Korp – the corpus infrastructure of Språkbanken

Lars Borin, Markus Forsberg, and Johan Roxendal

Språkbanken, University of Gothenburg, Sweden
{lars.borin, markus.forsberg, johan.roxendal}@svenska.gu.se

Abstract

We present Korp, the corpus infrastructure of Språkbanken (the Swedish Language Bank). The infrastructure consists of three main components: the Korp corpus pipeline, the Korp backend, and the Korp frontend. The Korp corpus pipeline is used for importing corpora, annotating them, and then exporting the annotated corpora into different formats. An essential feature of the pipeline is the ability to leave existing annotations untouched, both structural and word level annotations, and to use the existing annotations as the foundation of other annotations. The Korp backend consists of a set of REST-based web services for searching in and retrieving information about the corpora. Finally, the Korp frontend is a graphical search interface that interacts with the Korp backend. The interface has been inspired by corpus search interfaces such as SketchEngine, Glossa, and DeepDict, and it uses State Chart XML (SCXML) in order to enable users to bookmark interaction states. We give a functional and technical overview of the three components, followed by a discussion of planned future work.

Keywords: corpus, infrastructure, Swedish language resources

1. Introduction

Språkbanken[1] (the Swedish Language Bank) was established in 1975 as a national center with a remit to collect, process and store (Swedish) text corpora and other linguistic resources, and to make linguistic data extracted from the corpora available to researchers and to the public. Since then, Språkbanken has developed into a research unit whose work focuses on the development of linguistic resources and tools, and methodologies for using the resources in research in language technology and a number of other disciplines.

As a consequence of the typical research project life cycle, an increasing share of the available manpower in Språkbanken was being tied down in maintaining a plethora of old and new corpus search interfaces that had emerged organically over the years, with different kinds of functionality, corpora, annotations, and storage and search solutions.[2] However, in the last few years we have been in the fortunate position to revert this trend of fragmentation by a targetted effort aiming at the development of a *centralized corpus infrastructure*, where a crucial shift of perspective has been to put the corpus in the center, instead of the tool used to inspect the corpus.

An important aim informing the development of the corpus infrastructure has been a strong bidirectional connection to a lexical infrastructure (Borin et al., 2012) being simultaneously developed. In essence, this aim involves up-to-date lexical annotations together with search facilities enhanced by lexical information, and conversely, the use of corpus examples and statistics in the lexical infrastructure.

The corpus infrastructure, referred to as *Korp*,[3] has three main components: the *Korp corpus pipeline* for importing, annotating and exporting corpora; the *Korp backend* consisting of a set of web services for searching and retrieving information about the corpora; and the *Korp frontend*, a graphical web user interface. The corpus collection, at the time of writing, consists of 73 corpora with a total of 910M tokens and 57M sentences. The collection covers mostly modern written Swedish, where a sizable part (265M tokens) is a monitor corpus of Swedish blog text.

We will in the rest of the paper give a brief overview of the three components of the infrastructure, and conclude describing some planned future work.

2. Korp corpus pipeline

The Korp corpus pipeline is used for importing corpora, annotate them, and then exporting the annotated corpora into different formats.

The Korp pipeline requires an XML[4] document as input. An essential feature of the pipeline is the ability to leave existing annotations untouched, both structural and word level annotations, and to use the existing annotations as the foundation of other annotations. E.g., if someone has manually tagged a material with PoS tags, it would be unproductive not to keep and use that information in the annotation process. In addition, the pipeline has been carefully crafted to enable processing of large corpora, e.g., by allowing a corpus to be split into independent chunks that are processed in parallel. This could even be done in a distributed fashion, given that a grid of computers is available.

