# The Cornetto Database: Architecture and User-Scenarios

Piek Vossen[1] Katja Hofmann[2] Maarten de Rijke[2] Erik Tjong Kim Sang[2] Koen Deschacht[3]
[1]Faculteit der Letteren, Vrije Universiteit van Amsterdam, De Boelelaan 1105,
1081 HV Amsterdam, The Netherlands, p.vossen@let.vu.nl
[2]ISLA, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam,
The Netherlands, {khofmann,mdr,erikt}@science.uva.nl
[3]Katholieke Universiteit Leuven, Tiensestraat 41, B-3000 Leuven, Belgium,
koen.deschacht@law.kuleuven.be

## ABSTRACT

We outline the architecture of a semantic database for Dutch, developed in the Cornetto project, and its possible usage in language technology and information access applications. The database combines the Dutch part of EuroWordNet with the Referentie Bestand Nederlands. The resulting database is also aligned with the English WordNet and with a formal ontology. As such it represents a unique database with rich semantic and combinatoric information. There are many ways in which this knowledge can be exploited. In this paper we give an overview of possible application areas in order to inspire future research based on the Cornetto database.

## Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic networks; I.2.7 [**Natural Language Processing**]

## General Terms

Languages

## Keywords

Semantic databases, NLP applications

## 1. INTRODUCTION

Cornetto is a two-year Stevin project (project number STE05039) in which a lexical semantic database is built that combines WordNet with Framenet information for Dutch. The combination of the two lexical resources will result in a much richer relational database that may improve natural language processing (NLP) technologies, such as word sense-disambiguation, and language-generation systems. In addition to merging the WordNet and Framenet information, the database is also mapped to a formal ontology to provide a more solid semantic backbone.

The database will be filled with data from the Dutch part of EuroWordNet [25] and the Referentie Bestand Nederlands [14]. The Dutch WordNet (DWN) is similar to the Princeton WordNet for English, and the Referentie Bestand (RBN) includes frame-like information as in FrameNet plus much more information on the combinatoric behaviour of word meanings.

An important aspect of combining the resources is the alignment of the semantic structures. In the case of RBN these are lexical units (LUs) and in the case of DWN these are synsets. Various heuristics have been developed to do an automatic alignment. Following automatic alignment of RBN and DWN, this initial version of the Cornetto database will be extended both automatically and manually.

The resulting data structure is stored in a database that keeps separate collections for lexical units (mainly derived from RBN), synsets (derived from DWN) and a formal ontology (SUMO/MILO plus extensions [19]). These 3 semantic resources represent different view points and layers of linguistic and conceptual information. The alignment of the view points is stored in a separate mapping table. The database is itself set up so that the formal semantic definition of meaning can be tightened for lexical units and synsets by exploiting the semantic framework of the ontology. At the same time, we want to maintain the flexibility to have a wide coverage for a complete lexicon and encode additional linguistic information. The resulting resource will be made available in the form of an XML database.

The Cornetto database provides a unique combination of semantic, formal semantic and combinatoric information. This provides many new ways of exploitation in NLP applications. In this paper we give a (non-comprehensive) overview of possible application areas. We hope to mention some interesting points of view that may inspire future research.

The remainder of the paper is organized as follows. In Section 2 we describe the architecture and contents of the Cornetto database and specify the strategies and results for automatically aligning the word meanings of the two database in Section 3. In Section 5 we summarize the unique characteristics of the database that make it interesting for language-technology. Section 6 describes a number of scenarios in which the Cornetto database can and/or will be put to use. We conclude in Section 7.

## 2. ARCHITECTURE OF THE DATABASE

Both DWN and RBN are semantically based lexical resources. RBN uses a traditional structure of form-meaning

