DanNet – a Wordnet Project for Danish

Bolette Sandford Pedersen, Sanni Nimb
Center for Sprogteknologi,
University of Copenhagen
Njalsgade 80
DK-2300 Copenhagen, Denmark,
bolette@cst.dk

Jørg Asmussen, Nicolai Hartvig Sørensen
Lars Trap-Jensen, Henrik Lorentzen
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 Copenhagen, Denmark,
ja@dsl.dk

Abstract

This paper describes a recently initiated wordnet project for Danish called DanNet. The project is a collaborate project between a university institution and a literary and linguistic society.

Introduction

In 2003 and 2004, two official Danish reports forcefully underlined the need for a lexical-semantic wordnet for Danish. The reports coincided with the conclusion of two Danish projects, a large corpus-based paper dictionary of modern Danish (Den Danske Ordbog, henceforth DDO) and a computational semantic lexicon for Danish comprising descriptions of 10,000 concepts in the so-called SIMPLE model (Semantic Information for Multifunctional, Plurilingual Lexica, cf. Lenci et al. (2001), Pedersen & Paggio (2004), henceforth SIMPLE-DK). Altogether, these contributions convinced The Danish Research Councils that the time had finally come to support the construction of a Danish WordNet project, DanNet, with a grant of DKK 3m.

1 Background

DanNet is a collaborative project between a research institution, Center for Sprogteknologi, University of Copenhagen, and a literary and linguistic society, Det Danske Sprog- og Litteraturselskab under The Danish Ministry of Culture.

The plan is to exploit the DDO and SIMPLE-DK resources by automatically extracting and thereafter manually adjusting the genus

proximum information as well as other central semantic relations such as part-of relations, purpose relations and antonymy relations. From SIMPLE the ontological affiliation of the concepts can also be semi-automatically transferred to DanNet, especially with regard to concrete entities where the ontological frameworks of EuroWordNet and SIMPLE are more or less directly compatible. Former studies of the semi-automatic construction of wordnets on the basis of existing dictionaries are not all entirely optimistic (cf. among others Rigau & Aguirre (2002) and Véronis & Ide (1994)). However, as our starting points are rather elaborate and pre-adjusted for a consistent taxonomic description of the vocabulary, we believe that a re-exploitation strategy is definitely worth pursuing. Being corpus-based DDO also largely reflects sense distinctions that can be substantially verified in the corpus. Asmussen (2004) describes a statistically based method for extracting not only genus proximum (which from the beginning has been encoded in a separate, extractable data field), but also differentiating features from the definition field.

2 Coverage

The project is planned to run from 2005 through 2007, in which period the plan is to achieve a wordnet of approx. 40,000 concepts. The EuroWordNet base concepts will provide the starting point for the encoding, but apart from these, frequency will play the most important role. Since DanNet will contain only a subset of DDO (which includes approx. 100,000 sense descriptions and collocations), frequency will constitute the central criterion for extraction. Senses in DDO with a frequency below a certain threshold will therefore be discarded unless they are considered to play an important conceptual role in the taxonomy. The final distribution on word classes is not determined beforehand, but obviously nouns will constitute the largest represented word class by far.

3 Encoding tool

When applying a strategy of substantial re-exploitation of existing lexicographical data, the structure of the encoding tools becomes a very important efficiency parameter. We are developing an encoding tool that directly incorporates and makes immediately available the data from our existing resources, DDO and SIMPLE-DK. This approach implies that we cannot directly take over existing wordnet encoding tools like VisDic. Instead, we are developing a user interface on a MySql database. We are experimenting with different extraction methods, e.g. statistical significance tests of the definitions provided in DDO such as mutual information, in order to determine the most helpful and most reliable ones.

4 Further research aspects

The project encompasses several research aims, the most central one being an examination of the ontological status of the vocabulary. First of all, with the experiments of the SIMPLE project behind us, the result of which contains *very* granular semantic information, we generally see the need to reconsider the wordnet framework in several respects especially regarding non-concrete concepts, i.e. 2^{nd} and

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 329–330.

3^{rd} order entities according to Lyons. Not surprisingly, definitions of concrete entities in DDO seem more or less directly reusable, whereas non-concrete entities, in particular abstract entities, are much harder – if at all possible – to organise semi-automatically in a consistent taxonomical structure. We also need to study further whether part of the template information from SIMPLE can be taken over in DanNet without conflicting with the general structure of EuroWordNet. In other words, we strive towards a more guided, template-driven encoding than what is presented in EuroWordNet (Vossen (2005)). For both models, SIMPLE and EuroWordNet, it is the case that 3^{rd} order entities are described in a rather coarse-grained and rudimentary fashion. It is an additional aim of DanNet to examine this ontological category further partly on the basis of the definitional inventory provided by the DDO definitions – which to a large extent reflect the usage of the concepts in context – and partly from a more theoretical, formal ontological perspective.

References

Asmussen, J. (2004). 'Feature Detection – A Tool for Unifying Dictionary Definitions'. In: *Proceedings from 11th Euralex International Congress*, pp. 63–69. Lorient, France.

DDO = Hjorth, E., Kristensen, K. et al. (eds.). (2003-2005). *Den Danske Ordbog 1-6* ('The Danish Dictionary 1-6'). Gyldendal & Society for Danish Language and Literature.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villages, M., Zampolli, A. (2000). 'SIMPLE – A General Framework for the Development of Multilingual Lexicons'. in: T. Fontenelle (ed.) *International Journal of Lexicography Vol 13*, pp. 249–263. Oxford University Press.

Pedersen, B., Paggio, P. (2004). 'The Danish SIMPLE Lexicon and its Application in Content-based Querying'. In: *Nordic Journal of Linguistics Vol 27:1:p97–127.*

Rigau, German & Eneko Aguirre. (2002). 'Semi-automatic Methods for WordNet Construction'. Tutorial presented at *2002 International WordNet Conference*, Mysore, India.

Véronis & Ide (1994). 'Knowledge Extraction from Machine-Readable Dictionaries: An Evaluation'. In Steffens, P. (ed.) *Machine Translation and the Lexicon,* pp. 19–34. Springer-Verlag.

Vossen, P. (ed.). (2005). EuroWordNet General Document. University of Amsterdam.