# Survey on Swedish Language Resources

Kjell Elenius[1]

Eva Forsbom[2]

Beáta Megyesi[2]

[1]Speech, Music and Hearing
School of Computer Science and Communication, KTH
Sweden

[2]Department of Linguistics and Philology
Uppsala University
Sweden

2008-03-14

# Abstract

Language resources, such as lexicons, databases, dictionaries, corpora, and tools to create and process these resources are necessary components in human language technology and natural language applications. In this survey, we describe the inventory process and the results 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 an investigation sent to industry and academia, institutions and organizations, to experts involved in the development of Swedish language resources in Sweden, the Nordic countries and world-wide. This study is a result of the project called “An Infrastructure for Swedish language technology” supported by the Swedish Research Council´s Committee for Research Infrastructures 2007 - 2008.

# Table of Contents

# 1. Introduction

Research and development of language technology systems needs an infrastructure of publicly available and standardized basic resources for the Swedish language. 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. 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 for each language separately. 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 present the participating countries and organizations in our study. In Section 4 and 5 we summarize the results of the inventory. This involves a description of existing written, spoken and multimodal language resources, and is followed by a summarization of the needs for such resources. Lastly, in Section 6 we summarize all our findings. In general, the sections give overviews, while the appendices contain more detailed descriptions and/or data of the resources.

# 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 the existing Swedish language resources and needs, we developed a questionnaire inspired by previous surveys carried out for Arabic within the NEMLAR project (Nikkhou and Choukri, 2005). Below we give an overview of the content of the questionnaire.

## *2. 1 Questionnaire*

The entire questionnaire can be found in the Appendix A. As mentioned previously, the questionnaire 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, each dealing with different issues.

1. The first part of the questionnaire contains information about the person who filled out the questionnaire.
2. The second part aims to find out information about the actual organization, institution and/or individual, the number of employees, the main activity on research and language technology area, and the languages covered in the product/services.
3. The third part gathers information about existing Swedish language resource needs, divided in three subparts regarding written, spoken, and multimodal resources, respectively. In the case of written resources, the questions concern the need for lexical and terminological databases, corpora, grammars, semantic networks, annotation standards, needs for evaluation resources, and tools for processing written data. For spoken and multimodal data, we ask about the needs for speech types, database genres/types, speakers, bandwidths, annotation standards, needs for evaluation resources, and tools.
4. The fourth part deals with existing Swedish language resources. This section is also divided into written, spoken, and multimodal data, similarly to part three containing the same type of information. In addition, we also ask about other types of existing resources, and details about these, as well as available guidelines, standards, benchmarks, and validation. Here, we also give the possibility to leave comments on other resource needs and details about these.
5. The fifth part concerns with the acquisition of the 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. Also, there are questions about the reasons for not acquiring language resources.
6. In part six, general comments on the questions and/or on the resources can be left by the subjects.
7. Lastly, there is a possibility for the subjects to give us further suggestions on other contacts that might be interested in participating in the study.

## *2. 2 Survey*

In order to obtain a reliable survey, it is of great importance, that as many people as possible who work with Swedish language technology get an opportunity to participate in the inventory process. Therefore, we emailed 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 email 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 that 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), 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 can be found in the Appendix A. 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 were analyzed. We also added answers concerning details about resources from two previous investigations on language resources which were collected by Eva Strangert at DISC (Strangert, 2007) and by the ENABLER (European National activities for Basic Language Resources) project.

The answers of the web-based survey were automatically inserted into a MYSQL database. All answers acquired as text files were added to the database as well as the DISC and ENABLER data. A web-based interface made it easy to gather statistics on the answers.

There were a total of 57 answers to the survey and including the additional DISC and ENABLER surveys we received 71 answers in total. However, all statistics regarding our results were based on the 57 answers received to our questionnaire. Most results are given in tables as in the Table Example below. In the caption we see how many answered the question, i.e. how many responded to at least one multi-choice question or “pressed” at least one radio button. Also given is what percentage this corresponds to relative to all 57 answers. In the table we list all sub questions and the number of answers to each of them and also the percentage of answers relative to all 57 that answered our questionnaire. If there is a multi choice option “Other” in a question the answers to it are listed below the table.