*DIR 2007, March 28-29, Leuven, Belgium.*

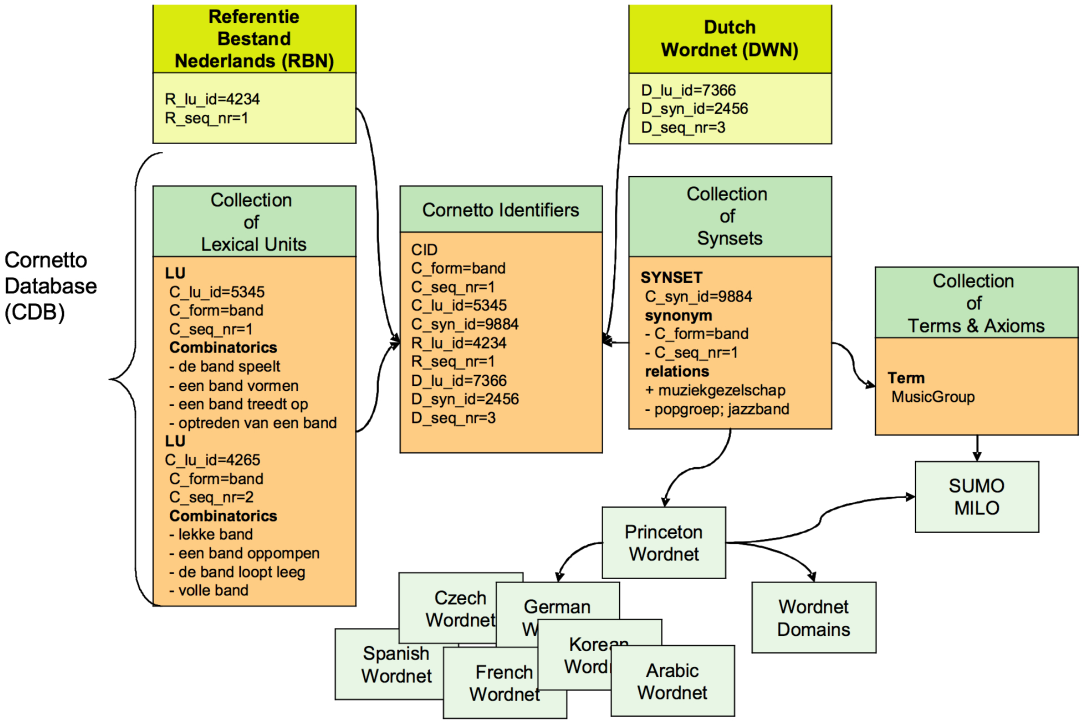

**Figure 1: Data collections in the Cornetto database.**

pairs, so-called Lexical Units [2].

The Cornetto database (CDB) consists of 3 main data collections:

1. Collection of Lexical Units, mainly derived from the RBN
2. Collection of Synsets, mainly derived from DWN
3. Collection of Terms and axioms, mainly derived from SUMO and MILO

The Lexical Units are word senses in the lexical semantic tradition. They contain all the necessary linguistic knowledge that is needed to properly use the word in a language. The Synsets are concepts as defined by Miller and Fellbaum [16] in a relational model of meaning. Synsets are mainly conceptual units strictly related to the lexicalization pattern of a language. Concepts are defined by lexical semantic relations. For the Cornetto database, the semantic relations from EuroWordNet [25] are taken as a starting point.

Outside the lexicon, an ontology will provide a third layer of meaning. The Terms in an ontology represent the distinct types in a formal representation of knowledge. Terms can be combined in a knowledge representation language to form expressions of axioms. In principle, meaning is defined in the ontology independently of language but according to the principles of logic. In the Cornetto database, the ontology represents an independent anchoring of the relational meaning in WordNet. The ontology is a formal framework that can be used to constrain and validate the implicit semantic statements of the lexical semantic structures, both the lexical units and the synsets. In addition, the ontology provides a mapping of a vocabulary to a formal representation that can be used to develop semantic web applications.

In addition to the 3 data collections, a separate table of so-called Cornetto Identifiers (CIDs) is provided. These identifiers contain the relations between the lexical units and the synsets in the CDB but also to the original word senses and synsets in the RBN and DWN.

Figure 1 shows an overview of the different data structures and their relations. The different data can be divided into 3 layers of resources, from top to bottom:

- The RBN and DWN (at the top): the original database from which the data are derived;
- The Cornetto database (CDB): the ultimate database that will be built;
- External resources: any other resource to which the CDB will be linked, such as the Princeton WordNet, wordnets through the Global WordNet Association, WordNet Domains, ontologies, corpora, etc.

The center of the CDB is formed by the table of CIDs. The CIDs tie together the separate collections of Lexical Units and Synsets but also represent the pointers to the word meaning and synsets in the original databases: RBN and DWN and their mapping relation.

Furthermore, the Lexical Units will contain semantic frame representations. The frame elements may have co-indexes with Synsets from WordNet and with Terms from the ontology. This means that any semantic constraint in the frame representation can directly be associated with the semantic information in the other collections. Any explicit semantic relation that is expressed through a frame structure in the Lexical Unit can also be represented as a conceptual semantic relation between Synsets in the WordNet database.

| **Strategy** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **Number of links** | 9,936 | 25,366 | 22,892 | 1,357 | 7,305 | 21,691 | 11,008 | 22,664 |
| **Average percentage correct scores** | 97.1 | 88.5 | 53.9 | 68.2 | 85.3 | 74.6 | 70.2 | 91.6 |

**Table 1: Number of links and average precision per strategy**

The Synsets in WordNet are represented as a collection of synonyms, where each synonym is directly related to a specific Lexical Unit. The conceptual relations between Synsets are backed-up by a mapping to the ontology. This can be in the form of an equivalence relation or a subsumption relation to a Term or an expression in a knowledge representation language.

Finally, a separate equivalence relation is provided to one ore more synsets in the Princeton WordNet.