The automatic annotation process of the Korp pipeline provides, at the time of writing, the following annotations: tokenization, sentence splitting, links to the lexical persistent identifiers, lemmatization, compound analysis, PoS/msd tagging, and syntactic dependency trees. The

[1] <http://spraakbanken.gu.se>

[2] When the development of the infrastructure described here was initiated, Språkbanken had on the order of ten different search interfaces to three different storage solutions for access to corpora in four or five different – mutually incompatible – format-annotation combinations. The maintenance of this variety of systems represented a major effort, with no discernible added value. The same was true for our lexical resources (Borin et al., 2012).

[3] The Swedish word *korp* means 'raven', hence the korp logo (see Fig. 4).

[4] Raw text data is trivially transformed into XML by embedding the text into an arbitrary XML tag.

links to the lexical persistent identifiers provide the strong bidirectional connection to the lexical infrastructure, i.e., the lexical persistent identifiers are the keys to a rich array of information categories present in the lexical resources.

After a corpus has been annotated, it is exported into the following formats: the backend format (see Sec. 3.); various corpus statistics; and a downloadable citation corpus, i.e., the corpus where the order of the sentences has been randomized, for IPR reasons.[5]

3. Korp backend

The Korp backend consists of a set of REST-based web services (Fielding, 2000) for searching in and retrieving information about the corpora. We are currently using Corpus Workbench (Christ, 1994) with the CQP query language, but we are at the same time exploring other technologies that will potentially deal better with data not naturally expressed as tables (e.g., hierarchical structures such as syntactic trees). In the case of using other technologies, e.g., an XML database and XQuery, we plan to keep CQP query language as an alternative query language, but translate it into the new query language in the backend.

The web services of the Korp backend are open to the rest of the world, which means that if a language technology researcher is in need of corpus material with up-to-date annotations, it will be readily available through the Korp backend. We are also planning to make the whole corpus pipeline available as a set of web services through the CLT Cloud (Forsberg and Lager, 2012), in order to make the latest developments of our tools generally available, and to be able to offer analyses of (shorter) modern Swedish texts on the fly.

4. Korp frontend

The Korp frontend[6] is the graphical search interface that interacts with the Korp backend. The interface has been inspired by corpus search interfaces such as SketchEngine (Kilgarriff et al., 2008b), Glossa (Nygaard et al., 2008), and DeepDict (Bick, 2009).

On the one hand, in building the frontend, we have been able to draw upon a couple of decades of experience of providing online corpus search and browsing solutions to Swedish linguists and others. This has informed the design of the basic functionality of the frontend. On the other hand, we have also made an effort to cater to the preferences of a modern, web-aware category of users. This user category expects, e.g., to be able to bookmark a corpus search in order to return to it later. For this kind of functionality, as a novel technical solution in this kind of application we have turned to State Chart XML (SCXML), an emerging W3C specification for control abstraction (Barnett et al., 2009), which provides a declarative syntax for state management. The use of SCXML enforces the identification of view states, which gives us a well-structured application together with the possibility to implement deep linking, i.e., to enable users to bookmark interaction states – combinations of corpus selections and searches – instead of only the page itself.

The search interface has three search views: *simple*, *extended*, and *advanced* and three independent result views: *KWIC*, *Statistics*, and *Word picture*.

All three views offer some common components: a corpus selection drop-down menu, a search history list, a link for switching between the single and parallel corpus search views, and links to a brief user guide and some other information about Korp.

Figure 4 shows a simple search in the KWIC result view, with a lemgram search of *gnaga (verb)* 'gnaw'. The lemgram, which is essentially an inflection table, is retrieved from the lexical infrastructure and combined with the lexical annotations of the corpora. The KWIC view gives a traditional sentence concordance view of the search, where the selection of a token brings up a side bar on the right hand side that displays the annotations, both structural and for the token itself. In the KWIC view, the context shown is the sentence in which the hit occurs. The choice of the sentence concordance format is motivated by IPR considerations; for most of the corpora, concordances are not shown in corpus order, and the restriction to single sentences prevents reconstruction of the texts by malicious software.

