Diabase: Towards a diachronic BLARK in support of historical studies

Lars Borin, Markus Forsberg, Dimitrios Kokkinakis

Språkbanken, Department of Swedish Language
University of Gothenburg
Box 200, SE-405 30 Gothenburg, Sweden
firstname.lastname@gu.se

Abstract

We present our ongoing work on language technology-based e-science in the humanities, social sciences and education, with a focus on text-based research in the historical sciences. An important aspect of language technology is the research infrastructure known by the acronym BLARK (Basic LAnguage Resource Kit). A BLARK as normally presented in the literature arguably reflects a modern standard language, which is topic- and genre-neutral, thus abstracting away from all kinds of language variation. We argue that this notion could fruitfully be extended along any of the three axes implicit in this characterization (the social, the topical and the temporal), in our case the temporal axis, towards a diachronic BLARK for Swedish, which can be used to develop e-science tools in support of historical studies.

1. Introduction

Språkbanken (the Swedish Language Bank <http://spraakbanken.gu.se>), 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. It can be argued that our central area of expertise – corpus linguistics – is among the oldest of the e-sciences, since the coining of the term "e-science" was still about a half-century in the future when language scholars began compiling and processing digital text corpora in the 1950s and 1960s. For method and tools development, corpus linguistics draws upon a number of fields, including a number of computer science subdisciplines. Nearly from the start, language technology has played a major role as tool provider for corpus linguistics, which in turn has contributed to and informed the advancement of language technology research.

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 (referred to as SHE below), and not just linguistic research (see section 2). More specifically, we are experimenting with applying language technology to historical textual sources in collaboration with literary scholars and historians, in the hope that we can in this way provide useful tools for these scholars' research. One concrete outcome of this work so far has been a named entity recognition system for 19th century Swedish literature, briefly described below in section 3.1.

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. These activities are described in section 3.2.

There are some natural ways in which these activities could be extended, (1) to cover more historical stages of the language, and (2) to provide richer functionality in support of SHE research. We discuss these in section 4.

The work that we are conducting in Språkbanken also fits into a larger picture, where the central concerns are with matters of research infrastructure and reusability of hard-earned resources. Specifically, in recent years, the concept of a BLARK (Basic LAnguage Resource Kit) has gained currency in our community (section 5), and we would like to propose here that this notion can usefully be extended both across languages and across time. The latter is our concern here, and in section 6 we discuss how our work in Språkbanken could contribute to a diachronic BLARK for Swedish.

2. Language technology for SHE research

In research in the social sciences, humanities and education (SHE), text – and speech, i.e., *language* – are central as both primary and secondary research data sources. In today's world, the normal mode of access to text, speech, images and video is in digital form. Modern material is born digital and older material is being digitized on a vast scale in cultural heritage and digital library projects.

This continuous increase of data in almost every field has afforded individual researchers the unique opportunity of having vast stores of searchable material close at hand. Along with this opportunity comes an equally unique challenge: to create the means whereby they can tap this great potentiality and engage it for the advancement of scientific understanding. The answer lies in the development of infrastructures, methodologies and designs – SHE e-science – that enable us to explore this immense repository of data in unprecedented ways, deriving fresh connections and novel facts.

The research questions are here akin to (very sophisticated and nontrivial) *information needs*, as this term is understood in information retrieval research. Just as in information retrieval, these information needs can be fulfilled only to a very limited extent purely manually, or even with low-level computational aids such as text search tools or search engines. For this reason, SHE researchers are turning to language technology as a way of overcoming this limitation. The ESFRI CLARIN[1] network is a European-level effort now starting to address this issue.

Consequently, the natural next step after digitization is the development of powerful tools to search, link, enrich, and mine the digitized data. Language technology holds central place in this endeavor, arguably even in the case of those cultural heritage collections which are primarily non-textual, since text is the pervasive medium used for metadata. This is nothing other than (an aspect of) SHE e-science.

In order to proceed successfully, an e-science research program requires three essential components:

1. research methodology and tool-developing disciplines that ideally feature leading-edge computer science, computer engineering and interaction design units;
2. disciplines that make tangible use of e-science methodology and tools in close collaboration with the developing disciplines mentioned above;
3. a viable infrastructure that supports these efforts.

Thus far the relatively new notion of e-science has been largely associated with the effort to create a sophisticated architecture that would allow for global collaborations among scientists and the sharing of computational systems, data collections and specialized experimental facilities. Perhaps the best know among these projects is the United Kingdom's *Globus approach to the Grid*, described by its inventors as "[a]n infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources" (Foster et al., 2001).