## 3. ALIGNING RBN WITH DWN

To create the initial database, the word meanings in the Referentie Bestand Nederlands (RBN) and the Dutch part of EuroWordNet (DWN) have been automatically aligned. The word *koffie* for example has 2 word meanings in RBN (*drink* and *beans*) and 4 word meanings in DWN (*drink*, *bush*, *powder* and *beans*). This can result 4, 5, or 6 distinct meanings in the Cornetto database depending on the degree of matching across these meanings. This alignment is different from aligning WordNet synsets because RBN is not structured in synsets.

For measuring the match we used all the semantic information that was available. Since DWN originates from the Van Dale database VLIS, we could use the definitions and domain labels from that database. The domain labels from RBN and VLIS have been aligned separately by first cleaning up the labels manually (e.g., *pol* and *politiek* can be merged) and then measuring the overlap in vocabulary associated with each domain. Domain labels across DWN and RBN do not require an exact match. Instead, the scores of the correlation matrix can be used for associating them. For other features, such as part-of-speech, we manually defined the relations across the resources.

We only consider a possible match between words with the same orthographic form and the same part-of-speech. The strategies used to determine which word meanings can be aligned are:

1. The words have one meaning and no synonyms in both RBN and DWN
2. The words have one meaning in both RBN and DWN
3. The words have one meaning in RBN and more than one meaning in DWN
4. The word has one meaning in DWN and more in RBN
5. If the broader term (BT) of a set of words is linked, all words which are under that BT in the semantic hierarchy and which have the same form are linked
6. If some narrow term (NT) in the semantic hierarchy is related, siblings of that NT that have the same form are also linked.
7. Words that have a linked domain, are linked
8. Words with definitions in which one in every three words is the same (there must be more than one match) are linked.

Each of these heuristics will result in a score for all possible mappings between word meanings. In the case of *koffie*, we thus will have 8 possible matches. The number of links found per strategy is shown in Table 1. To weigh the heuristics, we manually evaluated each heuristics. Of the results of each strategy, a sample was made of 100 records. Each sample was checked by 8 persons (6 staff and 2 students). For each record, the wordform, part-of-speech and the definition was shown for both RBN and DWN (taken from VLIS). The testers had to determine whether the definitions described the same meaning of the word, or not. The results of the test were combined and the result was a list of percentages of items which were considered good links. The averages per strategy are shown in Table 1.

The minimal precision is 53.9 and the highest precision is 97.1. Fortunately, the low precision heuristics also have a low recall. On the basis of these results, the strategies were ranked: some were considered very good, some were considered average, and some were considered relatively poor. The ranking factors per strategy are:

- Strategies 1, 2 and 8 get factor 3
- Strategies 5, 6 and 7 get factor 2
- Strategies 3 and 4 get factor 1

The ranking factor is used to determine the score of a link. The score of the link is determined by the number of strategies which apply and the ranking factor of the strategies

In total, 136K linking records are stored in the Cornetto database. Within the database, only the highest scoring links are used to connect WordNet meanings to synsets. There are 58K top-scoring links, representing 41K word meanings. In total 47K different RBN word meanings were linked, and 48K different VLIS/DWN word meanings. 19K word meanings from RBN were not linked, as well as 59K word meanings from VLIS/DWN. Note that we considered here the complete VLIS database instead of the DWN selection only. VLIS synsets that are not part of DWN can still be useful for RBN, as long as they ultimately get connected to the synset hierarchy of DWN.

## 4. EXTENDING THE DATABASE

Apart from including in the Cornetto database a wide variety of lexical information of which the combination had not been available in a single resource before, we also aim at incorporating entries for words which were missing in all of the present resources. We will use language technology techniques for suggesting new phrases and their relations, and include these in the database after they have been checked by a trained lexicographer.

