Green resources in plain sight:
opening up the SweFN++ project

Markus Forsberg
Språkbanken, Department of Swedish
University of Gothenburg, Sweden
markus.forsberg@gu.se

Abstract

SweFN++ is a project focused on the creation and curation of Swedish lexical resources geared towards language technology applications. An important theme of the project is **openness** and its realization as a lexical infrastructure.

We give a short overview of the project, elaborate on what we mean by openness, and present the current state of the lexical infrastructure.

1 The SweFN++ project

SweFN++[1] (Borin et al., 2010a; Borin et al., 2009) is a project conducted at Språkbanken. The objectives of the project are twofold: the creation of a new lexical resource: a Swedish framenet covering at least 50,000 lexical units built on the same principles as the English Berkeley FrameNet; a curation and integration of existing free lexical resources, and thereby reusing the valuable grammatical and semantic information painstakingly collected in these resources.

The core resource to which all other resources are connected is SALDO[2] (Borin and Forsberg, 2009; Borin et al., 2008), a large, freely available lexicon with morphological and semantic information. What makes SALDO suitable as a core resource is partly because of its size, but also because its morphological and sense units have been assigned persistent identifiers (PIDs).

The lexical information of a resource is linked to the sense identifiers of SALDO, which often have the effect that the ambiguity of a resource is explicated: many of the resources associate lexical information to Part-of-Speech tagged headwords, an information that is not always valid for all the senses of the current headword. Another way of expressing this is that the resource contains information requiring human intuition to be understood completely, an undesirable property for a language technology resource.

The linking of all resources to a core resource gives us a "super lexical resource" with a diversity of lexical information. This diversity of information may be used to improve the quality of its parts. For example, the lexicon developed in the EU-project PAROLE (1996-1998) contains syntactic valency information that can be mirrored against the semantic valency information in Swedish framenet, where an inconsistency indicates an error in one of the two resources. We are currently working on a unified test bench for expressing these kinds of dependencies.

SweFN++ also includes historical lexical resources, i.e., it has a diachronic dimension (Borin et al., 2010b). The starting point of the diachronicity is four digitized paper dictionaries: one 19th century dictionary (Dalin, 1853), and three Old Swedish dictionaries (Schlyter, 1887; Söderwall, 1884; Söderwall, 1953).

For computational purposes we need to associate morphological information to the headwords of the dictionaries, a work that has been begun in the CONPLISIT project for 19th century Swedish (Borin et al., to appear) and in a pilot project for Old Swedish (Borin and Forsberg, 2008).

Linking SALDO's identifiers to the entries of Dalin is relatively straightforward because of the closeness of the language varieties. The vocabulary differences are mainly in the compounds, e.g., a word like *bäfverhund* 'dog used for beaver hunt' would not find its way in a modern lexicon since beaver hunt is no longer pursued in Sweden, even though the meaning is still relatively transparent. In cases like this we link to the head of the compound, i.e., for *bäfverhund* it would be *hund* 'dog'.

The work on linking Old Swedish to SALDO is a much more challenging task that we just have

[1] http://spraakbanken.gu.se/swefn
[2] http://spraakbanken.gu.se/saldo

started to think about. An illustrative example is the Old Swedish word *bakvaþi* meaning 'fatal accident resulting from a sword being struck backwards without the striker looking in that direction beforehand'. Naturally, there is no modern variant of this word, and it is an open, empirical question where it is most beneficial to link.

2 Openness

An important theme of the project is **openness**. The theme is a philosophical stance — we believe that research should be carried out in the open to enable scrutinization and increased collaboration. It is, from our point of view, more valuable that anyone is allowed to download and inspect unfinished work today, and, at the same time, run the risk that it is confused with something more mature, rather than taking the safer, but less productive, road of publishing the "finished" product at the end of the project.

The work on openness up until now may be summarized into four goals:

1. *To make resources and related information accessible as soon as possible, preferably at day one.*

A project such as this has it main activity during its project time. This rather obvious observation has the effect that to enable the research community to influence and contribute to the project, access to the resources and tools must be provided as soon as possible, preferably at day one.

2. *To deliver development versions of the resources, tools and related information regularly.*

This goal is related to the first one, since the input of others is only relevant if they have access to up-to-date information. We mentioned the research community, but openness is actually just as important to enable coworkers sitting just a couple of offices away to get involved. Instantaneous updates would be preferred, but for technical reasons we settle for daily updates.

