Karp: 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

1. Introduction

Karp is the open lexical infrastructure of Språkbanken (the Swedish Language Bank)[1]. As of today, there are 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, defining a connected network of Swedish lexical information. There are also multi-lingual resources representing more than 30 languages, 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.

Karp has been designed to support the creation and development of lexical resources. There are three main components: a REST-based web service, a graphical user interface, and an authentication server for managing user access. 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, but the system also allows for unauthorized users to give suggestions that can later be approved by editors. Karp provides user support during editing, such as feedback on the formatting, the compliance to a standard, or similar. The editing functionality in Karp 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).

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. Moreover, the lexical resources developed at Språkbanken are published from day one. This is done to promote openness, since we believe that this is an important step towards increased scientific scrutiny and collaboration.

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).

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

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

Keywords: Swedish, lexical resources, research infrastructure, linked open data

2. References

Ahlberg, M., Borin, L., Dannélls, D., Forsberg, M., Toporowska Gronostaj, M., Friberg Heppin, K., Johansson, R., Kokkinakis, D., Olsson, L.-J., and Uppström, J. (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.

Borin, L., Forsberg, M., and Roxendal, J. (2012). Korp - the corpus infrastructure of språkbanken. In *Proceedings of LREC 2012, ELRA*, pages 474–478.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C., et al. (2006). Lexical markup framework (LMF). In *Proceedings of LREC 2006, ELRA*, volume 6, pages 233–236.

Lyngfelt, B., Borin, L., Bäckström, L., Forsberg, M., Leif-JöranOlsson, Prentice, J., Rydstedt, R., Sköldberg, E., and Uppström, S. T. J. (2014). Ett svenskt konstruktikon. Grammatik möter lexikon. In 37, N. H., editor, *Svenskans beskrivning 33*, pages 268–279, Helsinki.

[1] http://spraakbanken.gu.se/karp#?lang=eng
[2] http://www.blissymbolics.org/
[3] https://sweclarin.se/
[4] www.spraakbanken.gu.se
[5] http://anslag.rj.se/en/fund/50409
[6] http://spraakbanken.gu.se/swe/forskning/infrastruktur/karp/distribution