

AsIsKnown

**A semantic-based knowledge flow system for
the European home textiles industry**

Work package 3: Common sense ontology engineering

**Deliverable D11 “Reports on the tests and evaluation of common
sense ontology”**

Second version

Lead participant: IPP-BAS
Nature: Report
Dissemination level: PU
Delivery date: 21 PM



This document has been produced in the context of the AsIsKnown Project. The AsIsKnown project is part of the European Community's Sixth Framework Program for research and development and is as such funded by the European Commission. All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors view.



Context

WP 3	Common sense ontology engineering
Task 3.7	Testing and evaluation
Dependencies	This deliverable requires user requirement input from 3.2 and 3.4

Contributors: Kiril Simov (IPP-BAS) Petya Osenova (IPP-BAS)	Reviewers: Kiril Simov (IPP-BAS)
--	--

Approved by: Kiril Simov, Bulgaria as WP3 Head



Executive Summary

This report focuses on Tests and Evaluation of Common Sense Ontology. The testing and evaluation of the ontology is viewed as a process parallel to its exploration.

The testing part was concerned with the following tasks: aiming at an adequate common conceptualization within the system, which to serve all the users; multilingual access to the ontology via language specific lexicons; supporting the text mining module of the Trend Analyzer. With respect to the first task, a comparison was performed between the concepts and the categories, which were catalogued by the partners. As a result, the mismatches were detected and repaired (by adding new concepts or meanings, changing some previous ones etc.). With respect to the second and third task, a model was created to ensure the correct mapping among concepts and terminological lexicon items, and also – among concepts and the appropriate text chunks via the creation of annotation grammars. Our model benefited from the existing initiatives and best practices (SIMPLE, WordNet, LingInfo among others), but it also was innovative in many respects (the direction of workflow, the type of the ontology and the usage of different models for the various components).

The evaluation part included the following steps: first, manual annotation was done over part of the fashion magazine texts. Thus, a golden standard was prepared for future evaluation, and the insufficiency of the textile domain ontology was discovered. For the prognosis tasks of the Trend Analyzer the ontology had to be extended further. It was enriched with about 2400 new concepts. Second, some manual checks over the annotated documents were performed to discover inconsistencies. Third, a comparison was done between the manually and automatically annotated documents. This comparison showed the good coverage of the ontology. The only problem we encountered towards the evaluation task was the heavy apparatus of the upper part of the ontology (incomprehensible for the common user and burdened with redundant relations). The problem was solved by substituting the existing top ontology DOLCE with its lighter version DOLCE Ultralite, which was kindly provided by the ontology developers.

Table of Contents

EXECUTIVE SUMMARY	4
TABLE OF CONTENTS	5
LIST OF ABBREVIATIONS	6
1 INTRODUCTION AND PROBLEM STATEMENT	7
2 REPRESENTATION OF THE PARTNER'S INFORMATION	8
3 ONTOLOGY TO TEXT MAPPING	8
3.1 Ontology to Text Relation Model.....	8
3.2 Annotation of Magazine Articles	9
3.3 User Interface	11
4 CONCLUSIONS AND OUTLOOK	12
REFERENCES	13



List of Abbreviations



1 Introduction and Problem Statement

In this deliverable we discuss the test and evaluation of the common sense domain ontology in the area of Home textile done during the period from month 12 to month 21 of the project. As we have reported in the first version of the Deliverable D11, in our work we accept the ideas of [1] that evaluation of the ontology is part of the life cycle of its development, and therefore it takes place in parallel to this development. In this respect our work on testing and evaluation was guided by usage of the ontology within the project.

We used the ontology in three ways within the AsIsKnown System: (1) as a common conceptualization, which is able to accommodate the different conceptualizations of the domain used by the various users of the system (producers of home textile, interior designers, etc); (2) facilitating the human dialog with the system in different languages by mapping the ontology to the corresponding language lexicons; (3) supporting the text mining facilities of the Trend Analyser.

In order to be able to support the first task we compared the concepts within the ontology with the categories incorporated by the partners in AIKXML. AIKXML is a special XML language for description of home textile products jointly created by the partners within the project. We compared the first version of the ontology and the categories within AIKXML. On the basis of this comparison a list of correspondences was compiled between the categories and the concepts as well as a list of the missing categories. The missing categories will be added to the ontology before the end of the second year. The mapping between the ontology and the AIKXML will allow the usage of the entire ontology technology within the AsIsKnown System, namely - the inference and the mapping of the ontology to the lexicons.

