. Information on the web needs thus to be supplemented with *semantic* knowledge. This can be achieved by making the meaning of documents on the Web explicit and ontologies play a crucial role in this vision.

Ontologies allow for a representation of knowledge that enables inference to be performed obtaining thus new insights. Representing knowledge in the form of a conceptualization (i.e. ontologies) is crucial for the automatic processing of the information on the Web. However, ontologies can also enhance the management, distribution and retrieval of the learning material within a Learning Management System (LMS) and can thus play a relevant role in eLearning.

The aim of the LT4eL project ([www.lt4el.eu](http://www.lt4el.eu)) is to integrate the results of the research carried out in the Semantic Web area, as well as in the Language Technology area to enhance eLearning in order to develop innovative applications for education and training [2]. Our aim is to improve the retrieval of static and dynamic content by employing Language Technology resources and tools for the semi-automatic generation of descriptive metadata while semantic

knowledge is integrated to enhance the management, distribution and search of the learning material [3]. The integration of technology based functionalities will facilitate the construction of user specific courses, will allow direct access of knowledge, will improve the creation of personalized content and will support decentralization and co-operation of content management. The LT4eL project is rather innovative in this respect, even though some other approaches in this direction are emerging, as attested by [32] and [33].

Ontologies play a relevant role in the realization of these objectives since they can be employed to query and to navigate through the learning material supporting thus the learning process. The relevant concepts which are attested in the learning objects constitute the backbone of the ontology. Thus, a link is created between the learning material and its conceptualization which is represented by means of the ontology allowing for the creation of individualized learning paths. Ontologies allow for the possibility to develop a more dynamic learning environment with better access to specific learning objects.

In particular, in the LT4eL project, ontologies are employed in order to:

- improve the reuse of learning objects available within a Learning Management System;
- facilitate access to objects in various languages since the ontology plays the role of an interlingua which mediates at the conceptual level among language specific textual realizations of the concepts.

In the LT4eL project, we take two groups of users into account:

- Tutors/content providers who want to compile a course for a specific target group and who want to draw on existing texts, media etc.;
- Learners who are looking for contents which suit their current needs, e.g for self-guided learning.

Ontologies, however, can play a relevant role not only in improving the retrieval of learning material but they can also be exploited to facilitate social learning, if certain extensions are envisaged. Even though ontologies

provide a formal representation of the content of the learning material, which is crucial for its retrieval and its reuse they might not include the appreciation people express for different kinds of materials.

In order to support social and informal learning, it is thus necessary to create a link between the formal representation of a given domain in the form of ontologies and the informal descriptions produced by social tagging (i.e. folksonomies) [4]. It is through tagging that learners with similar interests and preferences can be identified.

Ontologies can have an impact in real life eLearning applications (and within the Semantic Web initiatives) if the problem of the knowledge acquisition bottleneck is solved. It is too costly to develop (domain) ontologies manually and therefore a semi-automatic approach should be envisaged. We believe, however, that the availability of techniques and tools in the Natural Language Processing area as well as in the Semantic Web area are providing a valuable contribution to the solution.

The structure of the paper is as follows: in the next section, we give an overview of the ontology creation process carried out in the LT4eL project as well as the roles of the language specific lexica and the semantic annotated objects in allowing for cross-lingual retrieval; section 3 discusses the semantic search engine and the way it facilitates the retrieval of learning material also across languages. Section 4 discusses the role of ontologies within eLearning. In section 5, we sketch a novel use of ontologies within eLearning to support social learning, while section 6 focuses on the automatic development of ontologies. The paper ends with some conclusions.

## II. ONTOLOGY DEVELOPMENT FOR ELEARNING

As already mentioned, the aim of the LT4eL project is to improve the retrieval and the usability of (multilingual) learning material within a Learning Management System. In order to achieve this objective, an ontology-based search functionality has been developed which is based on the following components:

- collection of (multilingual) documents which constitutes the corpus of learning objects;
- a (language independent) ontology that includes a domain ontology in the domain of the learning objects. A domain ontology consists of a set of concepts, belonging to the same domain, and various kinds of relations between the concepts;
- a lexicon for each of the languages addressed which comprises words or phrases that are mapped to concepts attested in the ontology;
- a collection of learning objects annotated on the basis of the concepts attested in the domain ontology.

In the rest of this section, the various components are described in detail and it will be discussed how the various components contribute to the development of the semantic search functionality.

### A. Collection of learning objects

The development of the semantic search functionality is based on domain specific corpora for the various languages addressed in our project, that is Bulgarian, Czech, Dutch, English, German, Polish, Portuguese and Romanian. 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 in the area of computing and the corpora include, mainly introductory texts and tutorials for word processing, HTML etc., texts which address 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 corpus (<http://www.calimera.org/>) because it is close to our domains. 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 educational texts. However, we assume that the corpora are large enough to build the functionalities developed in the LT4eL project (cf. the evaluation results in [3]).

