Swedish CLARIN Activities

Maia Andréasson
Lars Borin
Markus Forsberg
Språkbanken
Dept of Swedish Language
University of Gothenburg
first.last@svenska.gu.se

Jonas Beskow, Rolf Carlson
Jens Edlund, Kjell Elenius
Kahl Hellmer, David House
Centre for Speech Technology
School of Computer Science
and Communication
KTH
(rolf,kjell,davidh)@speech.kth.se

Magnus Merkel
NLP Lab
Dept of Computer Science
Linköping University
mme@ida.liu.se

Eva Forsbom, Beáta Megyesi
Language Technology Unit
Dept of Linguistics and Philology
Uppsala University
first.last@lingfil.uu.se

Anders Eriksson
Phonetics Unit
Dept of Philosophy, Linguistics
and Theory of Science
University of Gothenburg
anders.eriksson@ling.gu.se

Sven Strömqvist
Centre for Languages and Literature
Lund University
sven.stromqvist@ling.lu.se

Abstract

Although Sweden has yet to allocate funds specifically intended for CLARIN activities, there are some ongoing activities which are directly relevant to CLARIN, and which are explicitly linked to CLARIN. These activities have been funded by the Committee for Research Infrastructures and its subcommittee DISC (Database Infrastructure Committee) of the Swedish Research Council.

1 Introduction

CLARIN <http://www.clarin.eu> has two partners (Centre for Speech Technology, KTH and the Humanities Lab, Lund University) and a considerable number of members in Sweden, including the sites of the authors of this document.

However, the Swedish Research Council has decided not to allocate national funds for Swedish involvement in the ongoing preparatory phase of CLARIN, which means that any participation by Swedish members beyond that which is covered by EC funding to the two Swedish CLARIN partners must be covered by funds obtained elsewhere.

On the other hand, the Swedish Research Council has increased available funding for research infrastructure *in general*, and in fact Swedish CLARIN members have been able to secure project funding for some CLARIN-related activities in this way from the Committee for Research Infrastructures and its subcommittee DISC (Database Infrastructure Committee) of the Swedish Research Council.

CLARIN-related work in Sweden has been considerably aided by the fact that the Swedish language technology community is close-knit – with well-functioning channels and fora of communication and collaboration – and united in its recognition that the realization of the kind of infrastructure that CLARIN engagement requires is a costly endeavor which must be a collective undertaking involving the whole community.

In the next section we describe some of the ongoing CLARIN-related activities in Sweden, for which we have been able to secure funding by the Swedish Research Council.

2 Some CLARIN-related activities in Sweden

2.1 An infrastructure for Swedish language technology

In 2007, the Research Infrastructure Committee of the Swedish Research Council awarded a two-year

planning grant to a national Swedish consortium in language technology, with 7 partner institutions:

- University of Gothenburg (coordinating partner)
- Chalmers University of Technology
- KTH (Royal Institute of Technology)
- Linköping University
- Lund University
- The Swedish Language Council
- Uppsala University

The planning grant was awarded for a proposal entitled *An infrastructure for Swedish language technology*, with the aim of preparing a project proposal or project proposals for creating an integrated basic Swedish language technology research infrastructure, consisting of

1. a Swedish national corpus (*Svensk nationell korpus* – SNK);
2. a Basic LAnguage Resource Kit (BLARK) for Swedish.

The practical planning work has been carried out by two working groups, with researchers from Gothenburg and Linköping responsible primarily for the work on SNK, and researchers from KTH and Uppsala having worked mainly on the Swedish BLARK. The two groups have interacted constantly throughout the course of the work, both in physical meetings and by means of electronic communication, e.g., project reports and other documents have been collectively prepared using a project wiki.

The main tasks of the working groups have been:

- to make an inventory of and collect information about existing resources, their character, quality, and not least, availability for research and other purposes;
- to make a survey of the needs of the research community and industry;
- to collect information about similar initiatives – completed, ongoing and planned – in other countries, especially in Europe;
- on the basis of this information, to formulate a concrete funding proposal to VR/KFI, comprising a description of the SNK and the Swedish BLARK, together with an outline work plan and budget for creating the resources.