*Table Example. 56 or 98 % answered this question.*

| **Type of organisation** | **Answers** | **Percent** |
|---|---:|---:|
| Company | 31 | 54,4% |
| University | 20 | 35,1% |
| Public organisation | 5 | 8,8% |

# 3. Results

In this section, we will give an overview over the participating organizations, institutions, and universities who answered the questionnaire.

## 3.1 Information about the organizations

The 57 answers to the survey came from 43 different places: 28 different companies, 4 public organizations, 11 universities, and one individual without affiliation. Adding the DISC and ENABLER data gave us 71 answers to the resource questions in total, as mentioned above.

The following companies, universities, organizations, and individuals answered the questionnaire:

*Companies*

- Acapela Group
- Artificial Solutions
- ASIMUS AB
- CSC - Scientific Computing Ltd.
- DeLaval International AB
- ESTeam AB
- Fodina Language Technology AB
- FreeLanguage
- Inga-Beth Hinchliffe AB Språkman plc
- Katherine Stuart The Right Word
- Lexware Labs
- Lufkin Colleagues
- Master's Innovations Ab
- Mikro Værkstedet a/s
- Rehabmodul AB
- Scania CV AB
- Language Laboratory (Språklabbet)
- Språkservice Ingela Dellby/ Ninsun bok AB
- Södermalms talteknologiservice AB
- Textwerkstatt i Göteborg AB
- TransFalk AB
- Transmachina AB
- Viking
- Vocab AB
- Voice Provider Sweden AB
- Walters Lexikon
- Åkeson Språk & Energi HB
- Översätt Mera

*Organisations*

- Language Council of Sweden, Department of the official language authority, The Institute of Language and Folklore (Språkrådet vid Institutet för språk och folkminnen)
- Swedish Board of Agriculture (Jordbruksverket)
- Swedish Maritime Administration (Sjöfartsinspektionen)
- The Swedish Library of Talking Books and Braille, TPB (Talboks- och punktskriftsbiblioteket)

*Universities*

- Göteborg University
- KTH
- Linköping University
- Lund University
- Sahlgrenska University Hospital
- Stockholm University
- Umeå University
- University of Helsinki
- University of Tromsø
- Universität Leipzig
- Växjö University

*DISC-report*

- Computer Science and Engineering, Chalmers
- General linguistics, Department of linguistics, Göteborg University
- Språkbanken, Göteborg University
- Speech, Music and Hearing, School of Computer Science and Communication, KTH
- Language Technology Group, NADA, School of Computer Science and Communication, KTH
- Department of Computer and Information Science, Linköping University
- NLPLab, Linköping University
- Department of linguistics and phonetics, Lund University
- Computational linguistics, Department of linguistics, Stockholm University
- Department of linguistics, Stockholm University (general linguistics, phonetics, and sign language)
- General linguistics, Department of philosophy and linguistics, Umeå University
- Computational linguistics, Department of linguistics and philology, Uppsala University
- Linguistics, Department of linguistics and philology, Uppsala University
- Faculty of languages, Uppsala University
- The School of Mathematics and Systems Engineering, Växjö University

*ENABLER*

- Göteborg University (Dimitrios Kokkinakis)

### 3.1.1 Country of origin

The great majority of the answers arrived from Sweden while we collected 3 answers from Finland, one from each of Belgium/France, Denmark, Germany, Italy, Norway, Australia, and USA.

### 3.1.2 Type of organization

Most of the collected answers came from companies and universities, while a few arrived from public organizations.

*Table 1. 56 or 98 % answered this question.*

| Type of organisation | Answers | Percent |
|---|---:|---:|
| Company | 31 | 54,4% |
| University | 20 | 35,1% |
| Public organisation | 5 | 8,8% |

### 3.1.3 No of employees

Most answers originated from small companies with less than 10 employees but organizations with more than 100 employees were also frequent.  It is important to note, that subjects may have given the total number of employees, and not necessarily only those involved in language technology.

*Table 2. 53 or 93 % answered this question.*

| Number of employees | Answers | Percent |
|---|---:|---:|
| More than 100 | 18 | 31,6% |
| 50-99 | 4 | 7,0% |
| 10-49 | 7 | 12,3% |
| Less than 10 | 24 | 42,1% |

### 3.1.4 Main activity

Concerning main activity a majority is involved in research and software development. Many work with teaching or interpreting/translating/localization. Many companies work as language technology product vendors, content providers and in telecommunication.