In the SHE disciplines, the general reliance is on textual data and the primary concern is the development of methodologies and tools that will enable the researcher to extract, transform, correlate, mine and visualize this type of information in productive ways. To effectively deal with this problem, an SHE e-science must include more than infrastructural innovation (the improvement of broad human-to-human collaboration); it must also include fundamental research in language technology and interaction design (the improvement of specialized human-to-text and human-to-computer interaction).

Referring to the essential components listed above, in the cases discussed here, (1) is furnished by language technology and interaction design, (2) are the fields of literature studies and history (and linguistics). The infrastructural aspects of our work (3) will be discussed in sections 5 and 6.

3. Language technology in historical studies

Our research group is involved in several projects aiming at the development of language technology infrastructural solutions to be used in aid of historical studies, i.e., e-science for the historical disciplines. Two of these projects are presented in the following sections.

3.1. NER for 19th century Swedish fiction

Litteraturbanken (the Swedish Literature Bank, <http://litteraturbanken.se/>) is a national cultural heritage project financed by the Swedish Academy. It aims at making available online the full text of relevant works of Swedish literature, in 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.

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.

Literary onomastics is a field of inquiry where literature is seen through the names appearing in literary texts. Specific topics may comprise studies of the etymology or symbolism of names, those of how fictional names make the transition into the real world, or of the use and function of names and naming in the works of an individual author, a literary school, genre, or period (Alvarez-Altman and Burelbach, 1987; van Dalen-Oskam and van Zundert, 2004). Obviously, it is possible to mark up the names appearing in digital texts manually (Flanders et al., 1998), but this is a very time-consuming, and consequently costly, endeavor. Here language technology enters the picture in the form of named entity recognition technology. The aim of the work described in this section has been to explore how this technology could make Litteraturbanken more useful to literary scholars interested in investigating the use of names in literature.

Named entity recognition (NER) has numerous applications in a number of human language technologies. It has emerged in the context of information extraction (IE) and text mining (TM), which aim at the automatic retrieval of particular kinds of information from texts (IE) or the automatic discovery of new facts in texts (TM). In these activities, the automatic recognition and marking-up of proper names and some other related kinds of information – particularly words and phrases referring to human agents, time and measure expressions – have turned out to be a recurring basic requirement. At present, NER systems are geared toward a particular language and genre, and are often fine-tuned to a particular application domain as well. Thus, systems need to be adapted whenever NER is to be applied to other languages, genres and/or domains. This adaptation can be automatized to some extent, but it generally involves a fair amount of manual work. This is a one-time effort, however, and having accomplished it, we can then proceed to mark up unlimited amounts of text at virtually no addi-

[1] Common Language and Technology Research Infrastructure funded by the EC through ESFRI – European Strategy Forum on Research Infrastructures – for the period 2008–2010; <http://www.clarin.eu>.

tional cost, as opposed to manual markup, which incurs a cost more or less proportional to the amount of text processed and becomes virtually impossible when the volume of text grows beyond a certain limit.

Combined with suitable interfaces for displaying, searching, selecting, correlating and browsing them, we believe that the automatic recognition and annotation of named entities in Litteraturbanken can facilitate literary research, and even open new avenues of research. For instance, one notes in perusing the literature in this field that studies are generally very small-scale (e.g., Alvarez-Altman and Burelbach 1987). The strength of the computer lies – here as elsewhere – in its ability to process and correlate large amounts of information and to present this information in ways that make sense to people. Given the right kind of input and the right instructions, computers are capable of producing enlightening bird's eye views of large amounts of data, generalizations, as it were, over massive data sets. Frequency and other empirical quantitative information on linguistic phenomena obtained from language corpora have turned out to offer new interesting insights in linguistics (Ellis, 2002). This could well turn out to be the case in literary studies as well. We cannot know the answer for sure, unless we actually attempt to produce such quantitative data in a form that makes sense to literary scholars.

We also have reason to believe that historians, for example, could find this facility useful, insofar as these fictional narratives also contain descriptions of real locations, characterizations of contemporary public figures, and so on. Flanders et al. (1998, p. 285) argue that references to people in historical sources are of intrinsic interest since they may reveal "networks of friendship, enmity, and collaboration; familial relationships; and political alliances [. . .] class position, intellectual affiliations, and literary bent of the author". Regardless of what we think of literature as a historical source, it is easy to see that the same kind of questions could be put by a literary scholar to a work of fiction in order to explore the fictional world developed therein. The questions we ask about the real world and the fictional worlds of literature are often not very different in kind, and it should be no disadvantage if the tools that we develop for the one would be immediately applicable also to the other.

