Language Resources and Tools for Swedish: A Survey

Kjell Elenius[1], Eva Forsbom[2], Beáta Megyesi[2]
[1]Speech, Music and Hearing, KTH
[2]Dept. of Lingustics and Philology, Uppsala University
Sweden
E-mail: kjell@kth.se, evafo@stp.lingfil.uu.se, Beata.Megyesi@lingfil.uu.se

Abstract

Language resources and tools to create and process these resources are necessary components in human language technology and natural language applications. In this paper, we describe a survey of existing language resources for Swedish, and the need for Swedish language resources to be used in research and real-world applications in language technology as well as in linguistic research. The survey is based on a questionnaire sent to industry and academia, institutions and organizations, and to experts involved in the development of Swedish language resources in Sweden, the Nordic countries and world-wide.

1. Introduction

Research and development of Swedish language technology systems needs an infrastructure of publicly available and standardized basic language resources. These resources can be data or programs to process and use the data. A set of such basic resources is called a BLARK - Basic LAnguage Resource Kit (Krauwer, 1998). Examples of language resources are mono- or multilingual corpora or lexicons, grammars, benchmarks for evaluations, and tools for processing language data. A BLARK has to be created separately for each language. Several language resources exist for Swedish, but it is unclear to what extent and to what degree they are available. Therefore, there is a need to make an inventory and describe the existing language resources and how they are used. Also, it is necessary to survey the need of such resources for future development and usage. The goal of the present work is to prepare for the creation of an infrastructure for Swedish language technology.

This study is a part of a national venture to develop *An Infrastructure for Swedish Language Technology*, funded by the Swedish Research Council's Committee for Research Infrastructures 2007-2008. This venture is strongly supported by the language technology community in Sweden.

In section 2, we describe the inventory process of existing language resources for Swedish and the needs for these. In section 3, we summarize the results of the inventory, including a description of existing written and spoken language resources, followed by a description of the needs for such resources. Lastly, in section 4 we conclude the paper.

2. Method

The work on surveying existing basic language resources and the need for developing missing resources is carried out in three phases. In the first phase, we wish to get an overview of existing resources and find out what resources are needed. As the next step, we will use the information gathered to define what types of resources should be part of a Swedish BLARK, describe the existing resources in uniform metadata, and point out what type of resources that are missing. Lastly, the needed resources will be ordered according to their importance. The current study mainly concerns the first phase.

To make an inventory and collect information about existing and needed Swedish language resources, we developed a questionnaire inspired by previous surveys carried out for Arabic within the NEMLAR project (Nikkhou and Choukri, 2005).

The questionnaire, which is available on the web http://www.speech.kth.se/prod/blark/blark.html focuses on Swedish language resources and tools and covers the following resources:

- Language resources: mono- or multilingual corpora (spoken or written language), mono- or multilingual lexicons, terminology archives, grammars
- Standard resources (benchmarks) for evaluation
- Tools for processing language data: modules (e.g., part-of-speech taggers, parsers, text-to-speech converters), standards and tools for annotation, tools for searching and mining information from corpora

The questionnaire is divided into seven parts. The first part contains information about the person who filled out the questionnaire, while the second part aims to find out information about the actual organization, institution and/or individual. The third part gathers information about the Swedish language resource needs, and the fourth part focuses on the existing language resource for Swedish, both divided into written, spoken and multimodal data. The fifth part concerns the acquisition of Swedish language resources to find out how and from where the resources are acquired, and if there is a benefit from existing standard interchange formats when incorporating the acquired resources. In part six, general comments on the questions and/or on the resources can be left by the subjects. Lastly, there is a possibility for the subjects to give us further suggestions on other contacts.

In order to get a reliable survey, we e-mailed the questionnaire together with a cover letter to a large number of people who work with language resources in academia and industry, in Sweden and abroad. We used lists to reach as many experts, as possible, such as the Nordic computational linguist list (nodali) and the corpora list. We sent the questionnaire to all universities in Sweden who carry out research on language technology or computational linguistics, a large number of companies working with language technology products (participants at www.sprakteknologi.se and the partners of the Centre for Speech Technology, CTT at KTH), and institutions and organizations working with the Swedish language as professionals (such as networks for members of Swedish translation companies, language professionals). We also announced the survey on www.sprakteknologi.se and our project page.

