Applied Artificial Intelligence, 17:817–833, 2003

ISSN: 0883-9514 print / 1087-6545 online
DOI: 10.1080/08839510390225177

Taylor & Francis
Taylor & Francis Group

APPLYING LANGUAGE TECHNOLOGY TO ONTOLOGY-BASED QUERYING: THE ONTOQUERY PROJECT

PATRIZIA PAGGIO, BOLETTE S. PEDERSEN, and DORTE HALTRUP
Center for Sprogteknologi, Copenhagen, Denmark

This paper addresses the issue of how language technology resources and components can be applied in ontology-based querying. In particular, it presents the approach to text and query analysis adopted in the Danish research project OntoQuery, where shallow syntactic analysis and ontology-based parsing are combined in order to identify nominal phrases (NPs) and assign them a semantic description. Semantic descriptions are used by the search engine to match queries against texts in a database, and a ranking of the texts retrieved is produced based on a domain ontology. This is intended to be a general methodology applicable to texts from different domains, including those relevant to cultural heritage, although OntoQuery has chosen nutrition as its first target domain. The paper focuses on the language technology aspects of the methodology, the ontology-based lexicon inherited from the SIMPLE project, and the development of the domain-specific ontology. The methodology is partly implemented in a prototype.

ONTOLOGY-BASED QUERYING: WHY IT IS USEFUL

Future growth in the areas of digital libraries and online searches in databases available on the Web stress the need for content-based extraction of information, where the content is eventually represented by means of ontologies of the relevant domains. The collaborative Danish research project OntoQuery is developing a methodology for ontology-based information retrieval (IR) where linguistic analysis, shaped on the basis of an ontology, is applied to queries and texts such as online documents and encyclopedias.

The nutrition domain has been selected as a first test domain and is being modeled in our ontology from the perspective of a layman. In other words, the OntoQuery system is primarily intended to answer queries of the type one

Address correspondence to Patrizia Paggio, Center for Sprogteknologi, Njalsgade 80, DK-2300, Copenhagen S, Denmark. E-mail: patrizia@cst.dk

would expect from the general public and not necessarily from domain experts. For example, a nonexpert would not necessarily know which deficiency diseases are related to which vitamins (or lack of these), and would probably tend to query in rather general terms. The same type of nonexpert queries may also be posed in the cultural heritage domain. For instance, somebody generally interested in Danish 19th century painting may want to know about motifs in works by different painters. In our view, ontologies reflecting a nonexpert perspective are best built on the basis of text types intended for nonexperts, such as encyclopedias and textbooks. In OntoQuery, the Large Danish Encyclopedia has provided the corpus used both for selecting the vocabulary to be dealt with, and for determining the content relations between the terms to be included in the ontology.
Let us assume for example that a nonexpert user may pose the query in (1).

Hvilke sygdomme har at gøre med mangel på vitaminer i kosten? (1)
(Which diseases have to do with lack of vitamins in the diet?)

A system based on simple keyword recognition would probably retrieve texts containing exactly the words occurring in the query, possibly ignoring very frequent words such as *at* (to). This is in fact exactly what happens if the query is submitted to the popular search engine Google (www.google.com). The text at the top of the list produced by Google as a result for the query in (1) is shown below (the translation is ours):

...alkoholforbrug og ***mangel på*** *fysisk aktivitet ... hvordan* ***kosten*** *er sammensat ... vægttabet bevares* ***på*** *længere sigt ... en kost* ***med*** *et alt ... og* ***har*** *nogle ... til at* ***gøre*** *dem fede...* (2)

(alcohol consumption and ***lack of*** physical activity... how ***the diet*** is put together ... weight loss is maintained ***in*** the long run ... a diet ***with*** much ... and ***has*** some ... to ***render*** them fat ...)

The search was carried out using the Danish interface to Google, and asked the system only to look at Danish documents. Note that the words highlighted in the text retrieved are in exactly the same inflected form as in the query; in other words the system does not seem able to find, for instance, *sygdom* (disease) as well as *sygdomme* (diseases) as possible matches. At least for Danish, then, no stemming is used. Nor is any attempt made to find possible synonyms of the terms in the query. Seen from this perspective, texts containing the words *avitaminose* (vitamin deficiency) and *vitaminmangel* (vitamin lack), both of which are synonyms of *mangel på vitaminer* (lack of vitamins), should pop up on the screen. Note also that the text, at least

judging from the excerpt displayed, does not appear entirely relevant since it is apparently not about lack of vitamins, which is a central concept in the query, but is about lack of physical activity.

To see how the problems mentioned are not only relevant to scientific or technical texts, let us consider a query in the artistic domain:

Rom i dansk maleri fra 1800-tallet.
(Rome in Danish painting from the 19th century.) (3)

Presented with this query, Google retrieves at the top of its list the following answer:

Dyreetiske Råd...Knudsen, Johannes Døden i ***Rom****, ...* ***Dansk*** *kunst* ***fra 1800-tallet*** *i ... kunst, islamisk kunst* ***fra*** *... Symbolismen i* ***dansk*** *og europæisk* ***maleri*** *...*
(Committee for animal ethics... Knudsen, Johannes Death in **Rome**... **Danish** art in **the 19th century** in...art, Islamic art **from**... symbolism in **Danish** and European **painting**...) (4)

On closer inspection, the document turns out to be an alphabetically ordered list of books, in which the proximity of the word Rome and titles dealing with 19th century Danish art is totally coincidental. The document as a whole is irrelevant to the query. In fact, several of the documents retrieved in response to the query under consideration, and not shown here, display the same problem. They contain some of the relevant keywords; however, these are not connected in the text by a syntactic, let alone a semantic relation in spite of the fact that they occur not too far from one another. In the example just discussed, the words *Rom, dansk*, and *maleri* do not occur within the same nominal phase (NP) as they do in the query, in fact, not even in the same sentence. Therefore, the semantic relations that hold between them in the query are not represented in the document.

