Thinking Green: Toward Swedish FrameNet++

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

Språkbanken, Dept. of Swedish Language,
University of Gothenburg, Sweden
first.last@svenska.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 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,[1] including some domain-specific and multilingual initiatives (Dolbey et al., 2006; Boas, 2009; Uematsu et al., 2009; Venturi 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 & Nugues, 2006; Borin et al., 2007).

At the University of Gothenburg, we are now embarking 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 many-faceted lexical resource. Hence the name *Swedish FrameNet*++ (SweFN++).

Toward a 'green' FrameNet

As a result of almost half a century of work on Swedish linguistic resources and Swedish lexicography, our research unit is the proud owner of a number of digital linguistic resources of various kinds – including both data and processing resources – with various stages of coverage, and in various formats. When now planning the construction of a Swedish FrameNet, thinking green should be the order of the day, i.e., recycling these resources

[1] See <http://framenet.icsi.berkeley.edu/>.

should be a priority.[2] In addition, there are freely available suitable resources created elsewhere that can also be thrown into the pot.[3] If by this the resulting resource can become something more than a FrameNet, so much the better. This is the 'plus-plus' aspect of SweFN++. Below we describe briefly some of the existing lexical resources.

Resources at Gothenburg

Resources for the modern language

SALDO is the core lexicon of the SweFN++ to which all other information is to be merged.[4] It provides morphological and lexical-semantic information on 76,750 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 words' inflection, morphology and meaning. 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.

[2] We would also expect the data to be of an overall higher quality – at a more appropriate level of abstraction – than if we were to rely exclusively on large-scale automatic processing of corpus data (cf. Green et al., 2004).

[3] It is important to our goals that the resources be freely available and modifiable, i.e., under an Open Source or Open Content license, since we plan to make the resulting resource available under this kind of license.

[4] See <http:///spraakbanken.gu.se/sal/eng/>.

The Lexin/Svenska ord dictionary (21,000 entries) is a source of both semantic and syntactic information; the latter includes valence patterns with some basic information on selectional restrictions of arguments.

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.[5] 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.[6]

We have started the work on creating a morphological component for Old Swedish (Borin & 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 synomymy) to 5 (complete synomyms). The downloadable version contains all word pairs with a rating in the interval 3 to 5, almost 40,000 Swedish synonym pairs.[7] A Swedish-English dictionary – *Folkets lexikon* 'the People's Dictionary' – is now being constructed by the same method.[8]

The Swedish WordNet (SWN) is a lexical resource structured according to the principles of the English Princeton WordNet.[9] It contains around

[5]See <http://spraakbanken.gu.se/dalin/>
[6]See <http://spraakbanken.gu.se/fsvldb/>.
[7]See <http://lexikon.nada.kth.se/synlex.html> (in Swedish).
[8]See <http://folkets-lexikon.csc.kth.se/om.html> (in Swedish).
[9]See <http://www.lingfil.uu.se/swordnet_test/swordnet.php>.

26,000 lexemes of which 4,165 verbs and 21,888 nouns. However, SWN is currently distributed under a closed source license which effectively makes impossible its inclusion in our project. Hopefully this will change.

Swedish Wiktionary at present contains almost 43,000 entries (subdivided into senses).[10] 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.[11]

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 the FrameNet part of the resource, but realistically, in addition to the modern language, at least 19th century Swedish may be covered.

[10] See <http://sv.wiktionary.org/>.
[11] See <http://www.clarin.eu>.

Methodology for FrameNet development

Developing a resource like FrameNet for a language is in itself a valuable undertaking. Most languages are underdescribed, as anyone can tell who has come across claims about her native language or a language that she knows well in works on language typology, where authors by necessity must rely on secondary sources such as grammars and dictionaries.