The cover letter, explaining the aims of the project and giving details about the survey, was sent by e-mail and the questionnaire was made available on the Internet (http:www.speech.kth.se/prod/blark/blark.html) and as a text file downloadable from the same address. The users could choose between Swedish and English versions.

Once we had collected the answers in the first run, we sent out reminders to those that had not responded, and also contacted more people, recommended by the subjects.

After approximately 5 months, the inventory process was over and the answers of the web-based survey together with the answers acquired as text files were inserted into a MYSQL database, making it easy to gather statistics on the answers. Next, we will summarize our findings. The interested reader can find more detailed information about the survey and the results in our report (Elenius, et al., 2008).

3. Results

In total, we received 57 answers from 43 different places: 28 different companies, 4 public organizations, 11 universities, and one individual. The great majority of the answers arrived from Sweden while we collected 3 answers from Finland, one from Belgium/France, Denmark, Germany, Italy, Norway, Australia, and USA.

Main activity	Percent
Research	42%
Software development	39%
Teaching	28%
Interpreting; Translating; Localization	26%
Language technology product vendor	16%
Content provider	9%
Telecommunications	7%
Culture; Museum	4%
Minority language organization	4%
Other	19%

Table 1: Main activity.
57 or 100 % answered this question.

Table 1 shows the main activity of the 57 responding organizations.

Note that in all our result tables we give percentages of answers relative to the number that responded to the respective question. This number is given in the table caption as well as what percentage it corresponds to relative to all answers, 57.

Activities listed under *Other* are: dialog systems, multimodal systems, lexica, communication aids for people with disabilities, translation of medical texts, general text production, computer assisted language learning, text prediction, automatic summarization, text categorization, information retrieval and extraction, Swedish proof reading, language technology within education, phonology and phonetics.

The answers of the 55 organizations that answered the question regarding *Main language technology area* are shown in Table 2.

Main language technology area	Percent
Written technologies	51%
Language resources	49%
Machine and computer-assisted translation	33%
Search and knowledge mining	31%
Language learning	22%
Speech technologies	18%
Other	20%

Table 2: Main language technology area.
55 or 96 % answered this question.

The *Other* category includes: dialog systems, multimodal systems, translation, text production, language aids, building lexica, computer assisted language learning, text prediction, automatic summarization, text categorization, language and grammar checking, phonetics, and phonology.

3.1 Existing written language resources

We received at most 52 answers to the questions on existing written language resources. Most of the existing resources are Swedish monolingual resources, but as many of the subjects are in the translation business, bi- and multilingual resources also existed. Although we asked for resources that could fit into a Swedish BLARK, it is not clear that all reported resources could be made available to the public, whether for free or for a fee, or if they are programmatically available. Nor is the quantity and quality of all the reported resources known at this stage.

About two thirds of the subjects have some resources encoded in XML or SGML, and in an annotation standard such as the Text Encoding Initiative (TEI), XML Corpus Encoding Standard (XCES), or other standards.

A few monolingual lexical resources (lexica, term bases and semantically organized resources) of various sizes, levels of linguistic annotation and validation, and availability exist. For example, there exist lexical resources with morphological, pronunciation and semantic information, but, roughly speaking, the availability drops as the resource gets larger, more annotated and more validated.

Of the existing bi- or multilingual lexical resources including Swedish, the most reported other languages are Nordic languages, English, German, Japanese, Thai, and symbol languages, while the only specified specific domains are automotive technology and medicine. The Lexin series, for example, contains bilingual lexica from Swedish to major immigrant languages in Sweden (around 25).

When it comes to grammatical resources, 17 subjects said they have grammars, mostly language models, but also rule-based grammars.

Existing monolingual Swedish corpora include Swedish texts and transcribed speech from various genres and domains, Finland Swedish texts, non-native Swedish, and sign language.

Several subjects reported on bi- or multilingual (translation) corpora or translation memories involving Swedish. The following languages are specified as other languages: EU languages (and candidates), Hindi, Thai, Arabic, Persian, and Russian, while the most reported specific domains were EU texts, medicine, automotive literature, patents, and software documentation.

About two thirds of the subjects have specific or diversified genres, and sometimes balanced. Apart from news, documentation, reports, e-mail and chat, which we asked for, many also have texts from professional contexts, fiction, simplified/abridged texts, texts from teaching and writing, transcribed spoken language (mostly dialog), informative text, learner Swedish, web texts, historical texts, and rune texts.