The second and third tasks require a model of mapping between the ontology and language elements – lexical entries and textual chunks. We created such a model which reflects the last developments in the area of ontologies and lexicons. The model was published in [2] and it is also described below. We used the model in order to create annotation grammars for the textual part of the multimedia documents used by the Trend Analyser. The annotation then was checked manually for domain coverage. The results from this manual check were used for ontology extension in order to reach better abstraction over the content of the multimedia documents processed by the Trend Analyser. As a consequence from this, the ontology was enriched with about 2400 new concepts.

2 Representation of the Partner's Information

The main function of the ontology within the AsIsKnown system is to map the various representations of the users' data to a common conceptualization. We consider this function successful if the ontology provides appropriate concepts which incorporate the data of the project partners.

These data are summarised within the definition of AIKXML developed within the project. We annotated each identifier (or term) defined in AIKXML with an appropriate concept from the ontology. In this annotation we also tried to select the concept which is on the same level of granularity as the term in AIKXML.

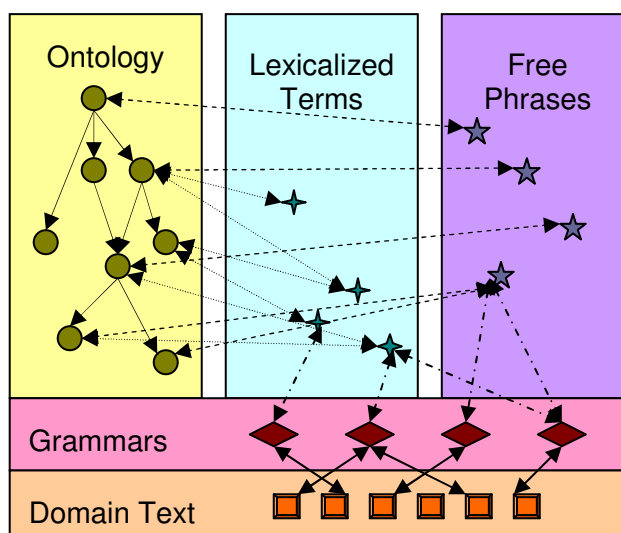
The result of this annotation identified the terms which were not covered by the ontology in its current version. The concepts necessary to cover these terms will be added to the ontology.

3 Ontology to Text Mapping

In this section we present work done on the evaluation of the ontology with respect to ensuring the dialog with the system and analysis of magazine articles for supporting text mining within Trend Analyser.

3.1 Ontology to Text Relation Model

In this subsection we present briefly the linguistic model of the annotation process adopted for the tasks within the project. We assume that the ontology is the repository of the lexical meaning of the language. Thus, we have 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 is overcome by allowing also non-lexical (fully compositional) phrases to be represented in the lexicon. The second problem is solved by the extension of the ontology. The lexicon items are then mapped to grammars. These grammars relate the lexicon to the text. Such a mapping is necessary as much as 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 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.

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 start 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.

3.2 Annotation of Magazine Articles

The initial annotation grammar for the concept annotation was covering only concepts from the home textile domain. The annotation with this grammar was very sparse as much as the magazine articles usually discuss complete interior designs in which the textile elements are only a fraction of the concepts mentioned in the articles. Thus, the concurrences of concepts were not frequent enough. The conclusion from this observation was that in order to support more interesting generalizations over the content of the articles we had to annotate more concepts in the text than the ones in the home textile ontology. This meant that we had to determine new domains related to the one of home textile ontology which to be covered by a new extended ontology. In order to do this we proceeded in the following way: (1) first, we annotated manually the magazine articles with terms from domains of architecture, interior, materials, etc; (2) then we extracted these terms from the articles, we mapped them to OntoWordNet (similarly to the way in which we constructed the home textile ontology), we have all related concepts from OntoWordNet to the ontology. In this way we added about 2400 new concepts.