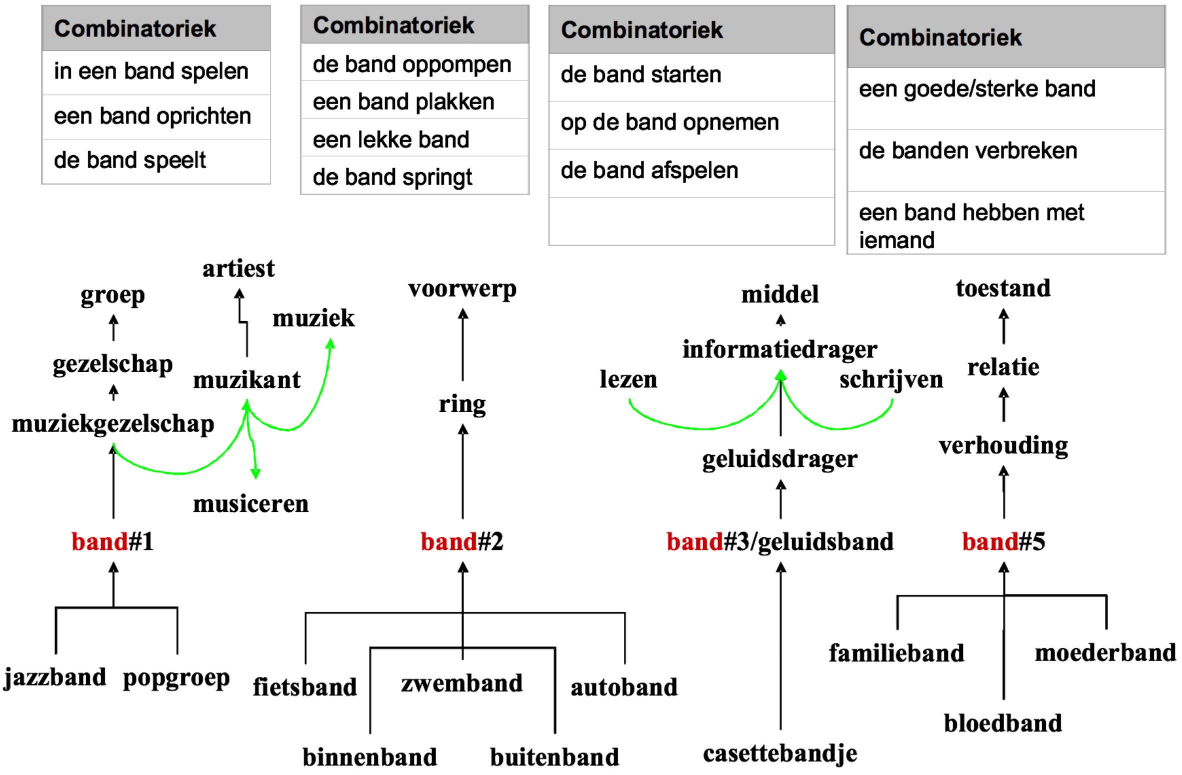

**Figure 2: Combinatorics and semantics combined**

We have evaluated three different techniques for proposing new elements for lexical relations. The first is a morphological method for predicting hypernyms: suggest the head of a compound noun as its hypernym: a *coffee cup* is a *cup*. The second method uses fixed lexical patterns for predicting hypernyms [9]: from a phrase like *animals such as goats*, we can derive that a *goat* is an *animal*. The third approach uses a combination of automatically derived lexical patterns for predicting relations from a text corpus [22].

We have evaluated the three approaches in an experiment in which they were applied for reconstructing the hypernymy tree of the Dutch WordNet. The morphological method performed best for this task: precision 54% and recall 33%. However, it is unclear how this technique can be used for deriving other relations. Fixed extraction patterns applied to the web was second best: precision 39% and recall 31%. Currently we use this approach in combination with the third method which extracts patterns from a text corpus based on a set of examples. This provides us with a flexible method which can be used for arbitrary relations and which can be applied to the largest available text corpus, the web.

## 5. LANGUAGE TECHNOLOGY FEATURES

The Cornetto database provides unique opportunities for innovative NLP applications. The Lexical Units contain combinatoric information and the synsets place these words within a semantic network. Figure 2 shows an example of this combination for several meanings of the word *band*: musical band, as a tube or tire filled with air, a magnetic band, and a relationship. The semantic network position of the word is depicted in separate WordNet fragments, relating the meanings to hypernyms, hyponyms and other related concepts. Above each fragment, we list the combinatoric information that is given in RBN for these different meanings. A musical band can be started and performs, a tube or tire can be inflated and fixed, can leak and explode, etc. Each of these examples illustrates a typical conceptual usage. Dutch speakers associate each of these examples with the correct meaning of the word *band*. These typical examples can be used for the disambiguation of occurrences in text. Moreover, the same contexts can also be used for other words related to these meanings. We can easily extend the examples of band as a tire/tube to the hyponyms *fietsband* and *autoband* and the examples of band as a relationship to the hypernym *verhouding* (affair) and *relatie* (relation).

Another example where combinatorics and semantic network relations are combined, relates to *drinks*. In Dutch, the preparation of drinks is usually referred to by the verb *maken* (to prepare). However, in the case of *koffie* (coffee) and *thee* (tea), another specific verb is used: *zetten*. So you typically use the phrase *koffie* *<u>zetten</u>* and *thee* *<u>zetten</u>* (to make coffee or tea) but you use the standard phrase *limonade maken* (to make lemonade) in Dutch. This example illustrates that conceptual combinations and constraints that are encoded in WordNet or in the ontology, do not explain the proper and most intuitive way of phrasing relations.

Finally, another important characteristics of the Cornetto database is its ontological basis. An ontology provides more fundamental distinctions between rigid and anti-rigid classes [8].