However, the activity itself of compiling a FrameNet for a new language can hardly anymore be characterized as pathbreaking research. From the point of view of linguistics, discovering new kinds of frames would certainly constitute a research contribution. From the point of view of language technology – our field of expertise – we believe that there is ample scope for methodological development, which may be of interest to the research community. There is a practical side to this as well: Even though we can draw upon highly qualified in-house lexicographical expertise, the resources – monetary and human – at our disposal are limited. Hence, we wish to find ways of conducting the work which will minimize human effort and focus it where it will be most useful. This will probably involve both arranging the workflow with respect to automated processing and human work and devising effective tools with good interfaces for the latter.

So far, we are aware of three approaches to semi-automatic frame and frame element (FE) acquisition for Swedish from corpora and other resources. These are:

1. cross-transfer of frames retrieved from parallel English Swedish corpora using automatic projection methods (Johansson & Nugues, 2005) (see also Pado & Lapata, 2005);
2. cross-transfer based on lexical units represented in the English frames and their possible equivalence in Swedish (Viberg, 2008);
3. direct acquisition from corpora that are semantically and syntactically annotated (Borin et al., 2007).

Awaiting a more thorough evaluation of these methods against the background of the available resources, we here offer a brief characterization of them, using as examples the verbs *gifta* 'to join in marriage (tr)' (the only regularly occuring form is the past participle *gift* 'married') and *gifta sig* 'to get married'.

Method 1: The Lund University frame list contains the following relevant entries:

gifta v: 5: Forming_relationships 4:Personal_relationship
gifta_sig v: 13: Forming_relationships
gifta_sig_med_sig v: 1: Forming_relationships

The first two correspond to entries in SALDO, and thus are plausible candidates. Next, we examine whether *Forming_relationships* and *Personal_relationship* are possible frames for the listed verb variants (the simplex and the reflexive verb). Access to a semantically and syntactically annotated corpus could provide answers to these questions. Some partial syntactic and semantic support can be also provided by data from the PAROLE and SIMPLE lexicons.

Method 2: A prerequisite for the lexical approach based on lexical transfer from English to Swedish is a digital bilingual dictionary supplying Swedish equivalents. The approach is far from simple as a single item can have many equivalents, as is the case with the verb *gifta* 'to join in marriage (tr)'. There are eight equivalents listed *förena, gifta bort, gifta sig, ingå äktenskap, gifta sig, äkta, viga, förena i äktenskap* in *The People's Dictionary*, which might be thought as potential lexical candidates for the frame *Forming_relationships*. From manual inspection of the equivalents it is clear that not all of them meet the semantic restrictions posed on the frame postulated for the verb *marry* as far as the roles of the core FEs are considered.

One of the disadvantages of this method is that it is biased toward creating a Swedish copy of the English FrameNet, rather than developing an original resource capturing the nature of the Swedish language. However, access to lists of semantically related words may be considered as helpful. The sets of semantically related Swedish words can be projected onto the Lund frame lists and follow the steps sketched above for method 1.

Method 3: This method aims at a direct acquisition of frames from Swedish corpora and relies on considerable linguistic pre-processing of corpora. Even if an extensively annotated corpus of general language is not available at this stage of the project, experiments with syntactic and semantic tagging of Swedish texts have indicated that this annotation can pinpoint relevant FEs and frames. Syntactic processing (Kokkinakis & Kokkinakis, 1999) is based on finite-state cascades applied to a pre-processed corpus annotated with part-of-speech and semantic labels. The parser is thus aware of shallow semantic annotations from both medical thesauruses (used in an earlier

pilot exercise) as well as a generic named entity recognition component. Semantic pre-processing results in improvement of accuracy for syntactic relation extraction (e.g. subject, object), which is relevant to the development of a FrameNet. This fact can be attributed to the decrease of coordination and structural ambiguity errors. Parsing accuracy is also enhanced by the fact that part-of-speech errors can be ignored since the semantic annotation have both higher priority, is more focused and has higher accuracy than the morphosyntactic one and has been extensively tested on general corpora. The semantic tags given by e.g. the partial parser, i.e., Human, Time, Place etc. correlate with the set of semantic types in FrameNet. However, this information needs to be subjected to further semantic interpretation and the semantic type is to be defined in terms of a particular semantic role. For example, in the case of the verb *marry*, this implies that the type Human needs to be narrowed to the semantic role Partner.

