The Bank of Swedish

Gellerstam, M., Cederholm, Y., Rasmark, T.

Språkbanken, Dept. for Swedish, Göteborg University Box 200, SE–405 30 Sweden {svemg,sveyc,svetr}@svenska.gu.se

Abstract

The Bank of Swedish is described: affiliation, organisation, linguistic resources and tools. A point is made of the close connection between lexical research and corpus data, the broad textual coverage from Modern Swedish to Old Swedish, the official status of the organisation and its connection to Göteborg University. The relation to the broader scope of the comprehensive Language Database of Swedish is discussed. A few current issues of the Bank of Swedish are presented: parallell corpora production, the construction of a Swedish morphology database and sense tagging of text corpora. Finally, the updating of the Bank of Swedish concordance system is mentioned.

The Bank of Swedish - a general background

The Bank of Swedish was established in 1975 as a national organisation with the purpose of collecting, processing and storing electronic texts for linguistic use. The organisation is financed by the Faculty of Arts at Göteborg University and has been developed within the Department of Swedish.

A comprehensive *corpus* material (some 75 million running words) has been collected from different types of texts: modern language (newspaper, fiction, official language etc.) but also historical texts from Old Swedish up to Strindberg and other 19th century writers. Today, these texts are available free of charge in concordance form on the Internet, used by some 20 000 users every month.

A corpus from the PAROLE project is annotated with *morphosyntactic tags* and available as an Internet option.

The Bank of Swedish has been heavily involved in lexicographic and lexicologic projects, some of which have resulted in standard Swedish dictionaries, notably the National Encyclopedia Dictionary (NEO 1995) and the Swedish Academy Glossary (SAOL 1998). The Bank of Swedish has also cooperated within The European Union in several corpus projects: PAROLE, a project developing and standardizing linguistic resources in a number of western European countries, the lexical continuation of the project (called SIMPLE) where semantic tags are being attached to a fixed lexical set of words, the AVENTINUS project developing a drug terminology database with terms in many languages and the linguistic resources network TELRI (Transeuropean Linguistic Resources Infrastructure) where a ftp centre of resources free of charge for academic purposes has been established.

Corpus building also comprises *alignment of texts* where a corpus is being built up and alignment service offered to faculty students, not least for the education of translators.

Lexical data are available over the net in the form of The Swedish Academy Dictionary, the big historical dictionary, not yet completed (SAOB 1893–; Cederholm 1996). Also, dictionaries of Old Swedish are available in the same way.

A *linguistic toolbox* is being developed which will include standard tools like tokenizers, lemmatizers, a morphosyntactic tagger and a parallell text aligner. Some of these tools are language independent.

The Bank of Swedish corpus resources in concordance form (tagged or untagged) are free of charge and available on the Internet. An internal service for aligning texts is being set up at the Faculty. Some lexical data are available today and there will be more to come if copyright problems can be solved. A payment system has been discussed but there is a strong wish from the users to make linguistic data available for free.

The Bank of Swedish as part of a comprehensive database of Swedish

The Bank of Swedish has been cooperating for a long time with the Swedish Academy, chiefly in the lexical field. Today, a more comprehensive Language Database of Swedish begins to emerge out of this cooperation, comprising lexical and grammatical resources together with corpus data (Allén forthc.). The lexical level includes a modern defining lexical database (60 000 entries), the Swedish Academy Glossary (the norm of Swedish spelling and inflection, 120 000 entries) and PAROLE/SIMPLE lexical Swedish data (30 000 entries). Other lexical sources of information are a morphology database and several historical dictionaries, the most elaborate being the database of the Swedish Academy Dictionary, containing about 450 000 entries... The language database also will include the recently published Swedish Academy Grammar (in four volumes) and a handbook of prosody.

Parallel corpora

The Bank of Swedish has been commissioned by the Göteborg University Faculty of Arts to coordinate the alignment of texts between Swedish and foreign languages of the faculty. The alignment is based on a model worked out by Ken Church and adapted for local needs by Pernilla Danielsson and Daniel Ridings (Danielsson & Ridings 1996).

Alignments have been worked out for English, German and French but Slavic languages will also be linked. The following (table 1) is an alignment of EU texts in Swedish and English version.