The close integration to the lexical infrastructure is seen in the *Related words* box, where semantically related words to the search term are displayed, which have been retrieved from the WordNet-like Swedish lexical resource SALDO (Borin and Forsberg, 2009). SALDO has a lexical-semantic structure which combines features of WordNet and Roget's *Thesaurus* (Roget, 1852), so that the related words are in the same semantic field as the search term, but do not necessarily stand in a classical lexical-semantic relation to it, or even have the same part of speech. Clicking a related word will initiate a new corpus search for this word.[7] The tight integration with the lexical infrastructure ensures that the related-words box contains only items actually found in the corpora.

The statistics result view gives a statistical overview of the search, with a row for every unique hit and a column for every selected corpus. There are also a couple of graphical diagram views available, in the form of a pie chart and a trend curve. Fig. 1 shows a partial statistics result view for the lemgram *gnaga (verb)* 'gnaw', showing its most frequent word forms and how they are distributed over a subset of the corpora. Fig. 2 exemplifies the use of a trend di-

[5] The randomization allows us to provide large amounts of language material in a form which does not preserve the original texts, yet may be useful for all purposes where no extra-sentential context is required. This format and the corpus statistics were developed for our corpora as part of our effort in the META-NORD project (<http://www.meta-nord.eu>; funded by the European Commission under the ICT PSP Programme, grant agreement no 270899), where they are being made available through META-SHARE.

[6] <http://spraakbanken.gu.se/korp>

[7] However, it should be mentioned that while the related-words box contains word senses, the corpus search is actually for all the lemgrams – form-level units – which may realize the sense in question, since we do not have word-sense annotated corpora. This of course means that a search for one sense of a polysemous or homonymous word will actually return hits for all the senses of the word, and not only that listed in the related-words box.

Träff	Totalt	Astra No…	Bloggmi…	Diabetol…	DN 1987
gnager	1,3 (1 196)		2,2 (583)		1,4 (7)
gnaga	0,7 (603)		1,0 (276)		0,2 (1)
gnagande	0,6 (514)	4,9 (1)	0,5 (143)		0,2 (1)
gnagde	0,4 (376)		0,4 (108)		0,4 (2)
gnagt	0,3 (310)		0,5 (125)		0,2 (1)
gnagas	0,0 (21)		0,0 (3)		
Gnager	0,0 (16)		0,0 (13)		
gnags	0,0 (16)		0,0 (1)		0,2 (1)

Figure 1: Korp frontend: *Statistics* result view

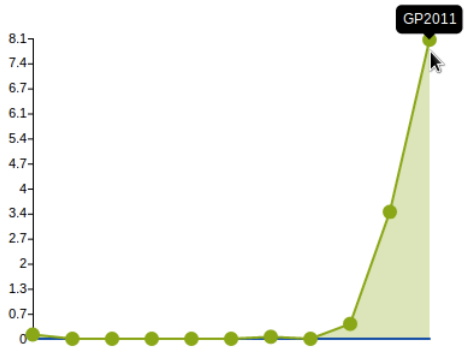

Figure 2: Korp frontend: trend diagram of *surfplatta (noun)*

agram, where we have searched for the lemgram *surfplatta (noun)* 'tablet (computer)' in eleven corpora, each corresponding to a yearly production of the newspaper GP in the years 2001 to 2011. The statistics in the diagram are in relative numbers (words per million), and we can clearly see that this word has exploded in use in the last few years.

The word picture result view gives an overview of selected syntactical environments of a word, e.g., for the lemgram *gnaga (verb)* 'gnaw' we get typical subjects, objects and adverbials. In Fig. 3 we can see that it is the 'conscience' (*samvete*) that typically gnaws 'in the back of the head' (*i bakhuvud*). Further down in the lists we see subjects such as 'beaver' (*bäver*) and objects such as 'bark' (*bark*).

