Automatic Ontology Construction for a National Term Bank

Bodil Nistrup Madsen & Hanne Erdman Thomsen
Copenhagen Business School, Denmark

Jakob Halskov
Danish Language Committee, Denmark

Tine Lassen
Roskilde University, Denmark

Keywords: ontology, knowledge extraction, knowledge structuring, data merging, knowledge dissemination, term bank

A prerequisite for continuous use and development of a national LSP in small countries like for example Denmark is free access to a term bank comprising domain knowledge in Danish and foreign languages. Domain specific knowledge goes beyond traditional dictionary information. In order to clarify and distinguish the meanings of domain specific concepts these must be described by means of characteristics and relations to other concepts, i.e. in the form of domain specific ontologies (concept systems). On the basis of these it is possible to develop consistent definitions that further understanding and correct use of terms. Terminology work that includes development of ontologies is a very labour-intensive task, and therefore most companies cannot afford this kind of work.

In our paper we present a project, the aim of which is to develop innovative and advanced methods for dynamic and automatic extraction of knowledge about concepts from texts and for automatic construction of ontologies. The project builds on and further develops the results of the CAOS project - Computer-Aided Ontology Structuring - which was carried out at Copenhagen Business School in the period 1998-2007. Terminological ontologies differ from other types of ontologies by comprising feature specifications and subdivision criteria. We have formalised subdivision criteria that have been used for

many years in terminology work, by introducing dimensions and dimension specifications. In the CAOS prototype, facilities for semi-automatic checking of inconsistencies were developed.

In the new project we will further develop facilities for automatic consistency checking, automatic changes to ontologies, automatic positioning of concepts and dynamic updating of the ontologies on the basis of the enriched information that they contain. To our knowledge no other systems have such capabilities.

In the project we will also develop methods for automatic merging of terminological data from various existing sources. In the process of bringing together data from different sources it is a big challenge to avoid double entries comprising the same concept, with varying formulation of the definitions and different translations. We are not aware of any existing term banks that have solved this problem. We will develop methods for automatic construction of ontologies on the basis of definitions from the various existing data sources and methods for automatic merging of entries based on the merging of these ontologies. Furthermore we will develop methods for target group oriented knowledge dissemination. Most other public term banks only offer restricted possibilities for setting up user specific search and presentation profiles.

Automatic Ontology Construction for a National Term Bank

Bodil Nistrup Madsen & Hanne Erdman Thomsen
Copenhagen Business School, Denmark

Jakob Halskov
Danish Language Committee, Denmark

Tine Lassen
Roskilde University, Denmark

Keywords: ontology, knowledge extraction, knowledge structuring, data merging, knowledge dissemination, term bank

Introduction

A prerequisite for continuous use and development of a national LSP in small countries like for example Denmark is free access to a term bank comprising domain knowledge in Danish and foreign languages. Domain specific knowledge goes beyond traditional dictionary information. In order to clarify and distinguish the meanings of domain specific concepts these must be described by means of characteristics and relations to other concepts, i.e. in the form of domain specific ontologies (concept systems). On the basis of these it is possible to develop consistent definitions that further understanding and correct use of terms. Terminology work that includes development of ontologies is a very labour-intensive task, and therefore most companies cannot afford this kind of work.

In our paper we present a project, the aim of which is to develop innovative and advanced methods for dynamic and automatic extraction of knowledge about concepts from texts and for automatic construction of ontologies. The project builds on and further develops the results of the CAOS project - Computer-Aided Ontology Structuring - which was carried out at Copenhagen Business School in the period 1998-2007. The project received funding by the Danish Research Council for the Humanities from 1998 to 2001.

In the project we will also develop methods for automatic merging of terminological data from various existing sources. In the process of bringing together data from different sources it is a big challenge to avoid double entries comprising the same concept in several entries, with varying formulation of the definitions and different translations. We are not aware of any existing term banks that have solved this problem. We will develop methods for automatic construction of ontologies on the basis of definitions from the various data sources and methods for automatic merging of entries based on the merging of these ontologies.

Furthermore we will develop methods for target group oriented knowledge dissemination. Most other term banks only offer restricted possibilities for setting up user specific search and presentation profiles.

Background

For a long period, many resources have been allocated to general language dictionaries, lexical databases and word nets. There is, however, a big need for domain specific knowledge within scientific, technical, economical and legal domains which can be made accessible by means of a Danish term bank. In 2008, the language committee of the Danish Government, issued a report, Sprog til tiden ('Language on demand'), in which the importance of a freely accessible national term bank is emphasised. In December 2009 the Danish Parliament encouraged the Government to analyse advantages and involved resources in establishing a Danish term bank and a national terminology centre, which can further the development of LSP and ensure knowledge sharing between research institutions and society.

Central concepts related to terminological ontologies

As an introduction to the description of the current project we present some central concepts related to terminological ontologies. A terminological ontology is a domain-specific ontology; c.f. the categorisation of ontologies in (Guarino 1998). Terminological ontologies differ from other types of ontologies by comprising feature specifications and subdivision criteria.

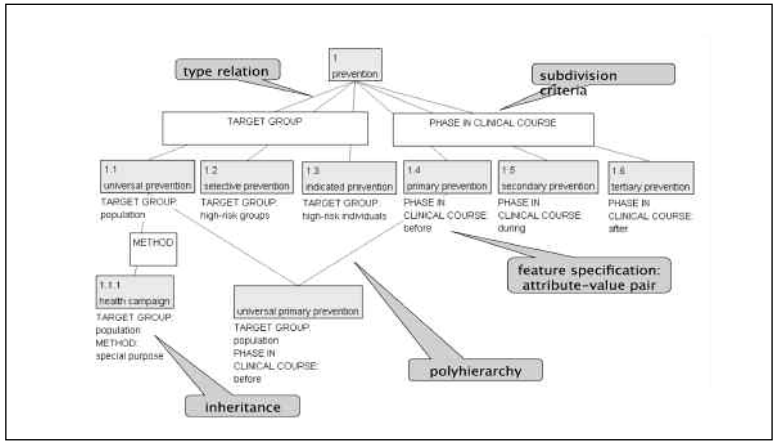
The term "terminological ontology" is a synonym for "concept system", which is used in terminology work, e.g. (ISO 704 2000).

