Semantic Search in Literature as an e-Humanities Research Tool: CONPLISIT – Consumption Patterns and Life-Style in 19th Century Swedish Literature

Lars Borin,[1] **Markus Forsberg,**[1] **Christer Ahlberger**[2]
[1]Språkbanken, Department of Swedish
[2]Department of Historical Studies
University of Gothenburg, Sweden
lars.borin@svenska.gu.se
markus.forsberg@gu.se
christer.ahlberger@history.gu.se

Abstract

We present our ongoing work on language technology-based e-science in the humanities, with a focus on text-based research in the historical sciences. Currently, we are working on the adaptation and integration of lexical resources representing different historical stages of Swedish into a lexical and morphological toolbox that will allow us to develop semantically oriented text search applications for historical research on Swedish text. We describe a semantic search prototype which was built using REST web services from this toolbox as components, and which has been evaluated by historians interested in using digitized 19th century novels as primary data for an historical investigation of the emerging consumer society in 19th century Sweden.

1 Introduction

Språkbanken[1] (the Swedish Language Bank), is a research unit at the University of Gothenburg in Sweden. It was established with government funding in 1975 as a national center with a remit to collect, process and store Swedish text corpora (i.e., large systematically compiled text collections). It also aims at making linguistic data extracted from the corpora and other linguistic resources, such as electronic lexicons and term lists, as well as the tools developed in-house for the purposes of linguistic processing of text, available to researchers and to the public.

Språkbanken's activities have traditionally been aimed at supporting (Swedish) linguistic research, but over the last few years we have become increasingly interested in the potential of the language technology tools and language resources that we develop and maintain in Språkbanken for forming key components in a general e-science infrastructure for the humanities, social sciences and education, and not just linguistic research.

Currently, we are working on the adaptation and integration of lexical resources representing different historical stages of Swedish into a lexical and morphological toolbox that will allow us to develop semantically oriented text search applications for historical research on Swedish text. More specifically, the work we are presenting here is a project named *CONPLISIT*, *Consumption patterns and life-style in Swedish literature*, which is a collaboration with historians at our university and with the literature digitization initiative Litteraturbanken (see section 2.3 below), and the aim of which is to develop semantic search tools for investigating the emergence of the modern consumer society in Sweden using contemporary literary sources (Ahlberger, 2009).

To this end, we are currently extending and merging two lexical resources, SALDO and Dalin, and connecting them to the search API of Litteraturbanken.

The rest of this presentation is organized as follows. In section 2, we describe the existing language resources and tools that we have utilized and in some cases enhanced in order to accomplish our goals. Section 3 contains a description of the prototype semantic search application for historical research. In section 4, we report on the first user test of the application. In section 5 we sum up our work so far and outline our plans for future work.

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

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 58–65

2 Existing language resources and tools

2.1 SALDO: a semantic lexical resource for present-day Swedish

SALDO (Borin, 2005; Borin and Forsberg, 2009; Borin et al., 2008; Borin and Forsberg, 2008), or SAL version 2, is a free modern Swedish semantic and morphological lexicon intended for language technology applications. The lexicon is available under a Creative Commons Attribute-Share Alike license and LGPL 3.0.

SALDO started its life as *Svenskt associationslexikon* (Lönngren, 1992) – 'The Swedish Associative Thesaurus', a so far relatively unknown Swedish thesaurus with an unusual semantic organization, reminiscent of, but different from that of WordNet (Borin and Forsberg, 2009). SAL has been published in paper form in two reports, from the Center for Computational Linguistics (Lönngren, 1998), and the Department of Linguistics (Lönngren, 1992), both at Uppsala University. Additionally, the headwords and their basic semantic characterizations have been available electronically, in the form of text files, from the very beginning.

The history of SAL has been documented by Lönngren (Lönngren, 1989) and Borin (Borin, 2005). Initially, text corpora were used as sources of the vocabulary which went into SAL, e.g., a Swedish textbook for foreigners and a corpus of popular-scientific articles. A small encyclopedia and some other sources provided the large number (over 3,000) of proper names found in SAL. Eventually, a list of the headwords from *Svensk ordbok* (SO, 1986) was acquired from the NLP and Lexicology Unit at the University of Gothenburg, and the second paper edition of SAL (Lönngren, 1992) contained 71,750 entries. At the time of writing, SALDO contains slightly over 104,000 entries, and new entries are added almost daily.

