Kiril Simov

Ontology-based Lexicon of Bulgarian

In contrast to morphological and syntactic processing semantic annotation based on domain ontology is still underdeveloped for Bulgarian. On the other hand, the prerequisites for an ontological annotation are already available. These are as follows: a morphosyntactic tagger for Bulgarian with more than 95% accuracy; a dependency parser with more than 84% accuracy; a general chunker and a named entity grammar. Semantic annotation is, therefore, the next logical step. We consider the following to be the minimal set of semantic resources:

- a lexicon for Bulgarian mapped to an upper ontology as a mechanism to cover the common lexicons in domain texts, and mapped to domain ontologies to cover domain terminology;
- an annotation grammar for Bulgarian, based on syntactic knowledge of Bulgarian and conceptual information from the ontology. It comprises grammar rules for recognition of lexical units in the text and rules for selecting the right interpretation in context;
- a corpus, manually annotated with ontology information in order to train the machine learning component for automatic word sense disambiguation — selecting the appropriate concept for a lexical unit in context.

In this paper, we will focus on the description of the lexicon. For that purpose we first present the structure of the domain ontology and then a model of `ontology-to-text` relations which facilitates the text annotation. The ontology-based lexicon for Bulgarian is viewed as part of the model. It is constructed in an incremental manner.

1 Introduction

There exist various types of lexicons. Some of them just register the lexemes and/or their wordforms, while others reflect valency or lexical relations, such as synonymy, antonymy, etc. There are also the so-called thesauri lexicons, which combine the formal and lexical aspects of the lexemes. The most popular thesaurus is the WordNet lexical database. On top of these types of lexicons there exist generative ones (in Pustejovsky's sense), which connect the words in the lexicon to some kind of linguistic ontology (e.g. EuroWordNet, SIMPLE lexicon). For the purposes of the semantic annotation of domain text another type of lexicon is needed, namely a lexicon mapped to a domain ontology. However, such a lexicon cannot exist in isolation from the more general lexicon

which covers non-domain words. The combination of the general and domain-specific lexicons is very important for the semantic annotation on domain level since the usage of the domain terms in the text depends on the semantics of the common words in the context. Semantic annotation is an enhancement in the area of language resources after the creation of morphologically and syntactically annotated corpora. The importance of semantic annotation became a hot topic within the Semantic Web initiative. Although much work is already done in this area, the term 'semantic annotation' is not yet well defined - see (Kiryakov et. al, 2005) and citations therein. In our work we think of a text as consisting of two types of (non-linguistic) information: (1) ontological classes and relations, and (2) world facts. The ontological part determines generally the topic and the domain of the text. The corresponding 'minimal' part of ontology implied by the text will be called 'ontology of the text'. The world facts represent an instantiation of the ontology in the text (here entities like beliefs, claims, attitudes, etc. are also included). Both types of information are called uniformly 'semantic content of the text'. Both components of the semantic content are connected to the syntactic structure of the text. Any (partial) explication of the semantic content of a text will be called semantic annotation of the text. Defined in this way, the semantic annotation could contain also some pragmatic information and actual world knowledge.

To support the semantic annotation we rely on an ontology-based lexicon. We assume that there is a domain ontology which is used in the process of annotation. A domain ontology comprises three layers: domain, middle and upper. The lexicon is mapped to the ontology. This mapping is based on relations between the meaning of the lexical units in the lexicon and concepts (relations and instances) in the ontology. Thus, we assume that the ontology contains the conceptual information necessary to model the word senses in the lexicon. The advantage of using an ontology is that the reflections of the conceptualization of the world become explicit.

The motivation for the construction of such a lexicon is the need for more precise semantic annotation. In order to ensure this, the lexicon has to provide more complex conceptual information than the one in computational lexicons like WordNet. The second requirement for the ontology-based lexicon is the coverage of the words in the text. The lexicon has to cover not only the domain terms, but also the non-specialized language. This is necessary for ensuring enough explicit knowledge for the application of word sense disambiguation methods based on statistics. Since the development of a general ontology to support all the lexical units in a language is an intractable problem, we construct the ontology in an incremental way from the upper ontology to the middle and domain specific ontology. Then lexical units are mapped to this ontology via two relations — `equality` and `subsumption`. The first is used when the appropriate concept for a meaning of some word is already represented in the ontology. The latter is used when such a concept is missing and only super-concepts are available.