Terminological ontologies as a basis for concept clarification

In figure 1 we present an example of an ontology for concepts related to disease prevention. This example is an extract from a more comprehensive ontology from the health care sector, illustrating only type relations, i.e. the green lines connecting the concepts, often referred to as ISA-relations. In terminological ontologies we use the terms "superordinate concepts", "subordinate concepts" and "coordinate concepts" instead of "hypernyms", "hyponyms" and "co-hyponyms". In figure 1, *universal prevention* is a subordinate concept of *prevention.*

From figure 1 it is clear that *universal prevention* is characterised by the intended target group, while *primary prevention* is characterised by the phase in the clinical course (even before there is a patient). Without this information, one might think that those two terms were synonymous, but an analysis of their characteristics, which are given below the concepts (e.g. [TARGET GROUP: population]), makes it clear that this is not the case.

Figure 1. Extract of and ontology for prevention



The characteristics of the concepts are presented as feature specifications in the form of attribute value pairs, e.g. [TARGET GROUP: population], cf. (Carpenter 1992). On the basis of these feature specifications, subdivision criteria are introduced (white boxes with text in capital letters) which illustrate that the three coordinate concepts 1.1-1.3 differ with respect to *target group,* while the three concepts 1.4-1.6 differ with respect to *phase in clinical course.* The subdivision criteria help the user to understand the meaning of the concepts, give a good overview and help the terminologist in writing consistent definitions. The definition of a concept is given by means of the position in the ontology and the characteristics. The ontology in figure 1 has been created using the concept modelling module i-Model of the terminology and knowledge management system i-Term ®, developed by the DANTERM Centre (the terminology centre) at the Copenhagen Business School. The concept modelling in i-Model is based on user input, and has no automatic consistency checking facilities.

Terminological ontologies implemented in CAOS

The principles of the terminological ontologies presented here have been developed in the research and development project CAOS - Computer-Aided Ontology Structuring - whose aim was to develop a computer system for semi-automatic construction of ontologies, cf. (cf. Madsen *et al.* 2004b; Madsen *et al.* 2005). CAOS was carried out by Bodil Nistrup Madsen, Hanne Erdman Thomsen and Carl Vikner at CBS, Dept. of International Language Studies and Computational Linguistics. The prototype includes an interactive graphical user interface which allows the user to build terminological ontologies on the basis of information entered while reading domain-specific texts. CAOS warns the user about inconsistencies and errors and informs users whenever they insert information that conflicts with the principles and constraints of the system.

In the CAOS prototype, facilities for semi-automatic checking of inconsistencies were developed. In the new project we will further develop facilities for automatic consistency checking, automatic changes to ontologies, automatic positioning of concepts and dynamic updating

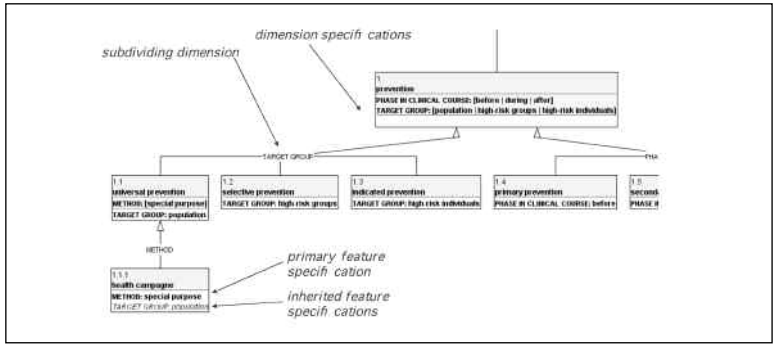
of the ontologies on the basis of the enriched information that they contain. To our knowledge no other systems have such capabilities.

In Figure 2 we present a part of the ontology from Figure 1, which is here constructed with the CAOS prototype.

Diagrams in CAOS are rendered in a UML-based notation. The type relations (ISA-relations) are represented by means of lines with arrow heads connecting the concepts. All types of concept relations can be used in CAOS. The system offers a set of concept relations organised in a taxonomy, cf. (cf. cf. (Madsen *et al.* 2002). It is also possible for the user to introduce user-defined relations.

The backbone of terminological concept modelling in CAOS is constituted by characteristics modelled by formal feature specifications, i.e. attribute-value pairs, as for example [TARGET GROUP: population]. This approach to modelling characteristics was proposed in (Madsen 1998), (Thomsen 1998) and (Thomsen 1999, cf. also (Carpenter 1992). The use of feature specifications is subject to principles and constraints described in detail in (Madsen *et al.* 2004b; Madsen *et al.* 2005). Subordinate concepts inherit the characteristics of superordinate concepts, e.g. health campaign inherits the characteristic: [TARGET GROUP: population] from the concept universal prevention.

Figure 2. Extract of an ontology for prevention modelled in CAOS



Polyhierarchy can be introduced, i.e. one concept may be related to two (or more) superordinate concepts. In figure 1 the concept *universal primary prevention* is an example of this. A very important principle in such cases is that the superordinate concepts of a concept inheriting characteristics from two (or more) concepts must always belong to two (or more) different subdivision criteria otherwise the ontology must be changed.

We have formalised subdivision criteria that have been used for many years in terminology work, by introducing dimensions and dimension specifications which form the basis for the facilities for semi-automatic construction of ontologies and for consistency checking. A dimension of a concept is an attribute occurring in a (non-inherited) feature specification of one or more of its subordinate concepts, i.e. an attribute whose possible values allow a distinction between some of the sub-concepts of the concept in question. A dimension specification consists of a dimension and the values associated with the corresponding attribute in the feature specifications of the subordinate concepts: DIMENSION: [value1| value2| ...], e.g."TARGET GROUP: [high-risk groups|high risk individuals]" in Figure 2.

The principles for constructing ontologies mentioned here are unique. No other ontology projects or systems make use of these principles that result in very precise descriptions of domain specific concepts. In the next two sections we describe our new project in more detail.

Terminological ontologies versus word nets and other types of ontologies

Lexical ontologies for general language, so called wordnets, which allow the user to navigate in a network of concepts, are being developed in many countries. A well known example of an electronic network is Princeton WordNet (http://wordnet.princeton.edu/), for which several graphical browsers have been developed. In Denmark a Danish wordnet, DanNet (http://www.wordnet.dk/), has been under development since 2005. DanNet is based on the Danish dictionary, Den Danske Ordbog (http://ordnet.dk/ddo).