The system we use originates from the work conducted in the Nomen Nescio project (Johannessen et al., 2005; Kokkinakis, 2004). This is a multipurpose Swedish NER system, which comprises a number of modules applied sequentially in a pipeline fashion. This system, which has been successfully applied in different domains (Kokkinakis and Thurin, 2007), was first extended to deal with written 19th-century Swedish language and evaluated on a small number of novels available in Litteraturbanken, work which we have reported on elsewhere (Borin et al., 2007). This evaluation prompted a number of modifications of the NER system, and a prototype of the modified system is now being incorporated as a regular feature of Litteraturbanken. All the hundred-plus books currently available as e-text in Litteraturbanken have been processed and the resulting named entity annotations added to the repository. Further, we have built an experimental name search and browsing application on top of the regular search interface.

Figure 1 illustrates a search for place names in Litteraturbanken. This can be seen as a generalized concordance function, i.e., it offers concordance lines for (types of) names, with links into the text locations themselves where the names appear and can be browsed in the text. We see the concordances for hits 1021–1040 out of 43,126 place names (one of 8 different NE categories available for search) automatically located and annotated by the NER system in the e-texts available in Litteraturbanken.

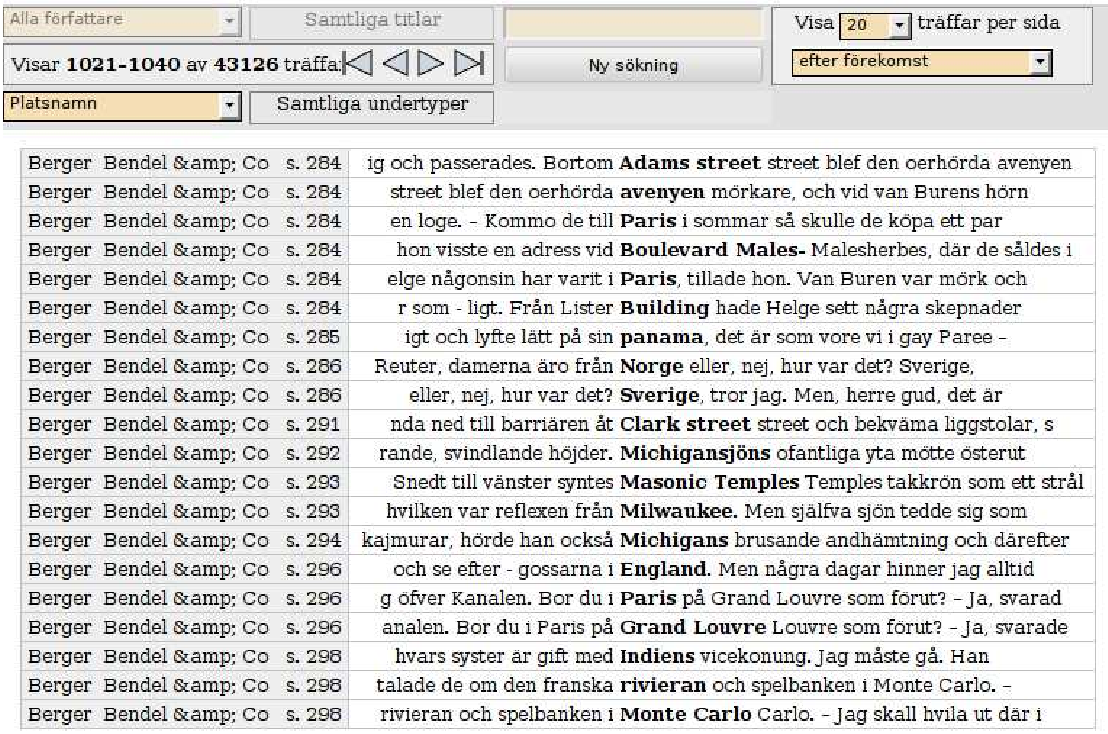
In the following section we provide a short description of the necessary adaptations made to our NER system in order to reach sufficient coverage on 19th century literary texts.

