The Development of a Complex-Structured Lexicon based on WordNet

Aleš Horák[1], Piek Vossen[2], and Adam Rambousek[1]

[1] Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno
Czech Republic
{hales,xrambous}@fi.muni.cz
[2] Faculteit der Letteren
Vrije Universiteit van Amsterdam
e Boelelaan 1105, 1081 HV Amsterdam
The Netherlands
Piek.Vossen@irion.nl

Abstract. The Cornetto project develops a new complex-structured lexicon for the Dutch language. The lexicon comprises information from two current electronic dictionaries – the Referentie Bestand Nederlands (RBN), which contains FrameNet-like structures, and the Dutch wordnet (DWN) with the usual wordnet structures. The Cornetto lexicon (stored in the Cornetto database) will be linked to English wordnet synsets and have detailed descriptions of lexical items in terms of morphologic, syntactic, combinatoric and semantic information. The database is organized in four data collections – lexical units, synsets, ontology terms and the Cornetto identifiers. The Cornetto identifiers are specifically used for managing the relations between lexical units on the one hand and synsets on the other hand. The mapping is first created automatically, but then revised manually by lexicographers. Special interfaces have been developed to compare the different perspectives of organizing concepts (lexical units versus synsets versus ontology terms).
In this article, we describe the background information about the Cornetto project and the implementation of necessary project tools that are based on the DEBVisDic tool for wordnet editing. The development of the Cornetto clients is a joint project of the Masaryk University in Brno and the University of Amsterdam.

Key words: Cornetto project; wordnet; DEB platform; DEBVisDic

1 Introduction

Cornetto is a two-year Stevin project (STE05039) in which a lexical semantic database is built that combines Wordnet with FrameNet-like 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-like 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 Wordnet [**?**] and the Referentie Bestand Nederlands [**?**]. 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 additional information on the combinatoric behaviour of words in a particular meaning.

Both DWN and RBN are semantically based lexical resources. RBN uses a traditional structure of form-meaning pairs, so-called Lexical Units [**?**]. Lexical Units contain all the necessary linguistic knowledge that is needed to properly use the word in a language. The Synsets are concepts as defined by [**?**] 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 Cornetto, the semantic relations from EuroWordNet are taken as a starting point [**?**].

Within the project, we try to clarify the relations between Lexical Units and Synsets, and between Synsets and an ontology. DEBVisDic is specifically adapted for this purpose.

In the next section we give a short overview of the structure of the database. The following sections give some background information on DEBVisDic and explain the specific adaptations and clients that have been developed to support the work of mapping the three resources.

2 The Cornetto Lexical Database

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

- Collection of Lexical Units, mainly derived from the RBN
- Collection of Synsets, mainly derived from DWN
- Collection of Terms and axioms, mainly derived from SUMO and MILO

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.

DWN was linked to WordNet 1.5. WordNet domains are mapped to WordNet 1.6 and SUMO is mapped to WordNet 2.0 (and most recently to WordNet 2.1). In order to apply the information from SUMO and WordNet domains to the synsets, we need to exploit the mapping tables between the different versions of Wordnet. We used the tables that have been developed for the MEANING project [**?**,**?**]. For each equivalence relation to WordNet 1.5, we consulted a table to find the corresponding WordNet 1.6 and WordNet 2.0 synsets, and via these we copied the mapped domains and SUMO terms to the Dutch synsets.

The structure for the Dutch synsets thus consists of:

Fig. 1. Cornetto Lexical Units, showing the preview and editing form

- a list of synonyms
- a list of language internal relations
- a list of equivalence relations to WordNet 1.5 and WordNet 2.0
- a list of domains, taken from WordNet domains
- a list of SUMO mappings, taken from the WordNet 2.0 SUMO mapping

The structure of the lexical units is fully based in the information in the RBN. The specific structure differs for each part of speech. At the highest level it contains:

- orthographic form
- morphology
- syntax
- semantics
- pragmatics
- examples

The above structure is defined for single word lexical units. A separate structure will be defined later in the project for multi-word units. It will take too much space to explain the full structure here. We refer to the Cornetto website [**?**] for more details.

