Swedish FrameNet++

The beginning of the end and the end of the beginning

Malin Ahlberg, Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, Karin Friberg Heppin, Richard Johansson, Dimitrios Kokkinakis, Leif-Jöran Olsson, Jonatan Uppström

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

At SLTC four years ago, we presented the then newly launched SweFN++ project (Borin et al., 2010). Now the project is approaching the end of its funding period,[1] and the day of reckoning draws nigh.

Below, we give an account of the accomplishments of the SweFN++project in relation to its original goals, and describe our future plans for the resources developed in the project.

The stated main goal of the project was to create an open-content – i.e., freely available and modifiable – integrated lexical resource for Swedish (called Swedish FrameNet++) to be used as a basic infrastructural component in Swedish language technology (LT) research and in the development of LT applications. To accomplish this, we set up four lower-level objectives:

(1) to build a Swedish framenet (SweFN) covering at least 50,000 lexical units (LUs), on the same principles as the English Berkeley FrameNet (BFN) and to be developed in collaboration with the BFN team at ICSI Berkeley;

(2) to integrate a number of existing free lexical resources, by harmonizing and merging them, thereby reusing their valuable manually defined linguistic information;

(3) to develop a methodology and workflow which makes maximal use of LT and other tools in order to minimize the human effort needed to build SweFN++; and

(4) to use the SweFN++ resource, especially the new SweFN component, in concrete LT applications.

The beginning of the end: current status[2]

Swedish FrameNet

In October 2014 Swedish FrameNet had over 34,000 LUs contained in close to 1,200 frames, and is thereby the world's largest framenet in terms of number of LUs.[3] In addition to that mentioned above, SweFN contains analysis of compound patterns in terms of frame elements being instantiated within compounds. This is unique to SweFN. SweFN also contains around 50 frames which do not yet exist in other framenets. Several of these frames describe nominal concepts, others are more fine-grained equivalents of frames in BFN, and a few have been created due to linguistic or cultural differences (Friberg Heppin and Toporowska Gronostaj, 2014).

The integrated lexical resource

Resource integration has turned out to be a many-faceted problem. Its technical side has been easily implemented: SALDO PIDs (assigned persistent identifiers) are used for sense linking, if needed in connection with SKOS (simple knowledge organization system) relations in order to handle non-isomorphisms between resources (e.g. words or word senses in the historical lexicons which have no counterpart in the modern language).

It is easy to achieve on the order of 80% correct sense linkages between resources automatically, simply because of the Zipfian distribution of word senses over lemmas in any lexical resource (Borin, 2010; Borin et al., 2013a). Interlinking of the most polysemous lemmas, which are also the most frequent ones in text, turns out to be a much slower and more laborious process. Work is still ongoing on utilizing the structure of the resources themselves, e.g., determining which SALDO sense should be chosen for a polysemous lemma in a Bring thesaurus class (Borin et al., 2014) based on the semantic distances (as determined by the SALDO topology) of the alternatives to other, monosemous lemmas in the class.

The SweFN++ macroresource now contains, wholly or in part, the following component resources (see <http://spraakbanken.gu.se/research/swefn/publications> for references to publications describing them in more detail):

- SALDO
- Swedish FrameNet
- Swesaurus
- Core WordNet
- IDS/LWT lists

[1] SweFN++ is funded by the Swedish Research Council (2011–2013; dnr 2010-6013) and by the University of Gothenburg through a strategic research grant to the *Centre for Language Technology* (2009–2015).

[2] The current state of the project can be viewed at the project homepage: <http://spraakbanken.gu.se/swefn>. SweFN statistics are available at <http://spraakbanken.gu.se/eng/resource/swefn>. The publications generated so far by the project are listed at <http://spraakbanken.gu.se/research/swefn/publications>.

[3] The number of frames is on a par with BFN, while SweFN has far fewer annotated corpus examples than BFN.

