Korp and Karp – a bestiary of language resources: the research infrastructure of Språkbanken

Malin Ahlberg, Lars Borin, Markus Forsberg, Martin Hammarstedt, Leif-Jöran Olsson, Olof Olsson, Johan Roxendal, Jonatan Uppström

Språkbanken, Dept. of Swedish, University of Gothenburg, Sweden

{malin.ahlberg,lars.borin, markus.forsberg, martin.hammarstedt, leif-joran.olsson, olof.olsson.2, johan.roxendal, jonatan.uppstrom}@svenska.gu.se

Abstract

A central activity in Språkbanken, an R&D unit at the University of Gothenburg, is the systematic construction of a research infrastructure based on interoperability and widely accepted standards for metadata and data. The two main components of this infrastructure deal with text corpora and with lexical resources. For modularity and flexibility, both components have a backend, or server-side part, accessed through an API made up of a set of well-defined web services. This means that there can be any number of different user interfaces to these components, corresponding, e.g., to different research needs. Here, we will demonstrate the standard corpus and lexicon search interfaces, designed primarily for linguistic searches: *Korp* and *Karp*.

Keywords: Swedish, corpora, lexical resources, research infrastructure.

1 The research infrastructure of Språkbanken

Språkbanken <http://spraakbanken.gu.se/eng/start> is a research and development unit at the University of Gothenburg, and a node in a cross-disciplinary research collaboration at the University of Gothenburg and Chalmers University of Technology formalized under the name of Centre for Language Technology <http://www.clt.gu.se>. Språkbanken was established with government funding already in 1975 as a national centre. The main focus of Språkbanken's present-day activities is the development and refinement of language resources and language technology (LT) tools, and their application to research in language technology, in linguistics, and in several other disciplines, notably text-based research in the humanities, social sciences, medicine and health sciences.

The larger context of these activities is the systematic construction of a research infrastructure based on interoperability and widely accepted standards for metadata and data. The two main components of this infrastructure deal with text corpora and with lexical resources. For modularity and flexibility, both components have a backend, or server-side part, accessed through an API made up of set of well-defined web services. This means that there can be any number of different user interfaces to these components, corresponding, e.g., to different research needs. Here, we will focus on the standard corpus and lexicon search interfaces *Korp* and *Karp* – designed primarily for linguistic searches – but the same backend web services are also used, e.g., in a corpus-driven grammar and vocabulary exercise generator (Volodina et al., 2012).

2 The search interface of Korp

The search interface of Korp (Borin et al., 2012b; <http://http://spraakbanken.gu.se/korp>) has been inspired by corpus search interfaces such as SketchEngine (Kilgarriff et al., 2008), Glossa (Nygaard et al., 2008), and DeepDict (Bick, 2009).

At first glance, the search interface of Korp is a concordance tool that displays search results in the standard KWIC (*keywords in context*) layout (figure 1), here with the example word *svininfluensa (noun)* 'swine flu'. This basic functionality is extended by various visualisations of statistical data, such as the basic table (figure 2) and and interactive trend diagram (figure 4) plotting relative frequency over time. Furthermore, the interface features so called word pictures that provides an overview of a selected set of syntactic relations for a word (figure 3). The purpose is to quickly gain an understanding of the contexts in which a word most commonly appears.

We are also working on increasing the diachronic coverage of the corpora, by including Swedish texts from the 19th century back to the 13th century. Ultimately, our goal is to develop tools for all types of text, at various levels of annotation, such as part-of-speech, morphosyntactic information, and dependency parses (Borin et al., 2010; Borin and Forsberg, 2011; Adesam et al., 2012). Our primary source material for Old Swedish (ca 1225–1526) comes from *Fornsvenska textbanken*,<http://project2.sol.lu.se/fornsvenska> a 3 MW collection of around 160 digitized texts, mainly from the 13th to the 16th century. Further, a 1 MW corpus of medieval letters from the Swedish National Archives <http://riksarkivet.se> is available. Work in progress concerns newspaper texts (17th–19th century) and a collection of law texts (13th century – present).

A number of issues are problematic for annotation of historical texts. For example, sentence splitting cannot be handled with standard tools, as sentence boundaries are often not marked by punctuation or uppercase letters. Compared to modern Swedish texts, the Old Swedish texts have a different vocabulary and richer morphology, show a more free word order, and Latin and German influences. Finally, the lack of a standardized orthography results in a wide variety of spellings for the same word.