3.1.1. Adapting the NER system to historical texts

Nineteenth-century Swedish spelling is noticeably different from today's orthography. This is because we are seeing two different standards divided by an intervening spelling reform, commonly dated to the year 1906, but actually encompassing a transition period of some decades around the turn of the 20th century. Our NER system was originally based on the modern orthography, and in principle we could have considered devising a new NER system specifically for 19th-century Swedish, thus regarding it as a language system in its own right. Instead, we have opted for extending the NER system so that it handles both orthographies – the pre-spelling reform orthography and the modern spelling – simultaneously, for three reasons. First, as already mentioned, there was a long transition period in which we will find texts using both spellings. In fact, the change was gradual also in the sense that some aspects of the modern orthography were adopted earlier than others, so that from our viewpoint we even find a mixture of older and newer spellings within the same text. Second, the extension comes with little risk of confusion, as there are few or no ambiguities introduced by allowing such orthographic doublets in our NER system, and consequently, its accuracy will not be compromised. Third, there is intrinsic methodological interest in being able to deal with spelling variation in texts (see section 4.1). The calculation of various metrics for measuring the amount of difference between two sequences of symbols, the so-called edit distance or similarity, between different spelling variants is the basic approach for dealing with the problem. This allows systems to cope successfully with the analysis of texts written in different time periods and even in different genres, than the one a system has been designed for. Our approach is thus based on the identification of spelling variants using an ensemble of standard similarity algorithms. After some empirical experimentation with such metrics, we decided on suitable threshold values. New candidate entities that passed such thresholds were manually inspected for the elimination of "false friends". In this way, we have complemented the NER system lexical component with the addition of a large number of entity variants.

3.2. Semantic search in 19th century text

Språkbanken is now embarking upon a new SHE e-science project together with historians at our university and Litteraturbanken, the aim of which is to develop semantic



Berger	Bendel &amp; Co	s. 284	ig och passerades. Bortom **Adams street** street blef den oerhörda avenyen
Berger	Bendel &amp; Co	s. 284	street blef den oerhörda **avenyen** mörkare, och vid van Burens hörn
Berger	Bendel &amp; Co	s. 284	en loge. – Kommo de till **Paris** i sommar så skulle de köpa ett par
Berger	Bendel &amp; Co	s. 284	hon visste en adress vid **Boulevard Males-** Malesherbes, där de såldes i
Berger	Bendel &amp; Co	s. 284	elge någonsin har varit i **Paris**, tillade hon. Van Buren var mörk och
Berger	Bendel &amp; Co	s. 284	r som - ligt. Från Lister **Building** hade Helge sett några skepnader
Berger	Bendel &amp; Co	s. 285	igt och lyfte lätt på sin **panama**, det är som vore vi i gay Paree –
Berger	Bendel &amp; Co	s. 286	Reuter, damerna äro från **Norge** eller, nej, hur var det? Sverige,
Berger	Bendel &amp; Co	s. 286	eller, nej, hur var det? **Sverige**, tror jag. Men, herre gud, det är
Berger	Bendel &amp; Co	s. 291	nda ned till barriären åt **Clark street** street och bekväma liggstolar, s
Berger	Bendel &amp; Co	s. 292	rande, svindlande höjder. **Michigansjöns** ofantliga yta mötte österut
Berger	Bendel &amp; Co	s. 293	Snedt till vänster syntes **Masonic Temples** Temples takkrön som ett strål
Berger	Bendel &amp; Co	s. 293	hvilken var reflexen från **Milwaukee.** Men självfa sjön tedde sig som
Berger	Bendel &amp; Co	s. 294	kajmurar, hörde han också **Michigans** brusande andhämtning och därefter
Berger	Bendel &amp; Co	s. 296	och se efter - gossarna i **England.** Men några dagar hinner jag alltid
Berger	Bendel &amp; Co	s. 296	g öfver Kanalen. Bor du i **Paris** på Grand Louvre som förut? – Ja, svarad
Berger	Bendel &amp; Co	s. 296	analen. Bor du i Paris på **Grand Louvre** Louvre som förut? – Ja, svarade
Berger	Bendel &amp; Co	s. 298	hvars syster är gift med **Indiens** vicekonung. Jag måste gå. Han
Berger	Bendel &amp; Co	s. 298	talade de om den franska **rivieran** och spelbanken i Monte Carlo. –
Berger	Bendel &amp; Co	s. 298	rivieran och spelbanken i **Monte Carlo** Carlo. – Jag skall hvila ut där i

Figure 1: Place name search in Litteraturbanken

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:

SALDO (Borin, 2005; Borin and Forsberg, 2009; Borin et al., 2008; Borin and Forsberg, 2008a), 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.