Around 60% have some tools for processing, segmenting, annotating, parsing, classifying, aligning, and checking written language. The most common tools reported are part-of-speech taggers, tokenizers, morphological segmenters, and sentence splitters. Some also have evaluation resources for word-sense disambiguation (SENSEVAL), syntactic analysis, part-of-speech tagging, base form reduction, named entities, translation and summarization.

3.2 Existing spoken language resources

Out of the 57 answers to our survey at most 11 or 19 % answered questions regarding spoken language resources. They cover all sorts of speech and environments, such as telephone, microphone and radio speech, read and spontaneous speech, dialogs and multi-party speech. The gender distribution seems to be rather balanced. Regarding ages, speakers from 20 to 60 years appear in majority.

The gender distribution is rather balanced. A few corpora contain child speakers, some more contain adolescent speakers, but the majority contains adult speakers although speakers over 60 become more rare. A few corpora have good dialect coverage while others reflect the local dialect of the recording organization. There also exist corpora with bilingual and immigrant speakers.

The largest databases contain telephone speech and thousands of speakers. They are commercially attractive and comparatively uncomplicated to record.

Regarding annotation standards XML and SGML are most common. Commercial companies mostly use the Nuance standard. Some organizations use their own annotation standards.

The most frequently used speech tools deal with speech recording, analysis, recognition and synthesis.

Multimodal language resources, speech and video, are still not very common, but there is a growing interest in them.

3.3 Needed written language resources

We received at most 52 answers to the questions on needed written language resources. Most of them needed Swedish monolingual resources, but bi- and multilingual resources are also asked for. In particular, resources for evaluation, (lexical and frame) semantic resources, and domain- or genre-specific resources seem scarce. Some of the lexical resources and corpora needed are under production, or planned within 2-5 years.

About two thirds of the subjects need the resources to be encoded in XML or SGML, and in an annotation standard such as TEI, XCES, standards from the ISO TC37/SC4 group, or exchange formats such as Translation Memory eXchange format (TMX) and Term Base eXchange format (TBX).

Two kinds of monolingual lexical resources (lexica, term bases and semantically organized resources) are detailed as follows: 1) A base lexicon of about 10,000-100,000 entries, covering general language, and containing information on frequencies, inflection, word formation, and lexical semantics. 2) A Swedish WordNet of about 10,000-50,000 entries.

As for bi- or multilingual lexical resources including Swedish, the most wanted other languages are EU languages and symbol languages, while the most wished for specific domains are (general and automotive) technology and medicine. The resources should in general preferably contain more than 50,000 entries, and also include semantic information.

When it comes to grammatical resources, mostly language models but also rule-based grammars are needed.

Several subjects also ask for a large balanced Swedish corpus corresponding to a national corpus. It should contain about 100 million words (2 millions per genre and roughly 10 millions of transcribed speech), and include linguistic annotation and rich metadata. The entire corpus need not be equally much linguistically annotated, a minimum being (automatic) part-of-speech annotation, and about 10% need to be syntactically annotated. A few subjects also need more specific Swedish corpora, such as annotated errors, educational texts, full texts aligned with extracts/abstracts, and also a corpus with sign language.

Bi- or multilingual corpora or translation memories involving Swedish are also needed. The most wanted other languages were specified as EU languages and Oriental languages, while the most wished for specific domains were EU texts, medicine, and automotive literature. The resources should preferably contain between 1 million and 10 million words per language, and also include syntactic and semantic information.

Most subjects need specific or diversified genres, and often balanced. Apart from news, documentation, e-mail, reports, and chat, which we asked for, many also wish texts from professional contexts, fiction, simplified/abridged texts, texts from teaching and writing, transcribed spoken language (mostly dialog), informative text, and learner Swedish.

Answers to our question *Which tools do you need for processing written data?* are shown in Table 3 below.

Tools needed for written data	**Answers**
Morfological segmenter	61%
Sentence splitter	56%
Part-of-speech tagger	56%
Tokenizer	54%
Clause splitter	51%
Normalizer	49%
Parser	49%
Formatter	44%
Lexical semantics analyzer	44%
Named entity recognizer	41%
Chunker	37%
Text/Genre classifier	34%
Word aligner	34%
Optical character recognition	32%
Formal semantics analyzer	32%
Term extractor	29%
Sentence aligner	29%
Generator	24%
Discourse segmenter	20%
Identifier of attitudinal expressions	20%
Other	22%

Table 3: Tools needed for written data.
41 or 72 % answered this question.