Ontologies covering specific domains are also developed, but they normally differ from terminological ontologies as defined in our work. Examples are UMLS, Unified Medical Language System (http://www.nlm.nih.gov/research/umls/) and SNOMED CT (http://www.ihtsdo.org/snomed-ct/), Systematized Nomenclature of Medicine-Clinical Terms, which has also been translated into Danish. In the period 2003-2006 the Aarhus School of Business, Aarhus University, worked on two projects: MEDVID and MEDTERM. MEDVID (http://www.asb.dk/article.aspx?pid=568#medvid) focused on knowledge sharing, dissemination and communication within the medico-technical and the medical scientific area. The project was a co-operation between companies, research centres and translation companies exchanging knowledge on language use, translation and communication in the medical area. The intention of the project MEDTERM was the development of a multilinguistic, internet-based dictionary supporting knowledge sharing within the medical field (http://www.asb.dk/article.aspx?pid=568#medterm).

These data collections have all been build manually, which is a very labour-intensive task, and to our knowledge methods for dynamic updating have not been developed.

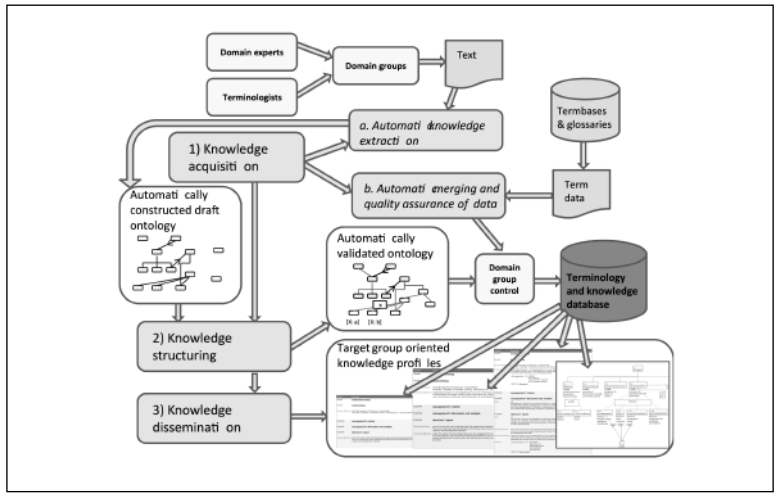
Aim of the project

The aim of our project is to develop methods for automatic knowledge extraction, automatic construction and updating of ontologies. In the project we will also develop methods for automatic merging of terminological data from various existing sources, as well as methods for target group oriented knowledge dissemination. The research carried out in the current project is a prerequisite for establishing a national Danish term bank which can ensure development and quality of Danish LSP. When the term bank has been established, it will form the basis for various other research projects.

Figure 3 gives an overview of the three subprojects of the project and the processes involved: 1) *Knowledge acquisition,* 2) *Knowledge structuring* and 3) *Kowledge dissemination.* In subproject 1) *Knowledge*

acquisition we will develop methods for a) *automatic knowledge extraction* and b) *automatic merging and quality assurance of data.* Below we describe the subprojects in more detail.

Figure 3. Processes involved in the three sub projects



Knowledge extraction

The aim of this subproject is to develop new advanced models of and methods for automatic extraction of concepts and information about conceps as well as a prototype which can automatically produce a draft version of a terminological ontology on the basis of an existing domain-specific text corpus, or on the basis of domain texts automatically collected from the Internet. Thus, the draft ontologies will contain subdivision criteria and characteristics as formal feature specifications on concepts.

State of the art

Several survey articles describe the state of the art in the field, e.g. Shamsfard and Abdollahzadeh (2003), Buitelaar *et al.* (2005) and Zhou (2007). Below some projects are described. None of the methods used in these projects are targeted towards constructing terminological ontologies, and thus, in order to be used in our project, they would have to be refined.

GlossOnt (Park 2004) describes a concept-focused ontology building method which is based on text mining technology. The relations are identified through different techniques; the ISA and synonym relations are detected primarily by applying (Hearst 1992), and other relations are identified through syntactic parsing, where verbs denote relations that relate the verbal arguments. **The SIABO Project** (http://siabo.org) focuses on engineering biomedical ontologies, and seeks to set up a novel so-called "ontological semantics" which maps phrases into nodes in a generative ontology. **OntoLearn** is a system for word sense disambiguation, used to automatically enrich WordNet with domain concepts and to disambiguate WordNet glosses (Hearst 1992). **Mo'K workbench** (Bisson *et al.* 2000) is a configurable workbench that supports the development of conceptual clustering methods for ontology building. **OntoLT** (Buitelaar *et al.* 2003) is a plug-in for Protégé (http://protege.stanford.edu) with which concepts and relations can be extracted automatically from linguistically annotated text collections. The plug-in performs interactive user validation of candidates and automatic integration of results into an OWL ontology. **Text2Onto** (Cimiano and Voelker 2005) is an ontology-learning framework that has been developed to support the acquisition of ontologies from text.

In the current project we will further develop and combine these methods and adapt them to computational terminology.

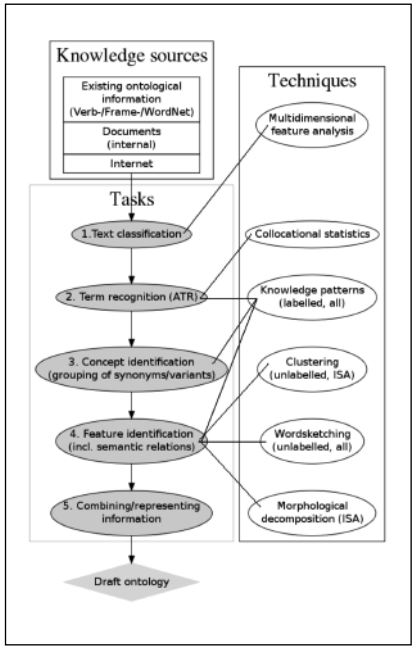
Scientific methods

A basic idea in this subproject is to investigate the possibilities of establishing and using groups of domain experts, who will contribute to knowledge acquisition and concept clarification. Among other things, we will implement knowledge extraction tools which are integrated in an interactive user interface, where the domain experts upload texts to a corpus collection. Methods for checking these texts automatically for their estimated contents of explicit knowledge, term richness and other quality indicators for LSP (Barrière 2006; Halskov et al. 2010) will be developed.