We place a special emphasis in this paper on the role of metonymy and regular polysemy. They are encoded as specific patterns extracted from a semantically annotated corpus and reflect the conceptual structure of the ontology. The lexicon is also connected to an annotation grammar which establishes a relation between the ontology and

the text. Other phenomena like metaphoric relations and near-synonyms are not treated at the moment in this version of the lexicon, but the model provides possibilities for extensions to represent them in future.

Although being part of the annotation process, the concept annotation grammar remains beyond the scope of this paper. When mentioned, it is only with the aim to highlight the lexicon within a context. More specifically, our idea is the following: the ontology delivers the concepts within the world, the lexicon stores the wordings of those concepts (both lexicalized and non-lexicalized), and the text disambiguates the concepts within a concrete discourse pattern.

The structure of the paper is as follows: the next section discusses related works; then the architecture of a domain ontology comprising an upper ontology, a middle ontology, domain specific part and linguistic knowledge mapped to them is explained; the fourth section presents the creation of the ontology-based lexicon of Bulgarian. It first discusses a model for domain lexicons used for semantic annotation. Then the model is extended to cover the general language lexicon. At the end the encoding of some special phenomena is presented; and the last section concludes the paper.

2 Related Works on Ontology and Lexicon

Ontologies and lexicons are artifacts reflecting the human abilities for representing, processing and managing linguistic and conceptual knowledge. As such, they allow for the combination of many different approaches. A recent overview of the relation between ontologies and lexicons is presented in (Hirst, 2004). The paper discusses the structure of lexical entries, the knowledge recorded in them and mechanisms for interrelation of the lexicon elements. Special attention is given to the definition of 'word sense', its conceptual structure, relations between senses and problematic cases. The main topics under discussion are near-synonyms, gaps in the lexicon, and linguistic categorizations that are not ontological. We treat these topics as follows.

First, we assume that the lexicon is based on the ontology, i.e. the word senses are represented by concepts, relations or instances. Near-synonyms are words that share the same central conceptual information, but differ in the additional information they provide to the semantic interpretation module, such as small changes in the denotation, different implications, speaker attitude, etc. Our model does not solve this problem completely. In it only the central part of the meaning of a word that can be represented. The additional parts of the meaning (context related variations) can be encoded as additional information in the lexical entry or as an extension of the ontology where it is appropriate — similarly to the model used in (Edmonds and Hirst, 2002). The problem of lexical gaps is solved by allowing the storage of free phrases. Similarly, gaps in the ontology (a missing concept for a word sense, for example) are solved by appropriate extensions of the ontology. Linguistic categorizations that are not ontological are not treated in our model.

As it was mentioned above, the construction of a Bulgarian ontology-based lexicon is motivated by the need to introduce more world knowledge into the semantic analysis

of texts. (Morris and Hirst, 2004) points out that most of the lexical relations necessary to determine the semantic content of lexical units are non-classical in contrast to the classical ones, i.e. `hyponymy`, `meronymy`, `antonymy`. The non-classical relations are specific to some classes of meanings, i.e. `made-of`, `used-for`, etc. In our case we assume that these relations are represented in the ontology. Thus, they are formally defined, can be used for the purposes of semantic inference and can be used for the representation of some language phenomena like polysemy, metonymy, etc.

Regarding the complexity and precision of a given ontology we follow the definition in (Guarino, 2000). It represents the following classification of ontologies:

- **Lexicon**: *Machine Readable Dictionaries; Vocabulary with natural language definitions*
- **Simple Taxonomy**: *Classifications*
- **Thesaurus**: *WordNet; Taxonomy plus related-terms*
- **Relational Model**: *Light-weight ontologies; Unconstrained use of arbitrary relations*
- **Fully Axiomatized Theory**: *Heavy-weight ontologies.*

The classification starts with a less formal and knowledge-poor ontology (hence — simple lexicons) and ends with heavily constrained theories about the world. Sometimes the first three elements of the classification are not considered as ontologies, because the ontological information is represented mainly implicitly. As it was pointed out to us by one of the reviewers this hierarchy shows the transition from lexicon to ontology. In our view such a transition supports the mapping between the ontology and lexicon. Our attempt is to move the current semantic lexicons for Bulgarian from the level of thesaurus to the level of light-ontologies (as a minimum).

Our approach draws in many respects on the work done on WordNet (Fellbaum, 1998), EuroWordNet (Vossen, 1998), SIMPLE (Lenci et. al, 2000). With WordNet-like lexicons — (Fellbaum, 1998) and (Vossen, 1998) — we share the idea of grouping lexical units around a common meaning and in this respect the term groups in our model correspond to synsets in the WordNet model. The difference is that the meaning is defined independently in the ontology. With the SIMPLE model (Lenci et. al, 2000) we share the idea to define the meaning of lexical units by means of an 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: Generative Lexicon — (Pustejovsky, 1995). Similar is the connection with EuroWordNet.