Subject		gnaga	Object		Adverbial	
1.	samvete 39		1.	— 879	1.	i bakhuvud 34
2.	oro 39		2.	hål 34	2.	i huvud 29
3.	oro^2 39		3.	bark 13	3.	i skalle 9
4.	bäver 19		4.	bark^2 13	4.	i hjärta 10
5.	tanke 28		5.	oro 17	5.	hela tiden 12
6.	råtta 16		6.	oro^2 17	6.	tid 14
7.	larv 16		7.	gång 24	7.	under bark^2 5
8.	tank 21		8.	ben 12	8.	under bark 5
9.	tank^2 21		9.	känsla 9	9.	samtidig 22
10.	tand 12		10.	bit^2 7	10.	i mage 10
11.	tvivel 8		11.	bit 7	11.	ständig 9
12.	husbock 4		12.	samvete 4	12.	i baktanke 3
13.	hela tiden 12		13.	sinne 4	13.	i inre 4
14.	misstanke 8		14.	bäver 3	14.	långsam 4
15.	tid 17		15.	sav 2	15.	från talg 2

Figure 3: Korp frontend: *Word picture* result view

In other words, a word picture is a statistical overview of a selected set of syntactic relations of a word. The annotations used in the word picture are produced by the dependency parser MaltParser (Nivre et al., 2007), and they are ordered by a collocation measure on selected dependency triples (word$_1$–syntactic dependency–word$_2$).

The extended view supports the building of complex queries graphically, in a similar fashion as Glossa, and the advanced view allows for querying using the CQP query language of Corpus Workbench.

5. Future work

Korp is under constant development by a team consisting of a handful of people. There is a long list of more or less technical issues which will be dealt with over the coming year or so, such as:

- a parallel corpus search interface;
- a better interface for specifying and viewing syntactic structures (dependency trees);[8]
- a text-format export function for KWIC concordances (at present only JSON can be exported);
- more statistics views, including one showing comparative statistics, e.g., a comparison of the usage of *hon* 'she' and *han* 'he' over a range of corpora;
- explicit reference to start and end of sentence in search expressions;
- better mechanisms for metadata filtering.

As a longer-term goal we are aiming to introduce *virtual corpora*. A virtual corpus is a sequence of sentences (or larger non-overlapping units) defined by a set of metadata constraints, which can be saved and used as if they constituted a normal corpus. The rationale behind this is to move away from corpora that are treated as units solely because of accident or convenience, rather than for theoretical reasons.[9] In this connection, the metadata constraints encompass also the linguistic aspects of the texts, so that a virtual corpus in this sense can be the result of a normal corpus search.

The present version of Korp was largely custom-built for our setup in Språkbanken, especially the close interaction between the annotated corpora and the SALDO lexical resource. There have been inquiries from other groups (outside Sweden) about the availability of the Korp software. Although the software is open-source, its adaptation to another language and another set of annotations is not straightforward. For this reason, we have started the development of a version of Korp which makes minimal assumptions about these things, and which will be configurable to a particular local situation wrt corpus and lexical resources as well as to the language(s) of the user interface. The first

[8]The current frontend has a poor man's visualization of the syntactic structures: when a word is clicked upon its syntactic head becomes highlighted.

[9]Even though the term "corpus" ideally should refer to a collection of texts compiled by careful consideration of the available text 'population' and the purposes of the corpus, in actual practice, the term is often used as a cover term for any collection of linguistic material, sometimes not even texts.

version of this 'vanilla' Korp is scheduled for release by early summer 2012.

The current infrastructure employs state-of-the-art annotations tools for modern Swedish text, but we are working on extending the set of annotation tools in various directions. While some tools have been developed in-house – e.g., the morphological analyzer – we are most happy to take advantage of good open-source tools developed by others. However, at present this has the consequences that (1) the tools are largely disconnected from each other, and do not really make up a processing pipeline (e.g., the Maltparser does not use the output of the morphological analysis); and (2) there is a noticeable 'text type effect' on the quality of annotations (Giesbrecht and Evert, 2009), particularly in the blog corpus.

