



Project no. 027391

Project acronym: LT4eL

Project title: Language Technology for eLearning

Instrument Specific Targeted Research Project

Thematic Priority Information Society Technology

D3.2c Ontology mapping and language vocabularies development

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]

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)	

D3.2c

Contents

- 1 Title
- 2 Summary
- 3 New WP3 functionality in cooperation with ILIAS interface changes
 - 3.1 Introduction
 - 3.2 AND/OR option
 - 3.3 Snippet Extraction
 - 3.4 Relevance score & Ranking
- 4 Lexicons Extension and Enrichment
 - 4.1 Maltese and Cross Lingual Information Retrieval
- 5 Concept Annotation Grammars Creation and Semantic Re-Annotation of LOs (wrt validation)
- 6 Scientific papers on the ontology, lexicons and annotation of learning objects

1 Title

D3.2c Ontology mapping and language vocabularies development

2 Summary

During the second year of the project we created lexicons aligned to the ontology in all languages of the project (except for Maltese, for which a lexicon was only created during the last six months of the project). On the basis of the lexicons as well as the lemmatization of the lexical items annotation grammars were created. Then they were used for the learning objects annotation. The annotated learning objects together with the ontology and the lexicons were relied on during the evaluation process of the new functionalities integration in ILIAS. Some new functionalities were added to ILIAS in order to provide the user with a better view on the results from the semantic search.

Using the model of ontology-to-text relation, developed during the second year, we extended the lexicons in all languages to cover the extension of the ontology. Consequently, the grammars were also extended on the basis of the lexicon extensions. This model reflects the relation between concepts and text by the means of the annotation grammars. In the model, we assume that the ontology represents the lexical meaning of the language. Thus, for each concept we searched for the lexical items and non-lexical phrases that represent the content of the concept. There are two problems here: (1) no lexical item for some of the concepts in the ontology, and (2) 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 the extension of the ontology. This strategy was successfully followed further in this phase of the project. The lexicon items were then mapped to the additional annotation grammars. The richer grammars reflected better the relation between the lexicon and the text.

3 New WP3 functionality in cooperation with ILIAS interface changes

3.1 Introduction

Based on the users' feedback from the first validation round, a set of additions and improvements to the

search process has been agreed on and implemented. Some of the planned, but not integrated at time additions, such as the AND/OR option, have been confirmed as necessary by the users' feedback. Other planned activities were the following ones: the boundary between the WP3 functionalities and the ILIAS user interface / WP4 integration. It was necessary, because, on the one hand, we were validating functionalities, while at the same time the users' expectations were influenced by the user interface of the chosen LMS.

A complete survey on the perspective of ILIAS can be found in the deliverable D4.1 Annex: M30 Changes. The most important improvements within WP3 services were the following ones:

- AND/OR search
- snippet extraction
- relevance score
- faster ranking

A more detailed description of each added functionality is given below:

3.2 AND/OR option

In version 1 of the WP3 services, the AND-option for search was only available for concepts that were selected from the ontology by the user. It was tested separately but had not been integrated in ILIAS yet.

In the new version, it was decided that an AND/OR option would be made available for the entered search terms as well as for selected concepts. In order to keep it simple for the users, no special syntax is needed: rather, a choice between AND/OR can be made by choosing a radio button. This choice holds equally to the entered words as well as to the selected concepts. Two points in the search procedure vary according to the AND/OR option:

1. Combination of multi-word terms.

- In case the OR option is selected, combinations for multi-word terms are created using several concatenators (e.g. if the words "computer" and "screen" are entered, the

combinations "computerscreen", "computer screen" and "computer-screen" are created and looked up, in addition to the individual words "computer" and "screen").

- With AND-search, no extra combinations of words are created and tried, since it could restrict the search more than the user wants. Suppose that a user wants documents about computers and screens, so she chooses the AND-option and enters computer and screen. So a document should be returned if those are both contained as a concept. Suppose there is a document D that satisfies this query. Now, if the system would create the multi-word combination "computer screen" additionally and take it into account when searching, because of the AND-option, D is not found anymore if it does not contain the concept for "computer screen", even if the original user query was already satisfied. It cannot be used as an alternative, since this would boil down to OR-search and disrespect the user's wish that both are contained. Thus, trying multi-word terms does not go together with semantic AND-search. If a user wants to use multi-word expressions in AND-search, the expression has to be put between quotes.

2. Combination of found concepts.

- In OR-search, a document is retrieved if it contains any relevant concept: a concept denoted by a search term, or a concept manually selected from the ontology, or a sub concept of them (query expansion).
- In AND-search, however, it is not possible to take the conjunctive of all those concepts. Certain concepts should be treated as alternatives, so that a retrieved document has to contain at least one of them. This concerns concepts that are found through the same search words, plus all their sub concepts. In the example of "computer" and "screen", for instance, two concepts are found for computer, i.e. {#Computer, #Personalcomputer}. A document satisfies the query, if it contains one of those two (or a subconcept of them, in case of query expansion), plus a concept denoted by

"screen" (or a subconcept of it). Thus, for each entered search term, a set of concepts is created, one of which should occur in the document. The AND-query as a whole is a set of those sets, thus a conjunction of disjunctions.

The OR-query can also be formulated in the same way: in this case, the outer set contains only one embedded set of concepts, which are thus treated as a disjunction.

3.3 Snippet Extraction

One more addition regarding the presentation of search results to the user is the selection of a Google-like text snippet for each retrieved document, which shows the context of the word or concept in the document. This feature was also requested by many test users. It is a small fragment of the document (or two discontinuous fragments, connected by three dots); in case of semantic search, it is selected around occurrences of the matching concepts. The concept annotation itself is removed, but the words that were annotated by one of the search concepts are marked with tags. This information can be used to highlight them when displaying. Words that were annotated by other concepts are not marked in the snippet.

As a result, a highlighted word can be different from the search term, e.g. "PC" occurs highlighted in the snippet, while the user entered "computer".

In order to select snippets efficiently, concepts as well as words are indexed with Lucene, so that the software does not need to read and process the entire XML-annotated document before selecting a snippet from it. Furthermore, snippets are extracted in batches, only as many at a time as are displayed on a page.

3.4 Relevance score & Ranking

Many users remarked that it was unclear how the search results were sorted. They were sorted by the normalised annotation frequency (contained number of annotated concepts that were used for search, divided by document length), but this was not visible. Also, this ranking was not very sophisticated.

To solve this, we introduced a relevance score, by which the retrieved documents are sorted. As the output of the WP3 services, it is a value between 0 and 1, which in ILIAS is presented to the user as a percentage, to indicate the estimated relevance. The value is an aggregation of two scores, reflecting the following aspects:

- The number of different *main search concepts* (concepts that are found on the basis of the entered terms, or concepts directly selected from the ontology) 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 normalized 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.

The relevance score for a document reflects the correspondence between the query and that document, independently of the other retrieved documents and independent of the total available set of documents. It might look logical to set the document with the highest normalised annotation frequency at 100% and the other retrieved documents proportional to it. However, in that way, the relevance score of other retrieved documents are very low if the result set happens to contain a document with a very high frequency of the searched concept. Thus, this is too susceptible for changes in the repository or query.

Instead, we base the score on an experimental *expected concept-token-ratio* per document. We use 0.005 (one matching concept per 200 tokens) at the moment. Obviously, this can result in a score above 1 for certain documents with very high annotation frequencies. To correct this, we do not cut them to 1, but rather use the following formula, that maps values between lower boundary B and infinity to values

between B and 1:

$$\text{corrected} = B + (1-B) * (S-B) / (S-B + 1-B)$$

where B is a chosen boundary value between 0 and 1, and S is the score to be corrected. We are currently using B = 0.7, which gives the following corrected scores (examples):

ORIGINAL	CORRECTED
0.8	0.775
1.0	0.850
1.5	0.918
3.0	0.965

Of course, the expected concept-token ratio is chosen such that the values are not above 1 in most cases. The corrected score is a factor in the final relevance score, which is a weighted average where the other factor is the normalized (between 0 and 1) number of main search concepts.

Combine multiple search methods

In contrast to the previous version of the search interface, users can now simultaneously use the various types of search (semantic, fulltext, keyword, definition search). Thus, the WP3 services were adapted to support a combination of search methods. In the resulting sorted list of search results, every document occurs at most once. In case a document is found by several methods, the respective relevance scores have to be joined.