With the LingInfo model — (Buitelaar et. al, 2006a), (Buitelaar et. al, 2006b) and (Romanelli et. al, 2007) — we share the idea that grammatical and context information also needs to be presented and linked to the ontology, but we differ in the implementation of the model and the degree of realization of the concrete language resources and tools.

Finally, we would like to mention the work within the Ontology Semantics (Nirenburg and Raskin, 2004) which is very similar to our model except that we use existing ontologies like DOLCE and we allow for an incremental construction of the lexicon.

3 The Structure of a Domain Ontology

Our work is based on a model developed for the annotation of domain concepts in a text. In this model we assume that the ontology is the starting point for the creation of the `ontology-to-text` relation. The structure of a domain ontology can be defined (at least) with respect to: (1) the collection of concepts represented in the ontology; and (2) the complexity of the concept definitions.

Independently from the methodology for ontology creation, the concepts represented in the ontology can be distributed on the following layers which reflect the generality of the conceptual information:

- **Domain layer**.

  At this layer we have the domain concepts and relations representing the main notions in the domain. These concepts and relations are used for solving different tasks, such as the representation of domain knowledge, the representation of common conceptualization for information exchange in the domain, the semantic annotation of domain texts, etc.

- **Upper layer.**

  The alignment of the domain layer to an upper ontology is an obligatory step in each ontology creation methodology. This alignment ensures several properties of the domain ontology: (1) its consistency with the design of the upper ontology; (2) inheritance of the knowledge represented in the upper ontology. The inheritance requires the imposition of more specific constraints reflecting the structure of the domain.

- **Middle layer.**

  This layer contains concepts and relations which are neither part of the upper layer, nor of the domain one, but play an important role for the alignment between them. For example, 'carpet' is in the domain layer for the Home Textile ontology and 'artifact' is in the upper layer, but the concept for 'covering' which is more specific than 'artifact' and more general than 'carpet' (defined as textile floor covering) is in the middle layer. This layer is the result of the ontology creation practice and depends on the coverage of the domain and the range of concepts defined in the upper ontology. In our view it is a useful instrument for transition from the domain to the upper layer.

An additional layer related to the conceptual information is the linguistic information represented by a lexicon and a grammar. This information is necessary in all cases where

the ontology interacts with natural language, for example in the analysis of texts, when navigating the ontology, in ontology based searches, etc.

- **Language layer.**

  It is assumed that the ontology with its three layers is language independent, formalized in some ontology representation language. In practice, such an ontology would be incomprehensible to humans and therefore has to be mapped to some linguistic resources. This mapping is required for at least two reasons: to allow to present the ontology to users who are not ontology engineers, and to support semantic analysis and retrieval of texts. Thus, as a minimum it is necessary to have a lexicon mapped to the concepts and the relations in the ontology. For example, the concept 'scanner' would have several possible wordings in English, such as 'scanner', 'image scanner', 'digital scanner'.

The following figure shows the structure of a domain ontology:

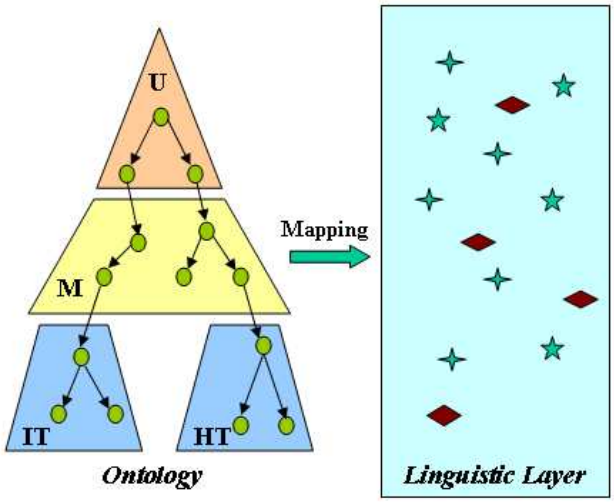

Fig 1. Domain ontology structure.