Using the model, presented in the previous section, we constructed an English lexicon which contains the corresponding terms to each concept. We started to construct the Bulgarian lexicon. The representation of the entries in the lexicons is as follows:

```
<entry id="entry-34">
  <owl:Class rdf:about="http://www.asisknown.org/AIKHT#Carpet">
    <rdfs:comment>a piece of thick heavy fabric (usually with nap or pile)
      used to cover a floor</rdfs:comment>
    <rdfs:subClassOf>
      <owl:Class rdf:about="http://www.loa-cnr.it/ontologies/OWN/OWN.owl#FURNISHINGS"/>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Class rdf:about="http://www.asisknown.org/AIKHT#FloorCovering"/>
    </rdfs:subClassOf>
  </owl:Class>
  <def>a piece of thick heavy fabric (usually with nap or pile) used to cover a floor</def>
  <termg lang="en">
    <term shead="1">carpet</term>
    <term>carpeting</term>
    <term>rug</term>
    <term>textile floor covering</term>
    <def>a piece of thick heavy fabric (usually with nap or pile) used to cover a floor</def>
  </termg>
  <termg lang="bg">
    <term shead="1">килим</term>
    <term type="nonlex">текстилно подово покритие</term>
    <def>Парче от дебело платно, което се използва за подово покритие.</def>
  </termg>
</entry>
```

Each entry comprises: (1) a fragment of the ontology which shows the concept (relation) for which terms are defined in the entry; (2) definition of the concept (relation) in English is given in order to facilitate the human understanding of the concept. This definition could be included in the ontology fragment, but it also might be missing; (3) Term group for each language is given. In the example two term groups are given – one for English and one for Bulgarian. Each term group consists of a list of terms and a definition in the corresponding language. Each term might be lexicalized or non-lexicalized. The latter case is marked-up by the attribute *type* with value *nonlex*. One of the terms in the term group is annotated as a representative for the group by the attribute *shead* with a value *1*.

Based on the English lexicon we have constructed annotation grammar which related the terms to their occurrences in the text. For the implementation of the annotation grammar we rely on the grammar facilities of the CLaRK System¹. 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)>
```

¹ <http://www.bultreebank.org/clark/index.html>

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 grammar started with the terms in the lexicon. 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 ambiguous cases were manually disambiguated. In future we envisage the development of rules for automatic disambiguation. Also, we envisage experiments with machine learning techniques for this task.

3.3 User Interface

As we mentioned above, one of the terms in the lexicon is marked up as a representative for the set of terms for the given concept. These representative terms are used when it is necessary to present the ontology to a user who is not an expert in ontologies. Besides concepts, similar sets of terms are linked to relations in the ontology. Thus, by translating the concepts and relations to these representative terms we could have the ontology (or a fragment of the ontology) represented in a given natural language. The main problem we have encountered here concerns the upper ontology. Sometimes the user will need to navigate the ontology above the domain part. This need imposes two problems: (1) the lexicon has to cover also the upper part of the ontology; (2) the upper part of the ontology has to be understandable for the user. The first problem is easy to solve as much as we could prepare a lexicon for it as well. The second problem is much more difficult, because it would require some simplification of the upper ontology which is not a trivial task. Fortunately, the authors of the upper ontology (DOLCE) which is used by us, already prepared a simplified version – DOLCE Utltralite. The new version uses smaller number of concepts and relations with more intuitive names. Thus, we decided to change the upper part of the ontology to the new version. In pursuing this task we need to construct a mapping between DOLCE Utralite and DOLCE. This process is ongoing at the moment. After finishing it, we will construct the necessary lexicon for the new upper ontology.

4 Conclusions and Outlook

In this report we present the tasks we have done or started within the period for testing and evaluation of the AsIsKnown Home Textile Ontology. We were concerned with the coverage of the ontology with respect to: (1) incorporation of the partners' data and (2) supporting the annotation of the magazine articles. This resulted in addition of new concepts in the ontology.

Also we initiated the process of substitution of the upper part of the ontology with a new simplified version. This is necessary in order to support better usability of the ontology.

References

- [1] Peter Haase, York Sure. 2005. *D3.1.2 Incremental Ontology Evolution-Evaluation*. Deliverable D3.1.2 EU-IST Project IST-2003-506826 SEKT.
- [2] Kiril Simov and Petya Osenova. 2007. Deliverable D11 "Reports on the tests and evaluation of common sense ontology" – first version. AsIsKnown Project.
- [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, Wim 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.