The open lexical infrastructure of Språkbanken

Lars Borin, Markus Forsberg, Leif-Jöran Olsson and Jonatan Uppström

Språkbanken, University of Gothenburg, Sweden
{lars.borin, markus.forsberg, leif-joran.olsson, jonatan.uppstrom}@svenska.gu.se

Abstract

We present our ongoing work on Karp, Språkbanken's (the Swedish Language Bank) open lexical infrastructure, which has two main functions: (1) to support the work on creating, curating, and integrating our various lexical resources; and (2) to publish daily versions of the resources, making them searchable and downloadable. An important requirement on the lexical infrastructure is also that we maintain a strong bidirectional connection to our corpus infrastructure. At the heart of the infrastructure is the SweFN++ project with the goal to create free Swedish lexical resources geared towards language technology applications. The infrastructure currently hosts 15 Swedish lexical resources, including historical ones, some of which have been created from scratch using existing free resources, both external and in-house. The resources are integrated through links to a pivot lexical resource, SALDO, a large morphological and lexical-semantic resource for modern Swedish. SALDO has been selected as the pivot partly because of its size and quality, but also because its form and sense units have been assigned persistent identifiers (PIDs) to which the lexical information in other lexical resources and in corpora are linked.

Keywords: lexicon, infrastructure, Swedish language resources

1. Introduction

Språkbanken[1] (the Swedish Language Bank) is research unit working on the development of (Swedish) linguistic resources and tools, and methodologies for using the resources in research in language technology and a number of other disciplines.

Språkbanken was established in 1975, and over the years has accumulated a number of lexical resources, but due to the normal format of research in our field, with funding granted mainly for shorter-term (two to four years) projects, many of these resources have lived a forgotten and slowly deteriorating life. However, in the last few years, Språkbanken has been able to start changing this negative trend through projects such as *SweFN++*[2] (Borin et al., 2010a; Borin et al., 2009), a project with the objective to create, curate, and integrate free Swedish lexical resources with the explicit goal of making them usable for language technology applications, and META-NORD,[3] a broad EC-funded European collaboration with the aim of upgrading and harmonizing language resources and tools and making them available across Europe. As part of this work, we are developing an open lexical infrastructure, presented in this paper.

2. The lexical resources

The infrastructure currently hosts 15 lexical resources for language technology use, some of which have been created from scratch using existing free resources, both external and in-house. For example, the lexical resource Swesaurus, a Swedish wordnet, is being built using not only in-house but also external resources, such as Synlex (Kann and Rosell, 2006), the Swedish Wiktionary,[4] and more indirectly, from semantic relations extracted from Princeton WordNet (Fellbaum, 1998) through links between SALDO (see below) and Core Princeton WordNet (Boyd-Graber et al., 2006).

The lexical infrastructure has one primary lexical resource, a pivot, to which all other resources are linked. This is SALDO[5] (Borin and Forsberg, 2009), a large (123K entries and 1.8M wordforms), freely available morphological and lexical-semantic lexicon for modern Swedish. It has been selected as the pivot partly because of its size and quality, but also because its form and sense units have been assigned persistent identifiers (PIDs) to which the lexical information in other resources are linked.

The standard scenario for a new resource to be integrated into the infrastructure is to (partially) link its entries to the sense PIDs of SALDO. This typically has the effect that the ambiguity of a resource becomes explicit: the bulk of the resources associate lexical information to PoS-tagged baseforms, information not always valid for all the senses of that baseform. This is natural since most of the resources have initially been created for human consumption, and a human has usually no problem dealing with this kind of underspecification. Some of these ambiguities can be resolved automatically – especially if information from several resources are combined – but in the end, manual work is required for complete disambiguation.

The infrastructure also includes historical lexical resources (Borin et al., 2010b), where the starting point is four digitized paper dictionaries: one 19th century dictionary, and three Old Swedish dictionaries. To make these dictionaries usable in a language technology setting, they need morphological information, a work that has been begun in the CONPLISIT project for 19th century Swedish (Borin et al., 2011) and in a pilot project for Old Swedish (Borin and Forsberg, 2008).

Linking SALDO to the historical resources is naturally a much more complex task than linking to the modern resources, especially when moving further back in time.

