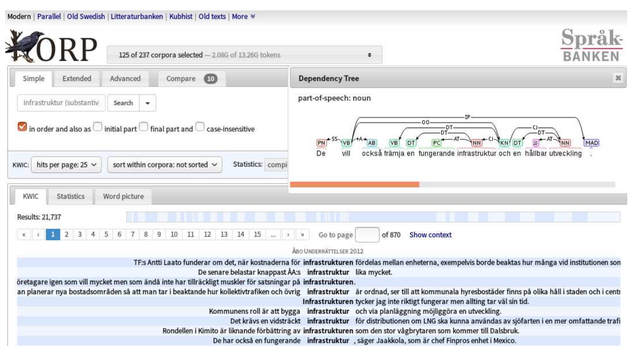
SPRÅKBANKENTEXT – a CLARIN B center at your service

The Språkbanken Text team
Språkbanken Text, Department of Swedish, University of Gothenburg sb-info@svenska.gu.se

SWE-CLARIN

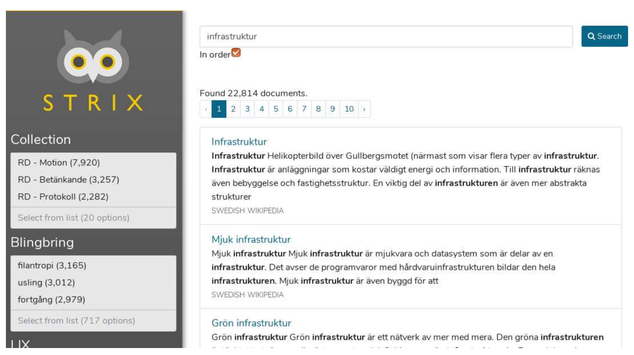
GÖTEBORGS UNIVERSITET

Our interfaces and tools



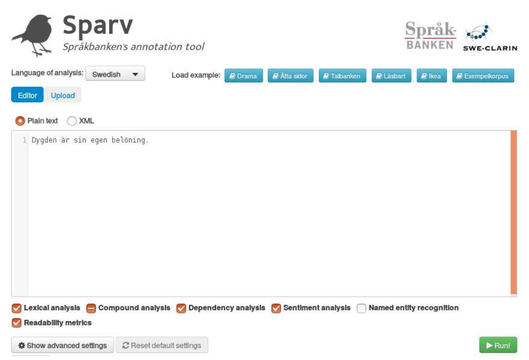
Korp – https://spraakbanken.gu.se/korp/

Korp is SB Text's oldest and most developed interface. It is mainly aimed at linguists, who can search about 15 billion words of contemporary and historical Swedish. Korp is open-source and has been deployed by a number of international groups (in Italy, Estonia, Denmark, Finland, Iceland and Norway) where it is used for an impressive array of languages.



Strix – https://spraakbanken.gu.se/strix

Strix is the newest kid on the block among SB Text's e-research tools. While Korp focuses on small linguistic entities such as words and sentences, the domain of Strix are *whole documents* and their *factual and other content* (rather than their linguistic form). Strix is thus expressly aimed at a broad community of humanities and social-science researchers.



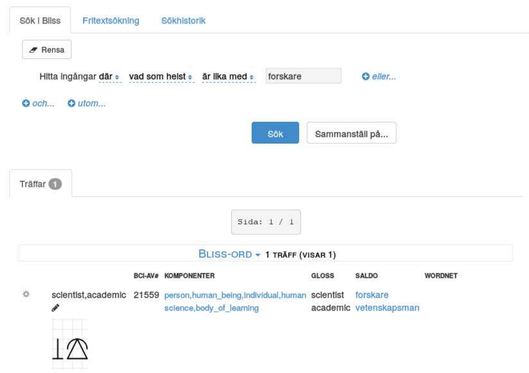
Sparv – https://spraakbanken.gu.se/sparv

Sparv provides an interface to the text import and annotation pipeline used by Korp and Strix. A user can upload their own texts and have them linguistically annotated and save the results for further offline processing. In the future, we plan to provide a facility for importing these texts into Korp and Strix as protected corpora.

CLARIN ERIC B centers

CLARIN's distributed network is made out of centres. The backbone of CLARIN is provided by technical centres, in particular *Service Providing Centres* or **CLARIN B-Centres**, for short. These units, often a university or an academic institute, offer the scientific community access to resources, services and knowledge on a sustainable basis. Therefore, there are strict criteria to become a CLARIN B-Centre: it should be based on a stable technical and institutional foundation. The Assessment Committee checks these requirements during an assessment procedure, while the technical coordination among the centres takes place in the Centre Committee.

(from https://www.clarin.eu/content/clarin-centres)



Karp – https://spraakbanken.gu.se/karp/

Karp is SB Text's environment for searching, browsing, editing and developing lexical resources and other formally structured linguistic datasets. The editing environment has been used to develop the Swedish FrameNet, as well as constructicons for Swedish and Russian. SB Text is currently working together with SB Sam (ISOF) on the development of a lexical editing environment for working with a number of Swedish minority languages.