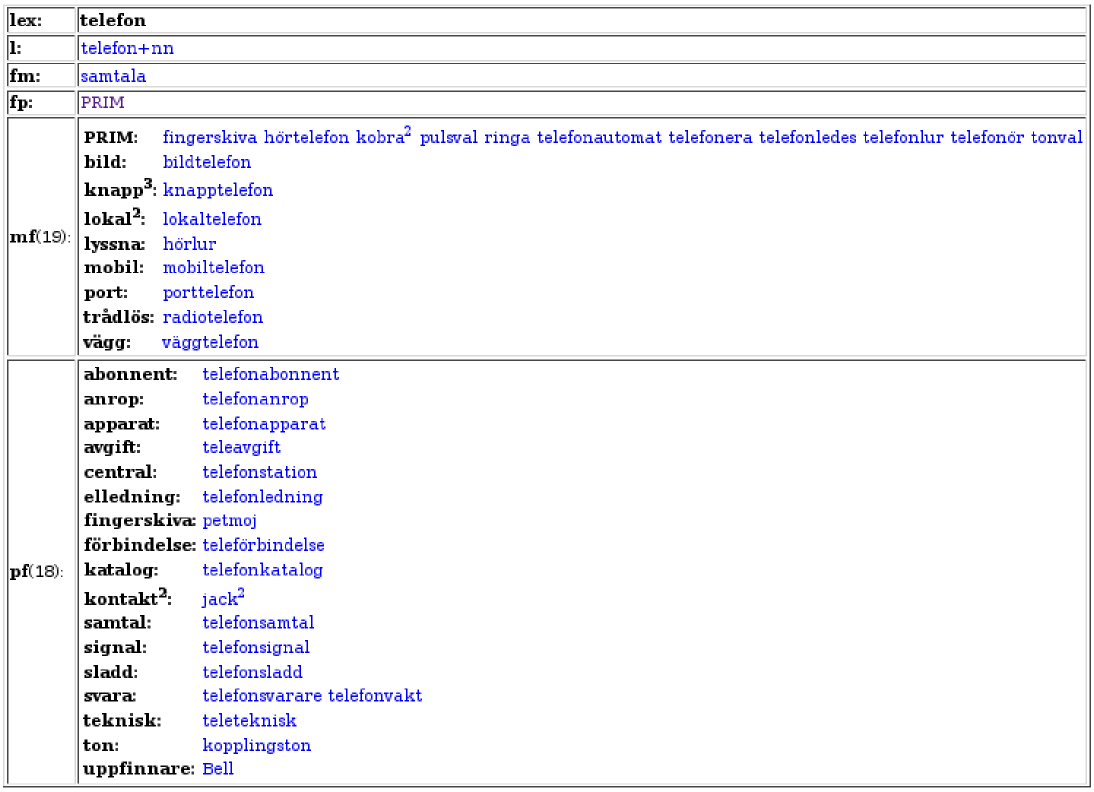
SALDO started its life as *Svenskt associationslexikon* (Lönngren, 1992) – 'The Swedish Associative Thesaurus', a so far relatively unknown Swedish thesaurus with an unusual semantic organization, reminiscent of, but different from that of WordNet (Borin and Forsberg, 2009). SAL has been published in 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 3000) 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 76,200 entries, the increased number being because a number of new words have been added, but also because a number of entries belong to more than one part of speech or more than one inflectional pattern.

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 2 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 2 (from <http://spraakbanken.gu.se/ws/saldo-ws/lid/html/telefon..1>).

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 of 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 web services, updated daily. Web services pro-



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:	teleförbindelse
	katalog:	telefonkatalog
	kontakt^2:	jack^2
	samtal:	telefonsamtal
	signal:	telefonsignal
	sladd:	telefonsladd
	svara:	telefonsvarare telefonvakt
	teknisk:	teleteknisk
	ton:	kopplingston
	uppfinnare:	Bell

Figure 2: Semantic neighbors (rendered in blue/non-bold) of *telefon* ‘telephone (n)’ in SALDO

vide clean interfaces and instant updates, but are restricted to small amounts of data because of network latency. Presently available web services include incremental full-form lookup, semantic lookup, compound analysis, and an inflection engine service. See <http://spraakbanken.gu.se/eng/saldo>.

A. F. Dalin compiled a Swedish dictionary (Dalin, 1853 1855) in the middle of the 19th century reflecting the Swedish language at that time. Dalin’s dictionary, which contains approximately 63,000 entries, has been digitized and published with a web search interface at Språkbanken: <http://spraakbanken.gu.se/dalin/>.

