# Human Language Technology Elements in a Knowledge Organisation System - The VID project

**Costanza Navarretta, Bolette Sandford Pedersen, Dorte Haltrup Hansen**

Center for Sprogteknologi, University of Copenhagen
Njalsgade 80, 2300 Copenhagen S - DK
{costanza,bolette,dorte}@cst.dk

**Abstract**

This paper describes how Human Language Technologies and linguistic resources are used to support the construction of components of a knowledge organisation system. In particular we focus on methodologies and resources for building a corpus-based domain ontology and extracting relevant metadata information for text chunks from domain-specific corpora.

## 1. Introduction

The amount of data available on intranets and/or on the internet has been increasing the last decade and there is a growing interest for developing methodologies and tools for sorting and organising relevant information. Several Danish business companies are beginning to realise the need for knowledge organisation systems that combine human language technology (HLT) with emerging technologies in the field of the semantic web in terms of semantically oriented metadata and/or domain ontologies. Investigating, developing and/or refining HLT techniques for acquiring and representing relevant parts of domain knowledge and corporate language is one of the main aims of the Danish VID project (VIden og Dokumenthåndtering med sprogteknologi – Knowledge and Document Handling with Language Technology). The project participants are the research institution Center for Sprogteknologi (CST), three large Scandinavian companies with high demands for the quality and efficiency regarding document production and two Danish technology companies specialised in search and knowledge organisation with the project role of technology providers. HLT comes into the project partly as a facility to semi-automate the *building* of parts of a knowledge organisation system on the basis of existent documentation, as well as a facility to be applied in *search* (Paggio et al., 2003) and *document production*.

This paper focuses on one aspect of the VID-project, the use of HLT to support the construction of components of a corporate knowledge organisation system exemplified by a case story. First we present the systems' relevant components, then we describe how HLT and existing linguistic resources are used to support the construction of a corpus-based domain ontology as well as the extraction of relevant metadata information from domain-specific corpora.

## 2. A case-story: linguistic-based components in a knowledge organization system

One of the companies participating in the VID-project is a consultancy company with offices in several Nordic countries. Maintaining and updating the company's standard documents require a lot of work, together with detailed knowledge about the working processes, the relevant domain(s), and the legislation in the relevant countries. The company wants to systematise and automate their document production and has therefore acquired a system for semiautomatically saving and producing standard documents which is currently being tuned to the company's needs. To use the system in an optimal way, the company has to systematically store knowledge about the content of their documents. The aim of constructing such a knowledge system is not only to make the document production and maintenance more effective, but also to increase the quality of the documents as well as the knowledge-sharing inside and in-between the company's departments. Because the quantity of standard documents is very large, it is important to be able to find relevant documents and/or text chunks in an easy and flexible way, preferably by natural language queries.

In the first phase of the project CST has worked with a scenario comprising the following knowledge modules:

- a lexical database containing terms from the relevant domains, in this case the patent and trademark domains, as well as general language words which are central to the actual domains and tasks;
- an ontology with concepts and relations covering relevant general language and domain-specific concepts (cf. section 3);
- a database of text chucks with corresponding metadata;
- document type definitions stating how standard documents are composed as different combinations of text chucks.

In addition to the document production component and the above listed knowledge components, an ontology-based search engine is foreseen which enables the user to search in the text chunks/standard documents and in the metadata which enrich them. The relation between the modules is illustrated in figure 1, while an example of search-and-query is given in section 3.4. As it can be seen in the figure we distinguish between words/terms and the concepts these words/terms represent. We treat words and terms as lexical entries which must be encoded in a lexical database. The database interacts with the ontology where the concepts are organised in a structured way. Metadata added to the text chunks are connected to lexical as well as to ontological information.