**rigidity** to what extent are properties of an entity true in all worlds? E.g., a person is always a "man" but may bear a Role like "student" only temporarily; "man" is a rigid property while "student" and "father" are anti-rigid.

**essence** which properties of entities are essential? For example, "shape" is an essential property of "vase" but not an essential property of the clay it is made of.

**unicity** which entities represent a whole and which entities are parts of these wholes? An "ocean" represents a whole but the "water" it contains does not.

These so-called identity criteria can be used to make a distinction between hyponyms in the Dutch WordNet that are roles and other hyponyms that represent disjunct types. Consider for example the hyponyms of *hond* (dog). The Dutch WordNet, like the English WordNet, lists as well dog-races such as *poedel* (poodle) and *Duitse herder* (German Shepherd) and dogs-in-roles such as *jachthond* (hunting dog) and *schoothond* (lapdog) as hyponyms. According to the OntoClean methodology the former are real types in the ontology but the latter are just dogs in a particular role. This means that a *poedel* (poodle) can never be a *Duitse herder* (German Shepherd) but any dog can become a *schoothond* (lapdog). This distinction is very important to guide expansion and co-reference of entities in texts.

# 6. USER-SCENARIOS

The Cornetto database will be useful for both text analysis and text generation, as well as for end-to-end applications. Text analysis is closely related to detecting word meaning in textual contexts but also to recognizing variation of reference, semantic entailments and applying simple reasoning. In most of these cases, the analysis is from text to concept, where we abstract from the surface form and represent information at the concept level.

In the case of text generation, we see the opposite. Information or implications need to be phrased properly. Since the Cornetto database does not only represent relations between concepts but also information on how to properly phrase these relations, we foresee that it can be very useful for summarization and translation, where lexical selection and phrasing play a major role.

The end-to-end applications that we believe will benefit from the Cornetto database are mostly concerned with information access, especially very focused forms of information access. In the next subsections, we will discuss some of the enabling technologies and applications that can benefit from this type of database.

## 6.1 Text Analysis: Word Sense Disambiguation

In the area of text analysis we envisage one short term application for the Cornetto database: word sense disambiguation. When using the Cornetto database in automated tasks, we face the problem of assigning the correct Lexical Unit from the database to the words in a given text (i.e., Word Sense Disambiguation, WSD). To develop a solution, we can start from techniques that are currently used for the assignment of WordNet synsets, for which the Senseval conferences provide a good benchmark.

Because of the limited amount of training data available to train disambiguation systems, research on weakly supervised or unsupervised techniques have recently gained importance. A common approach for these systems is to rely on the semantic relations which are defined in WordNet. For example, Carrol and McCarthy [1] use the synonymy relation to collect frequency data for every synset from untagged texts, while O'Hara et al. [20] use this relation to learn a Naive Bayes model for every synset, and Mihalcea and Faruque [15] use the hypernym/hyponym relation to determine a probabilistic model of senses which were not encountered in the training data.

In WordNet v2.1 the hypernym/hyponym relation organises the synsets in one hierarchical tree. Deschacht and Moens [3] have employed this fact to develop an efficient disambiguation system. For every level of the tree a probabilistic model, using Conditional Random Fields [13], is learned by training on the Semcor corpus [6]. Before assigning a synset to a given word, we traverse the tree from the top to that synset. This was done for every synset of the given word. The synset which has the most probable path was assigned. This is another approach to overcome the problem of limited training data.

By linking the RBN and the DWN, the Cornetto database offers a large collection of semantic relations and collocations which will both help in improving WSD systems. The collocations can be used as additional training data for a probabilistic model, which, enriched with the hypernym/hyponym relation of the DWN will result in a more accurate model (like in [20]). The semantic relations define constraints that significantly reduce the number of meanings of a word in a given context.

WSD systems are a necessary first step in many automated tasks using the Cornetto database. We will use them for instance in the applications described in the following paragraphs.

## 6.2 Text Generation

In the area of text generation we aim to use the Cornetto database in two domains: entity recognition in visual documents, and summarization.

### 6.2.1 Entity Recognition in Visual Documents

Since the advent of the Internet, an enormous amount of documents has been made available in electronic formats. The largest fraction of these documents is unstructured, such as texts, images and videos. The growth of available information is accompanied by a demand for more effective tools to search and work with this information. Moreover, there is a need to mine information from texts and images [21] when they contribute to decision making by governments, businesses and other institutions. The CLASS project is a project that aims to overcome some of these difficulties. It will develop learning methods that allow images, video and associated text to be automatically analysed and structured.

In the current state-of-the-art, object recognition in images is a difficult task to accurately perform. Therefore, there is an increasing interest in using the accompanying textual descriptions as a weak annotation of image content. Associated texts can for example include image-captions, video transcripts or web pages. Although an image and the associated text never contain precisely the same information, in many situations the text offers valuable information that helps to interpret the image.