When taking a simple average of the scores, we found that the final score can be undesirably low when only one of the search methods returns a good relevance score for a relevant document, especially in our case where three methods can be combined. In our opinion, a low score does not necessarily mean that the document is not relevant, but rather that there is no evidence for relevance, while a high score is based on positive evidence from the text or meta data. Therefore, we use a weighted average that favours the best score:

$$0.6 * \text{highest} + 0.3 * \text{middle} + 0.1 * \text{lowest}$$

Improving speed of ranking

In order to calculate the relevance score of a document with respect to a query, the annotation frequencies of search concepts have to be retrieved. In the first version of the services, for every retrieved document, the entire field with indexed concepts was retrieved from the database and the concepts relevant for the query were counted. This was too slow, taking around 30 seconds if the search result consisted of 100 documents. For the first validation experiments, a temporary solution was implemented. In the new version of the services, used for the second validation, a result object was introduced, which is used to store the relevant information about a retrieved document at the time of retrieval. Here, the annotation frequencies of only the concepts that match the query are stored; other annotated concepts are ignored. This way, the database does not need to be queried again during ranking.

4 Lexicons Extension and Enrichment

In the previous cycle the following information was stored within the lexicons: (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 schema was preserved, but concerning (2) an explanation of the concept meaning in the lexicon was also provided in the respective partner's language (for Bulgarian

lexicon - in Bulgarian, for Polish lexicon - in Polish, etc.). Thus the user could feel more comfortable when using the lexicon as an interface to the ontology.

During the last six months we have created new lexical items for the new added domain concepts, and also for the concepts and relations from the upper ontology. The new concepts in the domain part are near to 300. The concepts in the upper part of the ontology are more 270 and the relations are more 107. Thus, the effort for creation of the lexicons for the extension of the ontology and the upper part is comparable to the effort for the creation of the lexicons for the first version of the ontology. Another issue we have to mention is the impact of the size and the topics of the learning objects corpora in the various languages. While in the first cycles we relied on the pre-annotated keywords, now the feedback from the initial validation scenario tests had the main contribution. The partners reported the missing terms that were necessary for the tests. They added these terms into their lexicons. The adequate mappings and additions to the ontology were also made respectively. Most of the requested missing terms reflected the required levels of detailness within the ontology.

The creation of lexical items for the upper part of the ontology was necessary only for better and more comprehensive navigation over the ontology in each corresponding language. The domain part of the lexicon has more applications. On the one hand, it is also used for ontology browsing. On the other hand, it is a basis for the creation of the annotation grammars. Last, but not least, it supports the mono- and crosslingual semantic search.

During the last period of the project, a special attention was paid to the Maltese lexicon:

4.1 Maltese and Cross Lingual Information Retrieval

The point has already been made in previous reports that Maltese presents some very special difficulties compared to the other languages handled by the consortium. During the first three months of the project, we searched extensively for documents in Maltese related to ICT and eLearning. The few documents found (9 in total) were not rich in content (point form, summarised information or did not contain learning content) and not of the right format (without Maltese fonts or incorrect use of ICT terminology, specifically resorting to English terms rather than Maltese ones). The second difficulty encountered in the implementation of Maltese was the lack of linguistic tools available. Although an ongoing project was developing a part-of-speech tagger at the time of the project, this was not available during the timeline of LT4eL.

Although the problem with tools is on the way to resolution, problems with the kind of data required by LT4eL remain and are unlikely to be resolved for some time to come.

Without content and linguistic tools for Maltese, we nevertheless felt that some functionality for Maltese within LT4eL would be possible in the form of support for CLIR (Cross Lingual Information Retrieval), with query terms being entered in Maltese, and the results shown in another language within the project. This has been made a priority issue during the last months.

To achieve this aim, the translation of the main lexical entries from English to Maltese was carried out for the semantic dictionary. An example of one such translation is given below. To give an idea of what was involved, the figure below displays one particular entry for the term "Central Processing Unit" (as submitted by the student assistants who were helping with the work). The main Maltese term and various other terms can be clearly seen.

```
-----  
ID: id106  
Main Term: c.p.u.  
Main Maltese Term: CPU  
Term: central processing unit  
Term: central processor  
Term: cpu  
Term: mainframe  
Term: processor  
Other Maltese Terms: il-procossur centrali, il-procossur ewlieni  
Def: (computer science) the part of a computer (a microprocessor chip) that does most of the data processing  
the CPU and the memory form the central part of a computer to which the peripherals are attached.  
-----
```

When incorporated into the semantic lexicon, the entry became more complex, since it was mixed with

semantic information. The entry looks like this:

```
<entry id="idl06">
<owl:Class rdf:about="http://www.lt4el.eu/CSnCS#CentralProcessingUnit">
<rdfs:comment>(computer science) the part of a computer (a microprocessor chip) that does most of the data
processing; the CPU and the memory form the central part of a computer to which the peripherals are attached.
</rdfs:comment>
<rdfs:subClassOf>
<owl:Class rdf:about="http://www.loa-cnr.it/ontologies/WordNet/OWN#ELECTRONIC_EQUIPMENT"
</owl:Class>
<def>
(computer science) the part of a computer (a microprocessor chip) that does most of the data processing;
the CPU and the memory form the central part of a computer to which the peripherals are attached.
</def>
<termg lang="mt">
    <term shead="1">CPU</term>
    <term>il-processur centrali</term>
    <term>il-processur ewlieni </term>
</termg>
</entry>
```

The incorporation of this information into the semantic lexicon is sufficient to allow semantic searching through ILIAS using Maltese terms rather than English terms. The retrieval of documents will be carried out in English or any other language implemented within the project.

To conclude, the exercise we have carried out serves as a proof of concept for CLIR involving Maltese as the query language within the LT4eL framework. We should add that within the subject area that was chosen for the project, this would be an unnatural thing to do for the majority of Maltese speakers, who would be much happier querying computer-related material in English. However, as we have implied in earlier reports, the picture for data collections concerning other domains, such as law, medicine, and possibly religion, would be entirely different.

Although the Maltese language is without question the first language used in Malta, its usage, particularly in written form, tends to be domain-dependent. Future projects involving Maltese should take this fact into account when designing technology which is intended to reach ordinary Maltese speakers in realistic scenarios.

5 Concept Annotation Grammars Creation and Semantic Re-Annotation of LOs (wrt validation)

The extension of the ontology with new concepts required the creation of new entries in the lexicons for the languages of the project. Also, the creation of new annotation rules was performed. These rules were used for additional annotation of the learning objects.

The main problem with respect to the additional annotation was that interaction between the annotation that was already done on the previous stage of the project, and the new annotation. This interaction was based on the fact that some of the new terms for the new concepts contained some old terms for some old concepts. Thus, in order to have the new concepts annotated correctly we had to remove the old annotation.

For example, the new English term "Microsoft Word document" contains "Microsoft Word".

In order to deal with such cases we divided the new grammar rules in two grammars. The first one contained only the rules that interact with the old annotation. This grammar was applied over the old annotation. In cases where the new grammar rule recognised a chunk that covers an old annotation this old annotation was deleted automatically. The second grammar contained the rules that do not interact with the old annotation. Thus, the second grammar was applied outside of the old annotation.

Another improvement in this direction was the implementation of a simple automatic module for semantic disambiguation. It is statistical and aims at supporting search over LOs, which have not been pre-annotated.

The architecture of the annotation grammar applications will be further developed and evaluated within the thematically follow-up project LTfLL.

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

The following papers have been written, presented on conferences and published or are to be published. The full papers or the presentations are in the appendix. In the following, the bibliographical data are listed, together with a short summary.

- Simov and Osenova 2008: Language Resources and Tools for Ontology-Based Semantic Annotation. *OntoLex 2008 Workshop at LREC 2008*, pp. 9-13.

Short summary: This paper presents the resources and tools, which facilitate the ontology-based semantic annotation of domain texts, and subsequently - the semantic search. Some of these resources are language independent, such as the domain ontology. Some depend on the specific language: terminological lexicons, annotation grammars, sense disambiguation rules, relation annotation rules, gold standard corpus (used in the process of ontology creation). The combination of these tools defines ontology-to-text relation. Implementing different instantiations of this relation we could achieve semantic annotation of text with different granularity and for different tasks.

- Petya Osenova, Kiril Simov, Eelco Mossel 2008: Language Resources for Semantic Document Annotation and Crosslingual Retrieval. In: *Proc. of LREC 2008*.

Short summary: This paper describes the interaction among language resources for an adequate concept annotation of domain texts in several languages. The architecture includes domain ontology, domain texts, language specific lexicons, regular grammars and disambiguation rules. This is considered the preparatory phase for the integration of a semantic search facility in Learning Management Systems. The implementation and performance of this search are discussed in the context of related work as well as other types of searches. Also the results from some preliminary steps towards evaluation of the concept-based and text-based search are presented.