The flowchart in Figure 4 illustrates the different knowledge sources, tasks and techniques envisioned in the knowledge extraction part of the project. The approach uses statistics-based as well as rule-based extraction methods.

The textual input comes either from within the organisation (typically pre-classified) or from the Internet. Texts from the Internet will be extracted by applying automatic text classification algorithms to ensure that only domain-specific and relevant documents are extracted (task #1). All texts are linguistically annotated using standard natural language processing tools for automatic tokenisation, detection of sentence boundaries, part of speech tagging and lemmatisation.

Figure 4. From domain-specific text documents to draft ontology



In task #2, general language reference corpora (and possibly existing ontological resources) are used to automatically detect term candidates in the texts classified in task #1. In task #3, all variants of a particular term are combined into a concept, and arranged according to type relations, and associative relations (i.e. non-taxonomical relations) are added to the concepts in the form of feature-value pairs.

In task #4, the two techniques, “Clustering” and “Morphological decompostion” are used to structure the collected terms by means of a morphological analysis which recursively analyses the meaning units of

compound terms and constructs a "base ontology", based on **implicit** ISA-relations, (cf. Gillam *et al.* 2005).

The two pattern based techniques, “Knowledge patterns” and “Wordsketching”, will be adapted to identify associative relations (and other characteristics) expressed **explicitly** in the texts.

The initial stage of task #4 involves feeding the system with ‘seeds’, i.e. concept pairs, between which a particular relation type is known to exist. Based on a small number of seeds it is possible to identify a set of linguistic patterns that express the relation by searching for instances in a corpus or on the internet. The seed concept pairs are derived either from existing validated ontology resources, or are extracted from dictionary definitions, as proposed in e.g. (Park 2004). The method of extracting seed term pairs from dictionary definitions, however, requires a resource that has consistent and high-quality analytic definitions - a problem (Park 2004) does not mention.

Another approach to relation extraction that we will also apply is one used in (Park 2004). Verbs denote relations, which may be extracted based on the verbal expression alone. A problem with this approach is that the set of relations becomes extremely large, the size of the set being potentially as large as the number of verbs in the language. To avoid having such a large set, additional resources such as DanNet (Pedersen *et al.* 2009), could be applied. Here verbs are grouped into clusters with related meaning in a hierarchical structure which means that it is possible to identify more general relation types.

The technique known as Knowledge Patterns (KP) will be applied to tasks #3 and #4 as this technique is capable of identifying all sorts of semantic relations (not just ISA). Whether the knowledge source is the internet (in case of data sparseness) or internal documents, the KP technique often involves an iterative process known as Dual Iterative Pattern Relation Extraction (DIPRE).

The final goal is to generate a draft ontology by automatically combining and representing the extracted information about concepts, cf. task #5.

Novelty of the approach

While many existing ontology building systems extract type relations (ISA-relations) or unlabelled relations (Halskov 2007), some systems can in fact build ontologies that include a diverse set of semantic relations, e.g. GlossOnt (Park 2004). However, no system automatically builds terminological ontologies that adhere to the above mentioned principles.

A distinctive feature of our approach includes the automatic extraction of concepts and (associative) relations, which can be formalised as feature specifications. The ontologies will be based on the principles for terminological ontologies as described above. Further, the approach combines different sources of learning in that it extracts seed concepts both from existing ontological resources and from dictionary definitions.

Most tasks in Figure 4 build on, but further develop, existing methods, while in task #5 we develop entirely new methods.

The development of methods for knowledge extraction in the current project will to some extent build on the results achieved in (Halskov 2007) on semiautomatic expansion of existing terminological ontologies using knowledge patterns discovered on the internet (namely task #2 and parts of tasks #3 and #4), as well as (Lassen 2010). Halskov (2007) describes and evaluates methods for automatic identification and extraction of four different types of concept relations from untagged and uncategorized texts in the internet. Lassen (2010) describes and evaluates a machine learning method for identification of relation affinities, or preferences for particular ontological types of arguments for relations. Once such affinities have been discovered, they can be used in a rule-based relation annotation task.

Automatic merging and quality assurance of data from various sources

In this project, we will develop ontologies and definitions within selected pilot domains in combination with the development of methods and tools, and this research will result in terminological data within three important economic domains: organisations, taxes and auditing.

CBS also has access to existing terminological data in Danish and foreign languages within many other domains, which can be imported into a national term bank. These terminological data are comprehensive, and they originate from terminology thesis in Danish and various foreign languages within technical, economic and legal domains as well as from other research projects concerning e.g. climate, environment and IT. The idea is that the contents of the term bank should also be extended by means of import of the many term lists from the internet or from authorities, organisations and companies.

Two very complex problems exist in the process of converting and combining terminology data from different sources. One problem is that the data are likely to have different structures, be stored in different formats and be of varying quality. Another problem is that different entries from the different sources contain information about the same concept, but associated with different terms and definitions. We will refer to these double entries as 'false doublettes'. Such false doublettes reduce the usefulness and the quality of the term bank considerably. If a user has to go through a number of entries that differ to some large or small degree as an answer to a given query, the reliability and usefulness of the term bank is reduced. It is not possible for a user to choose the correct term if a query returns e.g. 25 definitions in random order, and a closer study reveals that there are in fact only 6 different concepts, which also means that there should only be 6 different definitions and entries. For this reason, it is very important to merge entries with concepts having the same meaning. Thus there is a very close connection between entry merging and quality assurance. Another central prerequisite for a successful merging of data is that a consistent subject classification is used.

Therefore, it is very important to do research in automatic ontology construction on the basis of existing term collections, and to develop methods for merging and quality assurance of term data from different sources.

State of the art

Merging of data from different sources is a problem which has not been solved in a satisfactory way by other term banks. In the Swedish Rikstermbanken (www.rikstermbanken.se) and IATE (www.iate.europa.eu) false doublettes have not been removed. In the EuroTermBank (www.eurotermbank.com) automatic entry compounding is carried out, but the result is not always optimal, since one entry may comprise definitions of completely different concepts.