References

Boas, H. C. (Ed.). (2009). *Multilingual framenets in computational lexicography*. Berlin: Mouton de Gruyter.

Borin, L., & Forsberg, M. (2009a). All in the family: A comparison of SALDO and WordNet. In *Proceedings of the 17th Nordic conference of computational linguistics (NODALIDA 2009).* Odense, Denmark: Kristiina Jokinen and Eckhard Bick.

Borin, L., & Forsberg, M. (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)* (pp. 9–16). Marrakech: ELRA.

Borin, L., Gronostaj, M. T., & Kokkinakis., D. (2007). Medical frames as target and tool. In *Frame 2007: Building frame semantics resources for Scandinavian and Baltic languages.* (pp. 11–18). University of Tartu.

Dalin, A. F. (1853–1855). *Ordbok öfver svenska språket. Vol. I—II.* Stockholm.

Dolbey, A., Ellsworth, M., & Scheffczyk, J. (2006). BioFrameNet: A domain-specific FrameNet extension with links to biomedical ontologies. In *CEUR workshop proceedings.*

Green, R., Dorr, B. J., & Resnik, P. (2004). Inducing frame semantic verb classes from WordNet and LDOCE. In *Proceedings of the 42nd meeting of the ACL* (pp. 375–382). Barcelona: ACL.

Järborg, J. (2001). *Roller i Semantisk databas* (Tech. Rep. No. GU-ISS-01-3). University of Gothenburg: Department of Swedish Language.

Johansson, R., & Nugues, P. (2005). Using parallel corpora for automatic transfer of FrameNet annotation. In *Proceedings of the 1st ROMANCE FrameNet workshop* (pp. 26–28). Cluj-Napoca.

Johansson, R., & Nugues, P. (2006). A FrameNet-based semantic role labeler for Swedish. In *Proceedings of Coling/ACL 2006.* Sydney: ACL.

Kokkinakis, D., & Kokkinakis, S. J. (1999). A cascaded finite-state parser for syntactic analysis of swedish. In *Proc. of the 9th european chapter of the association of computational linguistics (EACL).* Bergen: ACL.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., et al. (2000, December). SIMPLE: A general framework for the development of multilingual lexicons. *Lexicography*, *13*(4), 249–263.

Lönngren, L. (1989). *Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi* (Rapport UCDL-R-89-1). Centrum för datorlingvistik, Uppsala universitet.

Pado, S., & Lapata, M. (2005). Cross-linguistic projection of role-semantic information. In *Proceedings of HLT/EMNLP 2005* (pp. 859–866). Vancouver: ACL.

Schlyter, C. (1887). *Ordbok till samlingen af sweriges gamla lagar. (saml. af sweriges gamla lagar 13)*. Lund, Sweden.

Söderwall, K. F. (1884). *Ordbok Öfver svenska medeltids-språket. Vol I–III*. Lund, Sweden.

Söderwall, K. F. (1953). *Ordbok Öfver svenska medeltids-språket. Supplement. Vol IV–V*. Lund, Sweden.

Uematsu, S., Kim, J. D., & Tsujii., J. (2009). Bridging the gap between domain-oriented and linguistically-oriented semantics. In *Proceedings of the BioNLP 2009 workshop* (pp. 162–170). Boulder, Colorado, USA: ACL.

Venturi, G., Lenci, A., Montemagni, S., Vecchi, E. M., Agnoloni, T., Sagri, M. T., et al. (2009). Towards a FrameNet resource for the legal domain. In *Processing of legal texts.*

Viberg, Å. (2008). RIDING, DRIVING and TRAVELING. Swedish verbs describing motion in a vehicle in crosslinguistic perspective. In J. Nivre, M. Dahllöf, & B. Megyesi (Eds.), *Festschrift in honor of Anna Sågvall Hein* (pp. 173–201). Uppsala: Acta Universitatis Upsaliensis. Studia Linguistica Upsaliensia 7.