In the following we describe the extraction of information necessary for building an ontology for the relevant

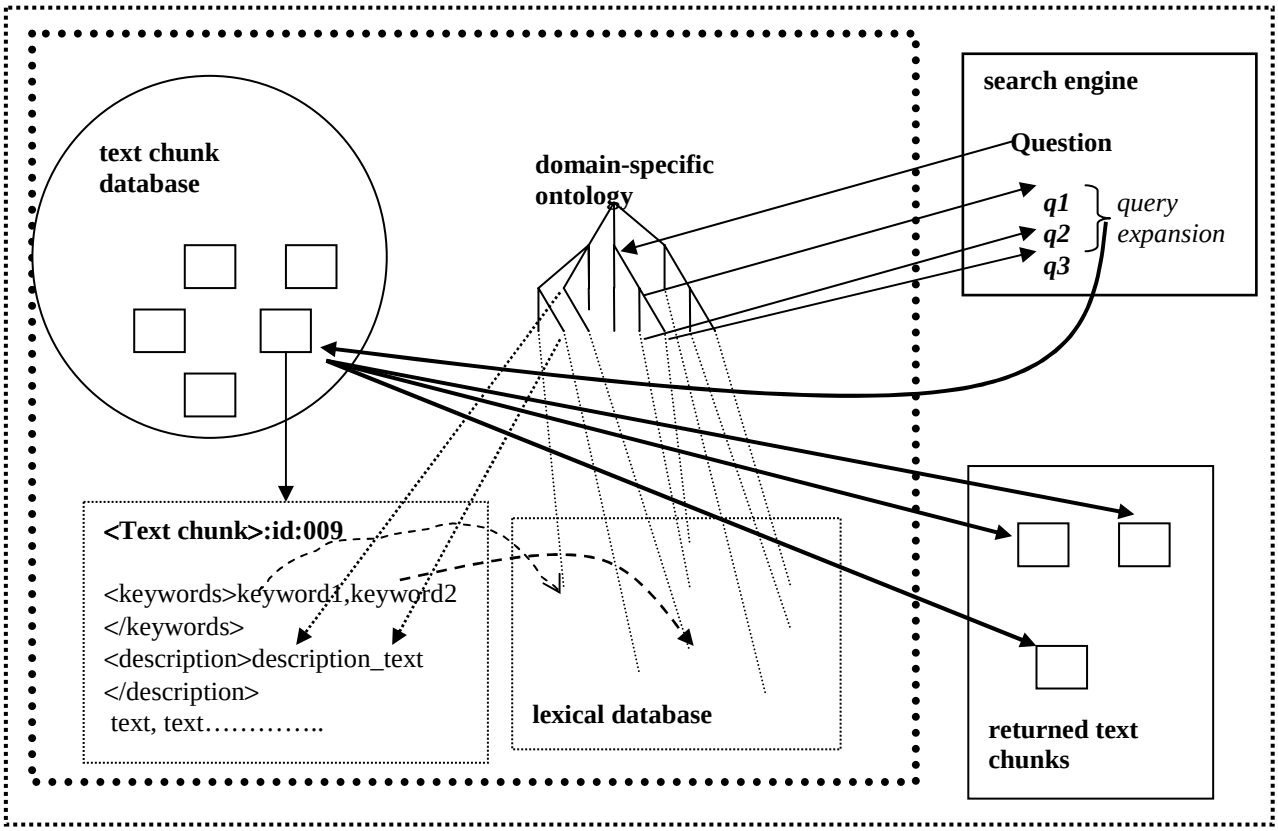


Figure 1: System modules

domains, and the production of content metadata for the text chunks.The definition of document types and text chunks, in contrast, is developed beforehand by the company.

## 3. Applying HLT techniques to construct a linguistic ontology and to define metadata

### 3.1 Vocabulary acquisition

The first step of building the domain-specific ontology consisted in the acquisition of the basic vocabulary. We started by identifying content words and then we distinguished general language words from terms in the two corpora. Content words were extracted from a normalized version of the corpora[2], then they were tagged and lemmatized using the morphology encoding in the Danish computational lexicon, STO (Braasch & Pedersen 2002). Simple terms were automatically identified from general language content words by comparing the extracted content words with a list of general language lemmas in the STO lexicon and marking the lemmas that did not occur in STO as term candidates, according to the methodology proposed by Jørgensen et al. (2003).

