Språkbanken's Open Lexical Infrastructure

**Malin Ahlberg, Lars Borin, Markus Forsberg,
Olof Olsson, Anne Schumacher, Jonatan Uppström**

Språkbanken
University of Gothenburg
{malin.ahlberg, lars.borin, markus.forsberg, olof.olsson2,
anne.schumacher, jonatan.uppstrom}@gu.se

Abstract

Karp is an open lexical infrastructure and a web based tool for searching, exploring and developing lexical resources. Språkbanken currently hosts a number of lexicons in Karp and on-going work aims at broadening the type of resources that can be developed in the system. This abstract gives a short overview of Karp's basic functionality, and describes some current projects and on-going work.

1. Overview

Karp, the open lexical infrastructure of Språkbanken (the Swedish Language Bank)[1], is under active development. As of today, there are more than 25, mostly Swedish, lexical resources available in Karp, including modern lexicons designed for LT use, as well as older digitized dictionaries. Most resources, including the historical ones, have been at least partially linked to a pivot resource, SALDO (Borin et al., 2013), defining a connected network of Swedish lexical information. There are also multi-lingual resources representing more than 30 languages (see Figure 3 on next page), as well as a lexicon for the ideographic writing system Bliss[2]. Karp is being developed in collaboration with Swe-Clarin[3], and we pay close attention to its standards and best practices.

The Karp system consists of three main components: a REST-based web service, a graphical user interface, and an authentication server for managing user access. A main function is to search in and query lexical data. The default setting in Karp is to search all lexicons, presenting all available information of any query grouped by resource. This may provide synergies, e.g. using a modern morphology to find entries in a historical resource. A lexicon chooser lets the user select the desired combination of resources to search in.

There are two ways of searching; the free text search function, which performs a search in all content of the selected lexicons, and the extended search function. The extended search may serve as a standard dictionary lookup, ie. looking up a keyword, but also lets the user construct more advanced queries, as in Fig. 1.

In addition to the search facility, Karp has also been designed to support the creation and development of lexical resources. Users can add, update and remove entries, and a revision history is kept for each resource. A resource may have a group of authorized editors, and the system also allows for unauthorized users to give suggestions that can later be approved by editors. In the editor mode, see Fig. 2, the different fields of the lexicon are presented. The graphical web interface provides user support during editing, such as feedback on the formatting, the compliance to a standard, or similar. The exact presentation can be specified in a configuration file for each lexicon since the kinds of data can differ considerably. A goal is to avoid the use of unstructured fields and informal conventions which are common obstacles for NLP use of lexicons. Karp has been used for lexicon creation since 2013, and the number of editable resources is steadily increasing. The editing functionality has been a central component in several projects, among them are the Swedish Framenet++ (Ahlberg et al., 2014) and the Swedish Constructicon (Lyngfelt et al., 2014). One current project features post processing of digitized resources, where the lexicographer manually proofreads a material and improves automatically marked-up data.

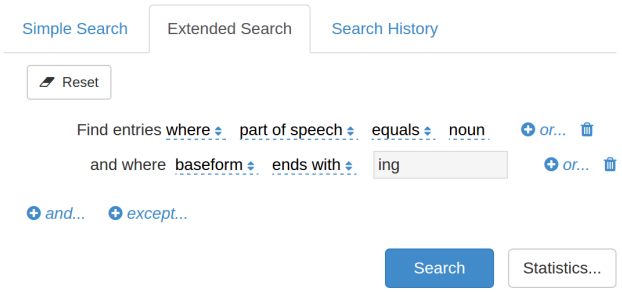

Figure 1: The extended search mode

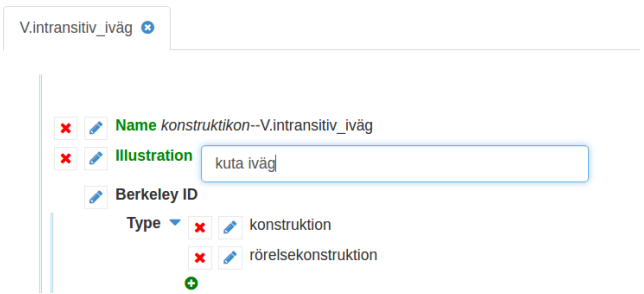

Figure 2: The editing interface

[1] http://spraakbanken.gu.se/karp#?lang=eng
[2] http://www.blissymbolics.org/
[3] https://sweclarin.se/

