



Project no. 027391

Project acronym: LT4eL
Project title: Language Technology for eLearning

Instrument Specific Targeted Research Project

Thematic Priority Information Society Technology

D3.2b Ontology mapping and language vocabularies development

Due date of deliverable: 30-11-2007
Actual submission date: 21-12-2007

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]

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

Ontology Lexicon Deliverable

Contents

- 1 Title
- 2 Summary
- 3 Concept to Text Mapping Model
- 4 Lexicons
- 5 Concept Annotation Grammars creation
- 6 Learning Objects annotation
- 7 Crosslingual search
- 8 References
- 9 Scientific papers on the ontology, lexicons and annotation of learning objects

1 Title

D3.2b Ontology mapping and language vocabularies development

2 Summary

During the first year of the project we created an English lexicon aligned to the ontology. For each domain class in the ontology we presented at least one term in the lexicon. This lexicon then was used for the annotation of the English learning objects.

In parallel to this task we performed a pilot study on the creation of lexicons and their alignment to the ontology in several languages - Bulgarian, Dutch, German and Romanian. We created 200 entries for each language.

On the basis of the work done during the first year we proceeded with the work during the second year performing the following tasks related to ontology mapping and lexicons:

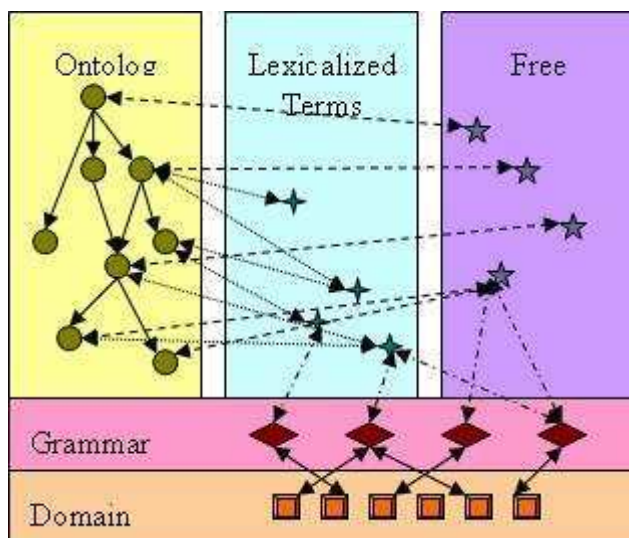
- Concept to Text Mapping Model creation;
- Lexicons creation;
- Concept Annotation Grammars creation;
- Learning objects annotation;
- Crosslingual search and integration within ILIAS.

In this deliverable we describe the work done with respect to the above mentioned tasks.

3 Concept to Text Mapping Model

In this section we present briefly the linguistic model of the annotation process adopted for the tasks within the project. For a detailed description see [1]. We assume that the ontology is the repository of the lexical meaning of the language. Thus, we started with a concept in the ontology and we searched for lexical items and non-lexical phrases that convey the content of the concept. There are two possible problems here: (1) there is no lexical item for some of the concepts in the ontology, and (2) there are lexical items in the language without a concept representing the meaning of the lexical item in the ontology. The first problem was overcome by allowing in the lexicon also non-lexical (fully compositional) phrases to be represented. The second problem was solved by extension

of the ontology. The lexicon items were then mapped to the grammars. These grammars related the lexicon to the text. This mapping was necessary as much as the lexical items and phrases from the lexicons allow for multiple realizations in the text and require some additional linguistic knowledge in order to disambiguate between different meanings of some lexical item or phrase. The following figure depicts the elements of the model.



We have been using the relations between the different elements for the task of ontology-based search. The connection from ontology via lexicon to grammars was relied on for the concept annotation of the text. In this way we established a connection between the ontology and the texts. The relation between the lexicon and the ontology was used for definition of user queries with respect to the appropriate segments within the documents. On this basis a multilingual search strategy was implemented within LT4eL project [2].

Our approach gains in many respects from such works as WordNet [3], EuroWordNet [4], SIMPLE [5]. In spite of the fact that we employ the experience from these projects (mapping to WordNet and Pustejovsky's ideas in SIMPLE), we also suggest an alternative for the connection between the ontology and the lexicons. Our model is very close to LingInfo model (see [6] and [7]) not only with respect to the mapping of the lexical items to concepts, but also with respect to the other language processing tools we connect to the ontology - the concept annotation grammars and concept disambiguation tools. As to WordNet and EuroWordNet, we differ in the direction of the workflow, i.e. we started from ontology to the lexicon, not vice versa. From SIMPLE we differ in using a domain ontology instead of a general linguistic ontology. From LingInfo model we differ in the fact that the three components (ontology, lexicon and grammars) are represented by different models. For a detailed discussion see [1].