The central semantic relation of SALDO is *association*, a "non-classical" lexical-semantic relation (Morris and Hirst, 2004). SALDO describes *all* words semantically, not only the open word classes. By way of illustration, figure 1 shows the semantic 'neighbors' (rendered in blue/non-bold) in SALDO of the word *telefon* 'telephone (noun)'. It is associated i.a. with words like *samtala* 'hold a conversation', *telefonledes* 'by phone', *pulsval* 'pulse dialling', *ringa* 'call (verb', *mobiltelefon* 'mobile phone', the proper name *Bell*, and many others, as shown in figure 1.[2]

We soon realized that in order to be useful in language technology applications, SAL would have to be provided at least with part-of-speech and inflectional morphological information – both entirely absent from SAL in its original form – and SALDO was created. The morphological component of SALDO has been defined using Functional Morphology (FM) (Forsberg and Ranta, 2004; Forsberg, 2007), a tool that provides a development environment for computational morphologies. It is a tool with a flexible language for defining morphological rules together with a platform for testing, which is used to fend off resource degradation during development. Furthermore, it has a rich export system, targetting around 20 formats, and supports both (compound) analysis and synthesis.

SALDO is, as one of its distribution channels, published as REST web services, updated daily. Web services provide clean interfaces and instant updates, but are restricted to small amounts of data because of network latency. Presently available web services include incremental fullform lookup, semantic lookup, compound analysis, and an inflection engine service.[3]

2.2 Dalin: a lexical resource for 19th century Swedish

A.F. Dalin: Ordbok öfver svenska språket (1850–1855) (Dalin, 1853 1855) is a dictionary that has been digitized at our research unit, and is the starting point for our work on 19th century Swedish.

The language of Dalin – Late Modern Swedish according to the traditional linguistic periodization of Swedish used by Swedish linguists – is close to the modern language of SALDO, where the differences are in minor spelling variations, such as the use of the letter <f> instead of <v>, some morphological differences, such as inflection of verbs in person and number, completely absent in Present-Day Swedish. Moreover, there is a difference in the vocabulary, since the vocabulary of a dictionary reflects of the society it was produced in, e.g., many of the words in Dalin have to do either with agriculture or with religious matters.

We have created a morphology for Dalin by adapting the morphological component developed for SALDO. With a comparatively small effort we

[2] From <http://spraakbanken.gu.se/ws/saldo-ws/lid/html/telefon..1>

[3] See <http://spraakbanken.gu.se/eng/saldo>

lex:	**telefon**
l:	telefon+nn
fm:	samtala
fp:	PRIM
mf(19):	**PRIM:** fingerskiva hörtelefon kobra[2] pulsval ringa telefonautomat telefonera telefonledes telefonlur telefonör tonval **bild:** bildtelefon **knapp[3]:** knapptelefon **lokal[2]:** lokaltelefon **lyssna:** hörlur **mobil:** mobiltelefon **port:** porttelefon **trådlös:** radiotelefon **vägg:** väggtelefon
pf(18):	**abonnent:** telefonabonnent **anrop:** telefonanrop **apparat:** telefonapparat **avgift:** teleavgift **central:** telefonstation **elledning:** telefonledning **fingerskiva:** petmoj **förbindelse:** telefförbindelse **katalog:** telefonkatalog **kontakt[2]:** jack[2] **samtal:** telefonsamtal **signal:** telefonsignal **sladd:** telefonsladd **svara:** telefonsvarare telefonvakt **teknisk:** teleteknisk **ton:** kopplingston **uppfinnare:** Bell

Figure 1: Semantic neighbors (rendered in blue/non-bold) of *telefon* ‘telephone (n)’ in SALDO

have been able to provide most of the entries with an inflectional pattern based on the inflectional information provided in Dalin. However, the inflectional information in Dalin is underspecified, which means that there are erroneous word forms that we need to weed out. Also, the remainder of the entries will require a considerably larger effort, since part of the 19th century inflectional patterns cannot be automatically or almost automatically created by adaptation of the modern inflections; this holds for *inter alia* many pronouns and strong verbs. These comprise central, high-frequency vocabulary, important in practical text processing applications.