Here on the left side we have two domain ontologies (IT for the domain of Information Technology for End Users and HT for the domain of Home Textile) aligned to the middle layer and the upper layer. The linguistic layer consists of lexical units, grammar rules, disambiguation rules, etc. The mapping between the ontology and the linguistic layer is the way to define the `ontology-to-text` relation. This relation supports the semantic annotation of text. It generally comprises a lexicon and a grammar.

We have used this structure of the ontology in three European projects - LT4eL, AsIsKnown and LTfLL. In each of them we have used as an upper ontology the DOLCE Ontology (Masolo et. al, 2002) for several reasons: (1) it is constructed on a rigorous basis which reflects the OntoClean methodology (Guarino and Welty, 2002); (2) it contains many useful relations and axioms which constrain the interpretation of the ontology; (3) it is represented in OWL-DL. For the middle layer we have used OntoWordNet (Gangemi et. al, 2003) - a version of WordNet aligned to DOLCE. The domain layer is created for each domain. The result of the three layers is a domain ontology with a better structuring of the concepts and relations. In addition, relations and axioms are inherited from the DOLCE upper part to the specific domain layer. The linguistic layer was implemented via domain lexicons, presented in the next section, and concept annotation grammar, described in (Simov and Osenova, 2007) and (Simov and Osenova, 2008).

4 An Ontology-based Lexicon of Bulgarian

In this section we present our work on a Bulgarian ontology-based lexicon. First we describe the structure of domain lexicons developed for the projects mentioned above. Then we extend their structure in order to overcome the problems we faced with respect to the annotation of domain texts.

4.1 Domain Lexicons

In order to support semantic annotation of domain texts we have defined an `ontology-to-text` relation based on three elements — domain ontology, domain lexicon and concept annotation grammar. The model of the domain lexicon is based on the assumption that the ontology has a central role in the definition of the `ontology-to-text` relation and the language information reflects the available conceptual information in the ontology. The mapping is directed from the ontology to the lexicon, then from the lexicon to the grammar and then to the text. For each concept (relation, instance) in the ontology the lexicon contains at least one lexical unit. This requires the lexicon to contain non-lexicalized (fully compositional or free) phrases as well[1]. Availability of different lexical units (lexicalized or not) for a given concept is used as a basis for the construction of the annotation grammar. This availability allows us to capture different wordings of the same meaning in a text. In general, a concept might have a few terms connected to it and a (potentially) unlimited number of free phrases expressing this concept in the language. Some of the free phrases receive their meaning compositionally regardless of their usage in a given text, other free phrases denote the corresponding concept only in a particular context. In our lexicons we register as many free phrases as possible in order to have better recall on

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

the semantic annotation task. In cases when a lexicalized concept is missing in the ontology we modify it. The model was used for the construction of lexicons in several languages for two domains — Information Technology and Home Textiles. When we have lexicons in several languages mapped to the same ontology we ensure a certain level of multilinguality. The following figure shows the mapping from the ontology to the lexicon:

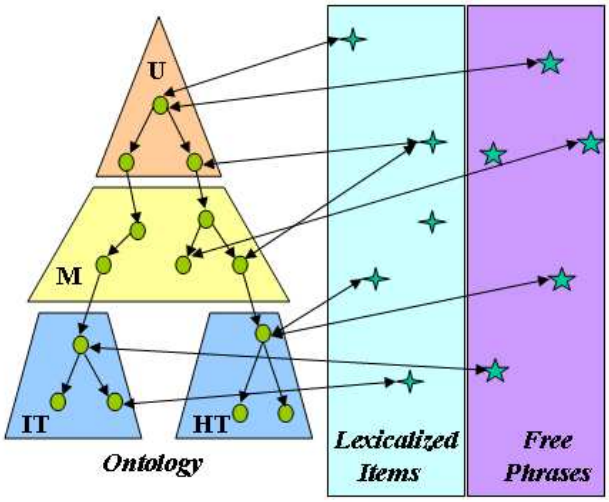

Fig 2. Ontology to lexicon mapping using `equality` relation.

The lexicon plays a double role within the three projects. 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 way natural for the user. For example, the concepts and relations might be named with lexical units employed 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 the sense that the ontology could be viewed through different terms depending on the context. For example, the colour names will vary from very specific terms within the domain of carpet production to more common words used when the same carpet is part of an interior design. Thus, the lexical entries contain the following information: lexical units (words or phrases), contextual information determining the context of the their usage, grammatical features determining their 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 user

roles. In the home textile domain the users are producer, retailer, interior designer, etc. In the eLearning domain they are student, teacher, tutor, administrator. Depending on the role, the ontology is filtered with the help of the lexicon, which focuses preferably on the user-specific words. For example, the carpet producer would be interested in the number expressions of the colours as presented in a standard, the various types of carpets and the detailed structure of the carpet. On the other hand, the interior designer would like to explore the relation of the carpet to other parts of the interior (wall, paintings, furniture, etc.).