4 Lexicons

Here (and in the next sections) we present the basic steps of linguistic processing necessary to provide ontology-based semantic search. We follow the relations described in the model presented in the previous section.

After the creation of the first version of the ontology, we have created an English lexicon aligned to the ontology. This lexicon was created on the basis of the keywords extracted from the learning objects (English keywords or the translation of the key words from other languages into English), WordNet (where it provides corresponding synsets) and terminological lexicons. Then the next step necessary to be done was to create corresponding lexicons in the other languages.

For this task we have created a DTD and XML templates for the lexical entries. Here we present (a part of) the DTD for the lexicon:

```

<!ELEMENT LT4ELLex (entry+)>
<!ELEMENT entry
  ((owl:Class|rdf:Description|rdf:Property), def,
  termg+)>
<!ELEMENT def (#PCDATA)>
<!ELEMENT termg (term+,def?)>
<!ATTLIST termg
  lang      (bg|cs|de|en|mt|nl|pl|pt|ro)    # REQUIRED
  >
<!ELEMENT term (#PCDATA)>
<!ATTLIST term
  type      (lex|nonlex)    "lex"
  shead     (1|0)           "0"
  gram      CDATA           #IMPLIED
  >

```

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

```

<entry id="id60">
  <owl:Class
rdf:about="http://www.lt4el.eu/CSnCS#BarWithButtons">
    <rdfs:subClassOf>
      <owl:Class
rdf:about="http://www.lt4el.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 types of information: (1) information from the ontology about the concept that represents 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 representative for the term set. This representative term is 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 <http://www.lt4el.eu/CSnCS#BarWithButtons>. The first term in the example is representative for the term set and it is marked with the attribute **shead** with value **1**. One of the shown terms is non-lexicalized: it has the attribute **type** with value **nonlex**.

Following this model, lexicons for all languages in the project were created.

Lexicons are used in two ways in the project. First, they provide the means for ontology presentation in the corresponding languages. Second, they are the basis for ontology annotation of LOs.

5 Concept Annotation Grammars creation

The next step was the creation of concept annotation grammars that recognise the usage of the terms in the text of the learning objects. Because the terms can have different word forms (plural for example), and for multi-word terms, their word order can vary or they can even be discontinuous, it was not possible to use the lexicons directly for concept

annotation. In order to ensure that the above variations of the terms in the text would be captured we converted the lexicons into concept annotation grammars.

For the implementation of the annotation grammar we relied on the grammar facilities of the CLaRK System (<http://www.bultreebank.org/clark/index.html>). The structure of each grammar rule in CLaRK is defined by the following DTD fragment:

```
<!------->
<!ELEMENT line (LC?, RE, RC?, RM, Comment?) >
<!ELEMENT LC (#PCDATA)>
<!ELEMENT RC (#PCDATA)>
<!ELEMENT RE (#PCDATA)>
<!ELEMENT RM (#PCDATA)>
<!ELEMENT Comment (#PCDATA)>
<!------->
```

Each rule is represented as a line element. The rule consists of regular expression (RE) and category (RM = return markup). The regular expression is evaluated over the content of a given XML element and can recognize tokens and/or annotated data. The return markup is represented as an XML fragment which is substituted for the recognized part of the content of the element. Additionally, the user could use regular expressions to restrict the context in which the regular expression is evaluated successfully. The LC element contains a regular expression for the left context and the RC for the right one. The element Comment is for human usage. The application of the grammar is governed by Xpath expressions which provide additional mechanism for accurate annotation of a given XML document. Thus, the CLaRK system was a good choice for the implementation of the initial annotation grammars.

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

6 Learning Objects annotation

The next step was to apply the concept annotation grammars to the actual text of the LOs. The result of this application was that each occurrence of a term (in any of its forms) was annotated with all the concepts from the ontology which were connected to the term.

Then a disambiguation step was performed. It was done manually with the support of the Constraint module in CLaRK. The task was to check each concept annotation for its correctness. When it was necessary, the annotation was deleted. In addition to ambiguous terms we also checked the annotation of some of the unambiguous terms in order to verify the grammars for overgeneration. The last task is still in progress.

In addition to that, we performed some manual investigation of LOs in order to find missing concept annotations.

The result of this step was a gold standard of concepts annotation of LOs in the languages of the project. This result was used to support the search for content and also it could be used in future for automatic disambiguation based on machine learning.

Software used during the creation of the Lexicons, Concept Annotation Grammars, Learning Objects annotation.

The IPPBAS provides the CLaRK system (<http://www.bultreebank.org/clark/index.html>) for lexicons creation and ontology annotation of the learning objects with appropriate resources. The choice of CLaRK was motivated by two reasons: (1) the system provides the necessary functionality for the corresponding tasks; (2) it is easy to implement the

necessary resources for the corresponding task. Also, there were no other systems well known by the majority of the partners for these tasks.

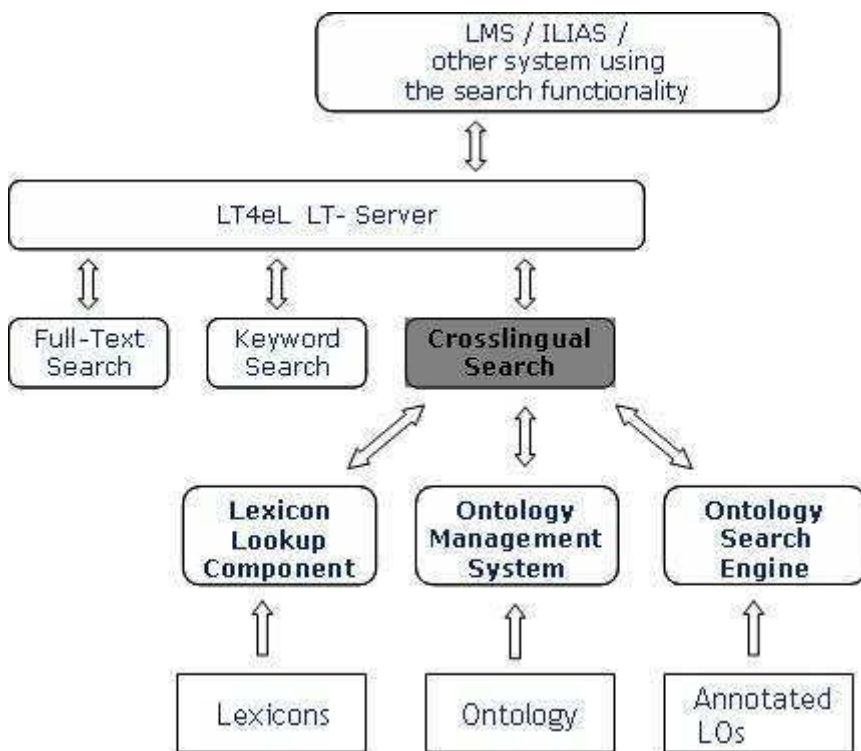
Needless to say, the partners were free to use other systems according to their experience.

7 Crosslingual search

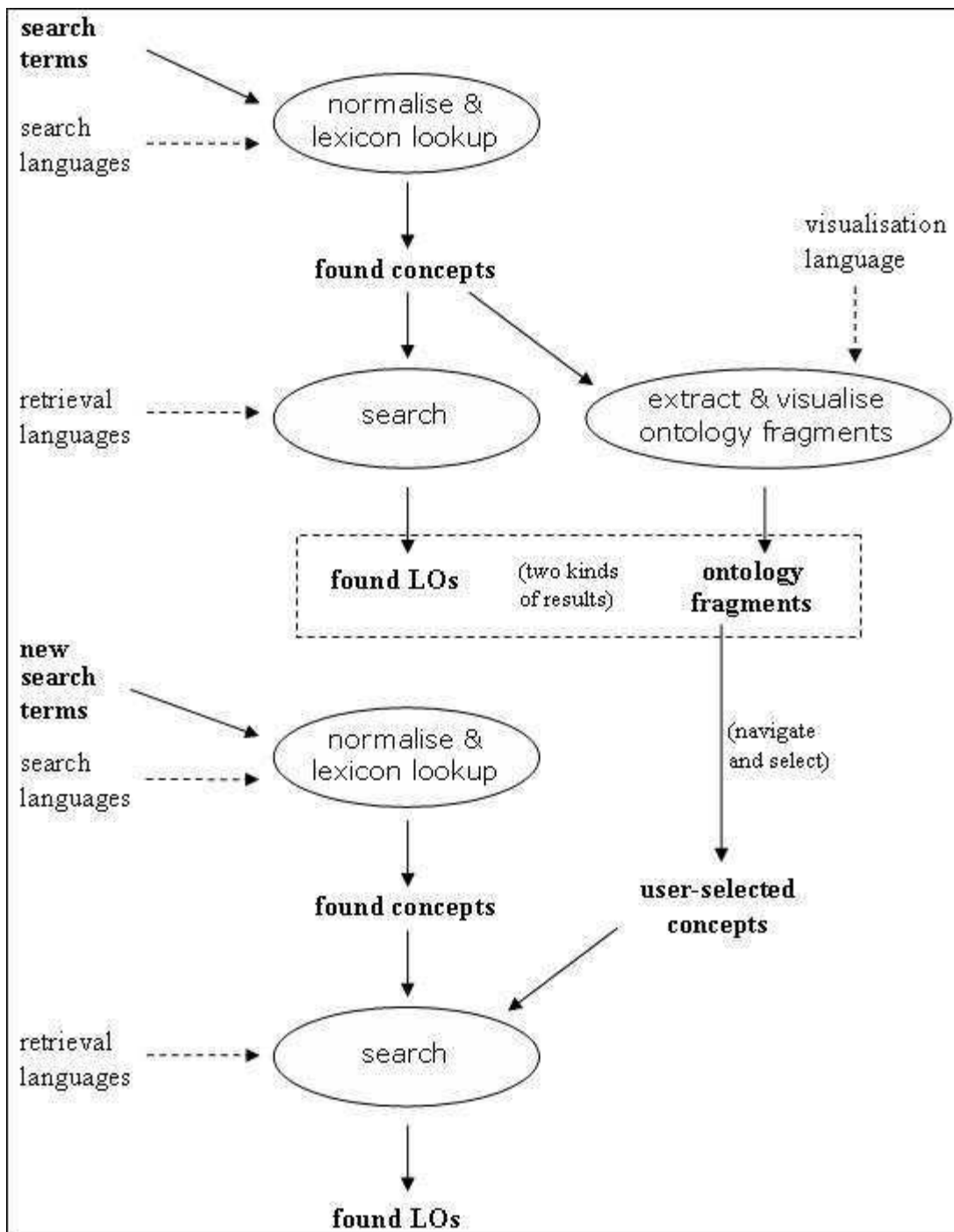
After the annotation of the learning objects in the described way they were used in the semantic mono and multilingual search.

▪ Functionality

The ontology, the lexicons and the concept annotation of the LOs are the basis for the crosslingual search functionality. The architecture of the integration is presented in following figure. Each of the three resources at the bottom of the picture is handled by the corresponding software component. The search module combines them, handles the user input and options, and performs the search algorithm.