The texts which are part of our corpus are in different formats, namely PDF, DOC and HTML. We have transformed these texts into structurally and linguistically annotated documents which serve as input for the ontology building. Some preprocessing was necessary to unify the different formats. We used third party tools, 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. 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 reusable 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. Around 1000 keywords have been identified and marked in the texts. This information has been used for the creation of a language independent keyword extractor [5] as well as for the development of our domain ontology.

### B. The LT4eL domain ontology

The domain of the corpus assembled within the LT4eL project is that of *computing* and the main application of the domain ontology we have created is related to the indexing of the learning objects within this domain.

The ontology has been developed on the basis of the manually annotated keywords in the eight languages of the project which have been translated into English. These terms have been augmented with definitions which have been collected by searching the web, in this way, it has been possible to represent the various meanings of the terms. On the basis of these terms, the relevant concepts

have been created which constitute the backbone of the domain ontology (cf. figure 1).

The domain ontology has been mapped to an upper

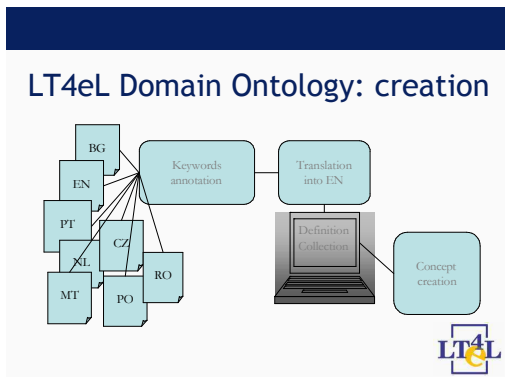


Figure 1: overview of the ontology creation process

ontology (in our case we used DOLCE – [6], [7]) in order to inherit the knowledge already encoded in the upper ontology including relations. In addition, we ensure an appropriate classification of the concepts with respect to concept meta properties that are defined in OntoClean [8] (i.e. the ontology creation methodology). The mapping to the upper ontology involved OntoWordNet [9] which is a version of WordNet [10] restructured in accordance to DOLCE (cf. figure 2).

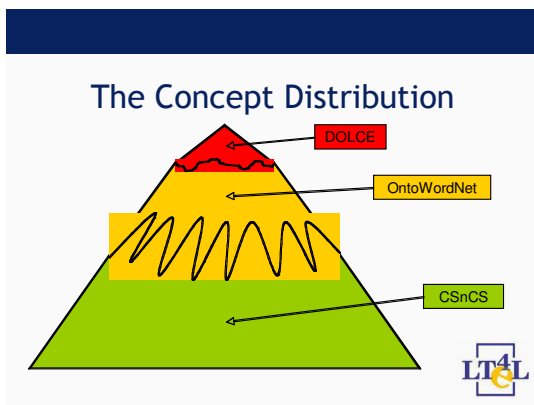


Fig.2 The LT4eL ontology

The ontology has been extended with additional concepts taken from:

1. the restriction on already existing concepts (for example, if a *program* has a *creator*, the concept for *program creator* is also added to the ontology);
2. superconcepts of existing concepts (if the concept for *text editor* is in the ontology, then we added also the concept of *editor* (as a kind of program) to the ontology);
3. missing subconcepts (if *left margin* and *right margin* are represented as concepts in the ontology, then we add also concepts for *top margin* and *bottom margin*).

4. the annotation of the learning objects has also played a relevant role in the extension of the ontology. If a concept is represented in the text of a learning object and it is relevant for the search within the learning material, we have added the concept to the ontology.

The current version of the ontology contains about 750 domain concepts, about 50 concepts from DOLCE and about 250 intermediate concepts from OntoWordNet.

### C. The LT4eL lexicon

For each language represented in the project, we have developed a lexicon on the basis of the existing ontology. The lexicons represent the main interface between the user's query, the ontology and the ontological search engine.

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

```
<entry id="id60">
<owl:Class rdf:about="lt4el:BarWithButtons">
<rdfs:subClassOf>
<owl:Class rdf:about="lt4el:Window"/>
</rdfs:subClassOf>
</owl:Class>
<def>A horizontal or vertical bar as a part of a window,
that contains buttons, icons.</def>
<term lang="nl">
<term sheaf="1">werkbalk</term>
<term>balk</term>
<term type="nonlex">balk met knoppen</term>
<term>menubalk</term>
</term>
</entry>
```

Each entry of the lexicons contains three types of information:

1. information about the concept from the ontology which represents the meaning for the terms in the entry;
2. explanation of the meaning of the concept in English;
3. a set of terms in a given language that have the meaning expressed by the concept.