On the basis of existing terminological data collections, ontologies may be constructed automatically on the basis of knowledge extracted from the definitions. Ideally, a definition comprises a reference to the nearest superordinate concept and the characteristic which distinguishes the concept from its coordinate concepts, and this has been exploited in various projects, such as Lexical Knowledge Base (Copestake 1992) and DanNet. In DanNet some type relations are established semi-automatically on the basis of definitions comprising a broader concept, which is not exactly the nearest superordinate concept (Nimb 2009).
In connection with merging of data from various terminological data collections, a merging of ontologies is also required. Many researchers work on ontology merging and matching (Barrasa *et al.* 2004). The methods are based on a comparison of term expressions, ontology structure and other information types in the ontologies, e.g. similarity (Bulskov 2006). However, none of these methods deal with terminological ontologies and thus it will be necessary to select and further develop relevant methods.

Scientific methods

In order to be able to import existing terminology collections in the term bank, procedures for handling terminology data from different

sources are to be developed. For the conversion of data, formats compatible with the ISO data exchange format TBX (ISO 30042 2008) will be used.

As an initial basis for the coupling of terms from different sources, a subject classification will be used. In this context it must be investigated how to cope with the problem that different sources use different subject classifications, often with different level of detail.

Another method of semantic mapping between terms from different sources representing the same concept is to evaluate the semantic similarity between the definitions. Several methods for comparison and semantic analysis of text strings exist, e.g. latent semantics (Landauer *et al.* 1998), techniques used in translation memories, automatic text summarization and term recognition techniques. In the project these methods will be further developed, and we will develop rules for automatic identification of potential coupling candidates with a view to subsequent manual treatment.

Finally ontologies will be automatically constructed on the basis of the imported definitions.

Novelty of the approach

The special approach to merging and quality assurance of terminological entries from different sources, among other things, a solution to the problem of false doublettes, will raise the quality of the contents of the term bank and ensure user-friendliness, and by using semi-automatic procedures for coupling of data, both initial and repeated tasks in connection with import of terminological data from different sources will be made easier. Other term banks have not solved these problems.

Knowledge structuring

The aim is to develop methods and a prototype that may be used for automatic validation and dynamic expansion of the draft ontologies that result from the automatic knowledge extraction.

State of the art

Since the mid 1990s, researchers and developers of ontology tools have described types of and criteria for the *evaluation of ontologies*, e.g. Guarino and Welty (2000a), Guarino and Welty (2000b) and Gómez Pérez *et al.* (2004). Typically, evaluation is performed to determine whether a particular ontology suits a particular purpose, or to decide which of a set of ontologies best suits the purpose. Some systems, however, also check the consistency of populated ontologies, i.e. they check the logical consistency of a data collection where data is structured according to a specific ontology.

According to (Suárez-Figueroa and Gómez-Pérez 2008), *ontology evaluation* refers to *the activity of checking the technical quality of an ontology against a frame of reference*, and three types of ontology evaluation are described: verification, validation and assessment.

- *Ontology verification:* the ontology is compared against the specifications which are defined before the ontology development

- *Ontology validation:* the ontology definitions are compared against the intended model of the world that one is attempting to conceptualise.

- *Ontology assessment:* the ontology is analysed in relation to the user requirements, such as usability, usefulness, abstraction quality etc.

More interesting for our project, however, are the *criteria* to automatically evaluate the internal logical coherence of an ontology. Gómez Pérez et al. also describe criteria for this kind of evaluation: Consistency, completeness, conciseness, expandability and sensitiveness (Gómez Pérez *et al.* 2004).

In the new project we will focus on the fulfilment of one of the above-mentioned criteria, namely *consistency*, more specifically

inferential consistency, and we use the term *knowledge validation* to cover the methods to obtain this kind of consistency, i.e. methods to avoid the three error types: semantic inconsistency, circularity and partition errors, (Gómez Pérez *et al.* 2004),the latter of which are related to the use of subdivision criteria.

In the CAOS project a prototype was developed that includes an interactive graphical user interface which allows the user to build terminological ontologies on the basis of information entered while reading domain-specific texts. The prototype makes use of **semi**automatic knowlege validation, i.e. users are warned whenever they insert information that conflicts with the principles and constraints of the system, e.g. constraints with respect to semantic inconsistency, circularity and partition errors. The current project will develop methods for **automatic** ontology construction and validation.

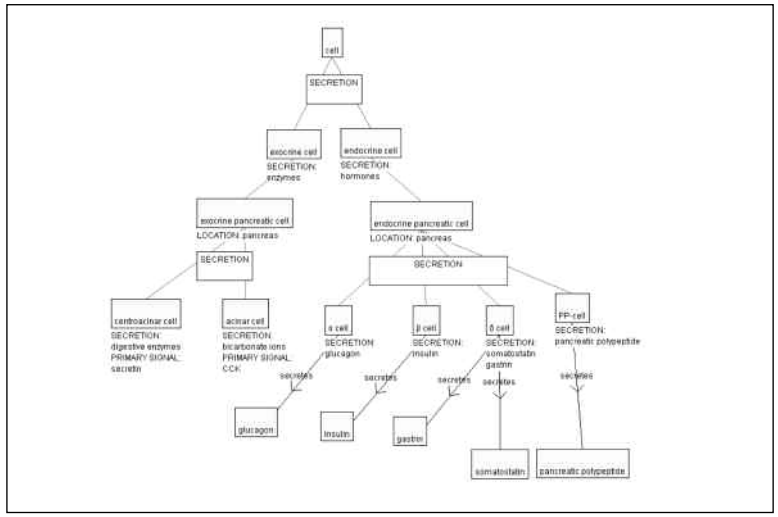
Scientific methods

To enable fully automatic validation of the draft ontologies resulting from the automatic knowledge extraction, the methods described above will be developed further as explained below.

1) Knowledge validation of an entire ontology: We will develop a new validation method to enable consistency control and correction of a whole ontology resulting from the automatic knowledge extraction instead of validation of one concept at a time as it is done in the interactive prototype.

2) Auto correction of ontologies: To enable automatic changes in the ontology enforced by the validation procedure, we will develop techniques for automatically positioning concepts in an existing type hierarchy by employing the characteristics registered for that concept. Such techniques have not previously been developed.