In this paper, we present an approach that relies on the identification of morphological, syntactic as well as semantic relations between words to enable systems to provide more refined answers to a query. More specifically, morphological processing can be used to neutralize different forms of the same word, synonyms can be identified, and syntactic chunking can help find syntactically meaningful groups of words. In fact, in OntoQuery, we are developing an ambitious methodology where ontological knowledge is also applied. To return to our nutrition example, texts containing words such as *beriberi* or *pellagra*, both subconcepts of *sygdom* (disease), are relevant answers to the query in (1), and should be among those retrieved. Looking at entire nominal phrases (NPs) rather than isolated terms, a text containing the

NP *antioxidanter i frugt og grøntsager* (antioxidants in fruit and vegetables) would also be relevant to the same query, since antioxidant and vitamin are closely related (they are in fact siblings in the OntoQuery nutrition ontology), and fruit and vegetables are usually part of the diet. The OntoQuery project is developing a methodology in which such similarities can be modeled and taken advantage of for text searching. The methodology is based on the availability of ontological knowledge and aims at improving the results provided by the keyword recognition strategy characteristic of most existing search engines.

Ontological knowledge plays a dual role in the project's methodology. First, it is used as a backbone to analyze texts and queries. The result of this analysis is a set of conceptual descriptions shaped according to the project formalism OntoLog (Nilsson 2001), and intended to represent the content of both texts and queries in a more meaningful way than lists of keywords. The analysis is—as we shall see—limited to nominal phrases. Second, the ontology is the term of reference used by the search algorithm to evaluate the relevance of the text descriptions stored in the system's text database to the query description.

This paper focuses on two important aspects of the methodology which constitute CST's main contributions: the reuse and extension of the SIMPLE ontology and lexicon for the purposes of IR and the use of various natural language processing (NLP) components to produce ontology-based analyses. Thus, after a brief introduction to the system's general architecture, the ontology underlying the system is presented and its relation to the Danish lexical resources provided by SIMPLE-DK explained. We also give an account of the NLP techniques implemented in the prototype in order to find, delimit, and analyze NPs in queries and documents.

A PROTOTYPE APPLICATION

As a test-bed of the methodology embraced by the project, a prototype application is being developed by the OntoQuery team at the Computer Science Section at Roskilde University (RUC) in cooperation with CST. It consists of two main components: a generator of conceptual descriptions and a search component. The generator, which implements the methodology described in this paper, is shown in Figure 1. First of all, it is in charge of processing new documents, in other words, it analyzes them to produce semantic descriptions and then storing these descriptions into the system's knowledge base. Furthermore, the generator also analyzes queries in order to produce semantic descriptions similar to those generated for the texts. It is then the job of the search component to compare the semantic description representing an incoming query with

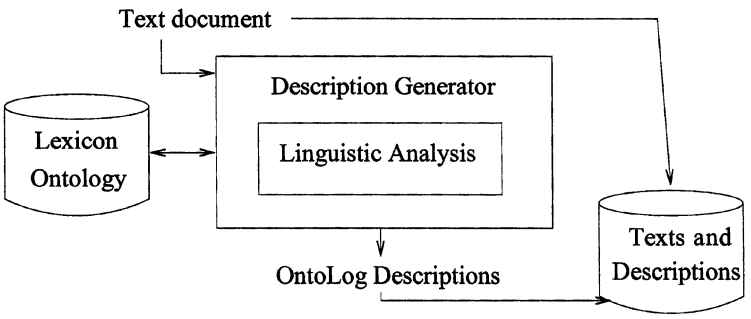

FIGURE 1. Description generator.

those stored in the knowedge base, and to find an appropriate ranking of the texts based on the results of this comparison. Informally, the measure of similarity used to rank the texts reflects the distance between the semantic descriptions they correspond to and the semantic description derived from the query, where this distance is determined with respect to the system's ontology. Since the focus of this paper is the text analysis process rather than comparison and retrieval, we refer to Andreasen et al. (2000) and Andreasen et al. (2002) for more details on the comparison strategy.