3. *To deliver resources with an open content license, to use open standards for the resources, and to use and produce open source tools*

These are necessary requirements to enable someone to make good use of the resources or to continue the work that the SweFN++ project now started.

4. *To make the resources and tools available through web service APIs*

Web services are convenient ways of making resources and tools available computationally, since they enable instantaneous updates and offers a straight-forward and platform-independent way of including new lexical information into existing systems.

Web services still suffer from network latency; batch processing using web services is only feasible for small materials. On the other hand, the network speed has increased drastically the last few years, so this will probably not be an issue in a not-so-distant future.

3 Openness in practice

We have started the work on a lexical infrastructure to reach the aforementioned goals. The infrastructure has three essential nuts and bolts:

- a versioning system: Subversion[3]
- a content management system: Drupal[4]
- an XML database: eXist-db[5]

The versioning system with anonymous access is our delivery channel for the lexical resources. The use of a versioning system has the advantage that not only the latest version of a resource is available but all of its history. Not to mention the added value of using a versioning system in a collaborative environment such as a research project.

It is not only the resources that are published on a regular basis, but also a set of HTML files that give up-to-date information about such things as change history, test bench output, and statistics. The use of a content management system greatly simplifies the publication of these files.

Many of the resources are developed in CVS format, but are published as XML files[6]. These XML files are every night imported into an XML database. The XML database also has good support for creating web services for the resources, which simplifies the work.

We have developed a simple search interface on top of these web services in the content management system. The interface and the web services is referred to with the collective name *SBLEX*.

[3] `http://subversion.tigris.org/`
[4] `http://drupal.org/`
[5] `http://exist.sourceforge.net/`
[6] We aim for the LMF standard, but have not yet decided on how to best encode all lexical information in LMF.

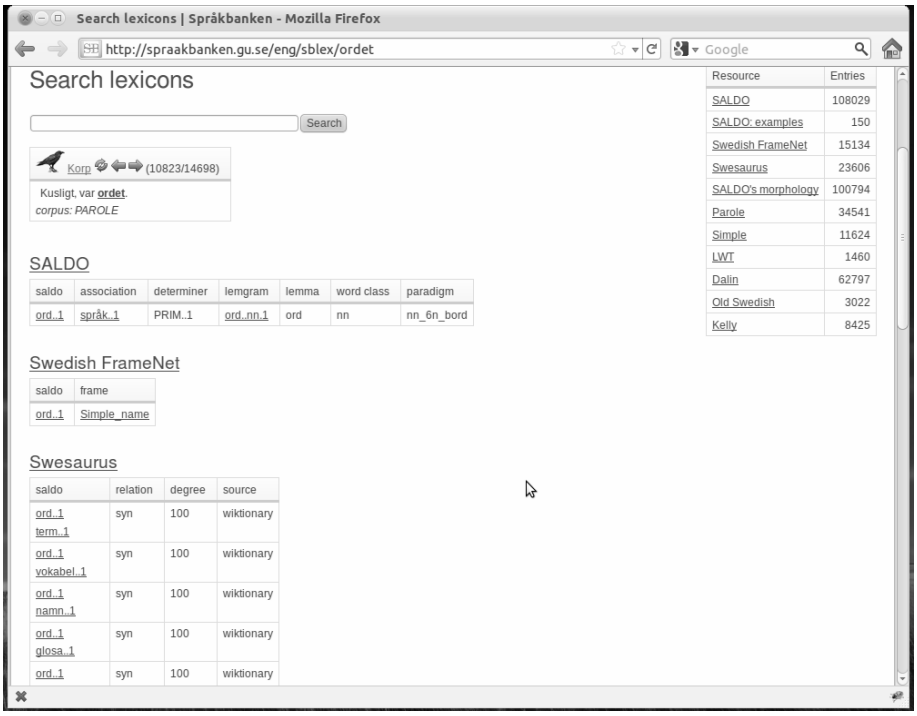

Figure 1: Searching for *ordet* 'the word' in SBLEX

Figure 1 shows a subset of the results when searching for *ordet* 'the word' in SBLEX. On the right hand side there is a table of the lexical resources in the system together with their number of entries. The first table is a random hit in our corpora material that has been annotated with SALDO identifiers, followed by information from the first three resources: SALDO, Swedish Framenet, and Swesaurus, a Swedish wordnet developed in the project.

Clicking on any of the resources in the table to the right moves us to the resource page, shown in Figure 2. All resources in SBLEX are downloadable from this page, together with XML schemata and CMDI metadata.