*Table 3. 57 or 100 % answered this question.*

| Main actitvity | Answers | Percent |
|---|---:|---:|
| Research | 24 | 42,1% |
| Software development | 22 | 38,6% |
| Teaching | 16 | 28,1% |
| Interpreting;Translating;Localisation | 15 | 26,3% |
| Language technology product vendor | 9 | 15,8% |
| Content provider | 5 | 8,8% |
| Telecommunications | 4 | 7,0% |
| Culture;Museum | 2 | 3,5% |
| Minority language organisation | 2 | 3,5% |
| E-commerce | 0 | 0,0% |
| Banking;Insurance | 0 | 0,0% |
| Other, please specify: | 11 | 19,3% |

### 3.1.5 Main language technology area

Main language technology areas are shown below.

*Table 4. 55 or 96 % answered this question.*

| Main language technology area | Answers | Percent |
|---|---:|---:|
| Written technologies | 28 | 49,1% |
| Language resources | 27 | 47,4% |
| Machine translation;Computer-assisted translation | 18 | 31,6% |
| Search and knowledge mining | 17 | 29,8% |
| Language learning | 12 | 21,1% |
| Speech technologies | 10 | 17,5% |
| Other, please specify: | 11 | 19,3% |

*Others specified were:*

– dialog systems
– multimodal systems
– translation

- text production
- language aids
- building lexicons
- computer assisted language learning
- text prediction
- automatic summarization
- text categorization
- language and grammar checking
- phonetics
- phonology.

### 3.1.6 Main products/services

To the question what the main products/services are, we got two types of answers. The majority answered the main activity, similar to the question about the main activity presented in Section 3.1.4, instead of specifying the main tool or the specific service they offer. In this section, we only present the main products in general that were identified. Specific resources or tools can be found in the Appendices together with other existing resources and tools.

*Resources:*

- Mono- and multilingual lexicon (6)
- Ontologies (2)
- Corpora (3)
- Parallel corpora (2)
- Parallel treebanks
- Database for dialog management

*Tools:*

- Corpus tools (3)
- PoS-Tagger (3)
- Morphological segmenter
- Lemmatizer
- Named entity recognizer
- Parser (2)
- Grammar rules

*Services/Areas:*

- Computer assisted language learning (2)
- Copytext
- Cross-lingual question answering
- Grammar checking (2)
- Language council
- Proofreading
- Machine Translation Systems (2)
- Speech recognition
- Technical documentation
- Text clustering (2)
- Text-to-speech-system (2)
- Text summarization (2)
- Translation (as service) (7)
- Tools for computer assisted translation (2)

- Text matching to find textual variation

### 3.1.7 Language coverage

The monolingual language technology products mostly deal with Swedish, but we can also find products covering Finnish, Sami, and Meänkieli which are official languages in Sweden.

*Table 5. 49 or 86 % answered this question.*

| **Monolingual language coverage** | **Answers** | **Percent** |
|---|---:|---:|
| Swedish | 46 | 80,7% |
| Finnish | 8 | 14,0% |
| Sami | 2 | 3,5% |
| Meänkieli | 1 | 1,8% |
| Jiddisch | 0 | 0,0% |
| Romani chib | 0 | 0,0% |
| Other, please specify: | 18 | 31,6% |

*Other monolingual languages*:

- English
- German
- Danish
- Norwegian (Bokmål and Nynorsk)
- Faeroese
- Sign language
- French
- Middle French
- Spanish
- Russian
- Swahili
- Xhosa
- Nepali

Among multilingual language products, Swedish constitutes a part in a great majority of cases together with the other Germanic languages.