Thus one planned direction of research aims at better integration of annotation tools – which has both theoretical and practical aspects – as well as the adaptation and retraining of tools on a wider range of text types. The former goal includes the aim of fully integrating multi-word units into the annotation, which involves all stages of the annotation pipeline, from tokenization to dependency parsing.

Another planned research topic is adaptation of the tools to enable annotation of historical Swedish text. The starting point of this work are digitized versions of a set of historical lexical resources, one 19th century Swedish dictionary, and three Old Swedish dictionaries. In this way we plan to create a diachronic corpus of Swedish ranging from the 13th century until the present day, with rich linguistic annotations, and language stages interlinked via the lexical infrastructure.

Finally, we are also working on ways to order and display the search results based on criteria of appropriateness, for instance to enable the population of the lexical resources with good examples. The task is to define what is meant by a good corpus example (Kilgarriff et al., 2008a) as well as an optimal set of examples for a particular word or phrase, and to find a technical solution to enable the ordering.

More generally, when corpora grow beyond a certain size, the concordance view – at one time devised as a convenient way of dealing with a text mass too large to browse in its original form – quickly becomes unmanageable: A search for the word *kvinna* 'woman' yields almost 450,000 hits in the full collection of Korp corpora. We are currently investigating how large numbers of hits can be grouped and presented in a way which will reflect their linguistic properties in a way which makes sense to the users, who are mainly linguists. At present the most promising approaches that we are pursuing have been inspired by work on word-sense disambiguation and word-sense induction (Navigli, 2009), which puts our work on this particular aspect of the Korp infrastructure right at the forefront of language technology research.

6. Acknowledgements

The research presented here was supported by the University of Gothenburg through its support of the Centre for Language Technology and through its support of Språkbanken (the Swedish Language Bank), by the Swedish Research Council through its funding of the project *Safeguarding the future of Språkbanken* (VR dnr 2007-7430), and by the European Commission through its support of the META-NORD project under the ICT PSP Programme, grant agreement no 270899.

7. References

Jim Barnett, Rahul Akolkar, R. J. Auburn, Michael Bodell, Daniel C. Burnett, Jerry Carter, Scott McGlashan, Torbjörn Lager, and No'am Rosenthal. 2009. State chart XML (SCXML): State machine notation for control abstraction, October.

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

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. Odense. NEALT.

Lars Borin, Markus Forsberg, Leif-Jöran Olsson, and Jonatan Uppström. 2012. The open lexical infrastructure of språkbanken. In *Proceedings of LREC 2012*, Istanbul. ELRA.

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *Proc. 3d Conference on Computational Lexicography and Text Research (COMPLEX)*, pages 23–32, July.

R.T. Fielding. 2000. *Architectural styles and the design of network-based software architectures*. Phd thesis, University of California, Irvine.

Markus Forsberg and Torbjörn Lager. 2012. Cloud logic programming for integrating language technology resources. In *Proceedings of LREC 2012*, Istanbul. ELRA.

Eugenie Giesbrecht and Stefan Evert. 2009. Part-of-speech tagging – a solved task? an evaluation of POS taggers for the web as corpus. In *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, San Sebastian.

Adam Kilgarriff, Miloš Husák, Katy McAdam, Michael Rundell, and Pavel Rychly, 2008a. *GDEX: Automatically Finding Good Dictionary Examples in a Corpus*, pages 425–432. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell, 2008b. *The Sketch Engine*, pages 297–306. Oxford University Press.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2/10):1–69.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Lars Nygaard, Joel Priestley, Anders Nøklestad, and Janne Bondi Johannessen. 2008. Glossa: a multilingual, multimodal, configurable user interface. In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)*, Marrakech. ELRA.

Mark Peter Roget. 1852. *Thesaurus of English Words and Phrases. Classified and arranged so as to facilitate the expression of ideas and assist in literary composition*. Longmans, Green & Co, London.