The linking of Dalin to SALDO has been done by analyzing modernized headwords of the entries in Dalin using SALDO. Connecting on the entry level is a first, over-generating, approximation, since the senses of Dalin is given within an entry. For the linking we were able to reuse an existing resource, as a manual spelling translation for the entries in Dalin had been made in an earlier project in our department and preserved in a database, which by a stroke of good luck still existed in one of our servers.

Since the vocabulary of Dalin reflects another time and another societal structure, many of its content words are not in SALDO. However, in a large number of cases they are compound words where the constituents of the compound are in SALDO. An example could be the headword *bäfverhund* with a modern spelling *bäverhund* ‘beaver dog’, meaning a dog used for hunting beavers, a word that would normally not find its way into a modern lexical resource – since the practice it refers to is no longer pursued – and adding the word to the modern resource would be unsatisfactory, since to a modern reader *bäverhund* would at most be a completely transparent compound, meaning ‘a dog in some way connected to beavers’, but without the conventionalized or lexicalized meaning the word had earlier.

In cases like this we instead choose another approach for linking the resources. Even though *bäverhund* is not in SALDO, both *bäver* ‘beaver’ and *hund* ‘dog’ are, so we use SALDO to do a compound analysis of *bäverhund* to *bäver+hund*, and link with respect to the head of the compound, i.e., *hund*.

The resulting entry for *bäfverhund*, which is analogous to the other linked entries in Dalin, is summarized in the table below. Every dictionary entry has been given an persistent identifier, here *bäfverhund..e.1*. We have the headword, its mod-

ern spelling, the inflectional information in Dalin, *m. 2.*, and the paradigm identifier it has been associated to, *nn_2m_ulf*. The paradigm identifier together with the headword defines the inflection table. Finally, we have the connection to the sense identifier in SALDO, *hund..1*.

headword	*modern*	*pos*	*paradigm*
bäfverhund	bäverhund	nn	nn_2m_ulf
gram. desc.	*saldo*		*id*
m. 2.	hund..1	bäfverhund..e.1	

Dalin contains many verb entries consisting of prefix plus verb written as one word, which in Present-Day Swedish generally correspond to phrasal verbs, i.e., verb plus separate particle/ adverb. E.g., *påspäda* 'add to', where *på* 'on, onto' is the particle, would in modern Swedish be a phrasal verb, *späda på*, or in its more common short form, *spä på*. We deal with these cases by allowing adverbs as the first constituent of a compound, given that the head of the compound is a verb. For example:

headword	*modern*	*pos*	*paradigm*
påspäda	påspäda	vb	vb_2a_ärfva
gram. desc.	*saldo*		*id*
v. a. 2.	spä_på..1	påspäda..e.1	

There is still much work to be done on the 19th century resource, but it is now in such a shape that we have been able to harvest the fruits of its creation, and the semantic search discussed in this article is one such fruit.

2.3 Litteraturbanken: a digital repository of classical Swedish fiction

Litteraturbanken[4] (the Swedish Literature Bank) is a national cultural heritage project financed by the Swedish Academy. It aims at making available online the full text of classical works of Swedish literature, in manually proofread digital versions of critical editions intended to be suitable for literary research and for the teaching of literature. There is also abundant commissioned ancillary material on its website, such as author presentations, bibliographies, and thematic essays about authorships, genres or periods, written by experts in each field. Currently, Litteraturbanken holds a bit less than 300 works in a fully searchable XML format (a variant of TEI P5), and about as much again in facsimile (image) and pdf format.

[4] <http://litteraturbanken.se/>

Similarly to many other literature digitization initiatives, most of the works in Litteraturbanken are such for which copyright has expired (under Swedish law this means that more than 70 years must have passed since the death of the author). At present, the bulk of the texts are from the 18th, 19th and early 20th centuries. However, there is also an agreement with the national organizations representing authors' intellectual property rights, allowing the inclusion of modern works according to a uniform royalty payment scheme.