We could summarize the connection between the ontology and the lexicons in the following way: the ontology represents the semantic knowledge in the 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 aimed at representing in our lexicons only the most frequent and important words and phrases. Here is an example of an entry:

```
<entry id="entry-34">
 <owl:Class rdf:about="http://www.asisknown.org/AIKHT#Carpet">
   <rdfs:comment>a piece of thick heavy fabric
                 used to cover a floor</rdfs:comment>
   <rdfs:subClassOf>
     <owl:Class rdf:about=http://www.asisknown.org/AIKHT#FloorCovering/>
   </rdfs:subClassOf>
 </owl:Class>
 <def>a piece of thick heavy fabric used to cover a floor</def>
 <termg lang="en">
   <term shead="1">carpet</term>
   <term>carpeting</term>
   <term>rug</term>
   <term type="nonlex">textile floor covering</term>
   <def>a piece of thick heavy fabric used to cover a floor</def>
   <gramline>reference to finite state grammar</gramline>
 </termg>
</entry>
```

Each entry in the lexicons contains the following 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; (3) a set of lexical units (in domain lexicon we call them terms) in a given language representing the concept; and (4) relation to grammar rules. The concept part of the entry provides the minimum information necessary for a formal definition of the concept. The English explanation of the concept meaning facilitates human understanding. The set of terms stands for different wordings of the concept in the respective language. One of the terms is a representative for the term set. Note that this is a somewhat arbitrary decision, which

might depend on the frequency of term's usage or on the expert's intuition. This representative term will be used where only one of terms from the set is called for, for example as an item of a menu. In the example above we present the set of English terms for the concept 'carpet'. One of the terms is non-lexicalized — the attribute `type` with value `"nonlex"`. The first term is representative for the term set and it is marked-up with the attribute `shead` with the value `"1"`. The elements `gramline` provide links to linguistic features of the terms like lemmatized variants of the terms, implementation as regular expressions to be compiled as finite state automata.

The second component of the `ontology-to-text` relation, namely the concept annotation grammar, is ideally considered to be an extension of a general deep grammar of a given language which is adopted in 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 the recognition of the term. The annotation with grammatical features and the lemmatization of the text are considered a preprocessing step. The disambiguation rules exploit the local context in terms of grammatical features, semantic annotation and syntactic structure, and also the global context such as the topic of the text, the discourse segmentation, etc. Currently we have implemented chunk grammars for several languages. We have implemented a very simple disambiguator which uses an unigram model. The annotation grammar is implemented within the CLaRK System (Simov et. al, 2001).

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

4.2 The Extended Model

The main problem with the model of the `ontology-to-text` relation, described in the previous section is the fact that the lexicon is mapped only in its domain part to the ontology. Thus, the annotation of domain texts with domain concepts is very sparse. For example, in the IT domain we have annotated 8 concepts within 100 tokens (with 14.8 tokens per sentence = 1.19 concepts per sentence at average). This sparse annotation blocks possibilities for using better methods for word sense disambiguation. This holds when the lexical units in the domain lexicon are ambiguous among themselves or with respect to the lexical units from the general lexicons. For example, the concepts 'key-of-keyboard', 'key-of-database' and 'key-for-door' have the same wording in English ("key") and the last one is not from the domain ontology. Therefore, we need a much better semantic annotation than one which just uses the domain terms and grammar constructed on their basis.

To achieve such a better annotation we consider two tasks to be solved: (1) to ensure better coverage of the text with conceptual information and (2) to exploit a better disambiguation model using this information. For the first task we envisage two interconnected solutions: (1) to improve the annotation grammar, and (2) to provide an interaction with the general lexicon. The second task is not discussed here.

The annotation grammar can be improved in several directions. Most prominent seems to be the usage of the syntactic structure and co-referential relations in order to distribute the domain knowledge to general lexical units or phrases in the text. For example, if a document is about 'desktop publishing system' very often this concept can be referred with expressions like 'publishing system' or 'system'. Similarly, predicates impose semantic restrictions on their arguments. The interaction with the general lexicon can be achieved via connecting of the domain lexicon to lexicons like WordNet. The interaction between the two solutions is as follows. First, the words in the text receive their annotation from the domain and the general lexicons. The grammar ensures additional distribution of the domain annotations via the co-referential relations and syntactic structures. Finally the disambiguation module selects the right annotations. In order to support this processing the domain terms and the common words in the context have to share the semantic annotation. In order to ensure this we have to augment the general lexicon with appropriate semantic information.