There is at present no morphological analysis module for this historical language variety, but one could be developed fairly quickly on the basis of the modern Swedish morphology module developed for SALDO (the morphology of 19th century and modern Swedish differ in details, but agree in general structure). This will allow text access on the level of lexical entries, rather than text words, already a gain for researchers working with the 19th century texts in Litteraturbanken. In addition, with a limited amount of automatic processing and manual work, most entries in Dalin’s dictionary could be linked to the corresponding entries in SALDO, which would provide access to the semantic thesaurus structure of SALDO, enabling a kind of semantic search, which we believe will be very useful for the kind of research that we wish to support. This would also provide access to the texts through the modern spelling, an additional decided advantage. This work has now been initiated in Språkbanken.

4. Further plans

Given the work that 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.

4.1. A diachronic Swedish lexical resource

In addition to SALDO and Dalin (section 3.2), there is a third digital historical lexical resource for Swedish available to us, in this case for the historical stage of the language known as Old Swedish (1225–1526). There are three major dictionaries of Old Swedish: Söderwall (1884) (23,000 entries), Söderwall supplement (Söderwall, 1953) (21,000 entries), and Schlyter (1887) (10,000 entries). All have been digitized by Språkbanken. There is much overlap, so that we are actually dealing with less than 25,000 different entries/lemmas/headwords. On the other hand, compounds – when the parts are written separately – are not listed as independent entries, but as secondary entries under the entry of one of the compound members.

We have started the work on creating a morphological component for Old Swedish (Borin and Forsberg, 2008b), covering the regular paradigms and created a smaller lexicon with a couple of thousand entries.

The natural next step after linking up SALDO and Dalin would be to add the Old Swedish lexicon to this growing diachronic Swedish lexical and morphological resource. Hopefully, we will be able to start on this work within a year or two. Including the Old Swedish lexicon in the same way as we are doing this for Dalin's dictionary will probably be more difficult, however, since the distance between Old Swedish and the other two forms of the language is fairly great, something like that between modern English and Anglo-Saxon (Old English). This certainly holds for the grammar – morphology and syntax – of the language, and even more so for the semantic information encoded in the SALDO lexical resource. It will be a difficult but hopefully rewarding endeavor to work with the lexical semantics of Old Swedish.

An additional complication in dealing with older stages of Swedish deserves mention. Some earlier periods in the history of the Swedish language are characterized by a marked lack of orthographic standardization, especially the Old Swedish period, but this is also largely true of the Modern Swedish period (from 1526 onwards), before a nationwide standard spelling was accomplished in the 19th century. In order to deal with pre-standardization textual material using modern language technology, we must find viable methods for dealing with the spelling and other variation exhibited in these texts (Borin and Forsberg, 2008b), and in order to reuse existing technology, we need principled methods for adapting modern language technology tools, such as part-of-speech tagging, to historical data sources (Rayson et al., 2007; Pilz et al., 2008). See further section 6.

4.2. Tracking entity references in text

Our NER system as described in section 3.1 does not pick out only proper names of people, but actually attempts to locate and annotate all human entities in the texts. Information about entities gathered in the course of analysis of the full text of a single literary work by combining entity mentions facilitates the collection and aggregation of entity profiles, for instance person profiles of the main characters. In the same way, identification of place and time expressions allows us to locate the characters in time and space. Working towards the integration of richer entity descriptions in the form of identifying and attaching meta-labels to entities in the text will allow more semantically-oriented exploration of texts by computer. This means that not only the proper names as such but also pronoun mentions and other linguistic indicators (Bontcheva et al., 2002) closely related to an entity such as family mentions (*hennes make* 'her husband') can be identified and integrated into structured formats that will allow a user exploration of, e.g., social networks of characters.

4.3. Diachronic grammatical resources

Traditionally, grammar is considered to comprise morphology and syntax. The lexical resources that we develop in Språkbanken come bundled with morphological descriptions and the machinery for performing morphological analysis on texts, as well as full-form expansion of lexical entries. In the same way that the analysis and other facilities of SALDO are now available through various web services, the whole diachronic lexical resource will be made available in this way, as each component reaches a mature enough stage.

More sophisticated text analysis functionality will require technology for full or partial syntactic analysis, i.e., so-called parsers or chunkers. To meet this demand while still staying within our areas of strength, we are now planning to initiate work on a Swedish FrameNet (Fillmore et al., 2002; Borin et al., to appear), building in part on the SALDO work and in part on our long experience in corpus linguistics. In this way, we should be able to forge a bridge from the lexical databases which we have already developed, to syntactic analysis systems. We believe that by continuing to place the lexical resources at the center of our work, we will be able to benefit also in the grammatical domain from the work already carried out on harmonizing lexical resources for different historical stages of Swedish, i.e., the hypothesis is that substantial parts of the frame semantic specifications in the modern Swedish FrameNet will carry over also to the lexical items in Dalin's dictionary, using the (semantic) links independently established between SALDO and Dalin, and possibly further to the Old Swedish lexical resources (but see section 4.1).