Basic tools such as morphological segmenters, sentence splitters, part-of-speech taggers, and tokenizers are the most wanted tools. Interestingly, the most wanted tools are also reported as already existing. This might have to do with poor quality or poor reusability of the existing tools, or that the subjects misunderstood the question in section 3.1 as "resources that you use (and need)", not "resources that you have (and have the right to distribute)", and for these reasons wanted the resources included in the BLARK.

3.4 Needed spoken language resources

At most 16, or 28 %, answered our questions regarding needs for spoken language resources.

Recorded speech is indispensable for the speech technology field. Although read speech is still the mostly required, there is a growing and marked interest also regarding spontaneous speech, dialog speech and multi-party speech.

As to speakers, the major demand is for adult speech and equal need of both genders. There is, however, also an interest in child and adolescent speech as well as speech from elderly people. Some declare a wish for second language speakers of Swedish. Good dialect coverage is also mentioned.

Regarding speech quality, the greatest need is for telephone speech, reflecting the needs of telephony based voice response companies, but also wideband and radio speech were brought up.

Pronunciation lexica and language models are naturally required. XML is preferred for annotation.

Table 4 lists the answers to the question *Which tools do you need for processing speech data?*

Tools needed for spoken data	**Answers**
Orthographic labeling of speech	64%
Text-to-speech	64%
Speech recording	57%
Phonetic labeling of speech	50%
Linguistic labeling of speech	50%
Speech synthesis with augmented control	50%
Checking of recording	43%
Automatic speech analysis	43%
Automatic phonetic segmentation	36%
Speech recognition-a couple of thousand words	36%
Pragmatic labeling of speech	29%
Speech recognition - few words	29%
Speech recognition - dictation	29%
Speech response with prerecorded speech	21%
Speaker recognition	7%

Table 4: Tools needed for spoken data.
14 or 25 % answered this question
.

Tools for measuring speech recognition performance and the quality of voice controlled services were also brought up, as well as tools for the evaluation of dialogs.

Our study shows that the demand for multimodal resources, speech and video, definitely is growing. Thus, there also exists a demand for tools for the annotation of gestures, face mimics and eye movements.

4. Conclusion

We gave a brief summary on our study investigating existing written, spoken and multilingual language resources and tools for Swedish, and the need for these, collected from industry, organizations and academia. We can conclude from the 57 answers that although many resources exist for Swedish, there is a need for freely available standardized resources and tools to be used both by industry and academia.

Acknowledgements

We would like to express our gratitude to Rolf Carlson for his engagement, support and wise advice throughout the project. We are also grateful to the members of the Swedish Graduate School of Language Technology who actively supported our project. A special thanks goes to all subjects who participated in the survey. The project was supported by the Swedish Research Council's Committee for Research Infrastructures 2007-2008.

References

Elenius, K., Forsbom, E., and Megyesi, B. 2008. *Survey on Swedish Language Resources.* Report. Dept. of Speech, Music and Hearing, CSC, KTH and Dept. of Linguistics and Philology, Uppsala University. http://stp.lingfil.uu.se/blark/swe-blark-survey-2008.pdf

Krauwer., S. (1998). ELSNET and ELRA: *A common past and a common future.* ELRA Newsletter, 3(2). http://www.elda.org/blark/fichiers/elsnet&elra.doc

Nikkhou, M. and Choukri, K. 2005. *Report on Survey on Arabic Language Resources and Tools in the Mediterranean Countries*. http://www.nemlar.org/Survey-questionnaires/index.htm.

Strangert, E., 2007. Vetenskapsrådets kartläggning av språkteknologiska databaser och framtida behov. http://sprakteknologi.se/dokument/disc-rapport-om-sprakteknologi.pdf

Strik, H., Daelemans, W., Binnenpoorte, D., Sturm, J., de Vriend, F. and Cucchiarini, C. Dutch HLT resources: From BLARK to priority lists. In *Proceedings of ICSLP, Denver*, pp. 1549 – 1552, 2002.