Figure 5. Ontology of types of cells



3) Treatment of characteristics and relations: In our previous approach, characteristics (registered as feature specifications) were distinguished from concept relations (registered as concept relation and related concept) (Madsen *et al.* 2004b). In Figure 5, it is illustrated for *α cell* (alpha cell) that a particular characteristic (in this case the characteristic of secreting *glucagon*) can be modelled either as a feature specification or as a relation to another concept. In a small terminology project, concepts outside the narrow domain are not included in the ontology, they will only exist as values in feature specifications. However, if these concepts are relevant for describing the domain they will be introduced as concepts in the ontology instead.

In the extraction module, which will be developed in subproject 1 a, attributes will not be distinguished from relations, and therefore associative relations will be registered as attributes and the related concepts as values. We will develop a new theory for distinguishing between characteristics and related concepts. The theory will be based on how central the values are within the domain.

4) Multiple values: The extraction tool is bound to deliver more than one concept for a particular relation to another concept (or more than one value

for a given attribute). Hence, in Figure 5, *δ cell* secretes both *gastrin* and *somatostatin*. But some relations may only be applied to a particular concept once. For example, no concept can have more than one instance of the relation HAS_LENGTH_IN_CM. This corresponds to the principle that for a particular attribute a concept can have at most one value. Therefore our earlier method must be modified to handle multiple occurences of the same relation on one concept. We will develop a method for distinguishing relations which can be applied only once from those that may apply more than once to a given concept, in order to enable knowledge validation.

5) Hierarchy of values: In the example of cells in Figure 5, further subdivisions of *exocrine cells* and *endocrine cells* are based on SECRETION. The problem is that e.g. the concept *centroacinar cell* inherits the characteristic [SECRETION: enzymes] from *exocrine cell,* but it is stated that it has the characteristic [SECRETION: digestive enzymes], i.e. another value for the same attribute. In this case it can be argued that the "new" value (digestive enzymes) is a specialisation of the value given at the higher level, and thus there is no conflict. In order to handle this, the technology we developed earlier will be enhanced to take into account a type hierarchy of values. This enhancement will take as point of departure the methods implemented e.g. in the Lexical Kowledge Base system (LKB) first developed by Ann Copestake (Copestake 1992) for lexical semantics and further enhanced for HPSG (Head Driven Phrase Structure Grammar) purposes.

Novelty of the approach

Researchers at CBS have earlier developed methods and a prototype for **semi**-automatic construction of terminological ontologies based on user interaction (Madsen *et al.* 2004b). In the current project we will develop facilities for automatic consistency checking, automatic changes to ontologies, automatic positioning of concepts and **automatic** and dynamic updating of the ontologies on the basis of the information that they contain. No other methods or systems exist for automatic construction and knowledge validation of terminological ontologies that comprise subdivision criteria and dimension specifications, which are crucial in the development of such ontologies.

Knowledge dissemination

This subproject will focus on presentation of data in the term bank. Traditionally, terminology and lexicography have been separate research fields with different approaches to compilation and presentation of data. However modern technology offers unlimited opportunities to meet the needs for several target groups in one database by offering the possibility of choosing between different presentations. The overall objectives of this subproject are to discuss and specify

- the extent to which the traditional lexicographical and terminological methods may be fruitfully combined allowing the presentation of concepts in one single database thereby contributing added value for a defined user group
- how a combination of the two research fields may create further opportunities towards developing principles for target-group oriented knowledge transfer.

Another research topic is to investigate how data from the term bank may be used in other electronic tools, such as machine translation systems. This interesting topic will, however not be addressed in this project.

In the project we will analyse possible user groups (e.g. experts, semi-experts and various groups of laymen) in order to target the structure and contents towards the communication-oriented and knowledge-oriented needs of different target groups.

State of the art

Traditionally, terminology and lexicography have been separate research fields with different approaches to compilation and presentation of data. Terminology work is concept-oriented (ISO 704 2000; Madsen 1999a; Madsen 1999b), which means that synonyms are registered in one entry in the database, while lexicography has been word-oriented, i.e. one dictionary entry comprises all meanings of an entry word.

Modern lexicographical methods focus on determining the functions of a given dictionary based on the mapping of types of (1) users, (2) user situations, and (3) user needs (Tarp 2008a; 2008b; 2009 and Bergenholtz and Tarp 2010), while, traditionally, the primary target group of terminology has been translators and domain experts. With the use of databases, however, the possibilities for presentation do not depend on the structure of the data collection, and thus it is possible to present data from a term base with a concept-oriented structure in a word-oriented user interface. Compared to the restrictions inherent in printed publications, modern technology offers unlimited opportunities with respect to volume. This has lead to new approaches within the field of LSP lexicography to meet user needs for encyclopaedic as well as lexical-semantic knowledge. As a consequence the two fields are converging.

Existing public term banks, such as the Swedish Rikstermbanken and the European IATE term bank, do not distinguish between different user groups. They both offer a list of search results with few information types or a full presentation of each entry. The search possibilities are rather restricted. The Rikstermbanken only offers the possibility of choosing between search in the field entry term or search in all fields.

The accounting dictionaries, Regnskabsordbogen, (Danish-English; Danish-Danish; English-English; English-Danish (http://www.ordbogen.com/ordboger/regn/index.php?dict=a007), which are commercial, web-based dictionaries, offer the possibility of choosing between three presentations depending on the purpose of the dictionary look-up (to understand accounting expressions, to produce a text or to look up a known sense/meaning of the expression).

Scientific methods

Relevant user groups will be identified, and on the basis of theoretical knowledge and experience with other term banks and electronic dictionaries we will develop user interfaces adapted to the known needs of the individual user groups, in order to carry out experiments with the prototype of the Danish term bank.

We will establish focus groups representing each user category, conduct focus group interviews about the needs of the users for searching in known electronic data collections, and the users will get an introduction to the Danish term bank prototype. The users will then be asked to solve various tasks by using the database, and their behaviour will be studied, e.g. by means of so called eye-tracking, by means of which the user's focus on the screen while looking for information can be registered. Expertise and eye-tracking devices are available at the Department of International Language Studies and Knowledge Technology of CBS. Supplementary interviews and experiments will be carried out if the observations indicate comprehensive adjustments.

Novelty of the approach

Existing publically accessible term banks have focused on making large amounts of data publically available without taking into consideration the different needs of the user groups. Therefore, systematic experiments with different user needs when using term banks have not been carried out. In the field of lexicography, there has been more focus on user group needs, both with respect to entry structure and the contents of the individual information categories, but the primary focus has been on elaboration of whole reference works for each user group. As opposed to this, the aim of this project is to develop one single term bank with different user interfaces and presentations of data adapted to the different user needs. Eye-tracking observations have to our knowledge not previously been used for optimisation of terminological or lexicographical data collections.