The figure below shows the data flow of the crosslingual search. Implicitly, it makes clear the user scenario of using the search. A more detailed description can be found in [2]. Although the retrieval is based on concepts from the ontology, we start with a free-text query for two reasons. First, we assume that the users, who are probably familiar with Google, want their results fast, with not too many intermediate steps. The simplest case is to type search words and click on search - this procedure is also used for full-text search in our system. Second, we use the entered search words to find a good starting point in the ontology, so that the user does not have to click his way through the ontology starting at the root.



Words in any of the covered languages can be entered and are looked up in the lexicon; the concepts that are linked to the matching lexicon entries are used for ontology-based search in an automatic fashion. Before lexicon lookup, the words are orthographically normalised, and combinations for multi-word terms are created (e.g. if the words "text" and "editor" are entered, the combinations "texteditor", "text editor" and "text-editor" are created and looked up, in addition to the individual words). For each of the found concepts, the set of all its (direct or indirect) subconcepts is determined, and is used to retrieve LOs. The use of these language-independent concepts as an intermediate step makes it possible to retrieve LOs in any of the covered languages, thus realising the crosslingual aspect of the retrieval.

When the found LOs are displayed, at the same time the relevant parts of the ontology are presented in the language that the user prefers. Now, in a second step, the user can select (by marking a checkbox) the concept(s) he wants to look for and repeat the search. If an entered word was ambiguous, the intended meaning can be explicated now by selecting the appropriate concept. Furthermore, by clicking on a concept, related concepts are displayed; navigation through the ontology is possible in this way, following the

ontological relations.

A list of retrieval languages (only LOs written in one of those languages will be found) is specified as an input parameter. The retrieved LOs are sorted by language. The next ordering criterion is a ranking, based on the number of different search concepts and the number of occurrences of those concepts in the LO. For each found LO, its title, language, and matching concepts are shown.

■ **Development process**

A prototype of the crosslingual search component was designed in month M15, implemented in M17 and M18, and integrated into ILIAS in M19 and M20. Based on feedback of partners in months M21-M22, combined with original ideas of the WP3 and WP4 partners, the following extensions and improvements have been implemented:

- Development of stand-alone GUI for semantic search independent of ILIAS: M19-M21
- Normalisation (wrt. diacritics and upper/lowercase) of search terms and lexicon entries, to increase recall of lookup in term-concept lexicon: M20
- Return matching concepts for each found LO: those search concepts, that occur in the LO: M20
- Merging of overlapping ontology fragments (current visualisation causes confusion if a concept appears in several fragments) (not integrated): M20
- Improvement of the search speed: M23
- Automatic concept query expansion: use the complete subconcept tree for each search concept: M23
- Include English concept descriptions if not available in user language: M24
- Extracted word form->lemma mapping from the LO corpus, use to improve lexicon lookup (integrated only in development version): M24