Ideally, each meaning of a lexical unit from the general lexicon has to be present in the ontology in order to use the model of `ontology-to-text` relation discussed above. Unfortunately, such an ontology does not exist yet. Thus, we have to use a smaller ontology and to change the implementation of the `ontology-to-text` relation.

On the basis of the gained experience within the projects mentioned above we conclude that there exists a relatively stable upper and middle part for each of the domain ontologies. Therefore, we think that a first step for the creation of an appropriate lexical resource for semantic annotation is the building of an upper-middle layer ontology. This step can provide the necessary semantic information for the tasks of word sense disambiguation. Such an ontology can be used in several ways: (1) for the representation of the general meaning of lexical units in a language; (2) as the basis for the construction of domain ontologies and lexicons; (3) to supply labels in the ontology comprehensible to speakers of human languages. For the concrete implementation of the Bulgarian ontology-based lexicon we use the ontology which results from extension of DOLCE and the upper part of OntoWordNet. This ontology was already used in the construction of the domain ontologies.

In the previous model we have used only the `equality` relation between the conceptual information in the ontology and the meaning of the corresponding lexical units. In this new lexicon this will not be possible because there will be insufficient concepts in the ontology. Thus, we separate completely the lexical information from conceptual one. In the lexicon each lexical entry contains linguistic information just for one lexical unit (representing one meaning). The linguistic information includes morphological and syntactic information. The conceptual information is presented via the relations `equality` or `subsumption`. When there is a concept in the ontology which is equivalent

to the meaning of the lexical unit, then only the relation `equality` is used to connect the lexical unit to the corresponding concept. If there is no equivalent concept for the lexical unit, the relation `subsumption` is used. In this case the lexical unit can be mapped to more than one concept from the ontology, because in the ontology multiple inheritance is allowed. When a lexical unit is mapped to several concepts, its ontology representation is a disjunction of these concepts. The requirement for the mapping via the `subsumption` relation is that the concepts used are the most specific ones available. The following figure depicts the addition of the new relation for the mapping between ontology and lexicon:

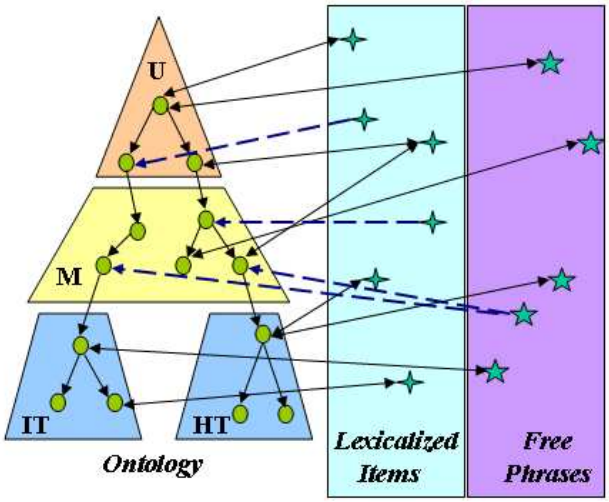

Fig 3. Ontology to lexicon mapping using `equality` and `subsumption` relations.

The dashed arrow represents the `subsumption` relation and the solid double arrow represents the `equality` relation. Concepts from the upper-middle layer are connected to two sets of lexical units — lexical units having a meaning represented by these concepts and lexical units having more specific meaning. The concepts from the domain layer are connected only to lexical units with equivalent meaning. Adding new domain ontologies will improve the precision of the mapping between the ontology and the lexicon.

The actual lexicon is under construction. It is based on several machine-readable dictionaries: a Morphological Dictionary, a Valence Dictionary and an Explanatory Dictionary of Bulgarian. The selection of the lexical units is done on the basis of constructing the lexicon aligned to the upper and middle parts of the ontology where

we encoded about 3000 lexical entries. The rest of the lexical units are selected on the basis of their ranking in a large Bulgarian reference corpus (72 million running words from the BulTreeBank text archive). The ranks are calculated on the basis of an automatic morphosyntactic analysis of the corpus and then lemmatization. For each lemma we consider the frequency in the corpus and in how many documents the lemma has occurred. The lexicon also contains the lexical entries for the two domain ontologies.