Resources and Tools for Ontology-Based Semantic Annotation

Kiril Simov and Petya Osenova

Linguistic Modeling Department,

Institute for Parallel Processing,

Bulgarian Academy of Sciences

25A Acad. G. Bonchev St.

Sofia 1113, Bulgaria

kivs@bultreebank.org, petya@bultreebank.org

Abstract

This paper presents the resources and tools, which facilitate the ontology-based semantic annotation of domain texts, and subsequently the semantic search. Some of these resources are language independent, such as the domain ontology. Some depend on the specific language: terminological lexicons, annotation grammars, sense disambiguation rules, relation annotation rules, gold standard corpus (used in the process of ontology creation). The combination of these tools defines ontology-to-text relation. Implementing different instantiations of this relation we could achieve semantic annotation of text with different granularity and for different tasks. The ideas are based on the empirical observations within two European projects.

1. Introduction

In this paper we present our work on defining of the *ontology-to-text* relation and its instantiations for several languages and two domains. This relation is important with respect to tasks, such as ontology annotation, ontology based search (or semantic search), information extraction, ontology learning and ontology browsing. The work described here was carried out within two European projects: LT4eL¹ (*Language Technology for eLearning*) and AsIsKnown² (*A Semantic-Based Knowledge Flow System for the European Home Textiles Industry*).

The relation *ontology-to-text* shows how the elements of ontology (concepts, relations, instances) are realized within the text of multimedia documents. Our model of the relation comprises four components: *ontology*, *lexicon*, *grammar*, *text*. The *ontology* is a domain one mapped to an upper part. The *lexicon* contains the terms (grouped on the basis of synonymy) and associated contextual information and grammatical features. The *grammar* contains the syntactic knowledge about the forms in which the terms might be realized in the text. It also contains some disambiguation information about the term in a certain context (in case it is ambiguous). The *text* is a description of a part of the domain in question for which we would like to explicate the ontological information. In the real life the situation is more complex, because the texts usually contain other means to represent the same concept (relation, instance) out of the terms in the lexicons. In order to handle such cases the grammar needs to contain also parts devoted to such phenomena as coreferential relations, metonymy, metaphorical usages, etc. In the actual realization of the relation in the two projects we started with annotation of concepts, but we will continue with relations and instances in a follow-up project. In the paper we will be discussing mainly concept annotation.

In many respects our model of the *ontology-to-text* relation is subsumed by more general and elaborated models, such as LingInfo (see (Romanelli et al., 2007), (Buitelaar et al., 2006a) and (Buitelaar et al., 2006b)). Main differences are

in: (1) the definition of the model the ontology-based model definition in LingInfo vs. XML-based resource representation oriented to particular processing tools in our work; and (2) coverage of the model the LingInfo covers all multimedia information objects like images, sounds, etc., while we focused only on the linguistic level, represented by texts. Thus, we might consider our work as an example of instantiation of some elements from the LingInfo model too. In future we envisage the incorporation of annotated images, since they - together with the texts - contribute to the better semantic search in a domain.

We would like also to stress that in practice there are many instantiations of the *ontology-to-text* relation. For example, see the ontology-based named entity annotation presented in (Kiryakov et al., 2004), among others.

The structure of the paper is as follows: first, we present in short the two projects and the role of ontology in them; then we present the ontology creation methodology employed in the projects (there are some differences in the two projects in this respect); in section 4 the elements of our model on the ontology-to-text relation are described; the last section polemizes the place of the current paper within other works, and concludes the paper.

2. The role of the ontology within the two projects

We had to construct domain ontologies for both abovementioned European projects. The main usage of these ontologies concerned, on the one hand, the annotation of domain texts for search purposes, and on the other hand, the connection among multilingual domain material or among the specific views of the various participants in the same domain. Let us point in short to the specificities of each project. The LT4eL project aims at demonstrating the relevance of the language technology and ontology document annotation for improving the usability of learning management systems (LMS) within the learning process. Thus, a semantic search module had to be created. This module built on concept annotated documents. With the help of the domain ontology sets of learning texts have been annotated in various subdomains and in eight languages (Bulgarian, Dutch, German, English, Czech, Polish, Portuguese and Romanian). The semantic

¹ <http://www.lt4el.eu/>

² <http://www.asisknown.org/>

search increased the precision and the speed in finding the most relevant documents for a topic.

The AsIsKnown project is developing an architecture of interrelated modules for speeding up the process of communication among agents in the textile industry. Here the challenge is not only the cross-lingual access to the system, but also the different communication preferences of the agents in this business area. The ontology was used as follows: in the annotation of fashion magazines for search and trend analysis; as an input-output communication system among producers, retailers and clients. The languages involved in the project are Bulgarian, English, French and German.

In addition to the semantic annotation and search the ontology has to support the communication with the user for query definition and result explanation. Thus, for example, it is necessary for the users to be able to navigate over ontology in a natural for them way. In our view this task has to be done via the natural language of the user.

3. Ontology creation

In this section we briefly outline the methodology for ontology creation used within the two projects. One of the main requirements for the methodology is that the initial version of the ontology is created from existing resources. The involvement of the domain experts in the process of the ontology creation is done at a later stage. In this way we attempt to maximize their contribution. Here we present the main steps of the methodology as it was applied within the project AsIsKnown.³

Processing of the standards and vocabularies in the domain

We consider standards in the domain as reliable sources of conceptual information. Being created by leading experts in the domain with the goal to facilitate the whole process of production and usage of the home textile, the standards can be viewed as "expert questionnaires" usually used in the process of knowledge acquisition. Thus, we expected to find definitions of the most important concepts and relations in the domain. The definitions also helped us to establish the main relationships between the extracted concepts. As a means for the extraction of the concepts and the relations we have been using a treebank constructed semi-automatically over the text of the standards. Then we inspected manually the analysis in order to identify the relevant knowledge. The result from this step was a list of (concept) *terms* (in English), a list of *relations* (relational terms), a list of triples - (*term1 relation term2*). These lists became the backbone of the ontology. The list of relations includes general ontological relations like *is-a*, *part-of*, etc. and domain specific relations. The extracted terms in many cases were equipped with a definition. These definitions had to reflect the triples for the term and the features of the relations.

Formalization of the terms

The next step is to define formal definitions of the extracted concepts and relations in OWL-DL. We have selected OWL-DL, because there exist implemented reasoners for it. For each term in the term list we constructed a class definition in OWL-DL. We did the same for each relational term. We also encoded the

³ The differences within the LT4eL project will be discussed later.

additional information in the definitions of the terms and the relations. The result of this step was an initial formal version of the ontology.

Link to an upper ontology

The establishing of the connection between the upper and the domain ontology helped us to check the consistency of the domain ontology with respect to the ontology construction methodology behind the upper ontology and to inherit the knowledge encoded in the upper ontology. Also the upper ontology provided general ontological information when it was required during the usage of the ontology. We selected DOLCE Ontology (Masolo et al., 2003) as upper ontology for several reasons: (1) it is constructed on rigorous basis which reflects the OntoClean methodology (Guarino and Welty, 2002); (2) it is represented in OWL-DL; (3) the authors of the ontology provide us comments and help on the alignment of the domain ontology to DOLCE. The alignment between the two ontologies is facilitated by OntoWordNet (Gangemi et al., 2003) - a version of WordNet aligned to DOLCE. OntoWordNet ensures more understandable concepts (more specific and closer to the domain) and the mapping between the concepts is easier. The result from this step is the better structuring of the initial lists of concepts and relations. Also relations and axioms were inherited from DOLCE to the domain ontology.

Evaluation by domain experts

The evaluation of the first version of the ontology has been done in two ways:

Practical evaluation

The ontology is evaluated in the process of incorporation and integration within the overall project architecture.

Expert evaluation

The ontology is reviewed by domain experts in the project. The review is mediated by questionnaires constructed on the basis of the already constructed first version of the ontology. Here is an example from such a questionnaire on carpets:

Nr.	Question	Answer	Comment
3.	What is the difference between Loop Column and Loop Row ?	a) <i>Loop Column</i> shows a product direction b) <i>Loop Row</i> shows a transverse direction	See 5.12 and 5.13, ISO 2424
4.	Does Tuft Column (a line of tufts essentially parallel to the direction of manufacture) consist of Tuft ?	Yes, if the meaning of Tuft is Cut Pile in this case (q.v. ISO 2424, 5.6)	The term <input type="checkbox"/> tuft describes a manufacturing technique too.