A funding proposal for an SNK/BLARK combination was submitted to VR/KFI in October 2008. The proposal is now being reviewed by international experts. The amount of funding needed for realizing the SNK and Swedish BLARK in parallel is estimated at 130 million SEK over 7 years. However, it is pointed out in the proposal, that pursuing the two separately would cost on the order of 50 million SEK more, i.e., there is considerable synergy in the proposal.

No doubt in large part as a result of the work in this planning project, the Swedish Research Council has listed language technology as one of a number of national research infrastructure areas of highest priority in its *Roadmap to research infrastructure*. This spring, a call will be issued for proposals by national consortia in exactly those areas. Thus, it seems there is a good chance that the two years of dedicated work laid down in this project might pay off.

2.2 Safeguarding the future of Språkbanken

Språkbanken (the Swedish Language Bank; <http://spraakbanken.gu.se>) at the University of Gothenburg provides an online service to the research community since 1975, whereby language resources (corpora and lexicons) are made available to the research community and the public. The resources are available free of charge on the internet through a number of search interfaces. Språkbanken possesses a unique combination of competences in the areas of Swedish text corpora, parallel text corpora, Swedish computational lexicons, and LT tools for the processing, annotation and presentation of text corpora, coupled with the kind of stable organization required for sustained large-scale corpus processing and presentation.

Språkbanken's resources are widely used for research and teaching, but also for other related purposes (for checking what is possible or good Swedish, as a reference in popular writings about language usage, etc.). In particular, a good number of PhD theses in Sweden and Finland have used Språkbanken as a data source.

Språkbanken has grown organically over the four decades of its existence. Many of the presently available corpora have been collected on Språkbanken's own initiative, and this is ongoing work; e.g., about 15–20 million words of press text are added annually. However, some of the corpora are the result of independent research

projects conducted by the NLP research group at Gothenburg or by groups at other Swedish universities. In principle, the same situation obtains for the lexicon resources. Tools for browsing and searching resources have been developed in concert with the creation of the resources themselves. This means that resources are stored in Språkbanken in several different formats, with varying amounts of added information. The use of different formats implies that idiosyncratic tools are required for browsing and searching each resource. A number of language technology tools are used with the resources, which have been developed or adapted in various research projects in the department. There are also tools that have been developed in collaboration with other groups, e.g. morphological processors for modern Swedish and Old Swedish which are being developed jointly with the Language Technology research group at Chalmers University of Technology. The conditions under which such research endeavors are undertaken have not in general been conducive to standardization and wider integration of these tools.

Generally, the kinds of research questions which can be addressed using a large text material such as that found in Språkbanken are heavily dependent on three characteristics of the material and the infrastructure in which it is embedded: (1) the character of the material itself (its representativity w.r.t. the language variety under investigation); (2) the annotations, markup and metadata that the material is provided with (and, more generally, which annotations, etc., are [formally] allowed by a given framework); (3) the level of access to the material, viz. (3a) inspection (search and presentation) access only: (3a1) restricted (individually [login] or by site [IP number]); (3a2) unrestricted; (3b) download access (or other in toto access): (3b1) restricted (individually [login] or by site [IP number]); (3b2) unrestricted.

The ideal would be to have fully representative corpora provided with the maximum possible amount of high-level linguistic annotations and rich metadata, which would be available both via sophisticated online user interfaces and for downloading. There is now an urgent need for integration of the (presently) diverse resources and tools in Språkbanken in a way that also takes into account international standardization work in the field of language (technology) resources. Thus, Språkbanken will be further developed in the following areas, broadly definable as those dealing with infrastructure components (1–5) and user interface/interaction components (6):

1. Standardization of storage and exchange formats;
2. Standardization of annotation, markup and metadata formats;
3. Addition of uniform linguistic annotations to all the corpora of contemporary Swedish;
4. Addition of metadata to existing resources;
5. Definition of a set of processing components and APIs (Application Programming Interfaces) for these components;
6. Development of a set of user interface components for selecting, browsing, searching, annotating, etc., Språkbanken's corpora and lexicons, as well as up- and downloading texts.