[1] <http://spraakbanken.gu.se>
[2] <http://spraakbanken.gu.se/swefn>
[3] <http://www.meta-nord.eu/>
[4] <http://sv.wiktionary.org>
[5] <http://spraakbanken.gu.se/saldo>

The promise is that a successful (but possibly partial) linking introduces the possibility to mirror the modern lexical-semantic relations onto the historical resources, so that, e.g., a WordNet-like resource for Old Swedish becomes available for use.

3. An open lexical infrastructure

The lexical infrastructure has two main functions: (1) to support the work on creating, curating, and integrating the lexical resources; (2) and to publish daily versions of the resources, making them searchable and downloadable.

A pervasive theme of the infrastructure is **openness**, which may be seen as a philosophical stance – we believe that research should be carried out in the open to enable inspection and increased collaboration. Openness pervades the infrastructure, in the use of open standards (see the next section) and open-content licenses, as well as the daily publication of not only the resources but everything else that is available in-house, such as formal test protocols and change history.

Below is an example showing the most recent change history at the time of writing for two of the lexical resources, where we can see that *Containers* and *Artifact* were the frames most recently worked on in the Swedish Framenet, and that the word *hästsova* ('sleeping while standing', literally 'horse-sleep') was the latest addition to SALDO.

SweFN	SALDO
2012-03-13 18:50: Containers Artifact	2012-03-13 21:14: hästsova
2012-03-13 12:50: Documents Containers	2012-03-13 13:54: skrämma_slag
2012-03-09 23:50: Artifact	2012-03-12 21:41: efterlysande[2] nos[2] efterlysning[2] efterlysa[2]

Yet another example, a snippet of a test protocol showing formal requirements that have been violated in the Swedish Framenet – here, senses appearing in more than one frame.

Global kontroll

1. Being_operational ∩ Suitability: fungera..1, funktionell..1
2. Cause_change ∩ Undergo_change: förändring..1, transformering..1
3. Suitability ∩ Usefulness: användbar..1, användbarhet..1

Global kontroll (Nya)

1. Amounting_to ∩ Intentionally_affect: göra..5
2. Appearance ∩ Cause_to_make_noise: klinga..3
3. Artifact ∩ Containers: punschglas..1
4. Cause_change_of_phase ∩ Change_of_phase: tina..4

An important requirement on the lexical infrastructure is that we maintain a strong bidirectional connection to the corpus infrastructure (Borin et al., 2012). This requirement includes ensuring up-to-date lexical annotations of the corpus material together with facilities to enhance the corpus search with lexical information,[6] and conversely, the use of examples and statistics from the corpora in the lexical infrastructure.

The daily publication of the resources is accomplished using the versioning system Subversion.[7] The resources together with accompanying information are published through Språkbanken's website using the content management system Drupal.[8] Moreover, the lexical resources are imported into the infrastructure and published via a web service used by the search interface (see Sec. 5.). The web service is open for others to use, and provides a convenient way of accessing the lexical information programmatically.

4. Upgrading the lexical resources to LMF

Since one of the main objectives of Språkbanken is to make the free resources accessible to the outside world, we strive to use standardized data models and file formats as far as possible. And thanks to our involvement in the project META-NORD,[9] we have been able to upgrade our lexical resources to Lexical Markup Framework, LMF, which is an ISO standard published in 2008 (ISO 24613) (ISO, 2008). The aim of the standard is to provide an intermediate format for lexical data exchange, and its specification captures a wide range of concepts, including description of morphosyntactic and lexical-semantic information.

Upgrading some of our more semantically orientated resources to LMF has turned out to be quite complicated, although still possible. The main obstacle has been that, even though the framework contains mechanisms for specifying semantic information, the model is based upon the assumption that lexical entries are formal entities expressing one or more senses, not semantic entities having one or more formal realizations. In more traditional lexicographic terms, LMF is geared towards *semasiological* rather than *onomasiological* lexical resources. This does not mesh well with the resources that contain mostly semantic information.

For example, the Swedish Framenet has the semantic frame as the natural conceptual unit, but to be able to fit the information into LMF, we had to split the frame into several entries. Moreover, the frame data is not enough to validate as LMF on its own, since it only refers to the sense PIDs of SALDO, so a choice had to be made whether to combine all fields in SALDO and Framenet into one bigger resource, or to duplicate some or all of the concerned SALDO entries in the Framenet LMF. We have settled for adding placeholders for the SALDO senses in the Framenet LMF, which are linked to the frame representations, thus

[6] For instance, in the corpus interface we provide a *Related words* box where semantically related words to the search term are provided from SALDO, but only such words that are actually present in the corpora as shown by the morphological analysis and disambiguation module in the corpus processing pipeline.