- Implementation of conjunctive search (AND-search) for user-selected concepts (not integrated): M24

8 References

- [1] Kiril Simov, Petya Osenova. 2007. Applying Ontology-Based Lexicons to the Semantic Annotation of Learning Objects. RANLP 2007 Workshop on Natural Language Processing and Knowledge Representation for eLearning Environments
- [2] Eelco Mossel. 2007. Crosslingual Ontology-Based Document Retrieval. In Proceedings of Workshop on Natural Language Processing and Knowledge Representation for eLearning Environments. RANLP 2007, Borovets, Bulgaria.
- [3] Christiane Fellbaum. 1998. Editor. WORDNET: an electronic lexical database. MIT Press.
- [4] Piek Vossen (ed). EuroWordNet General Document. Version 3, Final, July 19, 1999, <http://www.hum.uva.nl/~ewn>
- [5] Alessandro Lenci, Federica Busa, Nilda Ruimy, Elisabetta Gola, Monica Monachini, Nicoletta Calzolari, Antonio Zampolli, Emilie Guimier, Gaëlle Recourcé, Lee Humphreys, Ursula Von Rekovsky, Antoine Ogonowski, Clare McCauley, Wm Peters, Ivonne Peters, Robert Gaizauskas, Marta Villegas. 2000. SIMPLE Work Package 2 - Linguistic Specifications, Deliverable D2.1. ILC-CNR, Pisa, Italy.
- [6] Paul Buitelaar, Thierry Declerck, Anette Frank, Stefania Racioppa, Malte Kiesel, Michael Sintek, Ralf Engel, Massimo Romanelli, Daniel Sonntag, Berenike Loos, Vanessa Micelli, Robert Porzel, Philipp 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.

[7] Paul Buitelaar, Michael Sintek, Malte Kiesel A Lexicon Model for Multilingual/Multimedia Ontologies In: Proceedings of the 3rd European Semantic Web Conference (ESWC06), Budva, Montenegro, June 2006.

9 Scientific papers on the ontology, lexicons and annotation of learning objects

The following papers have been written, presented on conferences and published or to be published. The full papers are in the appendix of D3.1b. In the following, the bibliographical data are listed.

- Kiril Simov, Petya Osenova: Applying Ontology-Based Lexicons to the Semantic Annotation of Learning Objects. Proceedings of the RANLP 2007 workshop: Natural Language Processing and Knowledge Representation for eLearning Environments.
- Lothar Lemnitzer and Kiril Simov and Petya Osenova and Eelco Mossel and Paola Monachesi. Using a domain-ontology and semantic search in an eLearning environment. International Conference on Engineering Education, Instructional Technology, Assessment, and E-learning (CISSE-EIAE 07).
- Cristina Vertan, Paola Monachesi, Kiril Simov, Petya Osenova, Lothar Lemnitzer, Alex Killing and Diane Evans. Crosslingual retrieval in an eLearning environment. Proceedings of The 10th Congress of the Italian Association for Artificial Intelligence (AIIA 2007).
- Lothar Lemnitzer, Cristina Vertan, Alex Killing, Kiril Simov, Diane Evans, Dan Cristea, Paola Monachesi: Improving the search for learning objects with keywords and ontologies. Appeared in: Duval, Erik; Klamma, Ralf; Wolpers, Martin (Eds.) Creating New Learning Experiences on a Global Scale. Second European Conference on Technology Enhanced Learning, Lecture Notes in Computer Science, Vol. 4753, pp. 202-216.
- Eelco Mossel: Crosslingual Ontology-Based Document Retrieval. Proceedings of the RANLP 2007 workshop: Natural Language Processing and Knowledge Representation for eLearning Environments.
- Paola Monachesi, Lothar Lemnitzer and Kiril Simov *Language Technology for eLearning*. Poster presentation at KP7-congres. 13 February 2007, Den Haag. The Netherlands.

Accepted for presentation:

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