In the remainder of the paper, we discuss the principles and coverage of the system's lexicon and ontology—both derived from the Danish SIMPLE as described in the next section, and give an account of the linguistic analysis and the generation of semantic descriptions. Since the project's methodology is still evolving, we describe parts of the generator that are already in use, as well as enhancements with which we are experimenting. Figure 2 displays an example of results produced by the current prototype. The user can follow the analysis process by inspecting intermediate results produced by the various linguistic analysis steps. These results, as well as the techniques used to produce them, are discussed in "Using the Ontology." Here it will suffice to point out that given as input the query *Hvilke sygdomme har at gøre med mangel på vitaminer?* (Which diseases have to do with lack of vitamins), the system generates a semantic description consisting of two sets of concepts, corresponding to the meaning of the two NPs in the query as follows:

$$\begin{array}{l}(\textit{sygdom}),\ (\textit{mangel},\ \textit{vitamin}) \\ (\text{disease}), (\text{lack},\ \text{vitamin})\end{array} \tag{5}$$

A list of 270 texts matching this description in various degrees is produced (only the four top-ranking texts are shown in the figure.) None of these represents an exact match. However, the top-ranking texts are all relevant to the query since they contain sets of concepts closely related (via the *is-a*

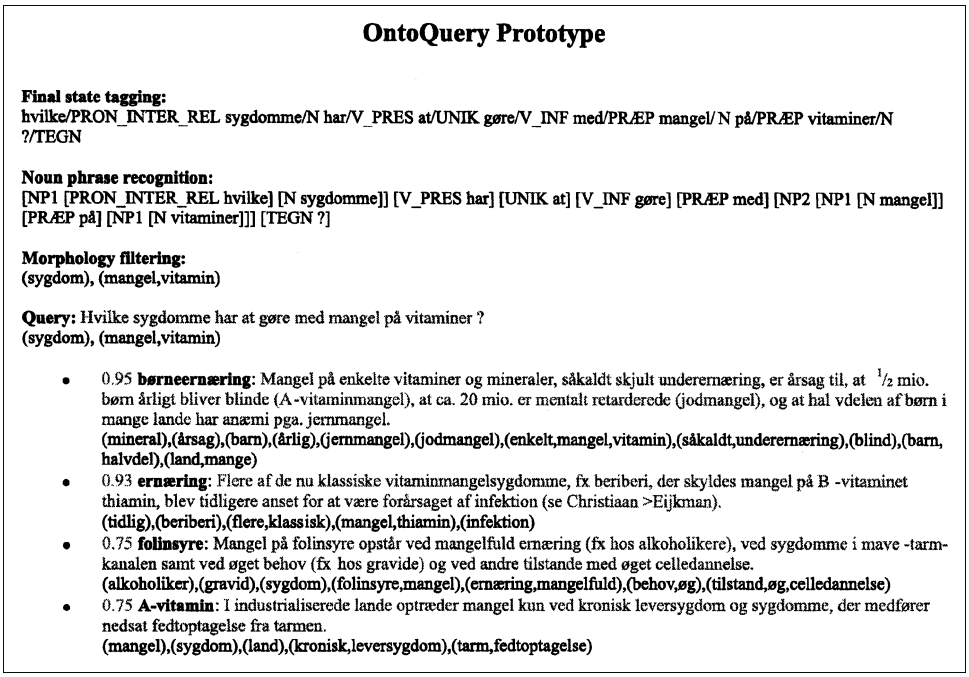

FIGURE 2. Search results for the query *Hvilke sygdomme har at gøre med mangel på vitaminer?*

relation in the ontology) to those in the query description; for instance, they contain subconcepts to *sygdom,* such as *anæmi* (anemia), *jernmangel* (iron deficiency), and *beriberi* (beriberi).

One can imagine a similar analysis with the art example presented earlier: *Rom i dansk maleri fra 1800-tallet* (Rome in Danish painting from the 19th century), where a conceptual analysis of the phrase would look as follows:

$$\begin{array}{l}(\textit{Rom},\ (\textit{dansk},\ \textit{maleri}),\ \textit{1800-tallet})\\ (\text{Rome},\ (\text{Danish},\ \text{painting}),\ \text{19th century})\end{array} \tag{6}$$

In other words, the phrase is represented as a composite concept rather than as a list of keywords, a fact which would ensure a better ranking of several of the irrelevant hits from Google seen earlier. In addition to this improvement, doing query expansion by means of the ontology would open up for hits containing the close synonym *billedkunst* (painting, literally: pictorial art), as well as the closely related superconcept *kunst* (art). Such a query expansion in Google (done manually) actually results in a very relevant first hit, namely:

Italiens og især Roms betydning for dansk kunst har ... gennem første halvdel af 1800-tallet...
(Italy's and especially Rome's influence
on Danish art ... during the first half of the 19th century) (7)

As already mentioned, the OntoQuery methodology is intended to be a general one, and the SIMPLE top-ontology and lexical entries contain concepts from several different domains, including the artistic domain mentioned here. However, the OntoQuery prototype is currently only being tested on nutrition texts. To return to nutrition then, the first preliminary tests (Pedersen and Paggio 2002) show an interesting tendency with regard to the way in which the retrieved texts pattern together. Typically, for each query, only few texts are retrieved with a score between 0.90 and 1.00. In contrast, a relatively large number of the retrieved texts have a score of 0.50 and below. From our point of view, the texts scoring between 0.90 and 0.95 are of particular interest: They do not contain exactly the same concepts as in the queries, but rather specifications of these (e.g., *nikotinamid* instead of *vitamin B*); however, they are still very good answer candidates to the queries put forward. These results are encouraging since they point to the fact that conceptually relevant texts are retrieved only because the system makes use of an ontology. In a system that does not rely on ontological knowledge, on the other hand, there is no guarantee that these texts would outrank less relevant documents containing a larger number of matching keywords or be found at all.

