Swedish FrameNet++

Lars Borin, Dana Dannélls, Markus Forsberg,
Maria Toporowska Gronostaj, Dimitrios Kokkinakis

Språkbanken, Dept. of Swedish Language,
University of Gothenburg, Sweden
first.last@gu.se

Access to multi-layered lexical, grammatical and semantic information representing text content is a prerequisite for efficient automatic understanding and generation of natural language. A framenet built along the lines of the original English Berkeley FrameNet (see <http://framenet.icsi.berkeley.edu/ >> is considered a valuable resource for both linguistics and language technology research that may contribute to the achievement of these goals.

Currently, framenet-like resources exist for a few languages, including some domain-specific and multilingual initiatives (Dolbey et al., 2006; Boas, 2009; Uematsu et al., 2009), but are unavailable for most languages, including Swedish, although there have been some pilot studies exploring the semi-automatic acquisition of Swedish frames (Johansson and Nugues, 2006; Borin et al., 2007).

At the University of Gothenburg, we have recently embarked on a project to build a Swedish framenet-like resource. A novel feature of this project is that the Swedish framenet will be an integral part of a larger lexical resource containing much other lexical information in addition to the framenet part, including information relating to older stages of Swedish. Hence the name *Swedish FrameNet++* (SweFN++).

As a result of almost half a century of work on Swedish linguistic resources and Swedish lexicography, our research unit is the owner of a number of digital linguistic resources of various kinds – including both data and processing resources – with various degrees of coverage, and in various formats. When now starting the construction of a Swedish framenet, recycling as much as possible of the content of these hard-won resources will be a priority.

In addition, there are freely available suitable resources created elsewhere that can also be thrown into the pot. Below we describe briefly some of the existing lexical resources.

Resources at Gothenburg

Resources for modern Swedish

SALDO is the core lexicon of the SweFN++ to which all other information is to be merged. It provides morphological and lexical-semantic information on about 88,500 entries (senses expressed by single words or multi-word units). The lexicon is an updated version of *The Swedish Associative Thesaurus* (Lönngren, 1989) remade into a fully digital resource and enhanced by Borin and Forsberg (2009a).

The SIMPLE and PAROLE lexicons for Swedish are lexical resources aimed at language technology applications, results of the EU projects PAROLE (1996–1998) and SIMPLE (1998–2000) (Lenci et al., 2000). SIMPLE contains 8,500 semantic units being characterised with respect to semantic type, domain and selectional restrictions. All the items are also linked to the PAROLE lexicon, which contains 29,000 syntactic units representing syntactic valence information.

The Gothenburg Lexical Database (GLDB) is a lexical database for modern Swedish covering 61,000 entries with an extensive description of their inflection, morphology and semantics. SDB (Semantic Database) is a version of GLDB where many of the verb senses have been provided with semantic valence information using a set of about 40 general semantic roles (Järborg, 2001) and linked to example sentences in a corpus. One goal of the work presented here will be to find effective ways of correlating framenet frame elements with these general semantic roles.

Historical resources

Dalin's dictionary (appr. 63,000 entries) reflects the Swedish language of the 19th century (Dalin, 1853 1855). It has been digitized and published with a web search interface at Språkbanken.

It is currently being linked on the sense level to SALDO as part of an eScience collaboration with historians interested in using 19th century fiction as historical source material. A morphological analysis module for this historical language variety is also being developed as part of this effort.

Old Swedish dictionaries There are three major dictionaries of Old Swedish (1225–1526): (Söderwall, 1884) (23,000 entries), Söderwall supplement (Söderwall, 1953) (21,000 entries), and (Schlyter, 1887) (10,000 entries). All have been digitized by Språkbanken.

We have started the work on creating a morphological component for Old Swedish (Borin and Forsberg, 2009b), covering the regular paradigms and created a smaller lexicon with a couple of thousand entries.

Resources from outside sources

The People's Synonym Dictionary is the result of a collaborative effort where users of a Swedish-English online dictionary have been asked to judge the degree of synonymity of a word pair (randomly chosen from a large set of synonym candidates) on a scale from 0 (no synonymy) to 5 (complete synonyms). The downloadable version contains all word pairs with a rating in the interval 3 to 5, almost 40,000 Swedish synonym pairs. A Swedish-English dictionary – *Folkets lexikon* 'the People's Dictionary' – is now being constructed by the same method.