The concept part of the entry provides the information for the formal definition of the concept. The English explanation of the meaning of the concept facilitates the human understanding. The set of terms represent different wordings of the concept in the given language. One of the terms is the one representative for the term set. Note that this is a somewhat arbitrary decision, which might depend on frequency of term usage or specialist's intuition. This representative term will be used where just one of 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 *lt4el:BarWithButtons*.

In the literature, various approaches have been proposed to carry out the mapping task between concepts and terms. Most of them consider the multilingual lexicons as starting point and then try to establish the connection to the concepts. Examples of such initiatives are WordNet [10], EuroWordNet [11], SIMPLE [12]. In our project, we have assumed an alternative approach to link the ontology to the lexicons which is very close to the LingInfo mode [13].

We have constructed the terminological lexicons on the basis of the formal definitions of the concepts within the ontology. In this way, we have avoided the complex task of mapping different lexicons in several languages, as was the case in EuroWordNet [11]. The main problems that might occur within this approach are that:

1. for some concepts there is no lexicalized term in a given language;
2. some important term in a given language has no appropriate concept in the ontology which should represent its meaning.

In order to solve the first problem we allow the lexicons to contain also non-lexicalized phrases which have the meaning of the concepts without lexicalization in a given language.

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 the construction of the regular grammar rules for annotation of the concept in the text which will be described in more details below.

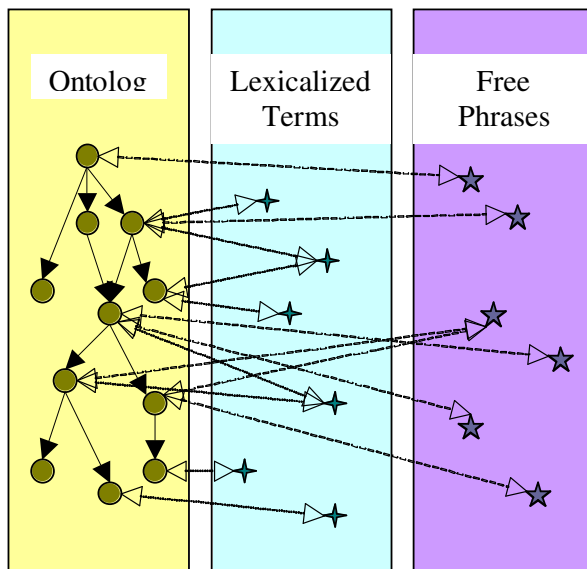


Figure 3: Relation between concepts of the ontology and

*lexicalized terms vs. phrases in a particular language*

In this way, it was possible to capture the different wordings of the same meaning which might occur in texts. The various mapping varieties are illustrated in figure 3.

The picture depicts the realization of the ontological concepts in a natural language. The concepts are language independent and they might be represented within a natural language as form(s) of a lexicalized term (or item), or as a free phrase. In general, a concept might have a few terms connected to it and a (potentially) unlimited number of free phrases expressing this concept in the language. In our lexicons, we have decided to encode as many free phrases as possible in order to have better recall on the semantic annotation task. In case of a concept that is not-lexicalized in a given language we require at least one free phrase to be provided for this concept.

In order to solve the second problem (i.e. an important term has no concept in the ontology), we modify the ontology in such a way that it contains all the important concepts for the domain. However, this solution requires a special treatment of the "head words" in the lexicons, because such phrases allow bigger freedom with respect to their occurrences in the text. Variability is a problem even with respect to the lexicalized cases and we have chosen to represent the most frequent (based on the learning objects we already processed) variants for each concept.

To conclude: in the LT4eL project, the ontology represents the semantic knowledge in form of concepts and relations with appropriate axioms while the lexicons represent the ways in which these concepts can be realized in texts in the corresponding languages. Clearly, the ways in which a concept could be represented in the text are potentially infinite in number. We could represent in our lexicons only the most frequent and important terms and phrases.

*D. Semantic annotation of learning objects*

Annotation grammars have been employed for the annotation of our corpus with ontology information. They can be considered as a special kind of partial parsing tool which for each term in the lexicon contains at least one grammar rule for the recognition of the term. For the implementation of the annotation grammar we rely on the grammar facilities of the CLaRK System. (<http://www.bultreebank.org/clark/index.html>)

The creation of the current annotation grammars started with the identification of the terms in the lexicons for the relevant languages. Each term has been lemmatized and the lemmatized form of the term was converted into regular expression of grammar rules. Each concept related to the term is stored in the return markup of the corresponding rule. Thus, if a term is ambiguous, then the corresponding rule in the grammar contains reference to all concepts related to the term.

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 he/she considers to be important for the classification of the learning object. However, in order to accurately link a learning object and/or its parts to the proper places in the conceptual space of the ontology, the *inline annotation* of the content of learning objects becomes an obligatory intermediate step in the meta-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 the texts of the corresponding languages. Additionally, rules for disambiguation between several concepts are applied when a text realization is ambiguous between several concepts. However, at the current stage of the project we carry out disambiguation manually.