THE ONTOLOGY AND THE LEXICON

The project's ontology currently builds on the SIMPLE ontology, which was developed in the EU-project SIMPLE (Semantic Information for Multifunctional Plurilingual Lexica) (Lenci et al. 2000). SIMPLE was completed in May 2000 and produced harmonized semantic lexicons for NLP for twelve of the European languages.

The language specific encodings in SIMPLE are performed on the basis of a unified, ontology-based semantic model representing an extended qualia structure based partly on Pustejovsky (1995) and partly on experience in previous lexical projects such as Genelex, WordNet (Miller et al. 1990), and EuroWordNet (Vossen 1999). A general design model is thus provided, allowing for the encoding of a large amount of semantic information such as ontological typing, domain information, semantic relations, argument structure, event structure, and selectional restrictions. The SIMPLE lexicons aim at meeting the demands of advanced language technology applications such as machine translation and content-based querying.

The SIMPLE top-ontology consists of approximately 130 concepts comprising concrete and abstract entities, properties, and events. It resembles the EuroWordNet top-ontology, only for events is it more detailed, comprising in all thirty-five event subtypes. SIMPLE constitutes a well-tested starting point for OntoQuery since it has been used for semantic lexicon building in different European languages. Furthermore, the project can benefit from the Danish SIMPLE lexicon, which consists of 10,000 Danish sense descriptions (see Pedersen and Nimb 2000 and Pedersen and Keson 1999), which are all directly grounded on the top-ontology.

In OntoQuery, the SIMPLE ontology has been extended with a domain-specific sub-ontology developed empirically on the basis of texts on nutrition from the Large Danish Encyclopedia (Nilsson and Oldager 2000). These texts also make up the coverage of the domain specific lexicon, currently consisting of approximately 1,000 lexical entries.

The OntoQuery ontology—founded on SIMPLE—applies orthogonal inheritance as a means of expressing a four-dimensional structure following Pustejovsky's four Qualia Roles: Formal, Agentive, Telic, and Constitutive role (Pustejovsky 1995). Most nutrition concepts however are so-called simple types (being natural substances) that only inherit from the Formal dimension in the hierarchy. For instance, the concept *vitamin A* displays the following inheritance structure:[1]

$$\begin{aligned} &\textit{a-vitamin}\ (\text{vitamin A})\ {}_{isa}\!\rightarrow \textit{fedtopløseligt vitamin}\ (\text{fat soluble vitamin}) \\ &\quad {}_{\text{isa}}\!\rightarrow \textit{vitamin}\ (\text{vitamin})\ {}_{\text{isa}}\!\rightarrow \textit{mikronæringsstof}\ (\text{micro nutrition substance}) \\ &\quad {}_{\text{isa}}\!\rightarrow \textit{næringsstof}\ (\text{nutrition substance})\ {}_{\text{isa}}\!\rightarrow \textit{natural substance} \\ &\quad {}_{isa}\!\rightarrow \textit{substance}\ {}_{isa}\!\rightarrow \textit{concrete entity}\ {}_{isa}\!\rightarrow (\textit{entity})\ {}_{isa}\!\rightarrow \textit{top} \end{aligned} \tag{8}$$

In contrast, a nutrition term like *kosttilskud* (dietary supplement) is a so-called unified type in that it inherits information from more than one

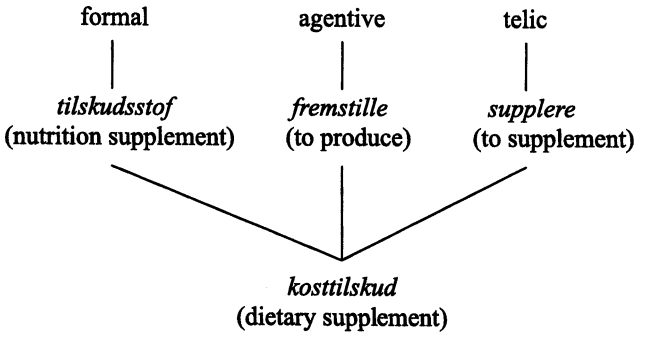

FIGURE 3. The concept *kosttilskud* as a unified type.

dimension, in this case three: its closest hyperonym *tilskudsstof* (nutrition supplement), its creation (as a man-made entity), and, finally, its function as a supplement to ordinary food. These additional dimensions are encoded by means of semantic relations as seen in Figure 3, where *kosttilskud* relates to the event *fremstille* (to produce) by means of a made-by relation and to the event *supplere* (to supplement) by means of a used-for relation.

Since SIMPLE is built as a multipurpose resource also suitable for machine translation and other LE applications, not all the semantic information it provides is applied in OntoQuery. Argument structure, event structure, selectional restrictions, and semantic relations other than hyponymy relations are not included in the current version of the OntoQuery prototype. However, the project's search methodology is being extended to take advantage of at least a subset of the semantic relations applied in SIMPLE (part-of, has-as-parts, used-for, made-by, located-in), as well as of argument structure and selectional restrictions for deverbal nouns. These information types are expected to play a role in the definition of a domain-relevant conceptual grammar capable of dealing with at least some types of lexical and semantic ambiguity, as discussed in some detail in Pedersen and Paggio (2002). Below we return to the role relations play in the project's methodology in connection with the generation of conceptual representations.