Besides the evaluation by domain experts the ontology is evaluated on the basis of annotation of a corpus of representative domain documents. In this way some adequate coverage of the ontology is ensured.

Documentation

In the process of construction of the ontology we keep track on the sources of each concept, relation, etc.

Lexicons and concept annotation grammar creation

This step is the creation of an instance of the

ontology-to-text relation for the given ontology. The actual model of the relation is given in the next section. In the two projects we had to create instances in several languages as it was mentioned above.

The methodology outlined in this section was successfully applied to the construction of both domain ontologies. The evaluation is still an on-going process. In case of LT4eL we did not have standards in the domain and this is why we started with the keywords annotated manually by the partners in the learning objects. Then for the keywords in the domain we collected definitions from different sources (terminological lexicons, Internet) and these definitions were the initial source for creation of the first version of the ontology.

4. Ontology-to-Text relation

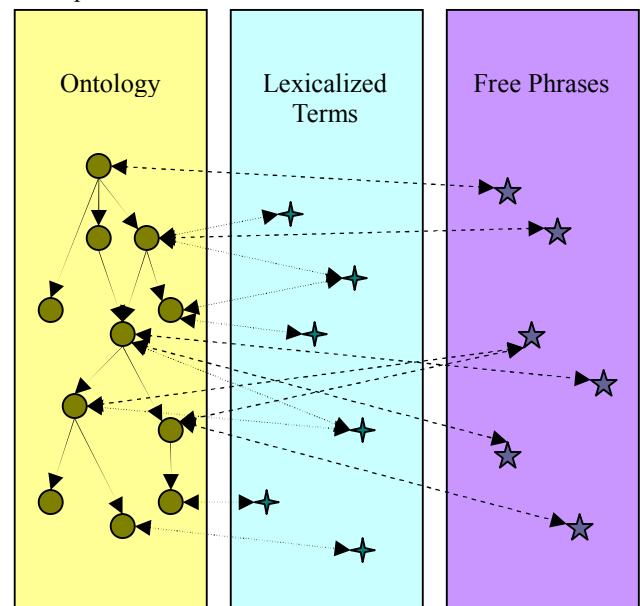
In this section we represent the two main components that define the ontology-to-text relation necessary to support the tasks within our projects. These components are: (terminological) lexicon and concept annotation grammar. The lexicon plays twofold role in our architecture. First, it interrelates the concepts in the ontology to the lexical knowledge used by the grammar in order to recognize the role of the concepts in the text. Second, the lexicon represents the main interface between the user and the ontology. This interface allows for the ontology to be navigated or represented in a natural for the user way. For example, the concepts and relations might be named with terms used by the users in their everyday activities and in their own natural language (e.g. Bulgarian). This could be considered as a first step to a contextualized usage of the ontology in a sense that the ontology could be viewed through different terms depending on the context. For example, the color names will vary from very specific terms within the domain of carpet production to more common names used when the same carpet is part of an interior design.

Thus, the lexical items contain the following information: a term, contextual information determining the context of the term usage, grammatical features determining the syntactic realization within the text. In the current implementation of the lexicons the contextual information is simplified to a list of a few types of users (producer, retailer, etc).

With respect to the relations between the terms in the lexicon and the concepts in the ontology, there are two main problems: (1) there is no lexicalized term for some of the concepts in the ontology, and (2) there are lexical terms in the language of the domain which lack corresponding concepts in the ontology, which represent the meaning of the terms.

The first problem is overcome by writing down in the lexicon also non-lexicalized (fully compositional) phrases to be represented. 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 annotation grammar. Having them, we might capture different wordings of the same meaning in the text. The picture below shows the mapping varieties. It depicts the realization of the concepts (similarly for relations and instances) in the language. The concepts are language independent and they might be represented within a natural language as form(s) of a

lexicalized term, 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 language⁴. Some of the free phrases receive their meaning compositionally regardless their usage in the text, other free phrases denote the corresponding concept only in a particular context. In our lexicons we decided to register 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.