CORPORA

(corpora resources held by the Bank of Swedish

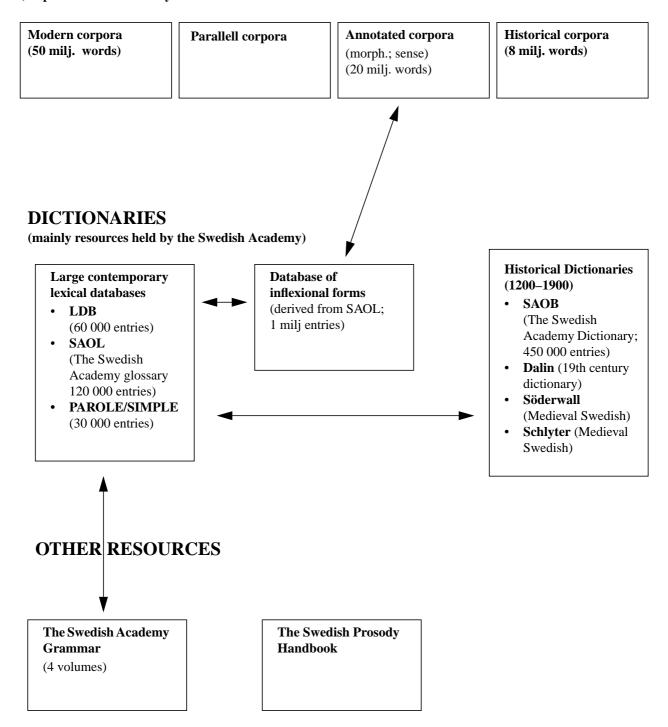


Figure 1: The Bank of Swedish as part of a comprehensive database of Swedish

Arbetsgruppen med inriktning på multmediala mjukvaror för utbildning skall titta närmare på språkinlärningsfrågorna.

För små och medelstora företag som har en internationell inriktning medför den nya tekniken stora fördelar

Tidigare åtgärder skall genomföras med särskild inriktning på de grupper som behöver speciell uppmärksamhet som till exempel ungdomar som söker sitt första arbete, långtidsarbetslösa och arbetslösa kvinnor.

Strukturfondernas inriktning kan komma att ändras mot bakgrund av detta, så att användningen av modern informations- och kommunikationsteknologi stimuleras.

Små och medelstora företag är en källa för att skapa arbetstillfällen, en källa till varierad inriktning av industriföretagen.

The task force addressing educational multimedia software will be looking into the language learning issues.

For the more globally oriented SMEs, the IS offers important advantages.

The above measures will be applied with particular emphasis on those categories requiring special attention, such as young people seeking their first job, the long-term unemployed and unemployed women.

There is a need to reorient the Structural Funds within the framework of the emerging IS to take account of this. Policies must be more oriented to stimulating the access and use of modern ICTs.

SMEs are a reservoir for the creation of jobs and a source of diversity in the industrial fabric.

Alignments will be used for purposes like contrastive language study, university teaching and language technology. From a theoretical point of view the aligned texts are invaluable for testing hypotheses about the relation between translations and the original. Hypotheses can be formulated about linguistic influence in translations, about the relation between "langue" and "parole", about translation strategies and style, typological research etc.

Today, the alignment procedure is as follows: people who want to have texts aligned prepare their texts according to a set of instructions (including automatic checking of spelling etc). After that, our alignment expert links the texts, the alignment is checked by the "customer", the program may be applied once more etc. At the same time, our experience will eventually result in an automatic device on the Internet for aligning texts: you send your texts to a specific website, the program tests the texts and returns them if they are not in accordance with the instructions. When everything is correct the alignment is carried out and you can collect the result.

The coordination of linking foreign languages to Swedish (and the other way round) will result in aligned corpora for many languages. The results will be published on the Internet and made available for research purposes.

An interactive procedure will also be carried out for other purposes, e.g. tagging of texts for various purposes, not only morphological tagging (or sense tagging) but also with tagsets for special purposes.

A Swedish Morphology Database

The Swedish Academy Glossary (12 edition) is the startingpoint for a morphological database of Swedish. The 120 000 words are defined according to inflection and word formation (mainly the shape of the juncture in compounds). The result is a morphological analyser and generator. The

database will be used to tag the texts of the Bank of Swedish. If you take the Swedish word *fisk* "fish", the generator will check the inflection class in the database (nr 12) and add the relevant inflectional forms to the word. The table also shows the connection between text and lexicon in that frequency can be added to the various inflectional forms (the tagset is the one used in the PAROLE project). In the case of the word *fisk*, all the inflectional forms happen to appear in the Bank of Swedish but this is rather an exception than a rule. Homographs will be tagged according to their possible lemmatic connection and the word will be disambiguated by aid of contextual signals. Words in the text that do not exist in the morphological database will be analysed with a special word formation analyser.

Table 2: Inflectional generation of Swedish FISK

form	morph.		SAOL-entry	class	POS-tag
fisk	fisk		fisk.1	12	NCUSNI
fisks	fisks		fisk.1	12	NCUSGI
fisken	fisk+en		fisk.1	12	NCUSND
fiskens	fisk+ens		fisk.1	12	NCUSGD
fiskar	fisk+ar		fisk.1	12	NCUPNI
fiskars	fisk+ars		fisk.1	12	NCUPGI
fiskarna	fisk+ar+na		fisk.1	12	NCUPND
fiskarnas	fisk+a	r+nas	fisk.1	12	NCUPGD
form	freq. POS-tag		tag	SAOL-entry	
fisk	1524	24 NCUSNI		fisk.1	
fisks	2	NCUSGI		fisk.1	
fisken	623	NCUS	SND	fisk.1	
		NCNF	PNI	fiske.1	l
fiskens	28	NCUS	SGD	fisk.1	
		NCNF	PGI	fiske.1	l
fiskar	502	NCUF	PNI	fisk.1	
		VMFF	PA	fiska.	1
fiskars	1	NCUF	PGI	fisk.1	
fiskarna	278	NCUF	PND	fisk.1	
		NCUF	PND	fiskar	e.1
fiskarnas	35	NCUPGD		fisk.1	
		NCUE	PGD	fiskar	e.1