USING THE ONTOLOGY: ANALYZING TEXTS AND QUERIES

As already mentioned, the concept ontology is used to derive semantic descriptions from both text databases and queries, and to determine the mutual closeness of these descriptions with respect to the ontology itself. Currently, only the concepts occurring in NPs are taken into consideration, whilst the remaining text is ignored.

To produce semantic descriptions, the project applies a number of NLP techniques. The idea that NLP may be helpful to Information Retrieval (IR) is not a new one. Already in the late 1960s and early 1970s, experiments were made to automatically identify references to thesaurus concepts and complex terms in documents (reported in Allan 2000). Experiments with the use of NLP, e.g., for query expansion and complex phrase detection, continued—although with scarce success—throughout the 1980s. In 1996, a special track was dedicated to NLP at the recurring Text REtrieval Conference (TREC-5) (see Strzalkowski et al. 1998). The four participating sites claimed up to 40% accuracy improvements due to the use of NLP techniques. Query expansion and nominal phrase (NP) identification seemed especially to have provided the most notable gains compared to term-only runs. Although it can be argued, as Allan does, that the same accuracy can be achieved by way of statistical means, and that NLP is still too error-prone to work well for large

texts, TREC-5 certainly contributed to a growing interest in the use of NLP in IR. In fact, basic NLP techniques such as tokenizing, stemming, and the use of stopword lists are fairly common in IR systems. Some systems also use more advanced but still relatively robust NLP, such as phrase identification and named-entity extraction. NLP has again come into focus in connection with the question answering track at TREC-8 (1999), as systems using NLP techniques actually outperformed statistics-based ones. In particular, Voorhees and Tice (2000) report that the best scoring system for the 50 byte limit task[2] was the *Textract* system developed at Cymfony, Inc. (Srihari and Li 2000), which uses named-entity tagging and shallow parsing to extract templates from queries and texts.

The originality of our approach consists in combining shallow syntactic analysis with the generation of ontology-based descriptions.

Delimiting Nominal Phrases

In addition to tokenizing, linguistic analysis consists of four steps: part-of-speech (POS) tagging, NP recognition, lemmatizing, and NP parsing. The first three steps are integrated in the prototype, whilst NP parsing is still under development.

The tagger is an implementation of Eric Brill's tagger (Brill 1995), customized for interaction with a text database and trained for Danish. The training corpus is the Danish PAROLE corpus, consisting of 250,000 tokens and semi-manually tagged with a tag set of 151 different tags. These were reduced to forty-three before training the tagger. Preliminary experiments indicated that a smaller tag set would provide a better analysis, especially with a relatively small training corpus. The results obtained are comparable with those reported by Brill, namely an accuracy of 96.5% (Brill 1995).

NP recognition is performed by the chunk parser "Cass" (Abney 1996), based on finite-state cascades—a very efficient technique that others have used with success in Information Extraction (IE) tasks (e.g., Ciravegna and Lavelli 1999 and Kokkinakis 2000). The grammar has been developed manually on the basis of the occurrence of various NP types found in the PAROLE corpus and the corpus built on the Danish encyclopedia. At first the NP recognizer finds NP chunks extending from the beginning of the constituent to its head, according to the definition of a chunk given by Abney himself. Then post-head prepositional phrases (PPs) are recognized and joined to the core NP chunk. More complex sentential modifiers, such as relative clauses, are not recognized nor is the PP attachment resolved. The extension of core NPs with PPs allows the search engine to work with relatively rich descriptions. However, because of the lack of PP resolution it also produces wrong analyses, which we rely on the conceptual grammar to weed out.

The output of NP recognition is then lemmatized, and words are replaced by concepts from the concept ontology. Currently then, each NP is represented as a set of concepts so that the semantic description, which is input to the ontology-based search engine, is a set of sets of concepts. However, the project is working on the development of a more fine-grained analysis of the ontological content of NPs, as will be explained next.

The results produced by the various NLP components for the query *Hvilke sygdomme har at gøre med mangel på vitaminer?* (Which diseases have to do with lack of vitamins) are shown in Figure 4. The figure illustrates the components that are already integrated in the prototype, but also the way in which an NP parser capable of generating more structured conceptual representations than those currently processed by the search engine will fit into the overall architecture. NP parsing is the subject of the next section.

Generating conceptual representations

As we have seen, at the moment the content of an NP is represented in the prototype as a set of sets of concepts. However, according to OntoLog, the algebraic logic supported by the search engine, it will ultimately be an algebraic term, or ontotype, associated with a node in a lattice of concepts.