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 the 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">
  <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>
  <termg lang="nl">
    <term head="I">werkbalk</term>
    <term>balk</term>
    <term type="nonlex">balk met knoppen</term>
    <term>menubalk</term>
  </termg>
</entry>
```

⁴ The presence of free phrases in the lexicon is also motivated by the fact that the lexicalization is not a discrete feature. There are many different degrees of lexicalization. Thus the free phrases are the extreme end of the scale.

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 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. 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*. 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 *thead* with value 1. In this way we determine which term to be used for ontology browsing if there is no contextual information for the type of users.

The second component of the ontology-to-text relation, the concept annotation grammar, is ideally considered as an extension of a general language deep grammar which is adopted to the concept annotation task. Minimally, the concept annotation grammar consists of a chunk grammar for concept annotation and (sense) disambiguation rules. The chunk grammar for each term in the lexicon contains at least one grammar rule for recognition of the term. As a preprocessing step we consider annotation with grammatical features and lemmatization of the text. The disambiguation rules exploit the local context in terms of grammatical features, semantic annotation and syntactic structure, and also the global context such as topic of the text, discourse segmentation, etc. Currently we have implemented chunk grammars for several languages. The disambiguation rules are under development.

For the implementation of the annotation grammar we rely on the grammar facilities of the CLaRK System (Simov et al., 2001). The structure of each grammar rule in CLaRK is defined by the following DTD fragment:

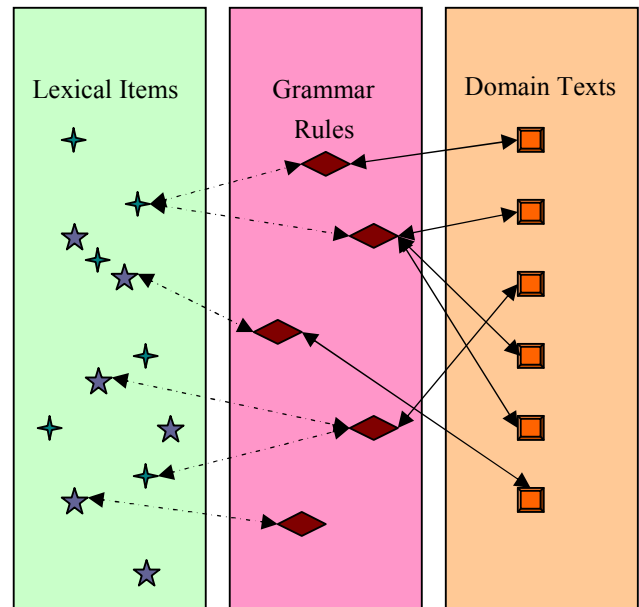
```
<!ELEMENT line (LC?, RE, RC?, RM, Comment?) >
<!ELEMENT LC (#PCDATA)>
<!ELEMENT RC (#PCDATA)>
<!ELEMENT RE (#PCDATA)>
<!ELEMENT RM (#PCDATA)>
<!ELEMENT Comment (#PCDATA)>
```

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 could 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 use. 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 grammar is a good choice for implementation of the initial annotation

grammar.

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 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 following picture depicts the relations between lexical items, grammar rules and the text:



The relations between the different elements of the models are as follows. A lexical item could have more than one grammar rule associated to it depending on the word order and the grammatical realization of the lexical item. Two lexical items could share a grammar rule if they have the same wording, but they are connected to different concepts in the ontology. Each grammar rule could recognize zero or several text chunks.

The relation ontology-to-text implemented in this way provides facilities for solving different tasks, such as ontology search (including crosslingual search), ontology browsing, ontology learning. In order to support multilingual access to semantic annotated corpus we have to implement the relation for several languages using the same ontology as starting point. In this way we implement a mapping between the lexicons in these languages and also comparable annotation of texts in them.

We have been using the relations between the various elements for the task of ontology-based search. The connection from ontology via lexicon to grammars is 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 is used for definition of user queries with respect to the appropriate segments within the documents. The annotation of texts in different languages on the basis of the same ontology could facilitate the definition of similarity metrics between such texts.

In AsIsKnown project we also exploited a domain independent partial grammar which supports the domain specific grammar providing additional context features.

5. Discussion and Conclusion

Our approach gains in many respects from such works as WordNet (Fellbaum, 1998), EuroWordNet (Vossen, 1998), SIMPLE (Lenci et al., 2000). The mapping between the language specific lexicons was facilitated by the ontology. Our model shares common features with other lexicon models: with WordNet-like Fellbaum, 1998; Vossen, 1998) lexicons we share the idea of grouping lexical items around a common meaning and in this respect the term groups in our model correspond to synsets in WordNet model. The difference in our case is that the meaning is defined independently in the ontology. With SIMPLE model (Lenci et al., 2000) we share the idea to define the meaning of lexical items by means of the ontology, but we differ in the selection of the ontology which in our case represents the domain of interest, and in the case of SIMPLE reflects the lexicon model. With the LingInfo model (Romanelli et al., 2007; Buitelaar et al., 2006a; Buitelaar et al., 2006b) we share the idea that grammatical and context information also needs to be presented in a connection to the ontology, but we differ in the implementation of the model and the degree of realization of the concrete language resources and tools.

In the paper we present a model for the ontology-to-text relation supporting semantic annotation. We assume the central role of the ontology on which all the other resources and tools depend. In future we envisage to implement an interaction with a general lexica and grammar. Some initial experiments are done by domain specific rules for exploiting the general analyses during domain semantic annotation. The model was successfully exploited in two EU projects for concept annotation and semantic search. The relation annotation requires in our view much more work on the level of general language processing in tasks like coreference resolution, metonymy patterns recognition, bridging relation annotation, etc. Some of these tasks require ontology based information and our model allows for ontology centered linguistic knowledge representation as much as knowledge in the lexicon and in the grammar is always related to the ontology. When it is necessary, information from general lexicons and grammar is transferred to the domain in an appropriate form. Thus we ensure interaction between general language processing tools and resources, and the domain specific ones.

6. Acknowledgements

This work has been supported by two FP6 European projects: LT4eL (Language Technology for eLearning) (FP6-027391) and AsIsKnown (A Semantic-Based Knowledge Flow System for the European Home Textiles Industry) (FP6-028044).

We would like to thank also the three anonymous reviewers for their valuable remarks and suggestions.

7. References

Buitelaar, P., Declerck, T., Frank, A., Racioppa, S., Kiesel, M., Sintek, M., Engel, R., Romanelli, M., Sonntag, D., Loos, B., Micelli, V., Porzel, R., and Cimiano, Ph. (2006a). 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.
- Buitelaar, P., Sintek, M., and Kiesel, M. (2006b). A Lexicon Model for Multilingual/Multimedia Ontologies In: Proceedings of the 3rd European Semantic Web Conference (ESWC06), Budva, Montenegro, June 2006.
- Fellbaum, Ch. (1998). Editor. WORDNET: an electronic lexical database. MIT Press.
- Gangemi, A., Navigli, R., and Velardi, P. (2003). The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. Meersman R, et al. (eds.), Proceedings of ODBASE03 Conference, Springer, 2003.
- Guarino, N., and Welty, C. (2002). Evaluating Ontological Decisions with OntoClean. Communications of the ACM, 45(2): 61-65.
- Kiryakov, A., Popov, B., Ognyanov, D., Manov, D., Kirilov, A., and Goranov, M. 2004. *Semantic Annotation, Indexing, and Retrieval*. Elsevier's Journal of Web Semantics, Vol. 1, ISWC2003 special issue (2), 2004. <http://www.websemanticsjournal.org/>
- Lenci, A., Busa, F., Ruimy, N., Gola, E., Monachini, M., Calzolari, N., Zampolli, A., Guimier, E., Recourcé, G., Humphreys, L., Von Rekovsky, U., Ogonowski, A., McCauley, C., Peters, W., Peters, I., Gaizauskas, R., and Villegas, M. (2000). SIMPLE Work Package 2 - Linguistic Specifications. Deliverable D2.1. ILC-CNR, Pisa, Italy.
- Masolo, C., Borgo, S., Gangemi, A., Guarino, N., and Oltramari, A. (2002). Ontology Library (final). WonderWeb Deliverable D18, December 2003. <http://www.loa-cnr.it/Publications.html>.