In current approaches, the text that accompanies an image is seen as a bag of words, thus ignoring that the text's discourse structure and semantics allow for a more fine-grained identification of what content might be present in the image. Deschacht et al. [4] test the feasibility of automatically annotating images by using textual information in near-parallel image-text pairs, in which most of the content of the image corresponds to content of the text and vice versa. First, entities are classified according to a semantic database. They use the English WordNet, employing the WSD system from [3]. To identify proper names, the system uses a Named Entity Recognition system, which recognizes persons, locations and organizations.

What type of entities are likely to appear in an image? For instance, "a dog" can appear in an image, while "a thought" can only appear indirectly. We will call this measure visualness, which is defined as the extent to which an entity can be perceived visually. To determine this visualness, we have used a method that was inspired by Kamps and Marx [12]. They use a distance measure defined on the adjectives of the WordNet database together with two seed adjectives to determine the emotive or affective meaning of any given adjective. They compute the relative distance of the adjective to the seed synsets "good" and "bad" and use this distance to define a measure of affective meaning. We take a similar approach in order to determine the visualness of a given synset. We have defined a similarity measure between synsets in the WordNet database based upon the hypernym/hyponym relation and using the information content of the synsets. Next, we have selected a set of seed synsets, i.e. synsets with a predefined visualness, and use the similarity of a given synset to the seed synsets to determine the visualness of every synset. We have found that this approach gives excellent results.

This experiment shows a great advantage of a semantic database: although the database was never intended to hold information about the visualness of a concept, it was very easy to implement a method that extracts this information. The same holds for the affective information that was extracted from WordNet in [12]. This exemplifies the wide range of applications that can benefit from a semantic database, which defines rather abstract relations between concepts and lexical units, but which can be used to extract very concrete information.

### 6.2.2 Summarization

Another application that is very important with regard to the explosion of information on the web is automatic text summarization. It aims at condensing text to its essential content and assists in filtering and selecting this information.

One of the core problems in automatic text summarization is the identification of (near) duplicate content. Moens et al. [17] for example, detect the most important content in a text by analyzing the discourse structure and patterns of thematic progression. After this, duplicate content is identified by statistical techniques that cluster the lexical and syntactic features of sentences. Sentences similar in content are put in the same cluster and the most representative sentence (medoid) of each cluster is selected.

This method suffers from the fact that similar content can only be identified when similar lexical units are used to express this content. Because of the huge variation in human language, sentences with the same content frequently use the different lexical units. Identification could be improved if we could map the sentences to a semantic space, for example one that used the Cornetto identifiers as representations. In this space, identical content will have an identical representation, thus greatly simplifying the problem of finding such content.

| Dutch | English translation |
|---|---|
| vervoer | transportation |
| kilometerheffing | "road pricing" |
| mobiliteit | mobility |
| openbaar | public |
| werkverkeer | "work-related traffic" |
| forensenforfait | commuter allowance |
| ondertunneling | tunnelling |
| overkappingen | coverings |
| heffing | charge |
| auto | car |
| wegennet | road network |
| snelwegen | highways |
| motorrijtuigenbelasting | car tax |

**Table 2: Semi-automatically extracted characteristic terms for the theme "Traffic jams."**

Another problem in which the Cornetto database could play an important role is sentence-level paraphrasing. Paraphrasing is the process where the content of a sentence is re-formulated using different words. To guide this process, we could use the collocations of the RBN, which contains common usages for every Lexical Unit.

## 6.3 End-to-End Applications

Within the short time-frame of the project, we plan to apply the Cornetto database in three end-to-end applications: in the VerkiezingsKijker search engine, for theme and query expansion, in cross-lingual news classification, and in dialogue systems for open domains.

### 6.3.1 Theme and Query Expansion in VerkiezingsKijker

VerkiezingsKijker.nl [11, 23] is an electoral search engine that provides access to the party programs of the Dutch political parties ("what the parties promise"), news ("how do parties and programs figure in the media?"), and blogs ("what do people think about political issues?").

Among other things, the visitors of VerkiezingsKijker can search the party programs using a "Thematic search" facility. Users can explore a hierarchically organized list of close to 200 themes. Under the hood, each of these is represented using a number of terms, each of which has been semi-automatically identified. E.g., Table 2 displays the terms associated with the theme "traffic jams."

In the current implementation of VerkiezingsKijker these expansion terms are identified by purely statistical means (followed by a manual sanity check). VerkiezingsKijker used a search engine to find candidate paragraphs for each topic simply using the title of the theme as search query; a trained expert was then asked to provide relevance feedback: mark the returned paragraphs as relevant or not relevant for the topic. For each topic, using the paragraphs manually annotated as relevant, the creators of VerkiezingsKijker collected the 15 most overused terms as characterizing the topic. Over-usage of terms was determined using the log-likelihood statistical test [5]. For most topics, 5 relevant paragraphs were found in the top 20 hits, making the task relatively easy.