Swedish Wiktionary at present contains almost 60,500 entries (subdivided into senses). Notably, for each sense there is a free-text definition provided. Definitions are rare in other free lexical resources, which makes Swedish Wiktionary interesting for our purposes.

The Lund University frame list Johansson and Nugues (2006) have performed several experiments in attempt to create a Swedish framenet automatically. One of their experiments has resulted in list of 17,844 Swedish lemmas annotated with the English frames they evoke. The data was produced through parallel corpora with classification accuracy of 75%.

Merging lexical resources

The available lexical resources are heterogeneous as to their content and coding. The resources have been developed for different purposes by different groups with different backgrounds and assumptions, some by linguists, some by language technology researchers – possibly with little linguistic background or none at all – and yet others in Wikipedia-like collective efforts. Thus one of the main challenges for SweFN++ is to ensure content interoperability not only among the lexical resources but also between the available tools for text processing and lexical resources to be used by various pieces of software, and to formulate strategies for dealing with the uneven distribution of some types of information in the resource (e.g., syntactic valence information at present being available for about one fourth of the entries). This is work that we have initiated quite independently of the SweFN++ plans, within the European infrastructure initiative CLARIN (See <http://www.clarin.eu>).

We envision the end product of this work as a diachronic lexical resource for Swedish, to be used in developing language technology tools for dealing with text material from all periods of the written Swedish language, i.e., from the Middle Ages onwards. It remains to be seen how much this can apply the framenet part of the resource, but realistically, in addition to the modern language, at least 19th century Swedish may be covered.

The current state of the project can be viewed at the project homepage: <http://spraakbanken.gu.se/swefn>. The content of the page is automatically updated daily, hence reflecting the project as-is. At the time of writing, the Swedish framenet contained 113 frames with 5,961 lexical units.

References

Hans C. Boas, editor. 2009. *Multilingual Framenets in Computational Lexicography*. Mouton de Gruyter, Berlin.

Lars Borin and Markus Forsberg. 2009a. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, NEALT Proceedings Series, Vol. 4 (2009), Odense, Denmark. Kristiina Jokinen and Eckhard Bick.

Lars Borin and Markus Forsberg. 2009b. Something old, something new: A computational morphological description of Old Swedish. In *LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 9–16, Marrakech. ELRA.

Lars Borin, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2007. Medical frames as target and tool. In *FRAME 2007: Building Frame Semantics resources for Scandinavian and Baltic languages.*, pages 11–18, University of Tartu. Nodalida.

Anders Fredrik Dalin. 1853–1855. *Ordbok öfver svenska språket. Vol. I—II.* Stockholm.

Andrew Dolbey, Michael Ellsworth, and Jan Scheffczyk. 2006. BioFrameNet: A domain-specific FrameNet extension with links to biomedical ontologies. In *CEUR Workshop Proceedings*.

Jerker Järborg. 2001. Roller i Semantisk databas. Technical Report GU-ISS-01-3, Department of Swedish Language, University of Gothenburg.

Richard Johansson and Pierre Nugues. 2006. A FrameNet-based semantic role labeler for Swedish. In *Proceedings of Coling/ACL 2006*, Sydney. ACL.

Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. SIMPLE: A general framework for the development of multilingual lexicons. *Lexicography*, 13(4):249–263, December.

Lennart Lönngren. 1989. Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi. Rapport ucdl-r-89-1, Centrum för datorlingvistik, Uppsala universitet.

C.J. Schlyter. 1887. *Ordbok till Samlingen af Sweriges Gamla Lagar. (Saml. af Sweriges Gamla Lagar 13).* Lund, Sweden.

Knut Fredrik Söderwall. 1884. *Ordbok Öfver svenska medeltids-språket. Vol I–III.* Lund, Sweden.

Knut Fredrik Söderwall. 1953. *Ordbok Öfver svenska medeltids-språket. Supplement. Vol IV–V.* Lund, Sweden.

Sumire Uematsu, Jin D. Kim, and Jun'ichi Tsujii. 2009. Bridging the gap between domain-oriented and linguistically-oriented semantics. In *Proceedings of the BioNLP 2009 Workshop*, pages 162–170, Boulder, Colorado, USA. ACL.