Our competence could be stronger in the area of syntactic analysis systems *sensu stricto*, as this has not been a focus of our research. There are other research groups in Sweden more experienced and better equipped in this regard, and we see a need for collaboration (which is ongoing in any case; see section 5). On the other hand, few other research groups in Sweden – or anywhere, for that matter – seem interested in the problem of parsing historical language varieties, let alone many such varieties in parallel, and in this respect, we will hopefully be able to make a genuine contribution to the field.

5. BLARK – Basic LAnguage Resource Kit

The need for a basic research infrastructure for language technology – and by implication for language-technology based SHE e-science – is increasingly recognized by the language technology research community and research funding agencies alike. At the core of such an infrastructure we find the so-called BLARK – Basic LAnguage Resource Kit. This label is normally used to refer to a core set of language resources and language technology tools deemed essential both to fundamental research in language technology and to the development of of useful language technology applications for a language. The definitions given of a BLARK vary somewhat, but normally they include at least:

- linguistically annotated text corpora
- speech databases
- tools for basic text and speech processing
- basic lexical resources
- tools for lingustic annotation of text (POS taggers, chunkers, parsers)
- text-to-speech and speech-to-text systems

An important aspect of the BLARK concept is that all resources and tools be interoperable, i.e., common (lossless) data exchange formats and tool APIs are necessary features of a BLARK. Preferably such formats should adhere to international standards in the field, e.g. those being prepared by ISO TC37/SC4 (*Language Resources Management*; <http://www.tc37sc4.org>) and to best practices, such as those being formulated in the framework of the ESFRI CLARIN initiative. Språkbanken is active in both initiatives.

Creating this infrastructure is beyond the means of a single research group or even a university. For example, language resources for even a single language are truly large-scale undertakings. In the Swedish Research Council planning project *An infrastructure for Swedish language technology* 2007–2008 (VR dnr 2006-6763), a national consortium consisting of seven partner institutions and coordinated by Gothenburg/Språkbanken estimated the cost for building a Swedish BLARK and a Swedish national corpus (SEK) to be on the order of 150 MSEK (about 15 MEUR), a figure that accords well with similar estimates or actual costs in other countries.

The survey conducted in connection with the planning project showed that many of the Swedish BLARK components are actually already in existence (Elenius et al., 2008); in theory, that is: Intellectual property rights restrictions and lack of interoperability need to be addressed before they can become generally available in the way understood to be part and parcel of the BLARK concept.

6. Diabase – towards a diachronic BLARK

The BLARK concept is useful for a number of reasons. First, it neatly summarizes a considerable amount of practical experience as well as concerted thinking about what kind of infrastructural support is needed in our field. Second, and this is the issue we will specifically be addressing here, when we try to see what a BLARK for a historical language stage would be like, we are led to reflect upon some assumptions underlying the BLARK notion itself.

As it is normally conceived of and presented, the BLARK assumes a modern standard language variety as the object of description, at least as far as the written language part of the BLARK is concerned, which is the part that we are competent to make judgements about. Part of the reason for this is certainly historical: The BLARK has been – and continues to be – informed more than anything else by language technology work on modern stable written standard languages.

However, modern linguistics increasingly recognizes *variation* as a fundamental and essential characteristic of human language. In this regard, research problems in the SHE domain – e.g., the study of history through textual primary sources – make up ideal interesting and challenging testbeds, where the robustness and the generality of existing language technology are subjected to the acid test of messy and multilingual reality, more so than in many other application areas, since they have to deal with, *inter alia*, historical, non-standardized language varieties in addition to a number of modern standard languages.

For our purposes here, we may discern three relevant axes of language variation:

1. by community (languages, dialects, sociolects)
2. by subject, purpose or medium (topics, genres)
3. by time (historical language stages)

The BLARK attempts to abstract away from all three; it can be thought of as reflecting a modern standard language, which is topic- and genre-neutral.

In principle, the BLARK idea could be extended along any of these three axes (or more than one, of course). The work described above can be seen as the first steps towards the development of *Diabase*, a Swedish BLARK extended along the diachronic axis. The methodologically interesting question in this regard is if we can find principled ways of doing this, and especially of coordinating it with the ongoing work on a (conventional) Swedish BLARK mentioned in the previous section.