The Litteraturbanken website provides a (string/ orthographic word) search function where search results are shown in a traditional concordance format, with links to the corresponding location in the digital full-text versions of the texts. However, there are also separate search (web service) APIs, both for displaying the results in the normal HTML format, and for retrieving a list of search hits for further processing.

3 Building a semantic search prototype

All the components needed for semantic search in 19th century literature are now at our disposal. A prototype was designed jointly by the language technology researchers in Språkbanken and the historians involved in the project. This was deliberately a low-stakes effort, since we would not like to put a lot of effort into a prototype which then would turn out not to be what the users desired. Our web service infrastructure supports this kind of rapid prototyping, which often becomes simply a matter of designing a couple of simple web pages acting as a kind of 'glue' joining REST web services which provide the needed functionality. Occasionally a new web service will have to be implemented, but it is our experience that in comparison to developing monolithic client-side applications, this web-service based approach allows very rapid construction and testing of quite sophisticated prototypes. Once the prototyping phase is over, it may turn out that this kind of architecture will not scale to meet production requirements due to, e.g., the inevitable network latency inherent in a web service setup. This is a separate problem, however, and its existence does in no way invalidate this approach to prototyping.

The semantic relation we decided to build a prototype around is the `md1` relation, which is the set of word senses at distance one from a target word sense in SALDO, i.e., semantically close senses.

The general idea behind the semantic search is the following: we look up up an input word form in the two morphologies of SALDO and Dalin. The SALDO senses associated to the input word form are collected, and all lexical entries in Dalin connected to these senses are listed.

The figure 2 shows the result of a search of the word form *soffa*[5] 'sofa', where there is only one sense identifier `soffa..1`,[6] connected to two entries in Dalin, `hvilosoffa..e.1` and `soffa..e.1`. Note that `hvilosoffa` has been connected to `soffa..1` through the compound analysis of its modernized form, *vilo+soffa* 'rest+sofa'. Clicking on either `soffa..e.1` or `hvilosoffa..e.1` gives a fullform search in Litteraturbanken.

There is only one `md1` for both *soffa* and *hvilosoffa*, given by the connection to `soffa..1`.

Figure 3 shows `soffa..1`'s `md1`,[7] where we can observe semantically related words such as *säng* 'bed' and *kanapé* 'couch', but also some erroneous ones due to the fact that the word *byrå* is homonymous, meaning both 'bureau' and 'dresser'. Clicking on any of the words in `md1` gives a fullform search in Litteraturbanken.

Figure 4 shows the result of clicking on *relaterade ord*[8] 'related words', which is a concordance search of the words in `md1`. The words are linked to the place in Litteraturbanken where they occur.

4 Semantic search in historical research: Practical evaluation

The historians involved in the project have been interested primarily in investigating changing consumption patterns during the first half of the 19th century. Sweden was during this period undergoing a rapid integration into the emerging world capitalist economy. A fundamental part if this integration process was a profound change in patterns of consumption among ordinary people. This change in consumption patterns is the focus of the historical research questions addressed in the CONPLISIT project. Earlier research on this topic has utilized sources such as probate inventories and tax registers. Today we have a reasonably good knowledge of the use of new commodities among social groups, sexes and age groups. On the other hand we still have very little knowledge of the contexts in which the new commodities were used. The methods and source material hitherto used in historical studies of emerging and changing consumption patterns simply has not allowed us to fully analyze and interpret the contexts of use of specific consumer items.

Hence, the research program of which our project forms a part has as one of its specific goals to develop new methods by the use of literature during the period 1830–1860 as a main source for understanding the context of consumption in this period. The prototype semantic search application is one such new method, which has undergone a small-scale formative evaluation conducted by two historians involved in the project. We now turn to the main impressions of this evaluation.

The historians initially searched for words like *porslin* 'porcelain, china', *spegel* 'mirror', *möbel* 'piece of furniture', *klocka* 'clock, watch' and so on. These items are example of the new way of living among average people. The novels also provide a context, which is how, why and by whom the new consumer items were used. The aim is to study the *new way of life* mirrored in contemporary literature and analyze the descriptions in the texts against what we know of actual consumption from other sources. Another important task for the research program is to reconstruct the new worldview and life-style of the emerging modern citizen.