3 The DEB Platform

The Dictionary Editor and Browser (DEB) platform [**?**,**?**] offers a development framework for any dictionary writing system application that needs to store the dictionary entries in the XML format structures. The most important property of the system is the *client-server* nature of all DEB applications. This provides the ability of distributed authoring teams to work fluently on one common data source. The actual development of applications within the DEB platform can be divided into the server part (the server side functionality) and the client part (graphical interfaces with only basic functionality). The server part is built from small parts, called *servlets*, which allow a modular composition of all services. The client applications communicate with servlets using the standard HTTP web protocol.

For the server data storage the current database backend is provided by the Berkeley DB XML [**?**], which is an open source native XML database providing XPath and XQuery access into a set of document containers.

The user interface, that forms the most important part of a client application, usually consists of a set of flexible forms that dynamically cooperate with the server parts. According to this requirement, DEB has adopted the concepts of the Mozilla Development Platform [**?**]. Firefox Web browser is one of the many

Fig. 2. Cornetto Synsets window, showing a preview and a hyperonymy tree

applications created using this platform. The Mozilla Cross Platform Engine provides a clear separation between application logic and definition, presentation and language-specific texts.

3.1 New DEB Features for the Cornetto Project

During the Cornetto project the nature of the Cornetto database structure has imposed the need of several features that were not present in the (still developing) DEB platform. The main new functionalities include:

- *entry locking* for concurrent editing. Editing of entries by distant users was already possible in DEB, however, the exclusivity in writing to the same dictionary item was not controlled by the server. The new functions offer the entry locking per user (called from the client application e.g. when entering the edit form). The list of all server locks is presented in the DEB administration interface allowing to handle the locks either manually or automatically on special events (logout, timeout, loading new entry, ...).
- *link display preview caching*. According to the database design that (correctly) handles all references with entity IDs, each operation, like structure entry preview or edit form display, runs possibly huge numbers (tens or hundreds) of extra database queries displaying text representations instead of the entity ID numbers. The drawback of this compact database model is in slowing down the query response time to seconds for one entry. To overcome this increase of the number of link queries, we have introduced the concept of *preview caching*. With this mechanism the server computes all kinds of previews in the time of saving a modified entry in special entry variables (either XML subtags or XML metadata). In the time of constructing the preview or edit form, the linked textual representations are taken from the preview caches instead of running extra queries to obtain the computed values.
- *edit form functionalities* – the lexicographic experts within the Cornetto project have suggested several new user interface functions that are profitable for other DEB-based projects like collapsing of parts of the edit form, entry merging and splitting functions or new kinds of automatic inter-dictionary queries, so called AutoLookUps.

All this added functionalities are directly applicable in any DEB application like DEBVisDic or DEBDict.

4 The New DEBVisDic Clients

Since one of the basic parts of the Cornetto database is the Dutch WordNet, we have decided to use DEBVisDic as the core for Cornetto client software. We have

Fig. 3. Cornetto Identifiers window, showing the edit form with several alternate mappings

developed four new modules, described in more details below. All the databases are linked together and also to external resources (Princeton English WordNet and SUMO ontology), thus every possible user action had to be very carefully analyzed and described.

During the several months of active development and extensive communication between Brno and Amsterdam, a lot of new features emerged in both server and client and many of these innovations were also introduced into the DEBVisDic software. This way, each user of this WordNet editor benefits from Cornetto project.

The user interface is the same as for all the DEBVisDic modules: upper part of the window is occupied by the query input line and the query result list and the lower part contains several tabs with different views of the selected entry. Searching for entries supports several query types – a basic one is to search for a word or its part, the result list may be limited by adding an exact sense number. For more complex queries users may search for any value of any XML element or attribute, even with a value taken from other dictionaries (the latter is used mainly by the software itself for automatic lookup queries).

The tabs in the lower part of the window are defined per dictionary type, but each dictionary contains at least a preview of an entry and a display of the entry XML structure. The entry preview is generated using XSLT templates, so it is very flexible and offers plenty of possibilities for entry representation.