[7] <http://subversion.tigris.org/>

[8] <http://drupal.org>

[9] <http://www.meta-nord.eu/>

complying with the LMF structure while causing a minimum of redundancy.

The LMF implementation of the Swedish Framenet generally follows the suggestions by Francopoulo (2005).

5. The search interface

As previously mentioned, one of the goals of the infrastructure is to make the lexical resources searchable, and to this end, we are in the process of developing a new search interface. The search interface is built upon a REST-based web service (Fielding, 2000) exposing an unified API for accessing all the lexical resources. The web service is based on the LMF representations of the resources, and it has been implemented using the XML database eXist-db.[10]

The resources are grouped in the interface into time periods corresponding to recognized stages in the history of the Swedish language, currently: modern (*Moderna*), 19th century (*1800-tal*), and Old Swedish (*Fornsvenska*). The time period categories define the number of separate result views and the structure of the lexical resource selector (see Fig. 1).

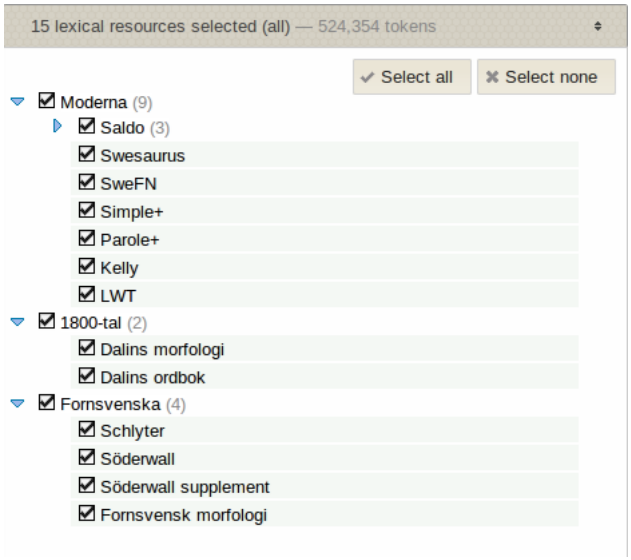


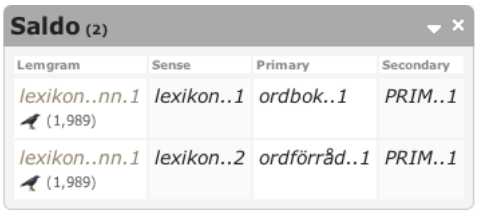
Figure 1: The lexical resource selector

The current version of the interface supports search using the morphology component of SALDO. A user can input either a word form, a lemgram (a form unit roughly corresponding to a complete inflectional pattern associated with a particular base form), or a sense unit, and the interface will render all information associated to all sense units related to the input. 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.

As an example, Fig. 2 shows the modern result view after a search for the lemgram *lexikon (noun)* 'lexicon'. In SALDO, the lemgram *lexikon (noun)* has two senses, 'dictionary' and 'vocabulary', and the information associated to these senses is shown in the result. The first table is a random occurrence of the word, provided by the corpus infrastructure, followed by information from the lexical resources: SALDO, Swesaurus, Swedish Framenet, Parole, Simple, and Kelly. In addition, we have five hits in the 19th century resources, and no hits in the Old Swedish resources.

The corpus infrastructure provides not only a random occurrence, but also information about the number of word forms of the lemgram *lexikon (noun)* (1 989 corpus occurrences). Furthermore, the inflection table is colored to indicate the existence of a word form in the corpus material; here, *lexikas* (the plural indefinite genitive form of *lexikon (noun)*) is marked as unattested in the corpora.

The interface shows the search results either exactly as the information appears in the resources, or in a more readable and aggregated manner, where redundant data are hidden. E.g., the SALDO information presented in Fig. 2 is actually aggregated from two resources, one semantic and one morphological, both having formal identifiers. The undecorated information about *lexikon (noun)*, retrieved only from semantic part of SALDO, is show below.



Clicking on *Download resources* at the top right moves us to the resource page, where all resources in the interface are downloadable.