- PAROLE
- SIMPLE
- Dalin's dictionary (19th c.)
- Old Swedish dictionaries
- Swedberg's dictionary (17th c.)
- Gothenburg Lexical Database
- the Lexin dictionaries
- Bring's thesaurus

For many of these, integration work is ongoing. Linking historical dictionaries to modern resources raises many intricate methodological problems (Andersson and Ahlberg, 2013; Ahlberg et al., 2014).

Tool and methodology development

Minimizing human effort needed to build SweFN++ requires advanced technical support and an efficient methodological approach. In this project we constructed Karp, the open lexical infrastructure (Borin et al., 2013b) to provide an adequate support to integrate, create and curate our modern and historical lexical resources. We adapted an expansion methodological approach which combines both manual and computational work to develop frames from BFN. Along with the improvement of Karp, we were able to switch to a more Swedish centered approach and enhance our lexical resources accordingly.

Karp combines 23 lexicon resources, uniquely organized around and interlinked to SALDO identifiers. It offers several search and editing functionalities to access lexical information from these interlinked resources. Lexical information may be accessed from Karp either through an interface or through webservices. The infrastructure is developed in parallel with Korp, the open corpus infrastructure of Språkbanken (Borin et al., 2012).

An essential tool in Karp is the SweFN editor used in the development and enhancement of SweFN. It integrates BFN and limited data from Korp. There are facilities to semi-automatically extract frames and frame information from BFN and select lexical units from SALDO, automatically extract sentences from Korp and manually select and annotate them for semantic structure. There is also support for adding useful information on the annotation of compounds, the domain or other information the developer wishes to emphasize regarding the Swedish language.

Use in applications

SweFN has been utilized in a semantic role labeling application making use of the semantically annotated sentences in the resource (Johansson et al., 2012). The system extracts the semantic roles of a given predicate within a given frame. The task is performed in two steps: *segmentation*, identifying the span of the semantic arguments and *labeling*, assigning semantic role lables to the given argument spans.

A major benefit of constructing a framenet that is formed on the basis of another, as in the case of SweFN and BFN, is the ability to build multilingual applications. This has been demonstrated in the automatic development of a large semantic grammar from valence patterns that are shared between the two framenets.[4] It has also been demonstrated in the development of multilingual natural language generation (NLG) applications such as tourist phrases and artwork descriptions (Dannélls and Gruzitis, 2014). A multilingual framenet approach to NLG has proven particularly relevant as the semantic and syntactic behavior of verbs vary depending on the target language, both in the constructions found and in their distribution.

Framenet resources may be very useful for studies in linguistics as well as for language studies. The online learning platform Lärka[5] for studens in linguistics and learners of Swedish as a second language uses SweFN as a resource for training semantic roles (Pilán and Volodina, 2014). As SweFN is built using the structure of BFN, but still having neccessary language specific solutions, it is an exellent resource for language teaching as it systematically demonstrates both similarities and differences between languages (Friberg Heppin and Friberg, 2012).

Some preliminary experiments have been conducted to investigate the possibilty of using semantic SweFN frames as variables in search queries. The searches were done using concepts rather than words as search keys and no prior annotation of the documents is neccessary (Friberg Heppin, 2013).

The end of the beginning: future prospects

There are already a number of ongoing fruitful interactions betwen the SweFN++ project and other projects in Språkbanken, e.g., *Digital areal linguistics*, *the Swedish Constructicon*, *Knowledge-based culturomics*, *MAÞiR*, and *KOALA*.[6] There are also a number of possible future applications and research envisaged using the Swedish FrameNet, such as enhancing automatic methods for semantic analysis of free text. One direction is the development of domain-specific framenet extensions, e.g. Dolbey et al. (2006), which can be used in possible application scenarios as means of achieving a higher level of text understanding (Fillmore and Baker, 2001). In this direction, event extraction can be seen as a representation of the semantic relationship between a frame-bearing lexical unit and its arguments (sentence constituents holding semantic roles) participating in the event. Event extraction can then be used to support deep knowledge acquisition and reasoning (Hogenboom et al., 2011). Developing better models for dealing with systematic polysemy is yet another non resolved area as well as addressing how to recognize and formalize Swedish verb meaning by finding and extracting syntactic and semantic patterns in text in line with the Corpus Pattern Analysis presented by Hanks (2013).