Perspectives

Ontologies are important in a term bank which aims at concept clarification and is a tool for both text production and translation.

Furthermore ontologies are relevant in many areas as they will lead to a much better foundation for

- development of classification systems and metadata taxonomies as a basis for efficient use of digital information in public and commercial web portals,
- automatic handling of large quantities of information by means of for example intelligent, ontology-based querying and document management systems,
- development / generation of consistent data models for large IT systems based on the clarification and consistent description of the underlying concepts of the IT system,
- development of software for semantic text control in order to ensure consistent and intelligible professional texts and to avoid fatal mistakes in e.g. user manuals,
- knowledge structuring and knowledge sharing in companies and organisations,
- obtaining interoperability between various knowledge sources in enterprises and organisations and developing formats for exchange of data.

A future Danish term bank will cooperate with the Danish Language Committee, the Southern Denmark University, Aarhus School of Business, Aarhus University, and term banks in the Nordic Countries.

Conclusion

Terminological ontologies are very useful tools for concept clarification. They form the basis for precise and consistent definitions of concepts within specific domains. In a term bank, terminological ontologies will help the user understand domain specific concepts and offer a solid foundation for choosing the right equivalent for translation purposes. In order to obtain a useful term bank, it is necessary that it covers a reasonable number of domains. However, the elaboration of terminology in several languages and the construction of domain ontologies is a very labour-intensive task. This is the motivation for the new project that aims at developing methods for automatic extraction of information about concepts and automatic construction of ontologies with a view to establishing a Danish term bank. In order to

obtain a reasonable amount of concepts in the term bank, existing terminology from different sources will be imported. In the project, methods for building ontologies on the basis of existing terminological data will be developed, and these ontologies will be used for merging entries from different sources and thus eliminating the problem that many entries comprise the same concept with different information, typically different definitions, which is confusing for the end user. In order to obtain user-friendliness the project will also develop methods for target group oriented knowledge dissemination.

Authors' Details

Bodil Nistrup Madsen (bnm.isv@cbs.dk), Hanne Erdman Thomsen (het.isv@cbs.dk),
Jakob Halskov (jhalskov@dsn.dk) & Tine Lassen (tlassen@ruc.dk)

References

Andreasen, T., Bulskov, H., Jensen, P. A., Lassen, T., Madsen, B., Nilsson, J.F., Szymczak, B.A., Thomsen, H.E. and Zambach, S. (2009) 'SIABO. Semantic Information Access through Biomedical Ontologies', in *Proceedings of the International Conference on Knowledge Engineering and Ontology Development. KEOD09.* Oktober 2009, Madeira, INSTICC.

Barrasa, J., Thanh Le Bach, J. E., Bouquet, P., De Bo, J., Dieng, R., Ehrig, M., Hauswirth, M., Jarrar, M., Lara, R., Maynard, D., Napoli, A., Stamou, G., Stuckenschmidt, H., Shvaiko, P., and Tessaris, S. (2004) *State of the art on ontology alignment.* KnowledgeWeb Consortium.

Barrière, C. (2006) 'TerminoWeb: A Software Environment for Term Study in Rich Contexts', *International Conference on Terminology, Standardisation and Technology Transfer (TSTT 2006),* Beijing.

Bergenholtz, H. and Tarp, S. (2010) 'LSP Lexicography or Terminography? The lexicographer's point of view', in *Future Trends in Specialised Dictionaries for Learners. A Festschrift in honour of Enrique Alcaraz Varó.* Pedro Fuertes Olivera.

Bisson, G., Nédellec, C. and Cañamero, D. (2000) 'Designing clustering methods for ontology building - The Mo'K workbench', in *Proceedings of the ECAI Ontology Learning Workshop 2000.*

Buitelaar, P., Cimiano, P. and Magnini, B. (2005) 'Ontology Learning from Text: An Overview', in Buitelaar, P., Cimiano, P. and Magnini, B. (eds) *Ontology Learning from Text: Methods, Evaluation and Applications/ Frontiers in Artificial Intelligence and Applications,* Volume 123.

Buitelaar, P., Olejnik, D. and Sintek, M. (2003) OntoLT: A Protégé Plug-In for Ontology Extraction from Text. *Proceedings of the International Semantic Web Conference, ISWC 2003.*

Bulskov, H. (2006) Ontology-based Information *Retrieva.* Ph.d.-afhandling, RUC.

Carpenter, B. (1992) *The Logic of Typed Feature Structures.* Cambridge, Mass: Cambridge University Press.

Cimiano, P. and Voelker, J. (2005) Text2onto - a framework for ontology learning and data-driven change discovery. *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB'2005),* Alicante, Spain.

Copestake, A. (1992) *The Representation of Lexical Semantic Information,* Doctoral dissertation, University of Sussex.

Gillam, L., Tariq, M. and Ahmad, K. (2005) 'Terminology and the construction of Ontology', in *Terminology*, 11(1) 55–81.

Gómez-Pérez, A., Fernández-López, M. and Corcho, O. (2004) 'Ontological Engineering', in *Advanced Information and Knowledge Processing,* London/Berlin/Heidelberg: Springer Verlag.

Guarino, N. (1998) 'Formal Ontology and Information Systems', in Guarino, N. (ed) *Proceedings of the First International Conference (FOIS'98) June 6-8,* Trento, Italy: Amsterdam IOS Press, 3-15.

Guarino, N. and Welty, C. (2000a) 'A Formal Ontology of Properties', in Dieng, R. and Corby, O. (ed), *Knowledge Engineering and Knowledge Management: Methods, Models and Tools. 12th International Conference, EKAW2000, LNAI 1937,* London, Berlin, Heidelberg: Springer Verlag, 97-112.

Guarino, N. and Welty, C. (2000b), 'Identity, Unity, and Individuation: Towards a Formal Toolkit for Ontological Analysis', in Horn, W. (ed) *August, 2000. Proceedings of ECAI-2000: The European Conference on Artificial Intelligence:* IOS Press, Amsterdam.

Halskov, J. (2007) *The semiautomatic expansion of existing terminological ontologies using knowledge patterns discovered on the WWW - an implementation and evaluation.* PhD thesis, Copenhagen Business School.