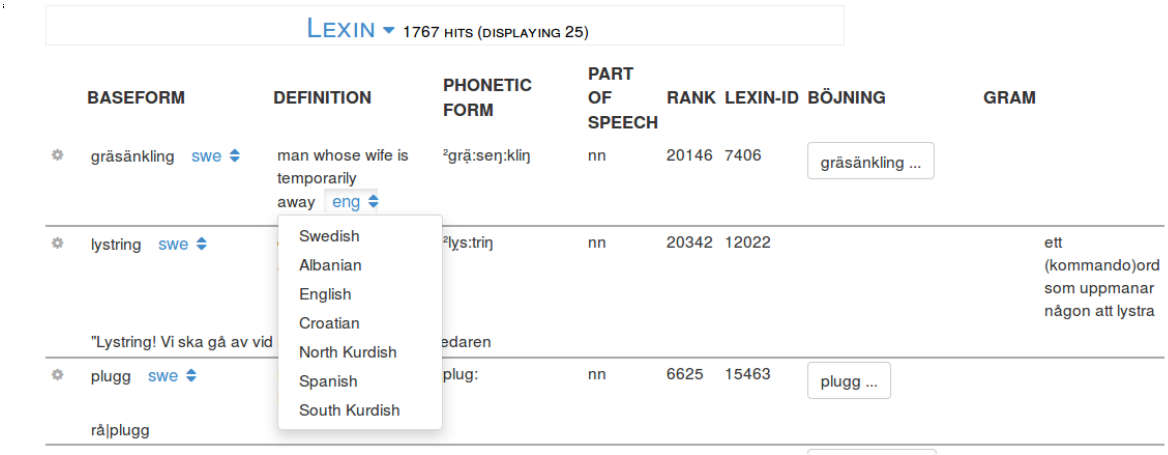

Figure 3: One of the multilingual lexicons in Karp

2. Standards and openness

Most lexical resources stored in Karp are exported every night to the Lexical Markup Framework format (LMF) and made downloadable from the homepage of Språkbanken[4]. LMF (Francopoulo et al., 2006) is an ISO standard published in 2008 that provides an intermediate format for lexical data exchange by combining designs and methods from many existing NLP lexicons. By using standardized data models, Språkbanken is actively contributing to improve the accessibility of its resources. The editing system in Karp publishes any updates instantly, which means that the lexical resources developed at Språkbanken are published from day one. This promote openness, since we believe that this is an important step towards increased scientific scrutiny and collaboration.

In accordance with our aims of openness, we strive to keep both the API and functionality generic and usable for other applications.

3. Karp in a wider perspective

Individual lexical resources accessible through Karp span a wide range of complexity, from simple bilingual word lists to the highly structured Swedish Framenet and Constructicon databases. In fact, we are increasingly seeing Karp as not only an infrastructure for lexical resources, but as a whole ecosystem for working with many kinds of structured data involving language as one component. A recently started project will utilize Karp for building an online biographical database of important Swedish women (SKBL[5]), and there are plans for developing a massively multilingual typological database on the basis of Karp, including not only lexical data but also structured grammatical features.

Another area of active development is to create tools allowing the editors to take advantage of the large amounts of linguistically annotated texts in Språkbanken's corpus infrastructure Korp (Borin et al., 2012). This is valuable for instance when annotating examples and writing sense definitions. It can further be used to compile statistics of genuine language usage, such as corpus frequencies of lexical entries, individual inflected forms or lemma co-occurrence statistics in dependency triples (e.g., nouns filling the subject slot of particular verbs).

4. Future plans

Current work focuses on other ways of exploring a resource. For traditional lexicons, this includes compilations of statistic lists and also interactive ways of exploring the data, moving between a macro perspective – the whole network of integrated lexicons – and the fine grained details. For some resources, such as the biographical database mentioned above, plotting geographical information, for instance birthplaces, on maps may give an additional understanding of the material.

Karp will also offer better support for lexicon modes as entry points for different user groups in order to enhance usability. Each mode will be optimized for one lexicon or a collection of lexicons with similar properties. The idea is that users can choose the mode that best suits their needs which will make the system more intuitive and swifter to use.

The support for multilingual lexicons will also be enhanced. Our current lexicons have language annotations at many different levels, ranging from form level to the lexicon as a whole. We need to provide search functionality fully supporting this and we also aim at developing the editing system to be functional and easy to use for multilingual resources.

The source code of Karp can be downloaded from Språkbanken's website[6] and is distributed under the MIT license.

References

Malin Ahlberg, Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, Karin

[4] www.spraakbanken.gu.se
[5] http://anslag.rj.se/en/fund/50409
[6] http://spraakbanken.gu.se/swe/forskning/infrastruktur/karp/distribution

Friberg Heppin, Richard Johansson, Dimitrios Kokkinakis, Leif-Jöran Olsson, and Jonatan Uppström. 2014. Swedish framenet++ the beginning of the end and the end of the beginning. http://www2.lingfil.uu.se/SLTC2014/abstracts/sltc2014_submission_33.pdf.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp - the corpus infrastructure of språkbanken. In *Proceedings of LREC 2012, ELRA*, pages 474–478.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. Saldo: a touch of yin to wordnet's yang. *Language resources and evaluation*, 47(4).

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Soria, et al. 2006. Lexical markup framework (LMF). In *Proceedings of LREC 2006, ELRA*, volume 6, pages 233–236.

Benjamin Lyngfelt, Lars Borin, Linnéa Bäckström, Markus Forsberg, Leif-Jöran Olsson, Julia Prentice, Rudolf Rydstedt, Emma Sköldberg, Sofia Tingsell, and Jonatan Uppström. 2014. Ett svenskt konstruktikon. Grammatik möter lexikon. In Nordica Helsingiensia 37, editor, *Svenskans beskrivning 33*, pages 268–279, Helsinki.