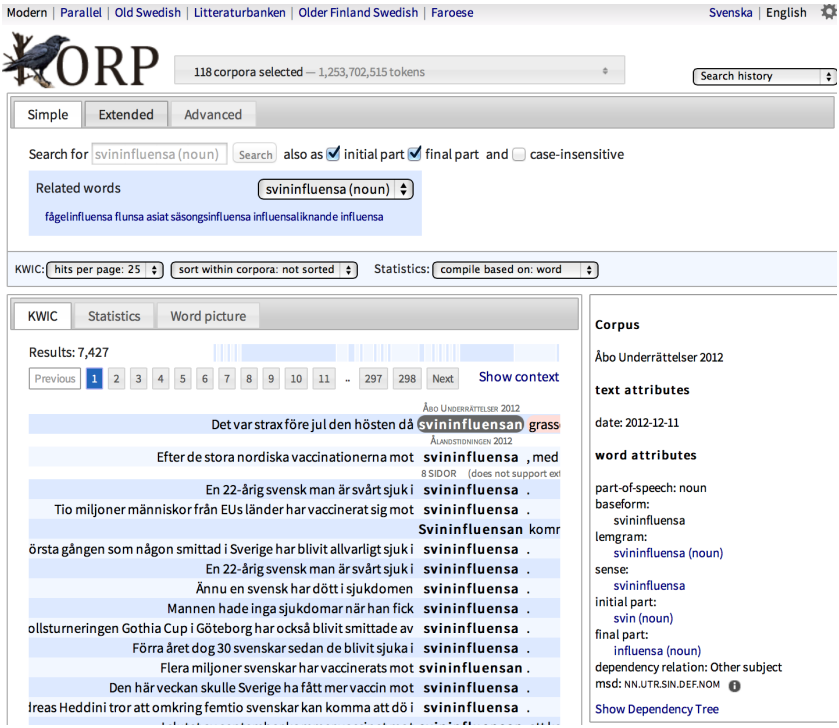

Figure 1: The Korp KWIC view of *svininfluensa (noun)* 'swine flu'

Hit	*Total*	Åbo Und…	Åbo Und…	Astra No…	Ålandsti…	8 SIDOR	Bloggmix…	Bloggmi
Σ	5.9 (7,427)	0.8 (1)			1.0 (1)	100.2 (68)		19.4 (8)
svininfluensan	3.0 (3,818)	0.8 (1)				47.1 (32)		9.7 (4)
svininfluensa	1.4 (1,785)				1.0 (1)	45.7 (31)		4.8 (2)
Svininfluensan	0.7 (878)					7.4 (5)		2.4 (1)
Svininfluensa	0.3 (416)							
svininfluensavaccinet	0.1 (95)							
svininfluensavaccin	0.0 (51)							
svininfluensans	0.0 (49)							2.4 (1)
svininfluensaviruset	0.0 (26)							
SVININFLUENSAN	0.0 (22)							
Svininfluensans	0.0 (17)							
svininfluensavirus	0.0 (16)							
svininfluensasprutan	0.0 (15)							

Figure 2: The Korp statistics of *svininfluensa (noun)* 'swine flu' (including compounds)

3 The search interface of Karp

The interface of Karp (Borin et al., 2012a; <http://http://spraakbanken.gu.se/karp>) supports searching and editing lexical resources. It currently hosts 21 lexical resources, some of which have been created from scratch using existing free resources, both external and in-house. The resources have been converted to the Lexical Markup Framework format (ISO, 2008) to ensure uniformity and interchangeability. The infrastructure has one primary lexical resource, SALDO (Borin and Forsberg,

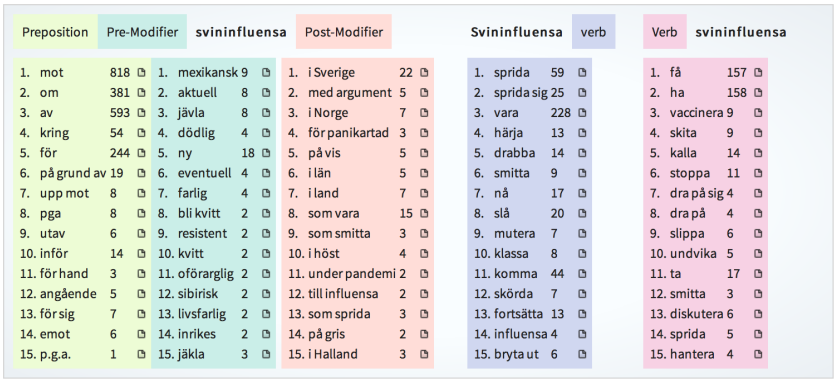

Figure 3: The word picture of *svininfluensa (noun)* 'swine flu'