4.3 Encoding of Special Phenomena

Including of a formal ontology in the lexicon construction provides many possibilities for using the knowledge, represented in the ontology and the services related to it, such as inference. In this section we present the encoding of some important phenomena for the task of word sense disambiguation: metonymy and verb frames representation. The metonymy covers also a substantial part of the cases of regular polysemy. For an overview on regular polysemy, its representation and importance of this representation see (Barque and Chaumartin, 2009). We assume that the patterns described by the authors can be represented as inference patterns in our model of lexicon to ontology mapping.

A general assumption in the treatment of the above mentioned phenomena is that the related word senses are already represented in the ontology. In this way, the lexical representation of the corresponding patterns (metonymical or frame) is done via appropriate mappings to corresponding concepts in the ontology. The application of such patterns for creation of new senses is not explored in this work.

Let us consider the case of metonymy in more detail. In general, metonymy is defined as a trope in which one entity is used to stand for another associated entity[2]. Our treatment follows the ideas of (Hobbs, 2003) who interpreted the metonymy by introduction of a function which relates the mentioned object with the intended one. The function is different for different cases of metonymy and it can be context dependent. In order to implement the same idea we assume that the function is determined by an inference over the ontology and the context. This function is a composition of relations from the ontology. We consider the representation of such compositions in the lexicon as an important device for facilitation of text annotation. Our view of these compositions is that they are very specific inference rules. In future we will investigate the possibility to encode the metonymy relations reported in the literature (like the ones presented in (Barque and Chaumartin, 2009)) as such special inference rules. Here we present the interpretation of two cases of metonymy.

Let us suppose that we have to annotate the sentence "She was wearing stripe." First we annotate 'stripe' as a kind of `property` and as such it is connected to 'cloth' via the `property-of` relation and 'cloth' is annotated as `material` and it is connected to 'clothing' via the `made-of` relation. The concept 'clothing' is of the relevant type for the object of the verb 'to wear'. Thus, the understanding of the sentence is something like: "She was wearing a clothing made from a textile with a stripe design." The composition

[2] `http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsMetonymy.htm`

of the corresponding relations is stored in the lexical entries for the corresponding lexical units. In the case of metonymy this is a better option, because the possible patterns are (potentially) infinite in number. Representing each metonymy usage as a separate meaning will result in many strange meanings for the lexical units. In this way we represent the most frequent metonymy uses as inference patterns and the actual inference is done during the analysis of the discourse where the lexical unit is used metonymically.

When the regular polysemy is an example of metonymy, we represent it in the same way. The different meanings are represented in the ontology as different concepts and these concepts are connected via appropriate relations. The main difference here is that for each of the meanings we construct a separate lexical entry. This means that during the analysis of the text we have to disambiguate between these senses. In some cases more that one of the senses is visible via one usage of the lexical unit. For example, in the sentence "This large book is very interesting." the word 'book' is used simultaneously as a `physical object` selected by 'large' and as an `information object` selected by 'interesting'.

The encoding of verbs is also very important for the task of semantic annotation. We assume that the appropriate information is represented in two ways: (1) in the ontology each verb is connected to an event concept related to the meaning of the verb. In the ontology all the participants (irrespective of whether they are considered to be arguments, adjuncts, etc.) are represented as such via appropriate relations; (2) the linguistic behavior is encoded in the lexicon as a set of frames. These frames determine the role of each participant in the given event. During the annotation the verbs are annotated with the frames from the lexicon and the corresponding relations are connected with appropriate phrases from the text. Some of them are left unconnected when the corresponding participant is not explicitly mentioned in the text.

Currently, we do not represent in the lexicon the relation between the literal meaning of a given word and its metaphorical meaning. In contrast to metonymy, metaphorical meanings are not always closely related in the ontology. They require a special kind of inference by analogy which differs in many respects from the inference necessary to deal with metonymy.

5 Conclusion

In this paper we presented the construction of an ontology-based lexicon for Bulgarian. This lexicon originates from the practical task of semantic annotation of domain texts. Our starting point was the mapping from domain ontologies to terminological lexicons. Due to the sparseness of the resulting concept annotation, the coverage was extended to the general lexicon. One suggestion we make within the model is that merging the upper part of the ontology with the WordNet middle layer would result in a reduced resource which, however, is more understandable to common users. In order to have a better coverage, we rely on two relations between lexical units and the concepts in the ontology: `equality` and `subsumption`. The first is used primarily for the domain

ontology and the second for the middle and upper part of the ontology. Additionally, we encode metonymy and verb frames in the lexicon in order to support a better text annotation.