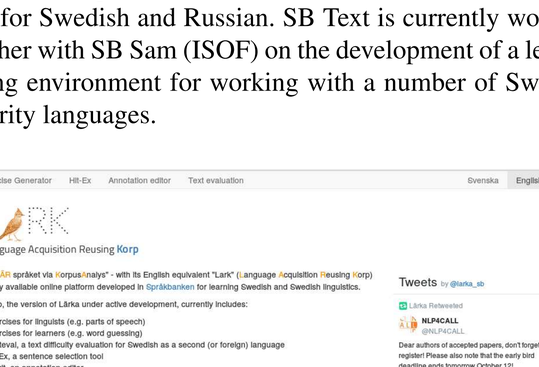
Our resources

Corpora

Most of our corpora are freely downloadable in a simple standarized XML format, which includes the full linguistic annotations and corpus metadata. For corpora where we do not own IPR, we provide scrambled "sentence sets", both through Korp and in the downloadable resources. The corpora represent many historical stages of Swedish and many different genres and text types, starting with the texts from the *Old Swedish Text Bank*, to vast amounts of modern social-media text. In addition, there are corpora of other languages, e.g. Somali and Faroese, and a number of parallel corpora.

Lexical resources

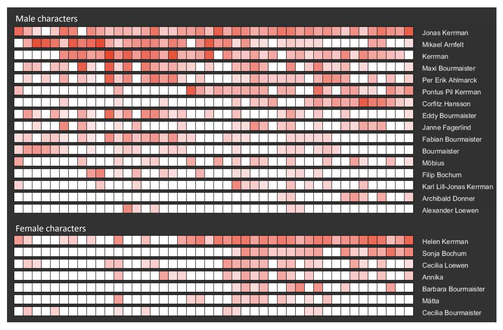
SB Text has a long history of developing and adapting *lexical resources* for use in language tools, all of which are freely available in their entirety under a CC-BY license. The modern lexical resources include SALDO, which is a large lexical-semantic network, a Swedish framenet, a Swedish sentiment lexicon, a Swedish Roget-style thesaurus, a small Swedish wordnet, a massively multilingual core vocabulary based on the word list of *Loanword Typology* project. There are also several historical lexicons, covering various stages of Swedish from the Old Swedish of the middle ages onwards.



Lärka – https://spraakbanken.gu.se/larka

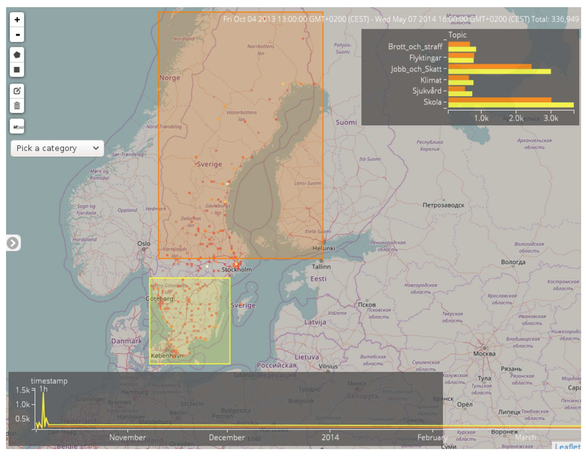
Lärka started out as a platform for automated corpus-based language and grammar exercises, but is now being developed into a tool for second-language learning research, allowing logging of carefully designed language learning exercises and systematic investigation of the effect of particular parameters of interaction and exercise design on learner progress.

Research that we support



Literary studies

Language technologies like *named entity recognition* can be applied to literary texts in order to generate "social networks" of the characters, or show the distribution of male and female characters over a literary text. Automated parsing and discourse analysis can uncover interactions among fictional characters, and also address the broader problem of investigating narrative structures. SB Text's language tools in combination with large digitized literature databases such as *Litteraturbanken* support sophisticated, language-structure aware versions of "distant reading".



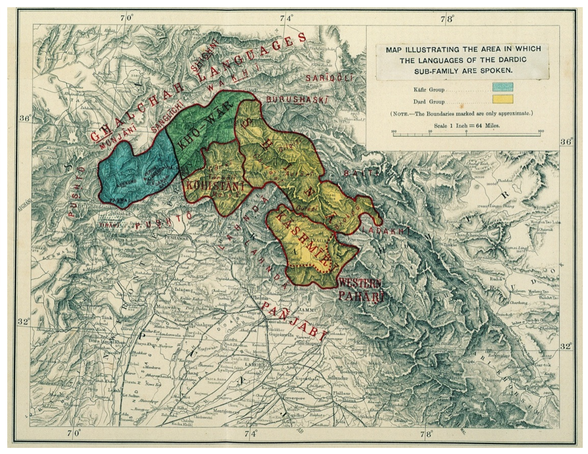
Public opinion in social media

There are many questions about public discourse in social media, about the demography and representativity of participation, whether the issues are the same as in traditional media, and whether public opinion formation processes now are fundamentally different as a result. Political scientists are eager to address these questions, but face the daunting challenge of dealing with the content of big and streaming textual data.

SKBL

Why is so little known of female artists, scientists, politicians, and so forth? This is one of the first questions which arises whenever any kind of history is about to be written. One of the reasons for this is that historiography relies on existing biographical dictionaries, most of which contain very few entries for women. The Biographical Dictionary of Swedish Women (SKBL) resolves this dilemma by providing free (CC-BY licensed) access to 1000 biographies of women who actively contributed to Swedish society.

(from https://skbl.se/en/about-skbl)



Large-scale comparative linguistics

A wealth of linguistic information relevant to large-scale typological and genealogical linguistic studies is laid down in traditional descriptive grammars. With the help of language technologies such as information extraction and parallel text alignment this information can be extracted automatically from digitized grammar texts and turned into formally structured linguistic databases.