From Table 2 we see that few of the terms associated with the theme "traffic jam" are in a hierarchical relation with the theme. Given the architecture of the Cornetto database as outlined in the previous section, this raises the following question: can we use the combinatoric information in the database to support and inform the theme-search term associations? One of the planned applications of the Cornetto database in the VerkiezingsKijker setting, then, is to use the combinatoric information for query expansion, comparing (and possibly complementing) the effectiveness of the information in the database for theme expansion with the effectiveness of statistically derived information.

Our second planned application of the Cornetto database in the VerkiezingsKijker setting also concerns query expansion, but then at the front end. At 45 pages the average party program is too long to be returned as a reply to a user query, which is what motivated VerkiezingsKijker's creators to offer passage-based access to the programs. The programs of participating parties were split into a total of 4618 pseudo-documents, making recall a serious issue. Indeed, a key problem for the search approach is dealing with synonyms, different words that have the same meaning. Currently, the system can only retrieve text snippets which contain words that have been specified in the queries and their morphological variants, such as *psychology* and *psychologist.* However, a query for a word like *baby* will not retrieve all relevant texts that contain the word *infant.*

A resource like the Cornetto database would be very useful to the VerkiezingsKijker project. The database contains synonymy relations which could be used for query expansions which most likely would improve the overall recall of the system. Useful relations are not limited to synonymy. Query expansion would also benefit from having access to hypernymy relations (word vs. semantic class) and meronymy relations (part-whole) [24].

### 6.3.2 Cross-Lingual Classification

A further usage scenario of the Cornetto database is in developing a cross-lingual text classification system. Irion Technologies has developed a classification system that is based on the TwentyOne System [10]. The system can perform very satisfactorily when trained properly. For each of the classification labels, sufficient training documents should be collected. Recently, this has resulted in a classification system that can assign IPTC classification labels to Dutch documents. IPTC is a world-wide and standardized press thesaurus.

The drawback of the current system is that it needs to be trained with documents in every language to create classifiers. Experiments have been carried out to derive a classifier from the Dutch IPTC classification system by translating the index to English. The translation process generates a lot of noise, but this is usually not a problem as many statistical classifiers are robust against noise. Wrong translations of word combinations do not effect the classification task as long as they do not overlap with word co-occurrences for other categories.

We foresee that the acceptable performance of the current system can be further improved by making the translation of the indexes more precise. We plan to realize this by first applying word-sense-disambiguation and then expanding concepts to English. The Cornetto database will be used to detect the proper meanings of words within coherent collections of documents, grouped by the IPTC thesaurus. Grouping, for example, all documents on *football* together, offers many possibilities to determine the relevant meaning within a semantic domain. The word meanings are then expanded to English WordNet synsets through the mapping from the database to WordNet.

### 6.3.3 Dialogue Systems for Open Domains

Question Answering (Q&A) systems are gaining a lot of attention in the field. They deliver more precision and more focused results compared to open text retrieval systems and can still handle large quantities of unstructured data, although some deep-analysis may be necessary. Dialogue systems are traditionally seen as more complicated systems that can only operate on closed worlds and use well-defined reasoning based on explicit ontological knowledge. However, there is a new field of intermediate systems that combine the robustness of Q&A systems with the interactivity of dialogue systems.

So-called clarification-dialogues can be used to interact with users on the process of retrieving information. Within these interactions, it is possible to resolve many semantic issues, such as vagueness, ambiguity of both questions and results. Within this process, the Cornetto database can be very useful. First of all, it can be used to detect different interpretation possibilities with respect to the question and the possible answers. For example, the system will first detect the different meanings of a word such as *cell* in the document collection. For this it should have applied some type of word sense disambiguation to the document collection, using the Cornetto database. Next, it will determine the intended meaning of the same word or a synonym in the query, again using the database. If it cannot decide on the meaning in the query (using the dialogue history), it can prompt the user for the possible meanings that are relevant to the text, e.g., *cell* in the biological meaning and in the context of power supply, but neither in the context of jails nor mobile phones.

Secondly, the system can provide intelligent feedback even if it does not know the words in the question. Assuming that the information is formulated using more general words, e.g., rules that apply to *vehicles* or *cars*, and the question is formulated using more specific words, e.g., *my Mercedes* or *my convertible*, the system can still provide relevant information if it can infer from the Cornetto database that these query words match the words and concepts in the documents. The system can then reply in a careful way: *I did not find specific information on Mercedes but I do have information on cars.* In this case, the question is within the domain but the general database is needed to determine that the question is related to the indexed domain.