6. Future plans

The work on the lexical infrastructure is just in its beginning, and in this section we present some of the future directions.

6.1. General

The infrastructure currently supports generation of simple statistics for all resources, and for some resources, such as SALDO and Swedish Framenet, much richer statistics and functionality such as generation of change history and formal verification of key properties of the resource. We are currently in the process of extending these functionalities to all resources and enriching them.

6.2. Metadata

Metadata for the lexical resources are used to populate the resource selector in the search interface, and to automatically generate information on the website of Språkbanken. We are planning to continue moving as much information as possible about the resources into the metadata, hence avoiding information duplication and ad-hoc solutions.

6.3. Editing

The resources are now edited outside the infrastructure, and then imported on a regular basis. This is not what we ultimately want. Instead, the infrastructure should support the

[10] <http://exist-db.org/>

creation of the lexical information in a more direct manner, so we are designing a distributed editing environment for our lexical resources that will not only support editing, but also provide methodological support, e.g., additional suggestions and consequence analysis, as well as providing on-the-fly 'intelligent' access to corpus examples by using state-of-the-art language tools. This is the topic of work in progress on our corpus infrastructure (Borin et al., 2012).

6.4. Formats

One goal of LMF is to ensure a consistent information model and terminology across different lexical resources. Acknowledging the need for specialized data fields, the LMF provides a feature structure facility that can be used to extend the format with features and corresponding values. The values should be chosen from, or registered in, a Data Category Registry, DCR (ISO 12620), such as ISOcat[11] in order to ensure not only internal consistency, but also interoperability with other, external resources. Work has begun to convert identifiers of our lexical resources to ISOcat registered categories.

6.5. Search interface

Språkbanken is the keeper of a multitude of older lexical search interfaces that have been developed for specific lexical resources. We are working on replacing all existing interfaces with the new one, which requires that the functions available in the old interfaces – or equivalent or better ones – should be available in the new interface.

6.6. Web service

The current web service will be extended to support all functions available in the older interface, and in addition to that, support more advanced searches, such as: selecting the information available in one resource but not in the other, or constraining the search with respect to non-word-features such as gender or part-of-speech.

The web service currently supports LMF (trivially) and JSON as the output format, but will soon support other formats, such as XHTML and ePub.[12]

7. Conclusion

We have briefly introduced the ongoing work on an open lexical infrastructure. It is still under active development, but is already a versatile tool for our work on the lexical resources.

The fact that we are publishing our work on a daily basis, so that fellow researchers are aware not only of what we are doing in general, but also of what we are doing *right now*, has proven a productive path to increased collaboration, especially since the infrastructure enables small contributions from colleagues who just want to learn more about a resource by working on it for a while.

8. Acknowledgements

The research presented here was supported by the Swedish Research Council (the projects *Safeguarding the future of Språkbanken*, VR dnr 2007-7430 and *Swedish Framenet++*, VR dnr 2010-6013), by the University of Gothenburg through its support of the Centre for Language Technology and through its support of Språkbanken (the Swedish Language Bank), and by the European Commission through its support of the META-NORD project under the ICT PSP Programme, grant agreement no 270899.

[11] <http://www.isocat.org>

[12] <http://idpf.org/epub/>

9. References

Lars Borin and Markus Forsberg. 2008. Something old, something new: A computational morphological description of Old Swedish. In *LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 9–16, Marrakech. ELRA.

Lars Borin and Markus Forsberg. 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.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2009. Thinking green: Toward Swedish FrameNet++. In *FrameNet Masterclass and Workshop*.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2010a. The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*.

Lars Borin, Markus Forsberg, and Dimitrios Kokkinakis. 2010b. Diabase: Towards a diachronic BLARK in support of historical studies. In *Proceedings of LREC 2010*.

Lars Borin, Markus Forsberg, and Christer Ahlberger. 2011. Semantic Search in Literature as an e-Humanities Research Tool: CONPLISIT – Consumption Patterns and Life-Style in 19th Century Swedish Literature. In *Proceedings of the Nodalida 2011*, Riga.

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

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet. In *Proceedings of the Third International WordNet Conference*.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, Mass.

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

Gil Francopoulo. 2005. Extended examples of lexicons using LMF (auxiliary working paper for LMF). Technical report, INRIA-Loria.