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* is used in the following ways: (1) as a step to metadata annotation of the learning objects; (2) as a mechanism to validate the coverage of the ontology; and (3) as an extension of the retrieval of learning objects where, except for the metadata, we could use also cocurrences of concepts within the whole LO or its subparts (paragraphs or sentences).

More specifically, the annotation was carried out by means of a version of the CLaRK System that includes the appropriate DTDs, layouts, grammar and constraints. The process implied two phases: (1) preparation for the semantic annotation and (2) actual annotation. The former refers to the compilation of appropriate regular grammars that identify the connection between the domain terms in some natural language and the ontological concepts. It also considers the construction of a DTD, layouts and support semi-automatic tools for assigning and disambiguating concepts, namely the constraints. In the annotation phase, the regular grammar finds the occurrences of terms in the text and it assigns all the possible concepts per term. As explained previously, the regular grammars were constructed automatically on the basis of the lemmatization of the terms in the lexicons. Thus, in some cases the grammar can under- or over-generate. The constraints, on the other hand, aim at making the annotation process more accurate. The constraints support the manual annotation.

### III. SEMANTIC SEARCH FOR CROSS-LINGUAL RETRIEVAL

One of the goals of the LT4eL project is to develop a search functionality which improves the accessibility to documents in a learning management system by exploiting semantic characteristics of search queries and

documents. In addition, it should work for several languages and it should enable users to find documents in various languages while using ontology representations or search queries in the user's language.

The search engine which we have developed builds on the data as described in the previous section, namely:

1. a collection of documents in several languages, covering one common domain or subject;
2. an ontology for this domain;
3. lexicons which provide language-specific terms for the domain concepts;
4. semantic annotation of the documents.

Another assumption is that there is a lexicon representing a user's native language and that there are documents in those languages which the user specifies as his second, third etc. language.

The basic idea of the ontology-based search (or *semantic search*) is that concepts from the ontology lead the user to those documents which are appropriate for his query. The search is most precise when the user directly selects concepts from the ontology. However, we want users to start with a free-text query for two reasons.

First, today's typical user is familiar with the Google search engine and is used to type one or more words and get results immediately. Two other kinds of search that work like that are part of the LMS, i.e. *full-text search* (all words from the text are considered when looking for a match for the query) and *keyword search* (matching with words that are assigned as keywords to the documents – to avoid confusion, we call the user input *search words/terms*, not keywords). Therefore, we invoke semantic search as soon as the user has entered his search words, and give first results together with the results of the other search methods.

Second, we consider it useful to provide the user with a starting point for finding the proper place in the ontology. The search words are used to find this place in the ontology. So in a second step, the user can navigate through the ontology and select concepts as the input for a more precise search.

It is also possible to start immediately from the ontology view; in this case, no starting point can be offered to the user, so navigation will start from the root of the ontology: the most general available concept.

Figure 4 shows how a part of the ontology can be presented to the user: the taxonomic structure (super/sub concepts) is made clear by indentation; concepts that have a different relation to a shown concept can be faded in by clicking on a button.

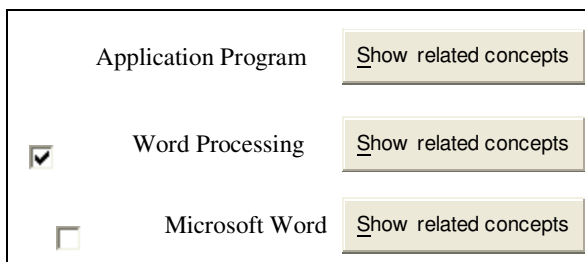


Figure 4 – Example of a possible representation of a browsing unit, where only taxonomical relations are present.

The semantic search procedure takes as input parameters: a) the language in which the query will be formulated (determines which lexicons to use for lookup); b) languages for which the user wants to see available documents; c) search terms given by user; d) concepts selected by user; e) a method for combining the concepts (“AND-search” or “OR-search”); f) two options indicating whether documents should be retrieved which do not contain the desired concept but, for the two options, respectively, a superconcept or a subconcept of it.

The data flow from the user query to the retrieved documents is as follows:

1. the search words are looked up in the lexicons of the chosen languages. Search words are normalised orthographically before lookup;
2. if lexical entries are found in the lexicon, they are matched to concepts in the ontology. These concepts are also used as starting points for ontology navigation which precedes selecting concepts;
3. documents in the desired languages are retrieved, based on the set of found concepts, while taking into account the *and/or* parameter, OR
4. concepts directly selected from the ontology are the basis for the search, again taking into account the *and/or* parameter.