General language words which occurred as elements of compounds were also marked as candidate terms, e.g. *extension* (extension) and *gebyr* (fee) were recognised from *extensionsgebyr* (extension fee).

Examples from the automatically generated term lists are given in figure 2 and are read as follows: total number of occurrences of the lemma in the corpus, the lemma itself, the POS tag (N for noun, EGEN for proper noun, ADJ for adjective) and in brackets the number of occurrences of each inflected form in the corpus.

142 nyhedsundersøgelse N (93 nyhedsundersøgelse/N, 45 nyhedsundersøgelsen/N, 4 nyhedsundersøgelsens/N_GEN)
111 EPO EGEN (111 EPO/EGEN) …
84 ansøgningstekst N (41 ansøgningstekst/N -, 43 ansøgningsteksten/N -)
62 præliminær ADJ (54 præliminær/ADJ, 8 præliminære/ADJ)

Figure 2: Extract from the term candidate list

The obtained lists of term candidates were given to the company experts, who evaluated and complemented them. Multi-word term candidates and collocation candidates were automatically extracted using pointwise mutual information of the tagged bigrams and trigrams in our corpus (Church and Hanks, 1989). A reduced tag-set exclusively indicating word class information was used, e.g. tags such as EGEN_GEN (proper-genitive) and V_PAST (verb in past form) were replaced with EGEN and V respectively. We especially focused on bigrams and trigrams with high mutual information[3] and consisting of subsequent nominals (proper nouns and/or common nouns) and on nominals followed by a

[2] In these phase Word-files were converted to simple text, tables, figures and lists were specially marked, felts to be filled in by the users of standard documents were changed to appropriate dummies and so on.

[3] Mutual information was calculated with the CMU-Cambridge Statistical Language Tool (Clarkson and Rosenfeld, 1997).

preposition and a nominal. This procedure identified multi-word terms, company names or organisations, addresses, countries, patent-related standards and general language collocations. Examples of these phrases are given in figure 3.

Burkina/EGEN Faso/EGEN
Eurasian/ADJ patent/N office/N
information/N disclosure/N document/N
den/PRON_DEMO ikke-registrerede/ADJ design/N
(the unregistered design)
EF/EGEN design/N (EC design)
skånefrist/N for/PRÆP design/N
(protective time-limit for design)

Figure 3: Automatically identified multi-words

Approximately one fourth of the proposed term candidates were removed from the list by the domain experts. Some of the terms which were added to the list were in the analysed corpora, but had been marked as general language content words, because they were encoded in the STO-lexicon. This was especially the case for legal words such as *ret* (law) and *domstol* (court). Finally, a group of terms were not contained in the corpora, but were added to the list by the company experts.

## 3.2 Structuring the extracted data

To model the extracted data into an ontology, we chose the standard W3C Ontology Web Language (OWL) (http://www.w3.org/TR/owl-ref/) while we used Protégé (version 2.0 with the owl plugin) as encoding tool. Protégé and the owl plugin are developed at Stanford University (http://protege.stanford.edu/).
We refer to the ontology as *linguistic ontology* for three reasons: (i) it is linguistically 'anchored' being produced primarily on the basis of text corpora, (ii) it is language specific at the lower levels, in this case Danish although mapped into a language-independent upper level-ontology (iii) it addresses linguistic problems like synonymy, synonymous expressions and polysemy. The ontology has been constructed combining bottom-up and top-down strategies, see i.a. (Hobbs 1984). The top-down strategy consists of organising the top levels of the ontology adapting the high level concepts from the SIMPLE (Semantic Information for Plurilingual, Multifunctional LExica) Core Ontology (cf. Lenci et al. 2001, Pedersen & Paggio, in press). Semantic relations between concepts are encompassed as a part of the SIMPLE model, but augmented further with a set of domain specific relations.
The lower nodes are established bottom-up on the basis of the term lists and the generated corpora. A company specific patent dictionary, which has been scanned in order to construct an electronic version, has also proven to be of considerable relevance and help for the ontology structuring. Term definitions from the patent dictionary include in most cases a genus proximum (closest superconcept) already present in the term list, a fact which proved useful during the basic structuring of concepts. All nodes in the ontology are revised by the term experts and further divided into ontological subtypes. The lower nodes are merged with the SIMPLE categories manually. A description of the ontology can be found in Petersen et al. (forthcoming).