Work is well underway in the project on all of these. One aim is to collaborate with other initiatives whenever feasible; thus, the corpus browser frontend Glossa developed by Tekstaboratoriet, University of Oslo, is now being adapted for use in Språkbanken. This work will be conducted jointly with Tekstlaboratoriet.

The CLARIN preparatory phase work is seen as so important by an institution such as Språkbanken – whose day-to-day activities will be profoundly influenced by the standards, recommedations, best practices, etc., which emerge from CLARIN preparatory phase work – that Språkbanken has decided to use part of the funding for this national project to participate in the preparatory phase of CLARIN; at the present time, this is one of the best ways of safeguarding the future of Språkbanken.

2.3 Spontal: Multimodal database of spontaneous speech in dialog

This section describes the ongoing Swedish speech database project, *Spontal: Multimodal database of spontaneous speech in dialog*. The project takes as its point of departure the fact that both vocal signals and gesture involving the face and body are important in everyday, face-to-face communicative interaction. Our understanding of vocal and visual cues and interactions in spontaneous speech is growing, but there is a great need for data with which we can make more precise measurements. Currently we have very little

data with which we can measure with precision such important aspects of human communication as the timing relationships between vocal signals and facial and body gestures, or how these gestures vary in spontaneous speech or in different speaking styles.

The goal of the Spontal project is the creation of a Swedish multimodal spontaneous speech database rich enough to capture important variations among speakers and speaking styles to meet the demands of current talk-in-interaction research. An important contemporary trend is the study of everyday spoken language in dialog which has many characteristics differing from written language or scripted speech. Detailed analysis of spontaneous speech can also be fruitful for phonetic studies of prosody and also reduced and hypoarticulated speech. The Spontal database will make it possible to test hypotheses on the visual and verbal features employed in communicative behavior covering a variety of functions. To increase our understanding of traditional prosodic functions such as prominence lending and grouping and phrasing, the database will enable researchers to study visual and acoustic interaction over several subjects and dialog partners. Moreover, dialog functions such as the signaling of turn-taking, feedback, attitudes and emotion can be studied from a multimodal, dialog perspective. In addition to basic research, one important application area of the database is to gain knowledge to use in creating an animated talking agent (talking head) capable of displaying realistic communicative behavior with the long-term aim of using such an agent in conversational spoken language systems. The database will be freely available for research purposes.

60 hours of dialog consisting of 120 half-hour sessions will be recorded. Each session consists of three consecutive 10 minute blocks. Subjects are told that they are allowed to talk about absolutely anything they want at any point in the session, including meta-comments on the recording environment and suchlike, with the intention to relieve subjects from feeling forced to behave in any particular manner. Subjects are informed about the time after each 10 minute block. After 20 minutes, they are asked to open a wooden box which contains objects whose identity or function is not immediately obvious. The subjects may then hold, examine and discuss the objects taken from the box, but they may also chose to continue whatever discussion they were engaged in or talk about something entirely different. The subjects are all native speakers of Swedish and balanced as to gender and whether the dialogue partners know each other or not. This balance will result in 15 dialogs of each configuration: 15x4x2 for a total of 120 dialogs. Currently (February, 2009), about 25% of the database has been recorded.

In the base configuration, the recordings are comprised of high-quality audio and high-definition video, with about 5% of the recordings also making use of a motion capture system using infra-red cameras and reflective markers for recording facial gestures in 3D. In addition, the motion capture system is used on virtually all recordings to capture body and head gestures, although resources to treat and annotate this data have yet to be allocated.