Halskov, J., Braasch, A., Haltrup Hansen, D. and Olsen, S. (2010) 'Quality indicators of LSP texts – selection and measurements. How to measure the terminological usefulness of a document from a particular domain in the task of compiling an LSP corpus', *Proceedings from LREC,* Malta.

Hearst, M. (1992) 'Automatic acquisition of hyponyms from large text corpora', in *Proceedings of the 14th International Conference on Computational Linguistics,* 539–545.

ISO 704. (2000) *Terminology work — Principles and methods.* Genève: ISO.

ISO 30042. (2008) *Systems to manage terminology, knowledge, and content - TermBase eXchange (TBX), ISO TC 37/SC 3/WG3.*

Johnson, C.R., Fillmore, C.J., Petruck, M.R.L, Baker, C.F., Ellsworth, M., Ruppenhofer, J. and Wood, E.J. (2002) *FrameNet: Theory and Practice.* ICSI Technical Report tr-02-009.

Kipper, K., Korhonen, A., Ryant, N. and Palmer, M. (2006) 'Extensive Classifications of English verbs', in *Proceedings of the 12th EURALEX International Congress,* September 2006, Turin, Italy.

Landauer, T.K., Foltz, P. W. and Laham, D. (1998) 'Introduction to Latent Semantic Analysis', in *Discourse Processes,* 25: 259-284.

Lassen, T. (2010) *Uncovering prepositional senses,* PH.d.-afhandling, RUC, Datalogi.

Madsen, B. N. (1998) 'Typed Feature Structures for Terminology Work - Part I', in Lundquist, L., Picht, H. and Qvistgaard, J. (ed) LSP - *Identity and Interface - Research, Knowledge and Society. Proceedings of the 11th European Symposium on Language for Special Purposes,* August 1997, Copenhagen: Copenhagen Business School, 339-348.

Madsen, B. N. (1999a) *Terminologi 1 Principper og metoder,* København: Gads Forlag.

Madsen, B. N. (1999b) *Terminologi 2 Øvelser og eksempler,* København: Gads Forlag.

Madsen, B.N., Sandford Pedersen, B. and Thomsen, H.E. (2002) 'The Role of Semantic Relations in a Content-based Querying System: a Research Presentation from the OntoQuery Project', in Simov, K. and Kiryakov, A. (eds) *Proceedings from OntoLex '2000, Workshop on Ontologies and Lexical Knowledge Bases,* Sept. 8-10 2000, Sozopol, Bulgaria, 72-81.

Madsen, B.N., Thomsen, H.E. and Vikner, C. (2004a) 'Comparison of Principles Applying to Domain Specific versus General Ontologies', in Oltramari, I.A., Paggio, P., Gangemi, A., Pazienza, M.T., Calzolari, N., Pedersen, B.S. *et al.* (eds) *OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments. ELRA, 2004,* 90-95.

Madsen, B. N., Thomsen, H. E., and Vikner, C. (2004b) 'Principles of a system for terminological concept modelling', in *Proceedings of the 4th International Conference on Language Resources and Evaluation,* Lisbon, 1: 15-18.

Madsen, B.N., Thomsen, H.E. and Vikner, C. (2005) 'Multidimensionality in terminological concept modelling', in Madsen, B.N. and Thomsen, H.E. (eds) *Terminology and Content Development, TKE 2005, 7th International Conference on Terminology and Knowledge Engineering,* Copenhagen, 161-173.

Madsen, B. N. and Thomsen, H. E. (2009) 'Terminological Concept Modelling and UML Diagrams', Accepted for publication in: *International Journal of Metadata, Semantics and Ontologies (IJMSO).*

Navigli, R. and Velardi, P. (2004) 'Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites', *Computational Linguistics,* 30(2): 151-179.

Nimb, S. (2009) 'The Semantic Relations of Artifacts in DanNet', in Pedersen, B., Braasch, A., Nimb, S. and Vatvedt Fjeld, R. (eds) *Proceedings of the NODALIDA 2009 workshop "WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies"* NEALT Proceedings Series, Northern European Association for Language, Vol. 7

Park, Y. (2004) *GlossOnt: A Concept-focused Ontology Building Tool,* Available online from Intelligence, American Association for Artificial [http://www.aaai.org/Papers/KR/2004/KR04-052.pdf] (accessed 5 July 2010).

Pedersen, B., Nimb, S., *et al.* (2009) 'DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary', *Language Resources and Evaluation 43(3),* s. 269-299.

Shamsfard, M. and Abdollahzadeh, B.A. (2003) 'The state of the art in ontology learning: a framework for comparison', in *The Knowledge Engineering Review,* Cambridge: Cambridge University Press, 18(4): 293-316.

Sprogudvalget (2008) *Sprog til tiden - Rapport fra sprogudvalget.* København: Kulturministeriet.

Suárez-Figueroa, C. and Gómez-Pérez, A. (2008) 'First Attempt towards a Standard Glossary of Ontology Engineering Terminology', in Madsen, B.N. and Thomsen,

H.E. (eds) *Managing Ontologies and Lexical Resources– Proceedings of the 8th International Conference on Terminology and Knowledge Engineering,* Litera 2008.

Tarp, S. (2008a) 'Kan brugerundersøgelser overhovedet afdække brugernes leksikografiske behov', in *LexicoNordica*, 2008, 15: 5-32.

Tarp, S. (2008b) *Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography* (Lexicographica : series Maior, Volume 134 ed.) Tübingen: Max Niemeyer Verlag.

Tarp, S. (2009) 'Reflections on lexicographic user research', in *Lexikos*, 2009, 19: 275-296.

Thomsen, H.E. (1998) 'Typed Feature Structures for Terminology Work - Part II', in LSP - *Identity and Interface - Research, Knowledge and Society, Proceedings of the 11th European Symposium on Language for Special Purposes,* August 1997, Copenhagen: Copenhagen Business School, 349-359.

Thomsen, H.E. (1999) 'Typed Feature Specifications for establishing Terminological Equivalence Relations', in *World Knowledge and Natural Language Analysis. World Knowledge and Natural Language Analysis,* 23: 39-57.

Zhou, L. (2007) 'Ontology learning: state of the art and open issues', *Information Technology and Management,* 8(3): 241-252.