The search engine returns a list of documents which semantically match the search word(s) or selected concepts. For each found document, in addition to already available information such as title and relevant metadata such as assigned keywords, the following information is made available:

a) “matching concepts”: all concepts that were the basis for search *and* match the document. This is a subset of all the concepts that relate to the document, and can include *main search concepts* (concepts that are found on the basis of the entered terms, and concepts directly selected from the ontology), but also super and subconcepts of those in case concept query expansion was invoked.

b) snippet: a small fragment of the document (or two fragments, connected by three dots), selected around occurrences of the matching concepts, which are marked so that they can be highlighted when displaying. If there are multiple matching concepts, the ones that occur more frequently are preferred. Furthermore, occurrences of different concepts close to each other in the text are

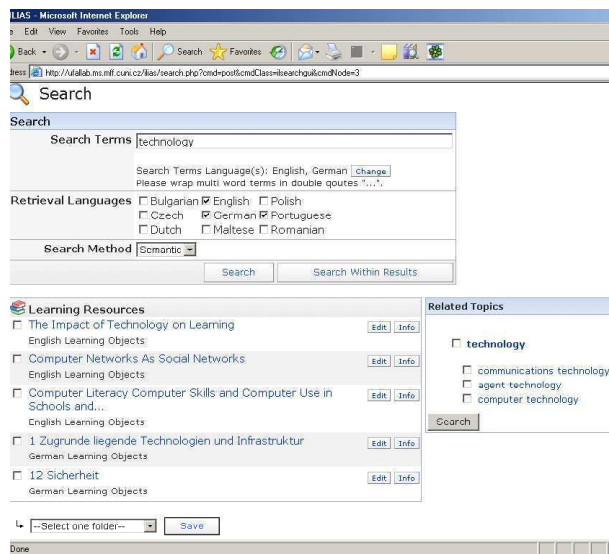


Figure 2 – User interface for search in ILIAS. In the upper part, search words can be entered and languages can be selected. In the lower left part, the resulting documents are listed. In the right part, concepts can be selected.

preferred. The idea behind this is that terms (or the concepts behind them) that describe a topic are likely to co-occur in a sentence or passage, both [14] and Google (<http://www.google.com/technology/whyuse.html>) use this notion for ranking (if such a passage is contained, the document is more relevant), while [15] and [16] apply it for passage retrieval. In our approach, it allows for selection of a snippet that is representative as a preview of the document to the user.

c) relevance score, by which the retrieved documents are sorted. It is a value between 0 and 1 which is an aggregation of two scores, reflecting the following aspects:

- the number of different *main search concepts* that match the document (excluding concepts that were automatically added by query expansion). This reflects how well the document matches the query;
- the *occurrence frequency* of the matched concepts: if they occur more often, they play a more important role in the document. The frequency is normalised for document length, to compensate for the fact that a short document cannot mention the concept as often as a long document but can still be very relevant. For this score, also the matched inferred (super/sub) concepts are taken into account, but with a lower weight than the main search concepts. Thus, the second part of the score reflects the relevance of the concepts to the document.

In the version of the search system that is currently being developed in LT4eL, users are able to choose which of the types of search they want to use simultaneously. The semantic search results will be

joined with the results of full-text search and keyword search, which also have a relevance score.

Figure 2 shows the user interface of the search engine as it is implemented in a prototype of the ILIAS Learning Management System ([www.ilias.de](http://www.ilias.de)).

#### IV. ONTOLOGIES AND ADDED VALUE FOR ELEARNING

The semantic search we have sketched in the previous section is currently subject to user-centred evaluations to reveal how well received it is in the context of the learning management systems where it is used to solve real tasks. Preliminary results show that both tutors and students highly appreciate the possibility to retrieve documents in several languages. Furthermore, they recognize that the added value of the ontology lies in the visualization of the relation among concepts. Thus, leading the way towards a more personalized learning path for students and a better way to reuse the learning material for tutors.

More specifically, students commented positively on the fact that related documents were retrieved, and the fact that semantic search gave access to the related topics provided by the concept browser. In addition, they stated that semantic search provides better results than full text search due to the ontological structure of the search. One gets better insight in the topic and thus a better view of things that one may not understand. If one doesn't exactly know what something means, the semantic search helps by displaying related topics. The majority of the students we have tested find very useful to be able to navigate to associated topics, and browse documents in this way. They remark that this is a good way to browse if one is uncertain on what one is looking for or if one doesn't know the right keywords. However, for a minority of students, browsing an ontology is a new activity and they prefer to search documents through full text search they are familiar in Google.

It should also be noticed that preliminary quantitative evaluation results show the superiority of the semantic search with respect to full text search for all the languages of the project [17].