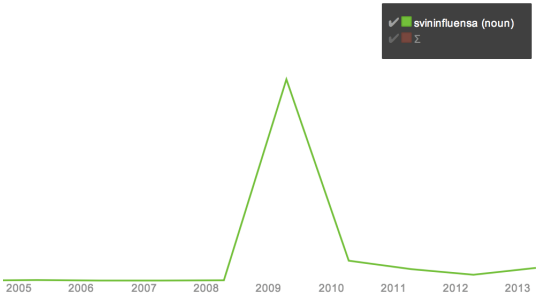

Figure 4: The trend diagram of *svininfluensa (noun)* 'swine flu'

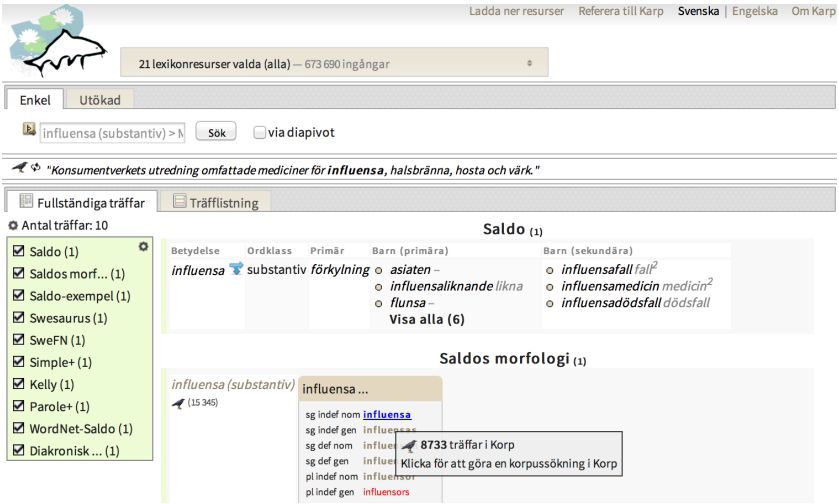

Figure 5: The search result of *influensa* 'flu'

2009), which acts as a pivot to which all other modern resources are linked. SALDO is a large freely available morphological and lexical-semantic lexicon for modern Swedish. Moreover, there is a

diachronic pivot resource with links between the modern and the historical morphologies.

In the simple search a user can input either a word form, a lemgram (a form unit), or a sense unit, and the interface will render all information associated to all sense units related to the input. E.g., figure 5 displays the search result of the lemgram *influensa (noun)* 'flu'. In the extended search the user can combine available filters from drop down boxes which are translated to SRU/CQL expressions. E.g. a word form as reglular expression and a certain part of speech type.

In addition, the interface supports full text search in the textual parts of the lexical resources, such as examples and definitions. The full text search, beyond extending the search capabilities, also makes the lexical information not linked to SALDO discoverable.

References

Adesam, Y., Ahlberg, M., and Bouma, G. (2012). *bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa...* Towards lexical link-up for a corpus of Old Swedish. In *Proceedings of LTHist 2012*.

Bick, E. (2009). A graphical corpus-based dictionary of word relations. In *Proceedings of NODALIDA 2009. NEALT Proceedings Series Vol. 4*, Odense. NEALT.

Borin, L. and Forsberg, M. (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. NEALT.

Borin, L. and Forsberg, M. (2011). A diachronic computational lexical resource for 800 years of Swedish. In *Language technology for cultural heritage*, pages 41–61. Springer, Berlin.

Borin, L., Forsberg, M., and Kokkinakis, D. (2010). Diabase: Towards a diachronic blark in support of historical studies. In *Proceedings of LREC 2010*.

Borin, L., Forsberg, M., Olsson, L.-J., and Uppström, J. (2012a). The open lexical infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 3598–3602, Istanbul. ELRA.

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

ISO (2008). Language resource management – lexical markup framework (lmf). International Standard ISO 24613:2008.

Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2008). The Sketch Engine. In Fontenelle, T., editor, *Practical Lexicography: A Reader*, pages 297–306. Oxford University Press, Oxford.

Nygaard, L., Priestley, J., Nøklestad, A., and Johannessen, J. B. (2008). Glossa: a multilingual, multimodal, configurable user interface. In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)*, Marrakech. ELRA.

Volodina, E., Borin, L., Loftsson, H., Arnbjörnsdóttir, B., and Leifsson, G. Ö. (2012). Waste not, want not: Towards a system architecture for icall based on nlp component re-use. In *Proceedings of the SLTC 2012 workshop on NLP for CALL, Lund, 25th October, 2012*, pages 47–58.