The historians found the semantic search to be a good way of finding alternative search words that one would normally not think of, and it provides an overview of the variation in vocabulary choice. In this respect, it saves time and effort. An example is the search word *soffa* 'sofa', where the word *ottoman* 'ottoman' is one of the suggested words. This word had not occurred to the historians, and moreover provided an interesting surprise, since a search showed, contrary to expectation, that it is mainly used by authors in the late 19th century.

However, in some cases, such as for the search words *porslin* 'porcelain, china' and *tapet* 'wallpaper', there are many proposed words that are hard to consider relevant in the context, which was particularly true for the word *tapet*, where many of the related words are compounds with the head *vägg* 'wall', a word normally unrelated to consumption.

[5] <http://http://spraakbanken.gu.se/ws/dalin-ws/fl/html/soffa>

[6] The HTML rendering hides the suffix `'..1'`.

[7] <http://http://spraakbanken.gu.se/ws/dalin-ws/md1/html/soffa..1>

[8] <http://http://spraakbanken.gu.se/ws/dalin-ws/lb/html/250/soffa..1>

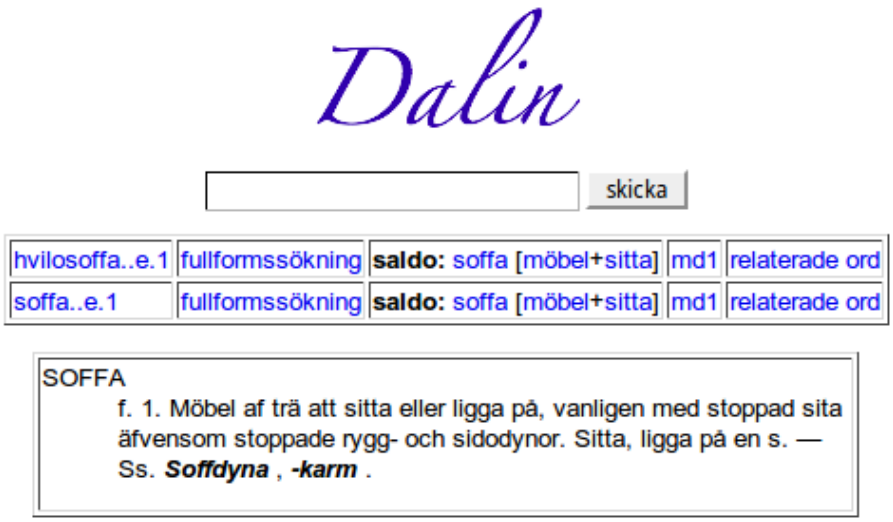

Figure 2: Word form lookup of *soffa* 'sofa'

uppmöblera	kanapé	bärstol	valkbord	tryckbord	tingsbord	tebord	svärtbord	strykbord	stigbord
spegelbord	skådebord	skärbord	skänkbord	ljusbord	kredensbord	kammarbord	hästskobord	herrskapsbord	färgbord
fällbord	friserbord	fortunabord	formbord	dambord	bröstbord	brädspelsbord	bord	bakbord	altarbord
tidningsbyrå	ottoman	kommissionsbyrå	klädesbyrå	divan	byrå	bord^2	affärsbyrå	adressbyrå	vändstol
väggbänk	vridstol	verkstol	verkbänk	vaskbänk	varpstol	varmbänk	valsstol	vaktbänk	understol
tvättstol	tvärbänk	torfbänk	sågbänk	svarfbänk	strumpstol	stol	sqvalbänk	spegelbänk	sofstol
soffstol	slipbänk	slagtbänk	slagtarbänk	skärbänk	skottstol	skjutstol	rörstol	ryggstol	rullbänk
rottingstol	reffelbänk	pressbänk	plantbänk	pinbänk	likstol	liggstol	kullerstol	korbänk	klädesstol
kardbänk	kalkbänk	häftstol	hvilostol	huggbänk	hjulstol	handstol	halmstol	gubbstol	gräsbänk
fogbänk	fallbänk	erkebiskopsstol	dufstol	dragstol	bönstol	bänk	brudbänk	borrstol	borrbänk
bordbänk	bokstol	biskopsstol	bibänk	bandstol	armstol	vaksäng	trädgårdssäng	säng	syskonsäng
sparrissäng	skogsäng	sjuksäng	plantsäng	paulunsäng	negliksäng	melonsäng	löksäng	kryddsäng	korgsäng
gurksäng	fältsäng	dödssäng	bröllopssäng	blomsäng	blomstersäng	bergsäng	utdragssoffa	soffa	hvilosoffa
skeppsbord									