2.4 SweDia 2000 – A Swedish dialect database

The SweDia database consists of recorded speech from 107 dialects representing the dialectal variation in Sweden and Swedish-speaking parts of Finland. The recordings were made in 1999 by a previous research project, SweDia 2000. Each dialect is represented by twelve speakers representing two generations with an equal number of male and female speakers. Research questions that may be addressed using the data are: What are the laws that govern language development and change? To what extent does internal structural coherence govern the development of dialects? The database has until now primarily been used by the SweDia group and a circle of researchers who have obtained personal copies on hard disks. The goal of the present work is to make the database available to a much wider circle by placing it on an internet server together with other language databases accessible via a common web-based interface. It should be possible to perform searches at syllable-, word- or word sequence levels. A first version of (nearly) the entire database already exists hosted on an IMDI-server at the Centre for Language and Literature at Lund University. The result of a successful search can, for example, be a sound file with the desired items and a time-aligned transcription. It should be possible to listen to it directly or download a file for further analysis. In its present form, only parts of the database

material are transcribed.

A part of the database that comprises informal interviews and semi spontaneous monologues will be simultaneously hosted on a server at Tekstlaboratoriet at the University of Oslo. This part of the database will be combined with data collected by the Scandinavian Dialect Syntax project.

To make the databases fully searchable they will have to be transcribed at the word level. This work is in progress and substantial parts of the material are already transcribed. Simple analysis tools will also be available. To the extent that it is possible they will be designed to run on-line. Additional tools will be offered for download.

2.5 Litteraturbanken

The project described in this section – *Litteraturbanken* (the Swedish Literature Bank; <http://litteraturbanken.se>) is different from the others described above, in that it has permanent funding by an independent private funding body, the Swedish Academy.

Litteraturbanken is a public digital repository of classical Swedish literary works in scientifically validated editions. It is slated to grow by approximately 100 novel-length works annually. The relevance to CLARIN of this endeavor is found in the following two circumstances:

1. The technical infrastructure of Litteraturbanken was developed by Språkbanken, which is also responsible for developing this infrastructure and maintaining the Litteraturbanken website in its servers. This means that the work on the technical solutions in Litteraturbanken is part of the work in the project decribed above in section 2.2;
2. Litteraturbanken is developed with the aim that it can serve as a primary data source for research in a number of disciplines in the humanities and social sciences (e.g., literature, various historical disciplines and sociology), using language technology tools, e.g., in the form of text mining.

3 Conclusion

Even though the Swedish Research Council has not set aside funds explicitly intended for CLARIN work, the projects described in the preceding section together represent a funding of 10.6 million SEK (about 1 million Euro), plus about 2.5 million SEK annually to Litteraturbanken. The resources being realized with this funding will be extremely valuable when CLARIN enters its permanent phase.

Acknowledgments

We gratefully acknowledge the following sources of funding for the work described or mentioned above.

The work in the CLARIN preparatory phase by the Centre for Speech Technology, KTH, and Centre for Languages and Literature, Lund University, supported by CLARIN.

The planning project *An infrastructure for Swedish language technology* 2007–2008 (a national collaboration, coordinated by Språkbanken, University of Gothenburg), by the Swedish Research Council's Committee for Research Infrastructures (VR dnr 2006-6763).

The project *Safeguarding the future of Språkbanken* 2008–2010 (Språkbanken, University of Gothenburg), supported by the Database Infrastructure Committe of the Swedish Research Council's Committee for Research Infrastructures (VR dnr 2007-7430).

The project *Spontal: Multimodal database of spontaneous speech in dialog* 2007–2009 (Centre for Speech Technology, KTH, supported by the Database Infrastructure Committe of the Swedish Research Council's Committee for Research Infrastructures (VR dnr 2006-7482).

The project *SweDia 2000 – A Swedish dialect database* 2008–2010 (Phonetics, University of Gothenburg), supported by the Database Infrastructure Committe of the Swedish Research Council's Committee for Research Infrastructures (VR dnr 2007-7432).

Litteraturbanken, supported on a permanent basis by the Swedish Academy.