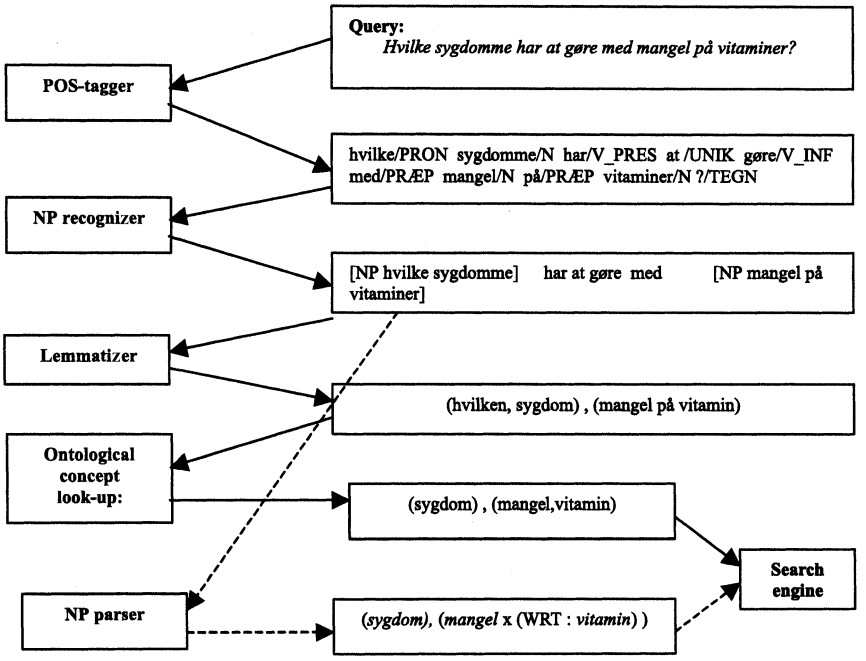

FIGURE 4. The linguistic components.

For example, the ontotype corresponding to the NP *mangel på vitaminer i kosten* (lack of vitamins in the diet) from the query in (1) will be the complex concept resulting from the combination (by means of the meet operator "x") of the atomic concept *mangel* with the two concepts *vitamin* and *kost* by means of the relations WRT (with-respect-to) and LOC (located-in).

$$(mangel \times (\text{WRT}: vitamin) \times (\text{LOC}: kost)) \tag{9}$$

The two relation-concept pairs (WRT: *vitamin*) and (LOC: *kost*), also called semantic roles, are valid restrictions of the concept *mangel*, and the resulting complex concept can thus be regarded as a valid subtype of *mangel*. Given this framework, our hypothesis is that a system based on typed feature structures like LKB (Copestake 1999), or the related PET (Callmeier 2000), can fruitfully be used to parse the NPs identified to produce ontological representations roughly equivalent to OntoLog descriptions. Typed feature structures support, in fact, the definition of concept ontologies and the expression of semantic selectional restrictions associated with the concepts in the form of conceptual feature types.

For example, a concept can be defined as a feature structure type with two valid attributes, a ROLES attribute taking as its value a set of semantic roles and a SELECTS attribute taking as its value a list of semantic roles as follows:

$$\begin{bmatrix} concept \\ \text{ROLES } set\_of\_roles \\ \text{SELECTS } list\_of\_roles \end{bmatrix} \tag{10}$$

This allows for the definition of simple types as well as unified types in the SIMPLE terminology (see "The Ontology and the Lexicon"). Below are a few lines of LKB code showing definitions for the concept *overvægt* (overweight), which is here modeled as a subtype of disease, in turn a phenomenon, as well as a state.

```
event := concept.
phenomenon := event.
state := event.                                    (11)
disease := phenomenon.
overvægt := disease & state.
```

A unified type like *kosttilskud* (dietary supplement), shown earlier in Figure 3, can also be modeled by a typed feature structure. The Qualia roles from

the SIMPLE lexicon specification will correspond to semantic roles as shown below:

$$\begin{bmatrix} \textit{kosttilskud} \\ \text{ROLES} \quad \begin{bmatrix} \text{MADE-BY } \textit{fremstille} \\ \text{USED-FOR } \textit{supplere} \end{bmatrix} \end{bmatrix} \tag{12}$$

where *fremstille* (produce) and *supplere* (supplement) are concepts in the event subontology.

The same kind of structure can be used to represent the content of our NP *mangel på vitaminer i kosten* (lack of vitamins in the diet), which in ontological terms is a complex concept as shown in (9). The corresponding feature structure representation is:

$$\begin{bmatrix} \textit{mangel} \\ \text{ROLES} \quad \begin{bmatrix} \text{WRT } \textit{vitamin} \\ \text{LOC } \textit{kost} \end{bmatrix} \end{bmatrix} \tag{13}$$

The semantic roles [WRT *vitamin*] and [LOC *kost*] correspond to the meaning of the two PPs. In particular, the attributes WRT and LOC, which stand for the semantic relations with-respect-to and located-in, are derived from the meaning of the prepositions as specified in the lexicon,[3] whilst atomic concepts correspond to the meaning of the nouns, also specified in the lexicon. For example in the entry of the word *mangel* (lack), the semantic content is specified to be the concept *mangel*, which in turn is defined in the ontology as a subtype of *state*. A simplified LKB lexicon is shown below:

$$\begin{aligned} &\text{lex\_indhold} := \text{lexeme} \ \& \\ &[\text{ORTH } \text{“mangel”}, \\ &\text{CATEGORY n}, \\ &\text{SEM } \textit{mangel}]. \end{aligned} \tag{14}$$