Figure 3: Words semantically related to *soffa* 'sofa'

190	begärligt förtärde sina kol. Ovanför en	ottoman	hängde en tavla, försedd med guldram	Kvartetten, som sprä...
191	bland kvinnotjusare att krypa upp på	soffan	, stiga in genom ramen och lägga sitt	Kvartetten, som sprä...
192	minnen, och satte sig gungande på	ottomanen	, varefter han fortsatte: – Den	Kvartetten, som sprä...
193	, på andra sidan bron. Där stodo gröna	bänkar	under träden, och jag erinrar mig	Kvartetten, som sprä...
194	av mig den höga hatten och lagt den på	bänken	bredvid mig, men alldeles plötsligt	Kvartetten, som sprä...
195	, och till sist satt jag ensam, satt i	sängen	och grät, så att det sjöng i resårerna.	Kvartetten, som sprä...
196	– herrar Backlund och Stoltz på någon	soffa	i en lämplig parkpassage, där folkfloden	Kvartetten, som sprä...
197	en väldig trasa avtorkade ett par små	bord	: – Avhöres något nytt, så meddelar vi	Kvartetten, som sprä...
198	Mussy hest fråga, där hon satt i sin	stol	och stötte med käppen i golvet. Då	Kvartetten, som sprä...
199	och lade sig på ett knä vid hennes	stol	, medan hon i barmen fiskade efter det	Kvartetten, som sprä...
200	. Doktorns höga gestalt syntes snart vid	sängen	, kring vilken anförvanterna rörde	Kvartetten, som sprä...
201	matsal. Det - ligt dukade	bordet	sken, rosorna doftade och kristal	Kvartetten, som sprä...
202	gick att intaga en hedrande plats vid	bordet	. När det omsider knackades i vinglaset	Kvartetten, som sprä...
203	åter stolsbenen på Häradshövdingens	stol	, då han med bakdelen sköt den ut på	Kvartetten, som sprä...
204	. Då han en afton, liksom nu, låg i	sängen	, medan tankarna travade omkring Guldkalven	Kvartetten, som sprä...

Figure 4: Related word search

The words that are proposed but get no hits in Litteraturbanken are in many cases especially interesting, since they define a set of words denoting items that may not have existed at the time, or were too uncommon, or just not interesting enough for the author. An example is *fickspegel* 'pocket mirror' that is mentioned in the literature, but only in the late 19th century. Since pocket mirrors where common already in the early 19th century, it raises the question why no writer mentioned them, especially in contrast to words such as *tapet* 'wallpaper' and *gardin* 'curtain', that show another pattern.

The function *relaterade ord* 'related words' that gives a search for all words in `md1` was judged by the historians as potentially very useful, but many irrelevant hits tended to drown the relevant ones. However, if it would be possible to remove some of the words before the search, it would make a good exploratory tool.

One of the strongest requirements by the historians was the addition of chronological information, i.e., enrichment of the metadata available for the works in Litteraturbanken, and the possibility to limit the search with respect to chronological information, as well as sort the search results chronologically by publication year.

5 Summary and future work

Summing up, in the CONPLISIT project so far we have accomplished the following:

- We have built a prototype semantic search application for 19th century Swedish text. This was possible to accomplish in reasonable time – even quite quickly – since we were able to reuse several existing language resources and tools through their standardized REST web service APIs, and only needed to provide simple HTML pages as 'glue' for the prototype application.
- We have conducted a small formative evaluation of this search application with historians involved in the project, who represent one intended end-user category. The results of the evaluation were encouraging, and indicated some directions for further development.-
- We would like to claim that we have been able to show how language resources and tools can be used with good effect for building new research tools for humanities scholars.