ISO. 2008. *ISO 24613:2008. Language resource management - Lexical markup framework (LMF)*. International Organization for Standardization, Geneva, Switzerland.

Viggo Kann and Magnus Rosell. 2006. Free construction of a free Swedish dictionary of synonyms. In *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, pages 105–110. Department of Linguistics, University of Joensuu.

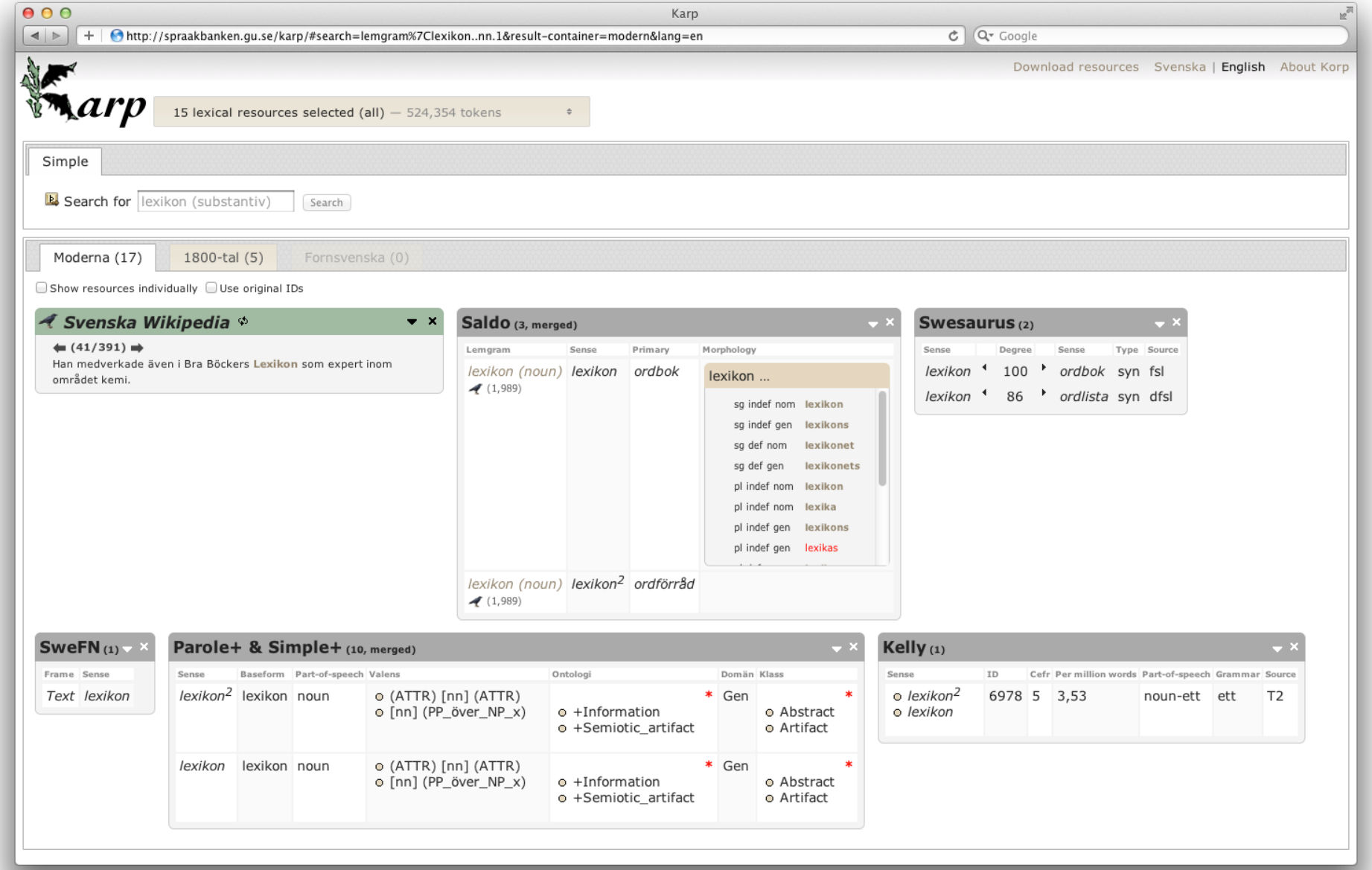


Figure 2: Searching for *lexikon* 'lexicon' in Karp