*Table 6. 34 or 60 % answered this question.*

| **Multlingual Language coverage** | **Answers** | **Percent** |
|---|---:|---:|
| Swedish | 29 | 50,9% |
| Finnish | 6 | 10,5% |
| Meänkieli | 1 | 1,8% |
| Romani chib | 1 | 1,8% |
| Sami | 1 | 1,8% |
| Jiddisch | 0 | 0,0% |
| Other, please specify: | 22 | 38,6% |

*Other multilingual languages:*

- English
- German
- Danish
- Norwegian (Bokmål and Nynorsk)
- French
- Spanish

- Russian
- Dutch
- Turkish
- Hindi
- Sign language and symbol languages such as Bliss, PCS, Rebus, and Piktogram

# 4. Information on existing language resources

See Appendix B for details on some of the existing resources, where details were given. In some cases, the meaning of “existing resources” and “resources that you have” were confused with “resources that you use” while we meant “resources that you own”, so that the resources reported are not actually owned by those who reported them. But, none the less, they do exist. These resources are listed in parenthesis. Furthermore, some subjects reported on resources not involving languages we asked for, but we did not include those resources in this report, as they will not be of interest to the Swedish BLARK, however interesting they are on their own.

## *4.1 Written language resources*

Numbers on existing mono and multilingual lexical databases, various types of corpora, terminological databases, grammars, and semantic networks are given in the tables below.

*Table 7. 32 or 56 % answered this question.*

| Your lexical databases | Answers | Percent |
|---|---:|---:|
| Monolingual | 22 | 38,6% |
| Bilingual | 15 | 26,3% |
| Multilingual | 9 | 15,8% |

*Table 8. 34 or 61 % answered this question.*

| Your corpora | Answers | Percent |
|---|---:|---:|
| Monolingual | 25 | 43,9% |
| Bilingual | 13 | 22,8% |
| Parallel | 11 | 19,3% |
| Translation memories | 9 | 15,8% |
| Enriched (annotated with e.g. tags, clusters) | 7 | 12,3% |
| Multilingual | 6 | 10,5% |
| Equivalent (non-parallel but same domain) | 6 | 10,5% |

*Table 9. 20 or 35 % answered this question.*

| Your terminological databases | Answers | Percent |
|---|---:|---:|
| Bilingual | 12 | 21,1% |
| Monolingual | 10 | 17,5% |
| Multilingual | 7 | 12,3% |

*Table 10. 17 or 30 % answered this question.*

| Your grammars | Answers | Percent |
|---|---:|---:|
| Rule-based | 13 | 22,8% |
| Language models | 7 | 12,3% |

**Semantic networks**

9 or 16 % had semantic networks (wordnets, thesauri, ontologies, etc.).

### 4.1.1 Genres

The distribution of the reported existing genres is shown in the table below.

*Table 11. 31 or 54 % answered this question.*

| Your genres | Answers | Percent |
|---|---:|---:|
| News | 13 | 22,8% |
| Documentation | 12 | 21,1% |
| Reports | 7 | 12,3% |
| Balanced | 7 | 12,3% |
| Chat | 1 | 1,8% |
| E-mail | 0 | 0,0% |
| Other, please specify: | 14 | 24,6% |

*Other genres:*

- Fiction (e.g. novels) (4)
- Texts on medicine (e.g. asthma) (3)
- Second language Swedish (3)
- Texts in immigrant languages (e.g. news) (2)
- Non-fiction (2)
- Text books (2)
- Children's literature
- Film subtitles
- Economy texts
- Historical texts
- Informative texts
- Texts on environment (e.g. Agenda 21)
- Texts on health
- Patent texts
- Rune texts
- School texts, young writer's writings
- Sentence fragments with errors and corrections
- Sign language
- Statements of government policy
- Technical manuals (e.g. automotive service literature)
- Think-aloud protocols
- Web texts
- Transcribed speech (from approx. 30 social activity types)
- Intercultural communication
- Links/databases
- Voice-controlled services via telephone

### 4.1.2 Annotation standards

*Table 12. 25 or 44 % answered this question.*

| Your annotations standards | Answers | Percent |
|---|---|---|
| XML | 19 | 33,3% |
| TEI | 6 | 10,5% |
| XCES | 6 | 10,5% |
| SGML | 2 | 3,5% |
| Other, please specify: | 7 | 12,3% |

*Other standards:*

- Corpus Workbench (CWB)
- Functional Morphology (FM) and Grammatical Framework (GF), convertible to other formats (XML, SQL, LEXC, GSL(Nuance))
- Göteborg Transcription Standard (GTS)
- Internal formats (2)
- Nuance' format for speech transcription
- Resource Description Framework (RDF)
- Ruby Annotation
- Tiger-XML
- Web Ontology Language (OWL)
- XML Corpus Encoding Standard (XCES, subset)

### 4.1.3 Resources for evaluation

- The following resources for evaluation exist:
- Analyzer and corpus texts for North Sami
- MaltEval (software for evaluation of dependency-based syntactic analysis)
- MT Quality Evaluation Toolbox (prototype for evaluation of translation quality and meta-evaluation of evaluation measures)
- Stockholm-Umeå Corpus (part-of-speech, baseform, named entities)
- Test data for baseform reduction and wordform generation
- Text summarization test data
- Translation test data
- Treebanks (2)

More information is available in Appendix C: Evaluation resources.