Given what we have already accomplished or are in the midst of carrying out, as well as the kinds of resources and expertise that we can bring to bear on the problem of language technology support for historical studies, there are some lines of research that are more natural than others for continuing our work:

- We have until now only tried one kind of semantic relation – that available already in SALDO – so a natural next step is to experiment with other relations. The lexical resources used in the CONPLISIT project are developed within the framework of a larger lexical project, Swedish FrameNet++.[9] As part of the larger project, we are also developing a Swedish wordnet. Because of the way our infrastructure is built, the wordnet relations will be only a web service away once the wordnet is in place.
- One of the strongest wishes on the part of the historians was to be able to search Litteraturbanken's works chronologically, by publication year. This requires that the corresponding metadata are added to Litteraturbanken, something which is on their agenda anyway.
- Obviously, user interface issues have taken second place in our work so far, and as the functionality of successive prototypes comes closer to fulfilling the basic requirements of the users, user interface design will become increasingly important.
- An interesting question is how to find good and convincing use cases where more sophisticated linguistic information would be required, e.g. named entity information or grammatical information such as part of speech or syntactic function.[10]

[9] See <http://http://spraakbanken.gu.se/eng/swefn>

[10] There are some indications in the literature that deeper linguistic processing of historical documents can be of help to humanities scholars (Rayson et al., 2007; Pilz et al., 2008).

On the other hand, and at the shallow end of the linguistic-processing scale, one anonymous reviewer pointed out the usefulness for large-scale investigations of historical texts of extremely knowledge-light approaches inspired by information retrieval technology, such as that of Michel et al. (2011), who investigate n-gram statistics in the enormous database collected in the Google Books project.

Acknowledgements

The research presented here was supported by the Swedish Research Council (the project *Safeguarding the future of Språkbanken* 2008–2010, VR dnr 2007-7430), the University of Gothenburg through its support of Språkbanken (the Swedish Language Bank), and CLARIN through its support of the CONPLISIT collaboration.

References

Christer Ahlberger. 2009. Consumption patterns and life-style in Swedish literature – novels 1830-1860 (CONPLISIT). CLARIN collaboration proposal, April.

Lars Borin and Markus Forsberg. 2008. Saldo 1.0 (svenskt associationslexikon version 2). Språkbanken, Göteborgs universitet.

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

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. In Joakim Nivre, Mats Dahllöf, and Beata Megyesi, editors, *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, number 7 in Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia, pages 21–32. Uppsala University, Department of Linguistics and Philology, Uppsala.

Lars Borin. 2005. Mannen är faderns mormor: *Svenskt associationslexikon* reinkarnerat. *LexicoNordica*, 12:39–54.

Anders Fredrik Dalin. 1853–1855. *Ordbok öfver svenska språket. Vol. I–II.* Stockholm.

Markus Forsberg and Aarne Ranta. 2004. Functional morphology. In *ICFP'04. Proceedings of the ninth ACM SIGPLAN international conference of functional programming*, Snowbird, Utah. ACM.

Markus Forsberg. 2007. *Three Tools for Language Processing: BNF Converter, Functional Morphology, and Extract*. Ph.D. thesis, Göteborg University and Chalmers University of Technology.

Lennart Lönngren. 1989. *Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi*. Centrum för datorlingvistik. Uppsala universitet. UCDL-R-89-1.

Lennart Lönngren. 1992. *Svenskt associationslexikon. Del I-IV*. Institutionen för lingvistik. Uppsala universitet.

Lennart Lönngren. 1998. A Swedish associative thesaurus. In *Euralex '98 proceedings, Vol. 2*, pages 467–474.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett and1 Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(176):176–182.

Jane Morris and Graeme Hirst. 2004. Non-classical lexical semantic relations. In Dan Moldovan and Roxana Girju, editors, *HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, pages 46–51, Boston. ACL.

T. Pilz, A. Ernst-Gerlach, S. Kempken, Paul Rayson, and Dawn Archer. 2008. The identification of spelling variants in English and German historical texts: Manual or automatic? *Literary and Linguist Computing*, 23:65–72.

Paul Rayson, Dawn Archer, A. Baron, J. Culpeper, and N. Smith. 2007. Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics 2007*, Birmingham. University of Birmingham.

SO. 1986. *Svensk ordbok*. Esselte Studium, Stockholm.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955