[4] <http://remu.grammaticalframework.org/framenet/>

[5] <http://spraakbanken.gu.se/larka/>

[6] <http://spraakbanken.gu.se/eng/research> and <http://spraakbanken.gu.se/eng/research/infrastructure>.

References

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of EACL 2014*, pages 569–578, Gothenburg. ACL.

Peter Andersson and Malin Ahlberg. 2013. Towards automatic tracking of lexical change: linking historical lexical resources. *NEALT Proceedings Series*, 18.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2010. Swedish FrameNet++. In *Swedish Language Technology Conference 2010*, Linköping.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, Istanbul. ELRA.

Lars Borin, Markus Forsberg, and Benjamin Lyngfelt. 2013a. Close encounters of the fifth kind: Some linguistic and computational aspects of the Swedish FrameNet++ project. *Veredas*, 17(1):28–43.

Lars Borin, Markus Forsberg, Leif-Jöran Olsson, Olof Olsson, and Jonatan Uppström. 2013b. The lexical editing sysem of Karp. In *Proceedings of the eLex 2913 Conference*, pages 503–516, Tallin.

Lars Borin, Jens Allwood, and Gerard de Melo. 2014. Bring vs. MTRoget: Evaluating automatic thesaurus translation. In *Proceedings of LREC 2014*, pages 2115–2121, Reykjavik. ELRA.

Lars Borin. 2010. Med Zipf mot framtiden – en integrerad lexikonresurs för svensk språkteknologi. *LexicoNordica*, 17:35–54.

Dana Dannélls and Normunds Gruzitis. 2014. Controlled natural language generation from a multilingual FrameNet-based grammar. In *Proceedings of the 4th Workshop on Controlled Natural Language (CNL 2014)*, volume 8625 of *Lecture Notes in Computer Science*, pages 155–166, Berlin. Springer.

Andrew Dolbey, Michael Ellsworth, and Jan Scheffczyk. 2006. BioFrameNet: A domain-specific FrameNet extension with links to biomedical ontologies. In *Proceedings of the Biomedical Ontology in Action Workshop at KR-MED*, pages 87–94.

Charles J. Fillmore and C. F. Baker. 2001. Frame semantics for text understanding. In *Proc. of the 2nd North American Chapter of the Assoc. for Computational Linguistics (NAACL)*, pages 1091–1094, Pittsburgh.

Karin Friberg Heppin and Håkan Friberg. 2012. Using FrameNet in communicative language teaching. In *Proceedings of the XV EURALEX International Congress*, Oslo.

Karin Friberg Heppin and Maria Toporowska Gronostaj. 2014. Exploiting FrameNet for Swedish: Mismatch? *Constructions and Frames*, 6(1):52–72.

Karin Friberg Heppin. 2013. Search using semantic framenet frames as variables. In *Proceedings of Sixth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2013), held at CIKM 2013 in San Francisco*, pages 25–28.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, Mass.

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. 2011. An overview of event extraction from text. In *Proc. of the Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE)*, pages 48–57, Bonn.

Richard Johansson, Karin Friberg Heppin, and Dimitrios Kokkinakis. 2012. Semantic role labeling with the Swedish FrameNet. In *Proceedings of LREC 2012*, pages 3697–3700, Istanbul.

Ildikó Pilán and Elena Volodina. 2014. Reusing Swedish FrameNet for training semantic roles. In *Proceedings of LREC 2014*, pages 1359–1363, Reykjavik. ELRA.