Sense tagging of a Swedish corpus

A semantic link has been established between lexical data and corpora in a project called SEMTAG (Järborg 1999). The tagging has been made manually. The results are important from at least two points of view:

- as an extension of the lexical information on sense level. The textual examples provide a whole range of contexts, specifying the contextual behaviour of a specific sense.
- as a gold standard för automatic sense disambiguation.

Towards a new system

Today, a new system for the Bank of Swedish is being developed. In redefining the overall goals, adaptability is a major concern. The Bank of Swedish is moving towards a service generally available both to the University and to the public. Typically this involves post-graduate students deliverering corpora to be tagged or aligned for their personal use. Later on these resources are incorporated into the Bank for the benefit of other scientists.

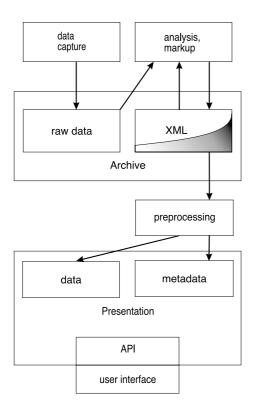


Figure 2: Configuration of a new corpora system

The flow of material is as follows (fig. 2). Text is captured and analysed linguistically and statistically. The standard formats used are XML and EAGLES standard annotation formats. Apart from the obvious need to initially store the raw text, the archive contains annotated corpora. Typically, a text is retrieved for additional markup several times. This demands from the presentation part to handle multiple lev-

els of description both as conditions for material selection (corpus definition) and as search criteria.

The XML material is preprocessed for incorporation in the presentation engine.

This involves separating the raw text from linguistic markup and text metadata. The markup and metadata is kept in a relational database

The presentation engine then integrates corpus positions retrieval from the database with the actual text retrieval (typically KWIC) from one or several plain files.

Apart from obvious enduser needs the user interface should enable developers to employ different methods for meta-data evaluation (like evaluating taggers, or trying different tagsets).

Preferably, the engine is deployed through a generalized application programming interface (API) which opens up for external users to define their own user interfaces.

The system architecture is modularized to both answer to these demands and to enable multileveled development.

Some of the design criteria:

- User-driven corpora definition. A corpus can be defined by any combination of criteria, like genre, sex, period etc. The only limit is the depth of tagging and meta-level description.
- Statistical treatment of the different corpora like comparison of frequency. This applies to data as well as metadata, and becomes metadata.
- Search based on lemmas (not only graphic words).
- Search based on morpho-syntactic, semantic etc. annotation.
- Comparison between annotated corpora (for evaluation of annotating methods).
- Search in aligned parallel corpora.
- The visual outcome of the presentation of texts will be enhanced by XML-dependent stylesheets.
- The search engine treats material of different tagging density in a uniform way.

The overall design goal can best be described as separating the analytical layer from the object of analysis.

References

Allén, Sture (forthcoming): The Nordic Languages. (In: Handbücher zur Sprach- und

 ${\it Kommunikations wissenschaft}, ed. \ Oskar \ Bandle)$

Cederholm, Y. 1996. A Historical Lexical Database of Swedish. The OSA Project. In *EURALEX* '96 *P*roceedings. pp. 65-72. Göteborg.

Danielsson, Pernilla & Ridings, Daniel (1996): *Parallell texts in Göteborg*. Research Reports from the Department of Swedish. Göteborg University: ISS 96:2

Järborg, Jerker (1999): *Lexikon i konfrontation*. Research reports from the Department of Swedish. Göteborg University: ISS: 99-6

Nationalencyklopediens ordbok. (NEO) 1995. Utarbetad vid Språkdata. Göteborgs universitet. Bra böckers förlag. Ordbok över svenska språket, utg. av Svenska Akademien. (SAOB) 1893–. Lund. (A–SVÄPA by 2000)

- Schlyter, C.J. 1877. *Ordbok till Samlingen af Sveriges gamla lagar*. Lund.
- Svenska Akademiens grammatik. 1–4. 1999. Norstedts ordbok.
- Svenska Akademiens ordlista (SAOL). 12 ed. 1998. Norstedts.
- Söderwall, K.F. 1884–1918. *Ordbok öfver Medeltids-språket. 1–2.* Suppl. 1925–1973 (Svenska Fornskrifts-Sällskapets Samlingar) Lund.