## 3.3 Integrating Dublin Core metadata and ontological information

The standard documents of the company consist of partly overlapping text chunks which are organized in the company's document production system. To facilitate the categorisation, searching and manipulation of text chunks, we assign metadata to each of the chunks. We use the Dublin Core Metadata Element set and focus here on the two types of metadata that deal with the content of a text, namely the DC.Subject and the DC.Description. The DC.Subject field contains the keywords of the text. We follow the Dublin Core Initiative's recommended best practice namely to select a value for the keyword from a controlled vocabulary, which in our case is the above described lexical database. The DC.Description field on the other hand contains an account of the content of the text here expressed by the salient NPs of the text. We assume that the described metadata coexist with other relevant metadata such as DC.Creator, DC.Publisher, DC.Date etc. so that search in each of the metadata fields or in a combination of these is possible.
Traditionally keywords are found by 1) removing high frequency words, 2) stripping suffixes and 3) detecting equivalents. We build upon the same approach but use linguistic methods instead of simple frequency calculation.
As described in 3.1 we process the texts by tokenizing, POS-tagging and lemmatizing them, ending up with lemma frequency lists. From these lists we extract nouns and look them up in the lexical database containing domain relevant entries. In this way the lexical database is the link between the domain ontology and the metadata.
The approach for dealing with text chunks is slightly different than the approach for dealing with bigger texts. In text chunks frequency doesn't play a role since most of the lemmas only occur once. Therefore it is crucial for the detection of keywords that they can be found in the lexical database. Besides assigning the ontological type (a concept) to each keyword, we can then abstract and assign wider keywords looking at the encodings of the ontology. Some of the concepts in the ontology have corresponding words in the lexical base, others, typically higher level nodes, have not.
Examples of keywords for the text in (1) are given in figure 4.
(1) *Endvidere beder vi Dem meddele os, om De ønsker at søge extension til Albanien, Letland, Litauen, Makedonien, Rumænien, Slovenien idet der i givet fald skal betales extensiongebyr for hvert land*
(Furthermore, we request that you inform us if you wish to extend the patent to Albania, Latvia, Lithuania, Macedonia, Romania, Slovenia, since an extension fee has to be paid for each country).

| **Keyword:** | **Wider Keyword:** |
|---|---|
| *Extensionsgebyr* (extension fee) | *Gebyr* (fee) |
| *Extension* (extension) | *CauseRelationalChange* |
| *Land* (country) | *Geopolitical* |
| *Slovenien* (Slovenia) | *Østland* (East European |

| | Country) |
|---|---|
| *Rumænien* (Romania) | *Østland* |
| *Makedonia* (Macedonia) | *Østland* |
| *Litauen* (Lithuania) | *Østland* |
| *Albanien* (Albania) | *Østland* |

Figure 4: Examples of keywords

From the automatically extracted keywords we derive the description used in the DC.description field by extracting the NPs where the keywords function as heads, as can be seen in the DC:description field in figure 5,[4] which contains some of the metadata connected to the text chunk in (1).

<dc:**title**>Extension to East European Countries </dc:**title**>
<dc:**publisher**>Patent Office</dc:**publisher**>
<dc:**creator**>Signe Holm</dc:**creator**>
<dc:**subject**>extension fee, extension, country, Slovenia, Romania, Macedonia, Lithuania, Latvia, Albania</dc:**subject**>
<dc:**description**> extent the patent to Albania, Latvia, Lithuania, Macedonia, Romania, Slovenia - extension fee</dc:**description**>
<dc:**date**>24-09 2003</dc:**date**>
<dc:**type**>text chunk for standard document </dc:**type**>
<dc:**language**>da</dc:**language**
<dc:**relation**>Standard document 5: Publish_newsletter_danish Standard document 7: EP_Publish_newsletter_danish
</dc**:relation**>