Finally, if real out-of-domain questions are asked, e.g., *Do you also have information on hotels in the area*, the system can use a general database such as Cornetto to infer that *hotels* are really different types of objects than *cars.* The system can then clearly reply: *We do not have information on hotels, only on cars.* Similarly, if words in the questions are really unknown, not only with respect to the index but also with respect to the Cornetto database, the system can rightly ask the user to explain the meaning of the word.

Dialogue systems can thus be made more natural and robust but also be more useful using a semantic database like Cornetto.

## 7. CONCLUSION

In this paper we have outlined the architecture of the Cornetto database and described possible usage scenarios. The database is currently being developed based on the existing lexical resources RBN and DWN. The information contained in these resources will be aligned with a formal ontology to provide a third structural layer in the database.

The Cornetto database will be a rich resource that combines lexical information in a unique way. This unique structure will allow future research in a variety of areas. We have described some of the possible application areas in text analysis and text generation. Furthermore, we have outlined a number of specific projects where the use of the Cornetto database is expected to have immediate benefits.

Additionally, as soon as the Cornetto database becomes available, we we will start adding new types of functionality to the database software itself, like providing users with related searches, using a variety of distance measures.

To conclude, with this overview of application areas we hope to have given ideas on how the unique structure of the Cornetto database could be used and to inspire future research and development.

## 8. ACKNOWLEDGMENTS

This research has been funded by the Netherlands Organisation for Scientific Research (NWO) via the STEVIN programme for stimulating language and speech technology in Flanders and The Netherlands. Maarten de Rijke was also supported NWO under project numbers 017.001.-190, 220-80-001, 264-70-050, 354-20-005, 600.065.120, 612-13-001, 612.000.106, 612.066.302, 612.069.006, 640.001.501, 640.002.501, and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

## REFERENCES

[1] J. Carrol and D. McCarthy. Word sense disambiguation using automatically acquired verbal preferences. *Computers and the Humanities*, 34 (1):109–114, 2000.

[2] D. Cruse. *Lexical semantics.* Cambridge, England: University Press, 1986.

[3] K. Deschacht and M.-F. Moens. Efficient hierarchical entity classification using conditional random fields. In *Proceedings of the 2nd Workshop on Ontology Learning and Population.* Sydney, 2006.

[4] K. Deschacht, M.-F. Moens, and W. Robeyns. Cross-media entity recognition in nearly parallel visual and textual documents. In *8th RIAO Conference on Large-Scale Semantic Access to Content (Text, Image, Video and Sound)*, 2007.

[5] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19 (1):61–74, 1993.

[6] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT Press, 1998.

[7] N. Guarino and C. Welty. Evaluating ontological decisions with ontoclean. *Communications of the ACM*, 45 (2):61–65, 2002.

[8] N. Guarino and C. Welty. Identity and subsumption. In R. Green, C. Bean, and S. Myaeng, editors, *The Semantics of Relationships: An Interdisciplinary Perspective.* Kluwer, 2002.

[9] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of ACL-92.* Newark, Delaware, USA, 1992.

[10] D. Hiemstra and W. Kraaij. Twenty-one in ad-hoc and clir. In E. Voorhees and D. K. Harman, editors, *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, pages 500–540. NIST Special Publication, 1998.

[11] V. Jijkoun, M. Marx, M. de Rijke, and F. van Waveren. Support for decision making: Electoral search. In *DIR 2007*, 2007.

[12] J. Kamps and M. Marx. Words with attitude. In *Proceedings of the 1st International Conference on Global WordNet*, pages 332–341. CIIL, Mysore India, 2002.

[13] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.

[14] I. Maks, W. Martin, and H. de Meerseman. *RBN Manual*, 1999.

[15] R. Mihalcea and E. Faruque. Minimally supervised word sense disambiguation for all words in open text. In *Proceedings of ACL/SIGLEX Senseval*, 2004.

[16] G. Miller and C. Fellbaum. Semantic networks of english. *Cognition*, October, 1991.

[17] M.-F. Moens, R. Angheluta, and J. Dumortier. Generic technologies for single- and multi-document summarization. *Information Processing & Management*, 41(3): 569–586, 2005.

[18] I. Niles and A. Pease. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, 2001.

[19] I. Niles and A. Pease. Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, pages 412–416, 2003.

[20] T. O'Hara, R. Bruce, J. Donner, and J. Wiebe. Class-based collocations for word-sense disambiguation. In *Proceedings of Senseval 3.* Barcelona, Spain, 2004.

[21] C. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Multimedia*, To appear.

[22] R. Snow, D. Jurafsky, and A. Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING/ACL 2006.* Sydney, Australia, 2006.

[23] VerkiezingsKijker. Electoral search engine, 2006. `http://verkiezingskijker.nl`.

[24] E. Voorhees. Using WordNet for text retrieval. In Fellbaum [6], pages 285–303.

[25] P. Vossen, editor. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages.* Kluwer, 1998.