Our future goals are to implement a system for automatic word sense disambiguation and for detection of metonymical uses in the text. The extension of the lexicon coverage is also one of our tasks. In addition, the general lexicon together with the ontology could be used for the creation of domain ontologies and lexicons. We also plan an annotation of a corpus with concepts from the middle and upper part of the ontology.

6 Acknowledgements

This work has been supported by three European projects: LT4eL (Language Technology for eLearning) (FP6-027391), AsIsKnown (A Semantic-Based Knowledge Flow System for the European Home Textiles Industry) (FP6-028044) and LTfLL (Language Technologies for LifeLong Learning) (FP7-212578).

I would like to thank Petya Osenova for the comments and the discussions on the earlier versions of the paper and to the two anonymous reviewers for their valuable comments and suggestions.

Literatur

Barque L. and Chaumartin Fr. R. (2009). *Regular polysemy in WordNet.* In this volume.

Buitelaar P., Declerck Th., Frank An., Racioppa St., Kiesel M., Sintek M., Engel R., Romanelli M., Sonntag D., Loos B., Micelli V., Porzel R., Cimiano Ph. (2006). LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies. In: *Proceedings of OntoLex06, a Workshop at LREC*, Genoa, Italy.

Buitelaar, P., Sintek, M., and Kiesel, M. (2006). A Lexicon Model for Multilingual/Multimedia Ontologies In: *Proceedings of the 3rd European Semantic Web Conference (ESWC06)*, Budva, Montenegro.

Edmonds Ph. and Hirst Gr. (2002). *Near-synonymy and lexical choice.* Computational Linguistics, Vol. 28:2, pp. 105–144.

Fellbaum Chr. (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. In: *Meersman R, et al. (eds.), Proceedings of ODBASE03 Conference*, Springer.

Guarino N. (2000). *Invited Mini-course on Ontological Analysis and Ontology Design.* First Workshop on Ontologies and lexical Knowledge Bases - OntoLex 2000. Sozopol, Bulgaria.

Guarino, N., and Welty, C. (2002). *Evaluating Ontological Decisions with OntoClean.* Communications of the ACM, 45(2): 61-65.

Hirst Gr. (2004). *Ontology and the lexicon.* In: Steffen Staab and Rudi Studer (editors), Handbook on Ontologies. Springer Verlag, Berlin, Germany. pp 209–229. `http://ftp.cs.toronto.edu/pub/gh/Hirst-Ontol-2003.pdf`

Hobbs J. R. (2003) *Discourse and Inference.* University of Southern California, Marina del Rey, California. Unpublished manuscript. `http://www.isi.edu/~hobbs/disinf-tc.html`

Kiryakov At., Popov B., Terziev Iv., Manov D., and Ognyanoff D. (2005). *Semantic Annotation, Indexing, and Retrieval.* Elsevier's Journal of Web Sematics, Vol. 2, Issue 1.

Lenci A., Busa F., Ruimy N., Gola El., Monachini M., Calzolari N., Zampolli A., Guimier E., Recourcé G., Humphreys L., von Rekovsky U., Ogonowski A., McCauley Cl., Peters W., Peters Iv., Gaizauskas R., Villegas M. *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.

Morris J. and Graeme Hirst Gr. (2004). Non-Classical Lexical Semantic Relations. In: *Proceedings of the HLT Workshop on Computational Lexical Semantics.* Boston, Massachusetts, USA. pp 46–51.

Nirenburg S. and Raskin V. (2004). *Ontological Semantics.* MIT Press.

Pustejovsky J. (1995). *The Generative Lexicon.* MIT Press. Cambridge, MA, USA.

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, 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: *Proceedings of the Corpus Linguistics 2001 Conference.* Lancaster, UK.

Simov, K. and Osenova P. (2007) Applying Ontology-Based Lexicons to the Semantic Annotation of Learning Objects. in: *Proceeding of RANLP 2007 workshop on Natural Language Processing and Knowledge Representation for eLearning Environments.* Borovets, Bulgaria. pp 49–55.

Simov, K. and Osenova P. (2008) Language Resources and Tools for Ontology-Based Semantic Annotation. In: *Proceeding of OntoLex 2008 Workshop at LREC 2008.* Marrakech, Morocco. pp. 9-13.

Vossen P. (1999). Editor. *EuroWordNet General Document. Version 3, Final, July 19, 1999.* http://www.hum.uva.nl/ ewn