### 4.1.4 Tools for processing language data

Of the 57 subjects, 34 (60%) had tools for processing written language. In case details were given, they can be found in Appendix B.

*Table 13. 34 or 60 % answered this question.*

| **Which tools do you have for processing written data?** | **Answers** | **Percent** |
|---|---:|---:|
| Part-of-speech tagger | 16 | 28,1% |
| Tokenizer | 14 | 24,6% |
| Morfological segmenter (stemmer, lemmatiser, compound analyser, etc.) | 13 | 22,8% |
| Sentence splitter | 11 | 19,3% |
| Normalizer (upper;lower case, numeric expressions, etc.) | 10 | 17,5% |
| Formatter (character encoding, file format, etc.) | 9 | 15,8% |
| Clause splitter | 8 | 14,0% |
| Optical character recognition | 7 | 12,3% |
| Parser | 7 | 12,3% |
| Named entity recognizer | 6 | 10,5% |
| Sentence aligner | 6 | 10,5% |
| Chunker | 5 | 8,8% |
| Generator | 5 | 8,8% |
| Word aligner | 5 | 8,8% |
| Term extractor | 4 | 7,0% |
| Lexical semantics analyzer (word sense disambiguation, etc.) | 3 | 5,3% |
| Text;Genre classifier | 2 | 3,5% |
| Identifier of attitudinal expressions (attitudes, opinions, feelings, etc.) | 1 | 1,8% |
| Discourse segmenter | 0 | 0,0% |
| Formal semantics analyzer (reference resolution, etc.) | 0 | 0,0% |
| Other, please specify: | 10 | 17,5% |

*Other tools:*

- Language recognizer (2)
- Dialog system modules
- Grammar checker
- Spell checker
- Part-of-speech tagger
- Dependency parser
- Word predictor
- Phrase aligner
- Semantic tagging support tool
- Text summarizer
- Tools for analyzing transcribed speech
- (Nuance' parser and pronunciation generator)
- (DAM-LR: Distributed Access Management of Language Resources)
- (Concordancer)

## *4.2 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, dialogues and multi-party speech. The gender distribution seems to be rather balanced. Regarding ages, speakers from 20 to 60 years seem to be in majority.

### 4.2.1 Type of spoken resources

*Table 14. 8 or 14 % answered this question.*

| Your speech types | Answers | Percent |
|---|---|---|
| Read speech | 4 | 50% |
| Spontaneous speech | 3 | 38% |
| Prompted speech | 3 | 38% |
| Dialogue speech | 3 | 38% |
| Multi-party speech | 2 | 25% |
| Other | 3 | 38% |

### 4.2.2 Speakers

Most speech is from adult speakers of ages 20 to 60 years. But there are also a significant amount of recordings for children, from age 4, and elderly people. There is no significant difference between the genders.

*Table 15. 7 or 12 % answered this question.*

| Your speakers | Answers | Percent |
|---|---|---|
| Male | 7 | 100,0% |
| Female | 6 | 85,7% |

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

### 4.2.3 Bandwidths

The databases with most speakers contain telephone speech. They are commercially attractive and comparatively simple to record, since the speakers can call from a distance.

*Table 16. 9 or 16 % answered this question.*

| Your bandwidths | Answers | Percent |
|---|---|---|
| Telephone (fixed, mobile etc.) | 6 | 66,7% |
| Wide-band microphone | 4 | 44,4% |
| Broadcast news | 3 | 33,3% |
| Other | 3 | 33,3% |

### 4.2.4 Database Genres/Types

Pronunciation lexica are essential for both text-to-speech synthesis and automatic speech recognition as well as language models.

