Cornetto Tools and Methodology for Interlinking Lexical Units, Synsets and Ontology

Ales Horák[1], *Isa Maks*[2], *Adam Rambousek*[1], *Roxane Segers*[2] *and Hennie van der Vliet*[2], *Piek Vossen*[2]

[1] Faculty of Informatics, Masaryk University, Botanická 68a, 60200 Brno, Czech Republic

[2] Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands

hales@fi.muni.cz, emaks@let.vu.nl, xrambous@fi.muni.cz, roxane.segers@gmail.com, hd.vander.vliet@let.vu.nl, p.vossen@let.vu.nl

Cornetto is a two-year project, funded by the Flemish-Dutch Taalunie in the Stevin-programme (project number STE05039). It produces a lexical semantic database for Dutch. The database combines Wordnet (Fellbaum 1998) with FrameNet-like information. The data is derived from two existing lexical resources: the Dutch Wordnet (**DWN**, Vossen 1998) and the Referentie Bestand Nederlands (**RBN**, Maks, Martin and Meerseman 1999). These two resources represent two different perspectives on word meaning. Whereas DWN takes synsets as a starting point, RBN takes lexical units as a starting point. Lexical Units are word senses in the lexical semantic tradition. They contain all the necessary linguistic knowledge that is needed to properly use a word in a particular meaning in a language, among which semantic-syntactic frames. Synsets are concepts as defined in a relational model of meaning. They are mainly conceptual units strictly related to the lexicalized synonyms of a language. Synsets are further defined by lexical semantic relations to each other. Outside the lexicon, an ontology provides a third layer of meaning in the database. The Terms in the ontology represent disjoint Types, organized in a Type hierarchy. A Type represents a class of entities that share the same essential properties. Types are disjoint in the sense that their members (instances) belong to only a single type. Terms can be combined in a knowledge representation language to form expressions of axioms, for example the Knowledge Interchange Format or KIF (http://logic.stanford.edu/kif/dpans.html). The Cornetto database (CDB) thus consists of 3 separate data collections:

(1) Collection of Lexical Units (LUs), mainly derived from the RBN
(2) Collection of Synsets (SYs), mainly derived from DWN
(3) Collection of Terms (TEs) and axioms, SUMO and MILO (Niles and Pease 2001, Niles and Terry 2004)

In Cornetto, the ontology represents an independent anchoring of the relational meaning in the Dutch wordnet. The ontology is a formal framework that can be used to constraint 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. The differences in perspectives in these collections lead to different distinctions of concepts. For the relation between synsets and the ontology this can be illustrated with the example of *water*. In the English wordnet and SUMO, it is represented by two concepts: the chemical element *H2O* and *bodies of water*. In the Dutch wordnet there is an additional third meaning: *water in its liquid form* (the most common appearance of water in our world). Furthermore, there are many hyponyms below this liquid concept that refer to the different usages of the substance in our daily life (in Dutch realized as compounds): *theewater* (water used for making tea), *koffiewater* (water used for making coffee), *bluswater* (water used for extinguishing fire). Whereas the chemical element is a disjunct type in SUMO, the other meanings of water and their specific usages in the Dutch wordnet are not. In the Cornetto database, this can be expressed by a direct mapping of the noun *water* in Dutch to the element (being a straight name for the Type) and a complex contextual mapping of the other meanings to the same Type. The lexicalizations in Dutch thus do not grant the introduction if new Types in the ontology (Fellbaum and Vossen 2007.). Nouns for bodies of water are used to refer to instances of H2O occurring in certain quantities and locations. Nouns for usages are used to refer to quantities of the same H2O playing a temporarily role. These relations are expressed through simplified relational expression from the synsets to the ontology.

The Cornetto database system allows you to define the relations and constraints between these different perspectives. It is a tool to study and encode these relations in a unique way. For storing and editing this complex database, we used the Dictionary Editor and Browser platform (Horák, Pala, Rambousek and Rychlý et al 2006). 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. We have developed four new modules. 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 fromCornetto project.

Acknowledgements

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.

References

Chaudhri, A.B., A. Rashid, and R. Zicari (eds). 2002. XML Data Management: Native XML and XML-Enabled Database Systems. Addison Wesley Professional.

Fellbaum, C. (ed). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fellbaum C. and P. Vossen, 2007. "Connecting the Universal to the Specific: Towards the Global Grid", In: Proceedings of The First International Workshop on Intercultural Collaboration (IWIC 2007), Kyoto, Japan, January 25-26, 2007

Horák, A., K. Pala, A. Rambousek, and P. Rychlý. 2006. New clients for dictionary writing on the DEB platform. In DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writings Systems, pages 17–23, Italy, Lexical Computing Ltd., U.K.

Maks, I., W. Martin, and H. de Meerseman. 1999. *RBN Manual*, Vrije Universiteit Amsterdam.

Niles, I., and A. Pease. 2001. Towards a Standard Upper Ontology. In: *Proceedings of FOIS* 2001, Ogunquit, Maine, pp. 2-9.

Vossen, P. (ed). 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.