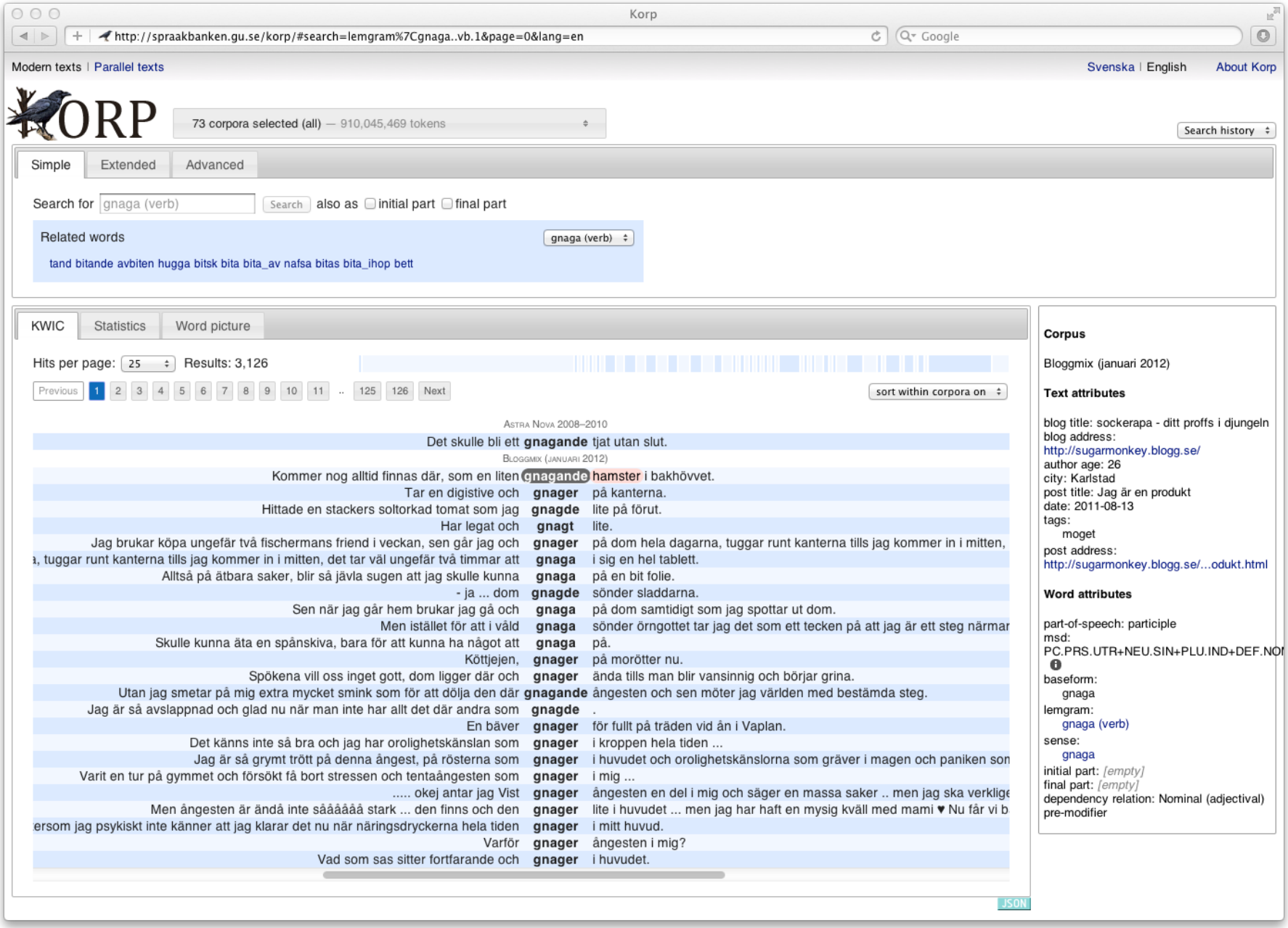

Figure 4: Korp frontend: *KWIC* result view