Figure 5 : Metadata for text chunk in (1)

### 3.4 An example of search-and-query

In the following we present an example illustrating how ontology, metadata and text chunks interact. Consider a case of adjustment of the legislation regarding patent fees. The employee asks for text chunks relating to *omkostninger ved patentansøgninger* (costs regarding patent applications) and the query engine applies the ontology for query expansion and expands from the concept 'Omkostning' (cost) via 'Gebyr' (fee) to the more specific 'Extensionsgebyr'(extension fee), 'Trykningsgebyr' (printing fee), 'Udstedelsesgebyr' (execution fee) etc.
When a match is found between a keyword and a concept (eventually expanded) from the query, the text chunk(s) tagged with the given keyword is (are) extracted, ready for the case officer to check and eventually update according to the adjusted legislation.

## 4. Evaluation and concluding remarks

The coverage and the quality of the data produced semi-automatically from the corpora have continuously been evaluated by the company experts. The results of these evaluations are promising and indicate that HLT is useful as a substantial support to the construction of knowledge organisation systems.
An evaluation of the constructed ontology as backbone of a semi-automatic document production system will be realised when the ontology is fully integrated in the document production system.

[4] In the example metadata are given in English.

Currently we are examining to which extent existing statistical clustering methods enriched with our linguistic resources, can support/validate the ontology building process. Furthermore we are extending the relations between concepts by using the syntactic and the semantic patterns encoded in the STO-lexicon. Finally we are testing some of the described methods for discovering metadata information on different domain and text types.


## Acknowledgements

The VID-project is funded by the Danish Research Councils. We would like to thank Lina Henriksen, Bart Jongejan og Bente Maegaard (CST) for their useful feedback and the companies participating in the project for a fruitful cooperation.


## References


Braasch, A. & Bolette S. Pedersen (2002). Recent Work in the Danish Computational Lexicon Project "STO", in *EURALEX Proceedings 2002*, Center for Sprogteknologi, Copenhagen.

Church, K.W. and P.Hanks (1989). Word association norms, mutual information and lexicography. In: *Proceedings of ACL 27*, pp.76-83.

Clarkson, P. and R. Rosenfeld. (1997). Statistical Language Modelling Using the CMU-Cambridge Toolkit. In *Proceedings of ESCA Eurospeech 1997*.

Hobbs, J. R. (1984). Sublanguage and Knowledge. Technical Note 329, SRI, California.

Guarino, N. & Welty, C. (2000). "Ontological Analysis of Taxonomic Relationships", in: A. Laender V. Storey (eds.*) Proceedings of ER-2000. The International Conference of Conceptual Modeling*. Springer Verlag.

Jørgensen, S.W., Hansen, C., Drost, J., Haltrup, D., Braasch, A., Olsen, S. (2003). Domain specific corpus building and lemma selection in a computational lexicon, In: *Corpus Linguistics 2003 Proceedings*, Lancaster.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villages, M. , Zampolli, A. (2000). "SIMPLE – A General Framework for the Development of Multilingual Lexicons", in: T. Fontenelle (ed.) *International Journal of Lexicography Vol 13*. pp. 249-263. Oxford University Press.

Paggio, P., B. S. Pedersen, D. Haltrup (2003) Applying Language Technology to Ontology-based Querying - The OntoQuery Project. Applied Artificial Intelligence Journal. Artificial Intelligence for Cultural Heritage and Digital Libraries, Vol. 17 Numbers 8-9:817-833.

Pedersen, B., C. Navarretta, D. Haltrup. (fortcoming) 'Building Business Ontologies with Language Technology Techniques-The VID project'. Submitted for review at *Ontolex 2004. Ontologies and Lexical Resources in Distributed Environments. LREC 2004 Workshop.*

Pedersen, B., Paggio, P. (in press) The Danish SIMPLE Lexicon and its Application in Content-based Querying, to appear in *Nordic Journal of Linguistics*.