More generally, we believe that with the addition of ontologies, the following eLearning requirements are improved:

- **Delivery**

While in *traditional learning* it is the instructor who determines the agenda, within *eLearning* this task is left to the student. In the LT4eL project we make a contribution in this respect. Learning objects are available within LMS, but they are linked to the ontology. Relevant concepts are annotated in the learning objects on the basis of the ontology allowing for semantic search. It will be thus possible to enable the construction of a user-specific course, by semantic querying for topic of interest.

- **Access**

While in *traditional learning* the progression of knowledge is linear, this is not the case within *eLearning* which allows for direct access to knowledge in whatever sequence makes sense to the situation at hand. The added functionalities allow the user to describe the situation at hand and perform semantic querying for the suitable learning material. Access to knowledge can be expanded by semantically defined navigation through the use of the ontologies that allow for some form of inference which can be used for semantic match like classification and taxonomic reasoning.

- **Personalization**

While in *traditional learning* content must satisfy the needs of many possible users, within *eLearning* content can be personalized and it is determined by the individual user's needs. A user searches for learning material customized for her/his needs. With the added functionalities developed within the project, the ontology is the link between user needs and characteristics of the learning material.

- **Authority**

While in *traditional learning* the content is centralized and it is selected from a library of materials developed by the educator, within *eLearning* the content is distributed since it comes from the interaction of the participants and the educators. The semi-automatic generation of metadata envisaged by the project, together with the ontology, allows for retrieval of both static (introduced by the educator) and dynamic (learner contribution) content within the LMS for an effective co-operative content management.

#### V. ONTOLOGIES FOR SOCIAL LEARNING

In the previous sections, we have shown how ontologies have been employed in the LT4eL project to improve the retrieval and the management of learning objects in a Learning Management System.

However, another possible use of ontologies which we sketch in this section is in the context of social learning. The amount of material that can be used for learning purposes is growing and it might go beyond text books, exercises or presentations originally developed by educational institutions. Due to the available technology, learners are also in the position to share the results of their learning activity through repositories similar in their functionality to YouTube (<http://www.youtube.com>), Flickr ([www.flickr.com](http://www.flickr.com)) or Del.icio.us ([del.icio.us](http://del.icio.us)). These repositories feature a strong social component and allow users to share and reuse the uploaded material as well as to comment on its quality. In this context, a formal representation of the content of the learning material, while crucial for its retrieval and its reuse might

not include the appreciation people express for different kinds of materials. In order to support social and informal learning it is thus necessary to create a link between the formal representation of a given domain in the form of ontologies and the informal descriptions produced by social tagging and folksonomies.

It is thus necessary to create the appropriate methodology to support social and informal learning through the development of services that are based on the interaction between a formal representation of domain knowledge and a social component which complements it. We envisage the need to create an infrastructure for knowledge sharing in which learners can develop a system of interoperable personal and community knowledge bases which is best formalized by means of an ontological layer. Ontologies present the right level of abstraction over the meaning in general domains (upper ontologies) as well as in concrete domain (domain ontologies). Through their formalization by means of standard ontology languages such as RDF(S) and OWL it is possible to employ inference mechanisms for searching and proposing to the learner the appropriate learning material. In order to extract the relevant domain knowledge Natural Language Processing techniques can be employed to identify terms (including their definitions) which are mapped into concepts as well as the relations among these concepts.

Furthermore, in order to include the social dimension into this system, the traditional bipartite model of ontologies can be extended leading to a tripartite model of users (actors), tags (concepts), and resources (instances of concepts) [18]. Additionally, the social tagging approaches can be integrated with the more formal conceptualization approaches based on ontologies. This can be achieved by grouping together highly related tags corresponding to elements in ontologies structured according to the relationships holding amongst those elements. Natural Language Processing techniques in combination with statistical techniques can be employed to this purpose. Tags can also be proposed to users on the basis of existing ontologies and/or the learning material being considered, leading thus towards a more formal social tagging.

This approach can lead to development of services to support social and informal learning by improving the creation of personalized content, by allowing for decentralization and co-operation of content management and by creating communities of learners on the basis of their learning needs and interests.

## VI. SEMI-AUTOMATIC DEVELOPMENT OF ONTOLOGIES

In the LT4eL project, a domain ontology has been employed to facilitate the retrieval of the learning objects across different languages. However, in order to be able to move from a prototype system to a real one, we should be able deal with different domains. While extending the dictionary is not problematic since existing terminological dictionaries could be adapted for the

purpose, building appropriate domain ontologies is not a trivial task. More generally, in order to be able to use ontologies within eLearning, we need to be able to create and populate domain ontologies semi-automatically on the basis of the available material.