- Romanelli, M., Buitelaar, P., and Sintek, M. (2007). Modeling Linguistic Facets of Multimedia Content for Semantic Annotation. In: Proceedings of SAMT07 (International Conference on Semantics And digital Media Technologies), Genova, Italy, Dec. 2007. pp 240-251.
- Simov, K., Peev, Z., Kouylekov, M., Simov, A., Dimitrov, M., Kiryakov, A. (2001). CLaRK - an XML-based System for Corpora Development. In: *Proc. of the Corpus Linguistics 2001 Conference*. Lancaster, UK.
- Vossen P. (1998). Editor. EuroWordNet General Document. Version 3, Final, July 19, 1999, <http://www.hum.uva.nl/~ewn>.

Language Resources for Semantic Document Annotation and Crosslingual Retrieval

Petya Osenova¹, Kiril Simov¹, Eelco Mossel²

¹Bulgarian Academy of Sciences, Acad. G. Bonchev St. 25A, 1113 Sofia, Bulgaria

²University of Hamburg, Department of Informatics, Vogt-Kölln-Str. 30, 22527 Hamburg, Germany

E-mail: petya@bultreebank.org, kivs@bultreebank.org, mossel@informatik.uni-hamburg.de

Abstract

This paper describes the interaction among language resources for an adequate concept annotation of domain texts in several languages. The architecture includes domain ontology, domain texts, language specific lexicons, regular grammars and disambiguation rules. This is considered the preparatory phase for the integration of a semantic search facility in Learning Management Systems. The implementation and performance of this search are discussed in the context of related work as well as other types of searches. Also the results from some preliminary steps towards evaluation of the concept-based and text-based search are presented.

1. Introduction

Given the huge amount of static and dynamic contents created for eLearning tasks, the major challenge for their wide use is to improve their accessibility within Learning Management Systems (LMS). The LT4eL project¹ tackles this problem by integrating semantic knowledge to enhance the management, distribution and retrieval of the learning material (Monachesi & Lemnitzer & Simov, 2006).

The semantic annotation has already become a key ingredient of Semantic Web. There is already a vast quantity of literature and initiatives, which approach this topic from various perspectives. For example, there was SAAW 2006 - the First Semantic Authoring and Annotation Workshop devoted on tools, standards and practice of semantic annotation. In this paper, we present a model of the relation of a domain ontology to text. In order to facilitate this relation we need to construct corresponding language resources including lexicons and grammars. Here we discuss the nature of these resources. Also we describe how they can be used for mono- and multilingual search.

The paper is structured in the following way: first we present our model on the relation between a domain ontology and domain text; then we present the ontology search based on this relation; the next section reports on a comparison between text-based search and ontology search; the last section concludes the paper.

2. Domain Ontology and Semantic Annotation

The ontology-based querying for content retrieval has

¹ <http://www.lt4el.eu/> □ the LT4eL (Language Technology for eLearning) project is supported by the European Community under the Information Society and Media Directorate, Learning and Cultural Heritage Unit.

been actively explored in recent years. Here we will mention the OntoQuery Project² among others. The differences with our project are as follows: OntoQuery contributed to the issues of the onto-based search in general. We focused on this kind of search for learning purposes. OntoQuery was designed mainly for Danish-speaking users, while our search aims at multi- and crosslingual retrieval. At the moment, our process involves previous semi-automatic processing (i.e. requiring some user intervention).

The domain of the learning corpus in the LT4eL Project is □Computer Science for Non-Computer Scientists□ It covers topics like operating systems; programs; document preparation □creation, formatting, saving, printing; Web, Internet, computer networks; HTML, websites, HTML documents; email, etc. The main application of the ontology is: the indexing of these domain documents with concept information and interconnecting the same information across different languages.

The initial stages of the ontology creation were supported by a core set of manually annotated keywords in the eight languages of the project (Bulgarian, Czech, Dutch, English, German, Polish, Portuguese, Romanian). In the next development some middle-placed concepts were added and some classes were expanded (e.g. various types of text editors). Let us explain these steps in more detail: the process of choosing keywords in a text is more or less subjective. We tried to handle this problem by exploring texts in various languages. Another problem is the usage of too specific or too broad terms. Thus, the middle ones were often left out (e.g. *human activity, resource, symbol, architecture, organization*). But they seem to be very important for tracing the connection with the top part of the ontology. The annotated keywords from the other languages were translated into English. Then by search on the Web we collected definitions for the keywords. The set of definitions for a keyword either highlight the various aspects of the meaning of the keyword or the relations between its meaning and other concepts. In the

² <http://www.ontoquery.dk/index.php>

ontology we keep the most representative definition and we keep the rest of the definitions in an additional corpus in order to consult them during the formalization of the concepts in the ontology. We have preferred definitions, which reflect basically the *is-a* relation. The other relations were encoded in the ontology (*part-of, used-for* etc.). The definitions were also translated into the other languages. This is important for the user who browses the ontology. After the determination of the keyword meanings we created concepts corresponding to them. These concepts became the backbone of the domain ontology.

The next step of the ontology development was to map the domain concepts to an upper ontology (in our case we used DOLCE (Masolo, C. et al., 2002a), (Masolo, C. et al., 2002b)) in order to inherit some knowledge already encoded in the upper ontology (relations, for instance) and to ensure the correct concept classification with respect to concept metaproperties defined in the ontology creation methodology (OntoClean (Guarino, N. & Welty, C., 2002). The mapping was facilitated by OntoWordNet (Gangemi, A. Navigli, R. & Velardi, P. 2003). The relations, inherited from the upper part, are very abstract. However, they were specified further with respect to the domain needs, or were used for consistency checks. Additionally, the ontology was extended with concepts from other sources like terminological lexicons and Wikipedia. At the moment, the domain ontology contains over 1000 domain concepts, about 50 concepts from DOLCE and about 200 intermediate concepts from OntoWordNet.

In order to use the ontology for semantic search over documents, we had to establish a connection between the ontology and the texts of the documents. We established this connection by three types of language resources and tools. The first type is the language specific lexicons aligned to the ontology. Each lexicon contains lexical items grouped by their meaning which is represented in the ontology. There exist various attempts to approach this mapping task. Most of them start from lexicons already existing for several languages, and then try to establish a connection among the concepts defined in these lexicons. Such initiatives were WordNet (Fellbaum, 1998), EuroWordNet (Vossen, 1999), SIMPLE (Lenci et al., 2000). 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 in connecting the ontology to the lexicons. Our model is very close to the LingInfo model (Buitelaar et al. 2006) 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. Thus, the other two language resources are (1) partial grammars which facilitate the mapping from the lexical items to their realization in the texts; and (2) disambiguation rules which solve the problem of ambiguity of the lexical items on the basis of the context of their usage in the texts. The partial grammars consist of

regular expressions that assign the appropriate concept label to a string sequence. The grammars were created semi-automatically on the base of the expressions in the lexicons. They were applied to the lemmas. Here is an example of such a rule for Bulgarian:

Regular Expression:

```
<"кодиране">,<"на">,<"знак">
```

Return Markup:

```
<Concept>\w<conl>
<c>lt4el:CharacterEncoding</c>
</conl>
</Concept>
```

The regular expression presents the sequence of the Bulgarian string [coding] [of] [sign]. The return markup expression assigns to it the domain concept *CharacterEncoding*. The disambiguation of the ambiguous cases was performed also semi-automatically. The implementation of the annotation grammars and disambiguation rules were implemented within the CLaRK System (Simov et al., 2001). A constraint was prepared to stop on the ambiguous cases only. Then a human expert differentiated among various possibilities. Here are some examples: The English term [help] is ambiguous between the concepts *HelpButton* or *HelpCommand*. The Bulgarian term “вмъкване” (inserting) might correspond to three concepts: *Import*, *InsertKey* and *InsertMenuItem*. Within the basic Bulgarian lexicon the ambiguous cases are about 5 % from all the mappings (62 rules with ambiguity out of 1092 altogether). In the next additions there were no such cases, because the concepts became more and more specific, and hence less inclined to ambiguity.

The model is graphically depicted in the next picture:

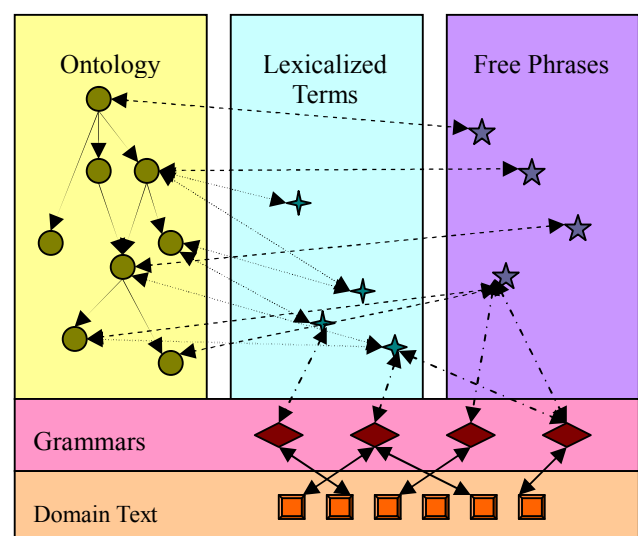


Figure 1: Model of ontology-to-text interrelations.

These mappings ensure the path from the ontology concepts to their string counterparts in the text. Let us

explain the presented architecture more explicitly. The concepts from the ontology were presented in the natural language either as lexicalized terms, or as freely combined phrases. Needless to say, one concept might correspond to only a lexicalized term, only a free phrase or both. All the language expressions were encoded in regular grammar rules, which mapped the string to the appropriate concept(s). The grammars were applied to the domain texts. In cases of ambiguity a human expert took a decision. The ontologically annotated texts became the base for the semantic search facility. Thus the relations, outlined in Figure 1, are a basis for (1) monolingual search in which the ontology inference is used for query expansion; and for (2) multi-language search in which, in addition to query expansion, the ontology is used as a mediator between the various languages.

3. Ontology-Based Semantic Search

Based on three resources as described above, i.e. the term-concept lexicons, the ontology and the concept annotation of the documents, we developed an ontology-based search engine having the following main characteristics, in accordance with the goals of the project LT4eL:

- 1) Domain concept matching: the accessibility to documents in a Learning Management System is improved by retrieving learning materials that have domain concepts in common with the search query. First, the lexicons are used to find domain concepts in the ontology; subsequently, the ontology is used for query expansion, which further improves recall.
- 2) Multilinguality: the functionality is multilingual - one implementation is used for all of the eight languages of the LT4eL project.
- 3) Crosslinguality: the search engine enables users to find documents in several languages at the same time, while using search terms or an ontology representation in one language according to the user's choice.

A presupposition for using the search is the availability of annotated learning material and a lexicon in at least one (or for crosslingual search at least two) of the languages the user knows. In the LMS, the user's choice of languages is a part of her profile. Furthermore, a requirement is that the topics of the documents are (at least partly) covered by the ontology.

The basic idea of the ontology-based semantic search is that concepts from the ontology lead the user to those documents that are appropriate for her query. The search will be most precise when the user directly selects concepts from the ontology. Some other approaches like Finkelstein et al. (2001), Lim et al. (2005) and Song et al. (2005) use a lexical ontology in an automatic way for query expansion, while the retrieval takes place on the basis of the expanded textual query.

Our approach is different in two respects. First, we allow two modes of ontology usage: an automatic mode, where users type search words, and a non-automatic ontology

navigation mode, where documents are retrieved after manually selecting concepts. Second, the ontology, together with the concept annotation of the documents, is used as an intermediate level between query and documents; no step back to a textual representation is involved. This allows for retrieval in multiple languages with one ontology, as well as for crosslingual retrieval.

Related approaches that also make use of concept annotation and retrieval by concepts are described by Kiryakov et al. (2003) and Vallet et al. (2005).

Search Procedure

The search procedure takes the following parameters:

- Language(s) of search query (determines which lexicons to use for concept lookup)
- Retrieval languages (for which the user wants to see available documents)
- Search terms (entered by user), or concepts (selected by user)
- Method for combining the concepts (AND/OR)
- Threshold for ontology-based query expansion

Learning objects are retrieved by means of the following steps:

- The search words are looked up in the lexicons of the chosen languages. Search words are normalized orthographically before lookup. In case the OR option is selected, combinations for multi-word terms are created using several concatenators (e.g. if the words "computer" and "screen" are entered, the combinations "computerscreen", "computer screen" and "computer-screen" are created and looked up, in addition to the individual words "computer" and "screen"). With AND-search, this step is skipped, since it would unintentionally restrict the search: it would require the result document to contain not only the concepts computer and screen, but also "computer screen".
- If lexical entries are found in the lexicon, their denoted concepts are taken as search concepts. These concepts are also used as starting points for ontology navigation. Alternatively, concepts directly selected from the ontology are the basis for the search.
- Documents in the desired languages are retrieved, based on the set of found or selected concepts, while taking into account the AND/OR parameter. If the number of retrieved documents is less than the threshold, the set of search concepts is expanded with their respective subconcepts and the search is repeated. Thus, documents are added that treat a topic at a more detailed level than was specified by the user. We did not experiment with dynamic ways of query expansion like Bonino et al. (2004) propose, where the number of levels and the ontology navigation direction (more general/more specific) is made dependent on the number of results. In our envisaged context, Learning Management Systems, the number of available documents can be very small and the collection biased, so for some queries, a short result list can be a "correct" result. In that case, the learner

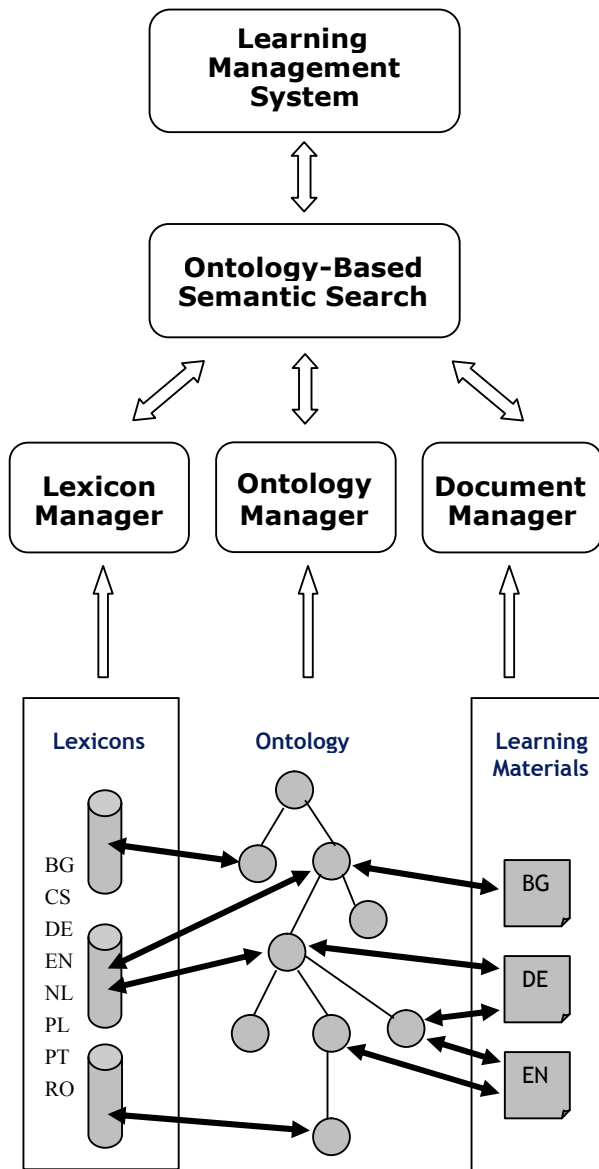


Figure 2: The architecture in which the search functionality is integrated. The lower part of the diagram shows the relationships between terms, ontology and documents.

with foreign language knowledge can look at relevant material in other languages.

- For each retrieved document, the following information is provided, so that the LMS can present it to the user in an appropriate way:
 - 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.
 - Relevance score, based on the occurrences of matching concepts.
 - Snippet: a part of the text selected around the

annotations of matching concepts.

The LMS can then display this information, together with some meta information such as the title of the document, its language and assigned keywords.

The next three subsections give more information on the use of the ontology in search, the relevance scores and the snippets, respectively. Figure 2 gives an overview of the architecture. The search functionality is being integrated into the Learning Management System ILIAS, in which the collected learning materials are stored, while Figure 3 shows an example of the user interface to the search functionality.

Ontology Interaction

Although it is our assumption that the search is most focused towards the desired topic if concepts are directly selected from the ontology, the reasons to allow a free-text query to initiate semantic search are two-fold. 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, and users might avoid semantic search if they think it is much more complicated than full-text search. 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.

In the list of retrieved documents, the concepts that match the query as well as the document are presented. By clicking on a matching concept, the user can switch to the screen for manual ontology navigation starting from that concept, and then select related concepts as the input for a more precise search. It is also possible to start immediately from the ontology navigation view; in this case, no domain concept can be offered to the user as a starting point, so navigation will start from the root of the ontology: the most general available concept.

Relevance score

For each document, a relevance score is calculated, by which the retrieved documents are sorted. It is a value between 0 and 1 that can be presented to the user as a percentage, to indicate the estimated relevance. The value 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 normalized 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.

We opted for a relevance score per document that reflects the correspondence between the query and the document independently of the other retrieved documents or the total available set of documents. It might look logical to set the document with the highest annotation frequency (relative to the document length) at 100% and the rest proportional to it, but we saw that in this way, the scores of the other retrieved documents were often too low and also too much influenced by changes in the repository. Instead, we base the score on an experimental *expected concept-token-ratio* per document. We use 0.005 (one matching concept per 200 tokens) at the moment. Obviously, this can result in a score above 1 for certain documents with very high annotation frequencies. To correct this, we do not cut them to 1, but rather use the following formula, that maps values between lower boundary B and infinity to values between B and 1:

$$\text{corrected} = B + (1-B) * (S-B) / (S-B + 1-B)$$

where B is supposed to be a value between 0 and 1, and S is the score to be corrected. We are currently using B = 0.7, which gives the following corrected scores (examples):

ORIGINAL	CORRECTED
0.8	0.775
1.0	0.850
1.5	0.918
3.0	0.965

Of course, the expected concept-token ratio is chosen such that the values are not above 1 in most cases. The corrected score is a factor in the final relevance score, which is a weighted average where the other factor is the normalized (between 0 and 1) number of main search concepts.

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 are joined with the results of full-text search and keyword search³, which also have a relevance score.

Similar to what Vallet et al. (2005) report about their combination of semantic and full-text search, we found that the final score can be undesirably low when only one of the search methods returns a good relevance score for a relevant document, especially in our case where three methods can be combined. In our opinion, a low score does not necessarily mean that the document is not relevant, but rather that there is no evidence for relevance, while a high score is based on positive evidence from the text or meta data. Therefore, we use a weighted average that favors the best score:

$$0.6 * \text{highest} + 0.3 * \text{middle} + 0.1 * \text{lowest}$$

³ In LT4eL, by *keyword search* we mean matching with words that are assigned as keywords to the documents.

Snippets

Comparable to the presentation of results in Google, we select a snippet of the text, which serves as a preview of the document. It is a small fragment of the document (or two discontinuous fragments, connected by three dots), selected around occurrences of the matching concepts. The annotation itself is removed, but the words that were annotated by one of the search concepts are marked with tags so 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 preferred. The idea behind this is that terms (or the concepts underlying them) describing a topic are likely to co-occur in one sentence or passage. Both Park et al. (2002) and Google (<http://www.google.com/technology/whyuse.html>) use this notion for ranking (a document is more relevant if it contains such a passage) while Hearst (1995) and Drori (1998) 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.

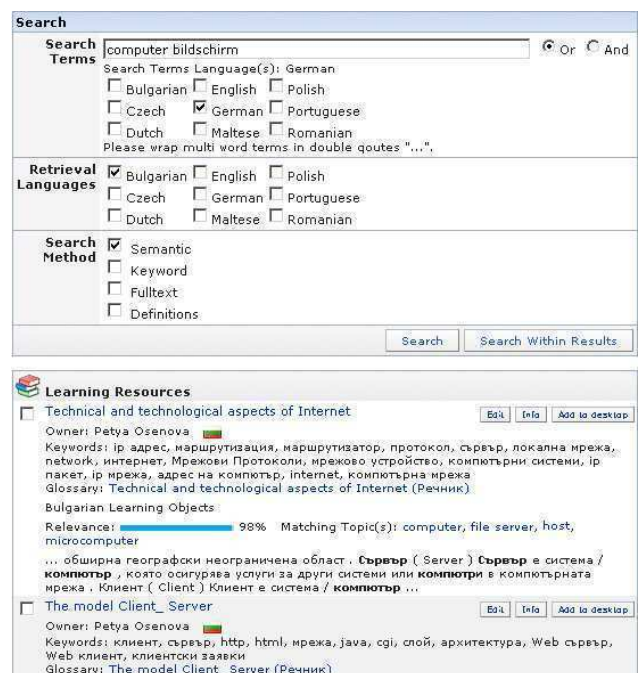


Figure 3: User interface making use of semantic search. Starting from the top, it shows the field for the search terms, OR/AND option, language of search terms, retrieval languages, search methods, and found documents. For the found documents, the interface shows: keywords, language, relevance score, matching concepts, snippet.

4. Evaluation

Within the project two types of evaluation were conducted: user-oriented and search-oriented. The user-oriented one referred to tests with students and tutors, based on various scenarios. In these tests all the languages were involved. The participants had to use different types of searches within LMS in order to perform their tasks. The general

conclusion was that, when handled properly, the semantic search was appreciated for its preciseness and fast results. On the other hand, the text search was the most familiar approach. For that reason, we performed an initial evaluation, based on text-search-like and concept-search like query.

In other words, we tried to evaluate the search function, comparing simple text search and semantic search. The basic task for this evaluation was as follows: two terms (which were also lexical entries in the lexicon of the language under investigation) have been chosen as parts of a query (the equivalents of the terms "program" and "slide" were taken in each language⁴). The combination of these two terms was used under the presupposition that they had unambiguous meanings in the domain texts. We encoded the question as two queries: a simple full text search (including just the lemmatized forms of the terms and the list of terms is extended with related terms which we suppose are known by the learner), and as a semantic search using the ontology to expand the query with all the subconcepts and just one superconcept. For example, the text query for English includes strings like: *program; software; editor; slide*. These basic forms are matched to the lemma annotation within the annotated documents. As it can be seen, this set does not include a sophisticated query expansion on this level. We kept the simple level, because we wanted to be closer to the search possibilities within the LMS⁵. However, in a more general plan, we envisage to make an experiment with advanced text queries. The semantic search query included more information, since it took all subconcepts of *Program* (among which some more specific - *NotePad*, *CorelDraw*, and some more general - *TextEditor*, *SortProgram*). Altogether the expanded queries consist of 96 concept names. These concept names are matched to the conceptual annotation in documents. Both types of queries were run over the document sets. This resulted in two sets of paragraphs, one for the text search and one for the semantic search. The conceptual annotation has been used to identify the paragraphs in these documents. The conjunction of the two result sets has been investigated by a researcher and each paragraph was rated as either relevant or irrelevant to the search. The retrieval results of both methods (text search and semantic search) have been weighted against the set of relevant paragraphs with the well-known measures of recall and precision. Both values have been combined in an F-measure. Similarly to the approach by which CLEF (<http://nlp.uned.es/clef-qa>) initiatives handle recall, we assumed that the sum of all found results equals the recall measure (i.e. that all the relevant paragraphs are among the retrieved ones). Of course, it is not quite true, but it serves well for our goal. The experiment was run for six languages: Bulgarian, Dutch, English, German, Polish and Portuguese. The

⁴ The interpretation of the query is "Which program do you use for the preparation of slides?"

⁵ We thank Pavel Smrř who pointed to us that there is a more advanced text search which could be used as a baseline. We are currently working on a new evaluation.

F-measures for both text search and semantic search are presented in Table 1. The gain is due to improvements in both recall and precision. It is significant for all languages. The gain is the lowest for Portuguese, because there were only a small number of returned documents. Also, there is visible variation between the languages.

Language	Text Search	Semantic search
Bulgarian	56,25	91,30
Dutch	47,50	94,12
English	27,96	79,42
German	36,00	59,26
Polish	12,50	50,00
Portuguese	28,67	33,33

Table 1: F-measures for full text search and semantic search in six languages.

Another factor that played a role in the results was the context: the narrower the context (e.g. sentences), the better the results, and vice versa. As has been said before, the conceptual search produces results only in those cases where the search words are in the lexicon and thus matched to concepts in the ontology. This has been the case in the evaluation example. In the case where the search word did not match a lexical item, the text search as well as the keyword-based search was used as a fallback strategy.

After repairing and extending the ontology, it is a subject to further user-centered evaluations to estimate how well the semantic search performs in the context of the Learning Management Systems real tasks and alternative search methods (text, keyword, definition). This part of the evaluation is still on-going.

5. Conclusion

In this paper, we described a complex architecture for relating the domain ontology to a multilingual collection of texts. We also presented the employed resources and their usage for the ontology-based search. The very first steps towards the evaluation of the search functionality were outlined, which indicates that the ontology-based search significantly improves the retrieval for several languages.

At the moment we believe that the semantic search outperforms the simple text search when querying with more general terms (in general when inference over ontology is used for query expansion and/or answer evaluation). In the case of queries based on the specific terms (which do not allow usage of inference) practically there is no difference between the two types of search. We envisage to test the semantic search against more advanced types of text search.

6. Acknowledgements

This work has been supported by LT4eL (Language Technology for eLearning) - European project of the European Commission under the Information Society and

Media Directorate, Learning and Cultural Heritage Unit (FP6-027391).

We would like to thank also the three anonymous reviewers for their valuable remarks and suggestions.

7. References

- Bonino, D., Corno, F., Farinetti, L., & Bosca, A. (2004). Ontology Driven Semantic Search. *WSEAS International Conference on Automation & Information (ICAI 2004)*, Venice, Italy.
- Buitelaar, P., Declerck, T., Frank, A., Racioppa, S., Kiesel, M., Sintek, M., Engel, R., Romanelli, M., Sonntag, D., Loos, B., Micelli, V., Porzel, R., Cimiano, P. (2006). 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*.
- Drori, O. (1998). The User Interface in Text Retrieval Systems. In: *SIGCHI bulletin*, ACM, New York, July 1998, 30 (3), pp. 26--29.
- Fellbaum, C. (1998). (ed). WORDNET: an electronic lexical database. MIT Press.
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E. (2001). Placing Search in Context: The Concept Revisited. In: *Proceedings of the 10th international conference on World Wide Web (WWW10, May 2001)*, Hong Kong. pp. 406--414.
- Gangemi, A., Navigli, R. & Velardi, P. (2003). The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. Meersman R, et al. (eds.), *Proceedings of ODBASE03 Conference*, Springer, 2003.
- Guarino, N. & Welty, C. (2002). Evaluating Ontological Decisions with OntoClean. *Communications of the ACM*, 45(2): 61-65.
- Hearst, M. (1995). TileBars: Visualization of Term Distribution Information in Full Text Information Access. In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI '95)*, pp. 56--66.
- Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., & Goranov, M. (2003). Semantic Annotation, Indexing, and Retrieval. In: *Proceedings of the Second International Semantic Web Conference (ISWC 2003)*, Sanibel Island, FL, USA.
- Lenci, A., Busa, F., Ruimy, N., Gola, E., Monachini, M., Calzolari, N., Zampolli, A., Guimier, E., Recourcé, G., Humphreys, L., von Rekovsky, U., Ogonowski, A., McCauley, C., Peters, W., Peters, I., Gaizauskas, R. & Villegas, M. (2000). SIMPLE Work Package 2 - Linguistic Specifications, Deliverable D2.1. *ILC-CNR*, Pisa, Italy.
- Lim, S., Park, S. & Lee, S. (2005). Document Retrieval Using Semantic Relation in Domain Ontology. In: P.S. Szczepaniak et al. (Eds.): *AWIC 2005, LNAI 3528*, Springer Verlag, Berlin Heidelberg, pp. 266--271.
- Masolo, C., Borgo, S. & Gangemi, A. & Guarino, N. & Oltramari, A. & Schneider, L. (2002a). The WonderWeb Library of Foundational Ontologies. WonderWeb Deliverable D17, August 2002. <http://www.loa-cnr.it/Publications.html>.
- Masolo, C., Borgo, S. & Gangemi, A. & Guarino, N. & Oltramari, A. (2002b). Ontology Library (final). WonderWeb Deliverable D18, December 2003. <http://www.loa-cnr.it/Publications.html>.
- Monachesi, P., Lemnitzer, L. & Simov, K. (2006). Language Technology for eLearning. In *Proceedings of EC-TEL 2006, Innovative Approaches for Learning and Knowledge Sharing, LNCS 0302-9743*, pp. 667-672.
- Park, E., Moon, S., Ra, D., & Jang, M. (2002). Web Document Retrieval Using Sentence-query Similarity. In: *Proceedings of the 11th Text Retrieval Conference (TREC-11)*.
- Simov, K., Peev, Z., Kouylekov, M., Simov, A., Dimitrov, M., Kiryakov, A. (2001). CLaRK - an XML-based System for Corpora Development. In: *Proc. of the Corpus Linguistics 2001 Conference*. Lancaster, UK.
- Song, M., Song, I., Hu, X., & Allen, R. (2005) Semantic Query Expansion Combining Association Rules with Ontologies and Information Retrieval Techniques. In: A. Min Tjoa and J. Trujillo (Eds.): *Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2005)*, LNCS 3589. Springer Verlag, Berlin Heidelberg, pp. 326--335.
- Vallet, D., Fernández, M., & Castells, P. (2005). An Ontology-Based Information Retrieval Model. In: Gómez-Pérez, A., Euzenat, J. (eds.): *The Semantic Web: Research and Applications: 2nd European Semantic Web Conference (ESWC 2005). Lecture Notes in Computer Science, Vol. 3532*. Springer Verlag, Berlin Heidelberg, pp. 455--470.
- Vossen, P. (1999). (ed). EuroWordNet General Document. Version 3, Final, July 19, 1999, <http://www.hum.uva.nl/~ewn>.