Note that no syntactic or semantic lexical restriction is part of the lexicon entry. Semantic selectional restrictions are expressed, in fact, not in terms of lexical restrictions, but rather as restrictions on possible combinations of concepts and relations and are thus part of the ontology. This is where the attribute SELECTS in the feature structure defining the generic type *concept* in (10) comes into play. The value of SELECTS is the list of possible semantic roles that the concept in question can be combined with. In the case of *mangel* (lack) then, the ontological definition may be:

$$\begin{bmatrix} \textit{mangel} \\ \text{SELECTS} \quad \left\langle \begin{matrix} [\text{WRT } \textit{natural-substance}] \\ [\text{LOC } \textit{concrete-entity}] \end{matrix} \right\rangle \end{bmatrix} \tag{15}$$

The list of semantic roles specified by the SELECTS attribute is used by the LKB parser (our experimental NP parser) to assign the correct semantic interpretation to complements and modifiers of the head noun via unification. In the example under consideration, the semantic role [WRT *vitamin*], corresponding to the meaning of the PP *på vitaminer* (of vitamins), will unify with the restriction [WRT *natural-substance*]. Unification is possible since the concept *vitamin* is a subtype of *natural-substance* as shown earlier in (8). The semantic role [LOC *kost*], which is the meaning of the PP *i kosten* (in the diet), unifies in turn with the restriction [LOC *concrete-entity*] as the concept *kost* is a subtype of *concrete-entity*.

The advantage of expressing semantic selectional restrictions as part of the ontology rather than in the lexicon, is that restrictions of specific concepts can be inherited from more generic ones. However, there may be cases in which a more specific concept will not inherit all the restrictions of its superconcept, and in which default inheritance must be blocked. An example is the compound term *vitaminmangel* (lack of vitamin), in which the WRT role is lexically saturated. Whether the strategy of moving restrictions from the lexicon (where they are expressed, e.g., in SIMPLE) to the ontology is a viable strategy will ultimately depend on how many possible concept combinations hold at a high level in the ontology, as opposed to the level of very specific concepts, as well as how many idiosyncratic cases we will find in which inheritance must be blocked.

To summarize, the linguistic resources currently developed for the LKB implementation consist of a concept hierarchy with associated selectional restrictions, lexical entries mapping words onto atomic concepts, and a simple NP grammar capable of analyzing the internal syntactic structure of NPs (e.g., a sequence [N PP*]) and of building the semantic representation compositionally on the basis of the semantic restrictions expressed in the relevant concept types.

By virtue of its being based on typed feature structures, then, a system like LKB makes it possible and easy to express semantic generalizations of varying granularity, i.e., at the level of top-ontology concepts as well as more domain-specific ones. The exact nature and number of such generalizations, however, is still subject to investigation and constitutes a challenge to be met by the project. In parallel with theoretical work aiming at developing a conceptual grammar with good coverage of the domain, we are also still in the process of choosing the best tool to implement an NP

parser in the project's prototype. While LKB is a good choice for experiments and quick grammar development, we need a faster tool to interact with the rest of the system. The PET system (Callmeier 2000), a platform for the development of constraint-based grammars employing a number of strategies to achieve compactness and efficiency and capable at the same time of processing grammars developed for LKB, seems to be a promising alternative.

CONCLUDING DISCUSSION

In this paper, basic NLP technologies are combined with advanced conceptual knowledge and an interesting test-bed for the reuse and extension of a multipurpose resource like the SIMPLE lexicon is provided. The paper describes work in progress and therefore extensive evaluation results are not yet available. However, we consider the results obtained so far encouraging as they show the potential usefulness of a methodology for ontology-based querying that relies not only on the semantic closeness of simple concepts in an ontology, but also on the semantic relations between concepts.

Already in the preliminary prototype, the identification of syntactic phrases like *mangel på vitaminer* (lack of vitamins) proves crucial to an appropriate ranking of the documents retrieved (see Figure 2). Syntactic phrases, NPs in our case, are meaningful chunks in that the concepts they contain are related to one another by a semantic relation. Therefore, identifying them means identifying complex concepts and is a prerequisite for finding synonyms and specifications of these. To return to our example, it is appropriate that texts containing the composite concept *mangel på vitaminer*, or synonyms or specifications of this, receive a higher scoring than documents where the concepts occur in isolation. The same ranking cannot always be obtained solely by measuring the closeness of words in a string. Consider thus one of queries discussed earlier, where the closeness of the words *mangel* and *på* in the string *mangel på fysisk aktivitet* (lack of physical activity) makes Google's search engine suggest as the top ranking text a document which is completely irrelevant to the query.

Extensive evaluation needs, of course, to be carried out to provide representative data in order to confirm the usefulness of the approach proposed here. Moreover, the project intends to further develop the methodology and the prototype by investigating and testing the importance of semantic relations in text querying. More specifically, we would like to explore to what degree identifying the relations that hold between concepts in NPs, thereby constructing more precise semantic descriptions of the content of queries and texts, improves querying results.

Let us consider the following examples:

a. *overvægtige brn* (overweight children)
b. *børn med overvægt* (children with overweight [problems])
c. *fede børn* (fat children) (16)
d. *børn med fedmeproblemer* (children with fat problems)
e. *børn der har fedmeproblemer* (children who have fat problems)

Our goal is for the system to be able to determine that in all of the NPs in (16), the same two concepts—*overweight* and *child*—are combined by the same semantic relation, i.e., the relation characterized by and irrespective of the lexical instantiation of the individual concepts or the syntactic realization of the relation. This would allow the system to recognize the semantic similarity of apparently rather different strings.