We believe, however, that the availability of techniques and tools in the Natural Language Processing area as well as in the Semantic Web area are providing a valuable contribution to the solution

In particular, the task can be achieved by employing a data driven approach and natural language processing techniques can be adopted to extract terms/concepts, definitions and relations from learning material. We can build on existing techniques which rely mainly on statistical analysis, patterns finding and shallow linguistic parsing ([19], [20], [21], [22], [23] for an overview). We want to emphasize the role that *linguistic information* can play in order to acquire knowledge from text (cf. [24], [25], [26] for similar attempts). More specifically, we want to exploit the implicit grammatical knowledge present in texts (i.e. morphological, syntactic and semantic information) to extract the knowledge which will constitute the building blocks of the domain ontology, that is *terms*, *synonyms*, *concepts* and ultimately *taxonomies* and *relations*.

In order to achieve this, we plan to build on results obtained within the LT4eL project in which a keyword extractor has been developed mainly to facilitate the semi-automatic generation of metadata, while definition extraction has been employed mainly for the creation of lexica [2].

In the context of semi-automatic ontology development, however, the keywords extracted could be used as first step towards the creation of concepts for a domain ontology related to the learning objects adopted (cf. also [27]). Similarly, definitions extracted from the texts employed could be used to define the concepts present in the ontology (cf. also [28], [29]). In addition, [30] and [31] are relevant for traditional methods of relation extraction from text in which NLP techniques are employed.

We believe that this approach on ontology learning, which is based on NLP techniques, can be integrated with more recent approaches which use dynamically selected ontologies as background knowledge in order to populate existing ontologies (cf. [31] as an attempt in this direction). More specifically, selected terms can be matched with concepts already present in ontologies dynamically selected from online ontology repositories such as Swoogle<sup>1</sup> or Watson<sup>2</sup> or Ontoselect<sup>3</sup>.

The integration of these two approaches has not been considered yet in the literature and constitutes an innovative aspect which might contribute to the solution of the knowledge acquisition bottleneck. Furthermore, the

<sup>1</sup> <http://swoogle.umbc.edu/>

<sup>2</sup> <http://watson.kmi.open.ac.uk/>

<sup>3</sup> <http://olp.dfki.de/OntoSelect/w/>

application of these approaches to eLearning represents a novel test case, which we leave for future research.

## VII. CONCLUSIONS

In this paper, we have addressed how ontologies, a key element of the semantic web vision, can contribute to eLearning and can enhance the learning process. We have taken the LT4eL project, as test case. In particular, the semantic search facility which has been implemented in the context of the project and which is now used by the ILIAS learning management system shows the potential of ontologies in the application domain of information retrieval. In this model, the central role is assumed by the ontology which determines the content of the other components of the model. Another advantage of our model is that it supports the work in a multilingual environment. The mappings between lexicons and ontology are performed with the aim that each term in each language is linked to a corresponding concept, and vice versa – each concept in the ontology is exemplified by at least one expression (be it lexicalized or a free phrase). Thus, the ontology itself as well as specific language lexicons are verified in a cross-lingual context. We have created lexicons and annotation grammars for all the languages in the project. The mapping between the language specific lexicons was facilitated by the ontology.

We have also sketched another potential use of ontologies in the context of social learning and we have made the claim that the technology in the area of Natural Language Processing and in the Semantic Web area is now ripe to allow for semi-automatic development of ontologies. We leave the assessment of this claim for future research.