SBLEX is a generic system: adding a new resource requires only that the resource is added to the versioning system in a compatible format together with a few pieces of additional information such as localization.

The fact that SBLEX is generic is both a strength, since a new resource is added with ease, and a weakness, since when assuming little about the resources, it is hard to create a search interface pleasing to the eye. The result of a search is not presented in a unified manner: every resource is listed separately in a tabular format. The weakness can be remedied by creating another interface that sacrifices the function that a new resource becomes visible instantly for the benefit of a more aesthetic and logical presentation of the search results.

4 Final remarks

We have presented SweFN++, a project focused on the creation and curation of Swedish lexical resources, and discussed its theme of **openness** and its realization as a lexical infrastructure.

Openness implies that all members of the SweFN++ project work in plain sight. This can be quite disconcerting at first, but we have experienced nothing but positive effects: we feel that the work has improved in terms of quality and relevance, and that the general interest of the project has increased.

The lexical infrastructure still requires work, especially when it comes to unifying essential functions such as testing and statistics; functionalities that today are supported by a set of ad-hoc scripts for individual resources. In the context of testing we are also adding the functionality of expressing dependencies between different resources to detect inconsistencies and to generate suggestions for new entries.

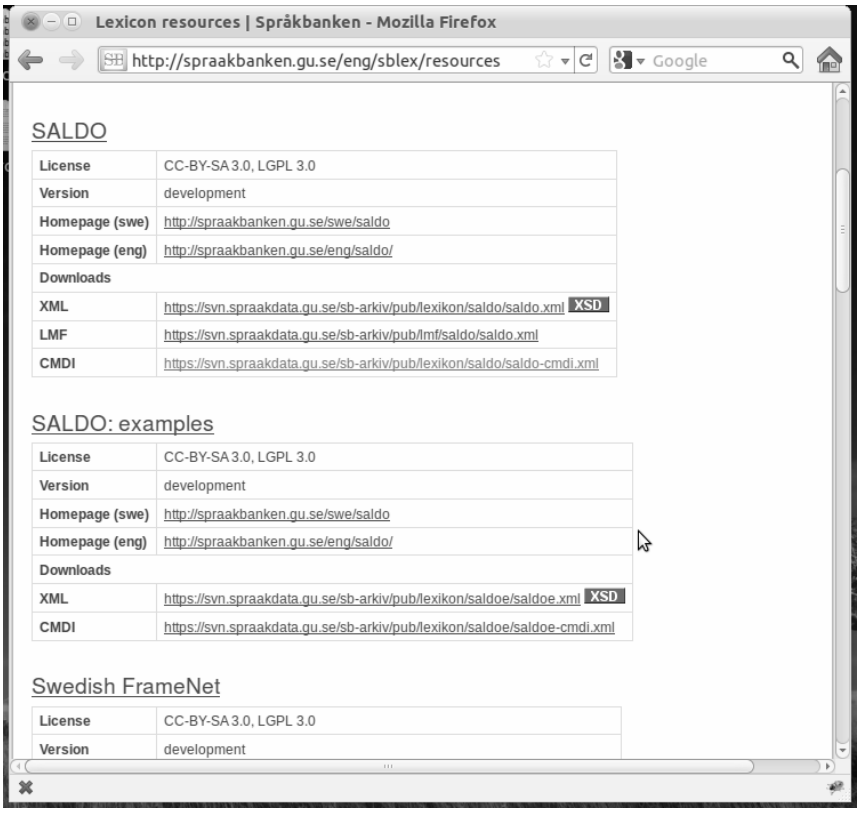

Figure 2: Download page for the resources

References

Lars Borin and Markus Forsberg. 2008. 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 and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. In Joakim Nivre, Mats Dahllöf, and Beata Megyesi, editors, *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, number 7 in Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia, pages 21–32. Uppsala University, Department of Linguistics and Philology, Uppsala.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2009. Thinking green: Toward swedish framenet++. In *FrameNet Masterclass and Workshop*.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2010a. The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*.

Lars Borin, Markus Forsberg, and Dimitrios Kokkinakis. 2010b. Diabase: Towards a diachronic blark in support of historical studies. In *Proceedings of LREC 2010*.

Lars Borin, Markus Forsberg, and Christer Ahlberger. to appear. Semantic Search in Literature as an e-Humanities Research Tool: CONPLISIT – Consumption Patterns and Life-Style in 19th Century Swedish Literature. In *Proceedings of the Nodalida 2011*, Riga.

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

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.