When working together with humanities scholars on devising sophisticated language technology-based e-science tools in support of historical studies, we have both practical and theoretical reasons for wishing to reuse as much as we can of the linguistic and other information painstakingly assembled in the modern Swedish linguistic resources and language technology tools that we have at our disposal. Developing linguistic resources and language technology tools for consecutive historical stages of one language is similar to solving this problem for closely related different languages, a problem which is addressed sporadically in the literature (e.g., Trosterud 2004; Scannell 2006), but which in our view deserves more attention.

In fact, both problems can be seen as instances of the more general problem of finding systematic and principled ways of dealing with language variation in language technology research and applications. Our chosen problem area of language technology-based SHE e-science, especially historical studies, will be one of the best ways of contributing to the investigation and hopefully eventual solution of this problem.

7. 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.

We would like to express our gratitude to the anonymous referees for their constructive remarks and insightful questions, which have contributed substantially to improving the presentation in this paper.

8. References

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

Grace Alvarez-Altman and Frederick M. Burelbach, editors. 1987. *Names in literature: Essays from* Literary Onomastics Studies. University Press of America, Lanham.

Kalina Bontcheva, Marin Dimitrov, Diana Maynard, Valentin Tablan, and Hamish Cunningham. 2002. Shallow methods for named entity coreference resolution. In *Proceedings of Traitement Automatique des Langues Naturelles, TALN*, Nancy.

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

Lars Borin and Markus Forsberg. 2008b. 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. Odense.

Lars Borin, Dimitrios Kokkinakis, and Leif-Jöran Olsson. 2007. Naming the past: Named entity and animacy recognition in 19th century Swedish literature. In *Proceedings of the ACL Workshop: Language Technology for Cultural Heritage Data (LaTeCH)*, pages 1–8, Prague. ACL.

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, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. to appear. The past meets the present in Swedish FrameNet++. In *Preceedings of Euralex 2010*.

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.

Kjell Elenius, Eva Forsbom, and Beáta Megyesi. 2008. Language resources and tools for Swedish: A survey. In *Proceedings of LREC'08*, Marrakech. ELRA.

Nick C. Ellis. 2002. Frequency effects in language processing. *Studies in Second Language Acquisition*, 24(2):143–188.

Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. 2002. The FrameNet database and software tools. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1157–1160, Las Palmas. ELRA.

Julia Flanders, Syd Bauman, Paul Caton, and Mavis Cournane. 1998. Names proper and improper: Applying the TEI to the classification of proper nouns. *Computers and the Humanities*, 31(4):285–300.

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.

I. Foster, C. Kesselman, and S. Tuecke. 2001. The anatomy of the Grid: Enabling scalable virtual organizations. *International Journal of Supercomputer Applications*, 15(3):200–222.

Janne Bondi Johannessen, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdóttir, Anders Nøklestad, Dimitrios Kokkinakis, Paul Meurer, Eckhard Bick, and Dorte Haltrup. 2005. Named entity recognition for the mainland Scandinavian languages. *Literary and Linguistic Computing*, 20(1):91–102.

Dimitrios Kokkinakis and Anders Thurin. 2007. Anonymisation of Swedish clinical data. In *Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME 07)*, Amsterdam.

Dimitrios Kokkinakis. 2004. Reducing the effect of name explosion. In *In Proceedings of the LREC Workshop: Beyond Named Entity Recognition – Semantic Labeling for NLP*, Lisbon. ELRA.

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.

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.

Kevin P. Scannell. 2006. Machine translation for closely related language pairs. In *Proceedings of the Workshop Strategies for developing machine translation for minority languages at LREC 2006*, pages 103–107, Genoa. ELRA.

C.J. Schlyter. 1887. *Ordbok till Samlingen af Sweriges Gamla Lagar. (Saml. af Sweriges Gamla Lagar 13).* Lund, Sweden.

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

Knut Fredrik Söderwall. 1884. *Ordbok Öfver svenska medeltids-språket. Vol I–III.* Lund, Sweden.

Knut Fredrik Söderwall. 1953. *Ordbok Öfver svenska medeltids-språket. Supplement. Vol IV–V.* Lund, Sweden.

Trond Trosterud. 2004. Porting morphological analysis and disambiguation to new languages. In *SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages*, pages 90–92, Lisbon. ELRA.

Karina van Dalen-Oskam and Joris van Zundert. 2004. Modelling features of characters: Some digital ways of looking at names in literary texts. *Literary and Linguistic Computing*, 19(3):289–301.