## REFERENCES

- [1] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific Am.*, May 2001, pp. 34–43.
- [2] P. Monachesi, L. Lemnitzer, K. Simov. Language Technology for eLearning. Proceedings of EC-TEL 2006, in Innovative Approaches for Learning and Knowledge Sharing, LNCS 0302-9743, pp. 667-672. 2006
- [3] L. Lemnitzer, C. Vertan, A. Killing, K. Simov, D. Evans, D. Cristea, P. Monachesi: "Improving the search for learning objects with keywords and ontologies". In `Proceedings of EC-TEL 2007', Springer LNCS.
- [4] Gruber, T (2005). "Folksonomy of Ontology: A Mash-up of Apples and Oranges". First on-Line conference on Metadata and Semantic Research (MTR05).
- [5] Lemnitzer L. and Monachesi P. (2008) Extraction and evaluation of keywords from Learning Objects – a multilingual approach. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2008).
- [6] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari and L. Schneider. 2002. The WonderWeb Library of Foundational Ontologies. WonderWeb Deliverable D17, August 2002.
- [7] C. Masolo, S. Borgo, A. Gangemi, N. Guarino and A. Oltramari. 2002. Ontology Library (final). WonderWeb Deliverable D18, December 2003
- [8] N. Guarino and C. Welty. 2002. "Evaluating Ontological Decisions with OntoClean." *Communications of the ACM*, 45(2): 61-65.
- [9] A. Gangemi, R. Navigli, P. Velardi. TheOntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. Meersman R, et al. (eds.), Proceedings of ODBASE03 Conference, Springer, 2003.
- [10] C. Fellbaum. 1998. Editor. *WORDNET: an electronic lexical database*. MIT Press.
- [11] P. Vossen (ed). "EuroWordNet". General Document. Version 3, Final, July 19, 1999
- [12] A. Lenci, F. Busa, N. Ruimy, E. Gola, M. Monachini, N. Calzolari, A. Zampolli, E. Guimier, G. Recourcé, L. Humphreys, U. Von Rekovsky, A. Ogonowski, C. McCauley, W. Peters, I. Peters, R. Gaizauskas, M. Villegas. 2000. SIMPLE Work Package 2 – Linguistic Specifications, Deliverable D2.1. ILC-CNR, Pisa, Italy.
- [13] P. Buitelaar, T. Declerck, A. Frank, S. Racioppa, M. Kiesel, M. Sintek, R. Engel, M. Romanelli, D. Sonntag, B. Loos, V. Micelli, R. Porzel, P. Cimiano "LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies". In: Proc. of OntoLex06, a Workshop at LREC, Genoa, Italy, May 2006.
- [14] E.K. Park, S.I. Moon, D.Y. Ra, and M.G. Jang. Web Document Retrieval Using Sentence-query Similarity. In Proceedings of the 11 Text REtrieval Conference (TREC-11), 2002.
- [15] M.A. Hearst. TileBars: visualization of term distribution information in full text information access. Proc. CHI'95, pages 56-66; May, 1995.
- [16] Drori, O., The User Interface in Text Retrieval Systems, SIGCHI bulletin, New York: ACM, July 1998, 30 (3), 26-29.
- [17] Lemnitzer L., Simov K., Osenova P., Mossel E., Monachesi P. (2007) "Using a domainontology and semantic search in an eLearning environment". In: Proceedings of The Third International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering. (CISSE 2007). Springer-Verlag. Berlin Heidelberg.
- [18] Mika, P. (2005) Ontologies are us: A unified model of social networks and semantics. 4th ISWC
- [19] Maedche, A., Staab S. (2000) Semi-automatic Engineering of Ontologies from Text. In: *Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering*, 2000.
- [20] Navigli, R., Velardi P. (2004) Learning Domain Ontologies from Document Warehouses and Dedicated Websites, *Computational Linguistics* (30-2).
- [21] Cimiano, P., Schmidt-Thieme L., Pivk A., Staab S., (2005) Learning Taxonomic Relations from Heterogeneous Evidence, *Ontology Learning from Text: Methods, Applications and Evaluation*, IOS Press.
- [22] Buitelaar, P., Cimiano, P., Magnini B., (2005) *Ontology from Text: Methods, Evaluation and Applications* Frontiers in Artificial Intelligence and Applications Series, Vol. 123, IOS Press.
- [23] Buitelaar, P., Olejnik, D., Hutanu, M., Schutz, A., Declerck, T., Sintek (2004) Towards Ontology Engineering Based on Linguistic Analysis. In: *Proceedings of LREC 2004*.
- [24] Pantel, P., Ravichandran, D., and Hovy, E.H. (2004). Towards Terascale Knowledge Acquisition. *Proceedings of Conference on Computational Linguistics (COLING-04)*. Geneva, Switzerland
- [25] Ravichandran, D., Pantel P., Hovy E., (2005) Randomized Algorithms and NLP: Using Locality Sensitive Hash Functions for High Speed Noun Clustering. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [26] F. Sclano and P. Velardi (2007). 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.
- [27] Judith Klavans and Smaranda Muresan. (2001). Evaluation of the DEFINDER System for Fully Automatic Glossary Construction. In: Proc. of AMIA Symposium 2001.

- [28] Fahmi, I. and Bouma, G.(2006), Learning to identify denitions using syntactic features, in R. Basili and A. Moschitti (eds), *Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications*.
- [29] Schutz, A. and Buitelaar, P. (2005). ReLExt: A Tool for Relation Extraction from Text in Ontology Extension. 4th ISWC, Galway,593-606
- [30] Specia, L., Motta, E. (2006) A hybrid approach for extracting semantic relations from texts. 2nd Workshop on Ontology Learning and Population at COLING/ACL 2006, Sydney 57-64
- [31] Sabou, M., d'Aquin, M., Motta, E. (2006) Using the Semantic Web as Background Knowledge for Ontology Mapping. International Workshop on Ontology Matching, Athens, GA.
- [32] J. Jovanovic, D. Gasevic, C. Brooks, V. Devedzic, M. Hatala (2007) LOCO-Analyst: A Tool for Raising Teachers' Awareness in Online Learning Environments. In `Proceedings of EC-TEL 2007', Springer LNCS.
- [33] A. Zoaq, R. Nkambou, C. Frasson (2007) Building Domain Ontologies from Text for Educational Purposes. In `Proceedings of EC-TEL 2007', Springer LNCS.