4.1 Cornetto Lexical Units

The Cornetto foundation is formed by Lexical Units, so let us describe their client package first. Each entry contains complex information about morphology, syntax, semantics and pragmatics, and also lots of examples with complex substructure. Thus one of the important tasks was to design a preview to display everything needed by the lexicographers without the necessity to scroll a lot. The examples were moved to separate tab and only their short resumé stayed on the main preview tab.

Lexical units also contain semantic information from RBN that cannot be published freely because of licensing issues. Thus DEBVisDic here needs to differentiate the preview content based on the actual user's access rights.

The same ergonomic problem had to be resolved in the edit form. The whole form is divided to smaller groups of related fields (e.g. morphology) and it is possible to hide or display each group separately. By default, only the most important parts are displayed and the rest is hidden.

Another new feature developed for Cornetto is the option to split the edited entry. Basically, this function copies all content of edited entry to a new one. This

way, users may easily create two lexical units that differ only in some selected details.

Because of the links between all the data collections, every change in lexical units has to be propagated to Cornetto Synsets and Identifiers. For example, when deleting a lexical unit, the corresponding synonym has to be deleted from the synset dictionary.

4.2 Cornetto Synsets

Synsets are even more complex than lexical units, because they contain lots of links to different sources – links to lexical units, relations to other synsets, equivalence links to Princeton English WordNet, and links to the ontology.

Again, designing the user-friendly preview containing all the information was very important. Even here, we had to split the preview to two tabs – the first with the synonyms, domains, ontology, definition and short representation of internal relations, and the second with full information on each relation (both internal and external to English Wordnet). Each link in the preview is clickable and displays the selected entry in the corresponding dictionary window (for example, clicking on a synonym opens a lexical unit preview in the lexical unit window).

The synset window offers also a tree view representing a hypernym/hyponym tree. Since the hypero/hyponymic hierarchy in Wordnet forms not a simple tree but a directed graph, another tab provides the reversed tree displaying links in the opposite direction (this concept was introduced in the VisDic Wordnet editor). The tree view also contains information about each subtree's significance – like the number of direct hyponyms or the number of all the descendant synsets.

The synset edit form looks similar to the form in the lexical units window, with less important parts hidden by default. When adding or editing links, users may use the same queries as in dictionaries to find the right entry.

4.3 Cornetto Identifiers

The lexical units and synsets are linked together using the Cornetto Identifiers (CID). For each lexical unit, the automatic aligning software produced several mappings to different synsets (with different score values). At the very beginning, the most probable one was marked as the "selected" mapping.

In the course of work, users have several ways for confirming the automatic choice, choosing from other offered mapping, or creating an entirely new link. For example, a user can remove the incorrect synonym from a synset and the corresponding mapping will be marked as unselected in CID. Another option is to select one of the alternate mappings in the Cornetto Identifiers edit form. Of course, this action leads to an automatic update of synonyms.

The most convenient way to confirm or create links is to use *Map current LU to current Synset* function. This action can be run from any Cornetto client package, either by a keyboard shortcut or by clicking on the button. All the

required changes are checked and carried out on the server, so the client software does not need to worry about the actual actions necessary to link the lexical unit and the synset.

4.4 Cornetto Ontology

The Cornetto Ontology is based on SUMO and so is the client package. The ontology is used in synsets, as can be seen in the Figure **??**. The synset preview shows a list of ontology relations triplets – relation type, variable and variable or ontology term.

Clicking on the ontology term opens the term preview. A user can also browse the tree representing the ontology structure.

5 Conclusions

We have just presented the design and implementation of new tools for supporting the work on the Dutch Cornetto project developing a new complex structure lexicon. The tools are prepared on top of the DEB platform, which currently covers in six full featured dictionary writing systems (DEBDict, DEBVisDic, PRALED, DEB CPA, DEB TEDI and Cornetto). The Cornetto tools are closely related to the DEBVisDic system which, within the Cornetto project, has shown the versatility of its design as well as has been supplemented with new features reusable not only for work with other national wordnets but also for any other DEB application.

Acknowledgments

The Cornetto project is funded by the Nederlandse Taalunie and STEVIN. This work has also partly been supported by the Ministry of Education of the Czech Republic within the Center of basic research LC536 and in the Czech National Research Programme II project 2C06009.