ACKNOWLEDGMENTS

The work presented in this article builds on work carried out by a large group of researchers. The OntoQuery prototype has been developed at the University of Roskilde under the leadership of Troels Andreasen, whereas the nutrition ontology has been developed mainly by Nikolaj Oldager at the Danish Technical University. The language technology components of the OntoQuery prototype have been devloped at CST by Dorte Haltrup and the authors with contributions from Hanne Thomsen and Bodil Nistrup Madsen from the Copenhagen Business School. Finally, the formal logic used by the search engine is the work of Jørgen Fischer Nilsson from the Danish Technical University, whereas the ideas concerning a conceptual grammar have been developed by Jørgen Fischer Nilsson in cooperation with Per Anker Jensen, Southern Danish University, and Carl Vikner, Copenhagen Business School.

NOTES

1. Note that top-ontology concepts (taken over from the SIMPLE top-ontology) are in English, whereas all other concepts are in Danish.
2. In the QA track, systems had to match user queries with relevant passages no longer than either 50 or 250 bytes.
3. Relations are ordered in a hierarchy and prepositions are, in fact, typically mapped onto sums of relations due to their semantic ambiguity.

REFERENCES

Abney, S. 1996. Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI'96 Robust Parsing Workshop*. Available from: http://www.sfs.nphil.uni-tuebingen.de/~abney/.

Allan, J. 2000. Natural Language Processing for Information Retrieval. Tutorial at ANLP/NAACL 2000. Available from: http://ciir.cs.umass.edu./ (last accessed June 2000).

Andreasen, T., P.A. Jensen, J.F. Nilsson, P. Paggio, B.S. Pedersen, and H.E. Thomsen. 2002. Ontological extraction of content for text querying. In *Proceedings of the International Workshop on Applications on Natural Language to Information Systems*, Stockholm, Sweden.

Andreasen, T., J.F. Nilsson, and H.E. Thomsen. 2000. Ontology-based querying. In *Proceedings from Flexible Query Answering Systems, Advances in Soft Computing*, pages 15–26. New York, NY: Springer Verlag. Available from: http://www.springer.de/cgibin/search_book.pl?isbn = 3-7908-1347-8.

Brill, E. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics* 21(4): 543–565.

Callmeier, U. 2000. PET: A platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering* 6(1): 99–107.

Ciravegna, F., and A. Lavelli. 1999. Full text parsing using cascades of rules: an information extraction perspective. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 102–109, Bergen, Norway.

Copestake, A. 1999. *The (New) LKB System: Version 5.2*, CSLI, Stanford. Available from: http://www.csli.stanford.edu/ ~ aac/lkb.html.

Kokkinakis, D. 2000. PP-attachment disambiguation for Swedish: Combining unsupervised and supervised training data. *Nordic Journal of Linguistics* 23(2): 191–213.

Lenci, A., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, and A. Zampolli. 2000. SIMPLE: A general framework for the development of multilingual lexicons. *International Journal of Linguistics* 13(4): 249–263.

Miller, G., R. Bechwith, C. Feldbaum, D. Gross, and K. J. Miller. 1990. An online lexical database. *International Journal of Lexicography* 3(4): 235–244.

Nilsson, J.F. 2001. A logico-algebraic framework for ontologies. In *Ontology-Based Interpretation of Noun Phrases*. P.A. Jensen and P. Skadhauge, eds. *Proceedings of the 1st International OntoQuery Workshop*, Syddnask Universitet, pages 89–103.

Nilsson, J, and N. Oldager. 2000. *Lidt om ontologier med skitse til ontologi om ernæringsviden*. Technical OntoQuery Report. Danish Technical University.

Pedersen, B.S., and B. Keson. 1999. SIMPLE: semantic information for multifunctional plurilingual lexicons: Some examples of Danish concrete nouns. In *Proceedings of SIGLEX 1999, ACL-Workshop*, pages 46–51, Maryland, USA.

Pedersen, B.S., and S. Nimb. 2000. Semantic encoding of Danish verbs in SIMPLE: Adapting a verb-framed model to a satellite-framed language. In *Proceedings from 2nd Conference on Language Resources and Evaluation, LREC*, pages 1405–1412, Athens, Greece.

Pedersen, B.S., and P. Paggio. 2002. *A Danish Semantic Lexicon and Its Application in Content-Based Querying*. Technical report. Center for Sprogteknologi, Copenhagen, to appear in CSTs Working Papers.

Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, MA: The MIT Press.

Srihari, R. and W. Li. 2000. A question answering system supported by information extraction. In *Proceedings of the 6th Applied Natural Language Processing Conference and of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, ACL*, pages 166–172, Seattle, Washington, USA.

Strzalkowski, T., F. Lin, and J. Perez-Carballo. 1998. Natural language information retrieval TREC-6 report. In *Proceedings of TREC-6*. Available from: http://trec.nist.gov/pubs/trec6/t6_proceedings.html (last accessed June 2000).

Voorhees, E.M., and D.M. Tice. 2000. The TREC-8 question answering track evaluation. In *Proceedings of TREC-8*. Available from: http://trec.nist.gov/pubs/trec8/t8_proceedings.html (last accessed June 2000).

Vossen, P. (ed.). 1999. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. The Netherlands: Kluwer Academic Publishers.