*Table 17. 8 or 14 % answered this question.*

| Your database genres/types | Answers | Percent |
|---|---|---|
| Pronunciation lexicon | 5 | 62,5% |
| Language model | 4 | 50,0% |
| Other | 1 | 12,5% |

### 4.2.5 Annotation standards

*Table 18. 5 or 9 % answered this question.*

| Your annotations standards | Answers | Percent |
|---|---:|---:|
| XML | 3 | 60,0% |
| SGML | 2 | 40,0% |
| SAM | 1 | 20,0% |
| NIST SPHERE/LDC | 1 | 20,0% |
| Other | 3 | 60,0% |

Annotation of speech is considered a time consuming task. This is especially true for spontaneous speech and annotation at the phoneme or syllable level. Thus it requires good tools. It seems that no annotation standard is predominant. Swedish Interactive Voice Response companies mostly use the Nuance speech recognizer and thus the Nuance annotation scheme. A few sites use their own transcriptions.

### 4.2.6 Resources for evaluation

This question resulted in only a couple of answers indicating that these resources are scarce.

### 4.2.7 Tools for processing speech data

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

*Table 19. 11 or 19 % answered this question.*

| Which tools do you have for processing speech data? | Answers | Percent |
|---|---:|---:|
| Speech recording | 9 | 81,8% |
| Text-to-speech | 8 | 72,7% |
| Checking of recording | 6 | 54,5% |
| Automatic speech analysis | 5 | 45,5% |
| Orthographic labeling of speech | 3 | 27,3% |
| Phonetic labeling of speech | 3 | 27,3% |
| Automatic phonetic segmentation | 3 | 27,3% |
| Speech recognition - few words | 3 | 27,3% |
| Speech recognition - a couple of thousand words | 3 | 27,3% |
| Linguistic labeling of speech | 2 | 18,2% |
| Speaker recognition | 2 | 18,2% |
| Speech synthesis with augmented control | 2 | 18,2% |
| Pragmatic labeling of speech | 1 | 9,1% |
| Speech recognition - dictation | 1 | 9,1% |
| Speech response with prerecorded speech | 1 | 9,1% |
| Other | 2 | 18,2% |

## *4.3 Multimodal language resources*

Multimodal language resources, speech and video recordings, are still not very common as may be seen from our results below, but there is certainly a growing interest in them.

### 4.3.1 Database Genres/Types

The 4 answers reported 2 language models and 1 pronunciation lexicon. There are also databases for Augmentative and Alternative Communication, AAC, needs as well as for the Swedish sign language.

### 4.3.2 Speakers

Only 4 speakers were reported: 2 female and 2 male.

### 4.3.3 Annotation standards

Standards reported were: 2 XML, 1 SGML and 1 SAM. Also other standards were mentioned: GTS, Göteborg Transcription Standard, MUMIN, the W3C's Resource Description Framework (RDF) and Web Ontology Language (OWL).

### 4.3.4 Resources for evaluation

There was no answer to this question.

### 4.3.5 Tools for processing multimodal data

Only 3 answered this multi choice question. But all asked for tools were reported: speech recording, checking of recording, orthographic labeling of speech, phonetic labeling of speech, linguistic labeling of speech, pragmatic labeling of speech, system for movement measurements. Also a couple of special tools were reported.

## *4.4 Other language resources*

Of the 57 subjects, 18 (32%) answered the question on whether they had other language resources than the ones given: no (11), and yes (7). All the resources specified were not actually "other" resources, but resources we asked for above, and the details of those resources are described under their corresponding headings. The "other" resources left are symbol databases, if they are not counted as multimodal resources, and a compound database, if it is not counted as a lexicon.

*Table 20. 18 or 32 % answered this question.*

| Do you have other language resources? | Answers | Percent |
|---|---:|---:|
| Yes | 7 | 12,3% |
| No | 11 | 19,3% |

## *4.5 Production of language resources*

Of the 32 that answered the question on where their language resources were produced, all used internally produced resources. Some also used resources produced by specific contracted vendors and resources distributed by data centres.

*Table 21. 32 or 56 % answered this question.*

| Do you use language resources | Answers | Percent |
|---|---:|---:|
| produced internally? | 32 | 56,1% |
| produced by specific contracted vendors? | 12 | 21,1% |
| distributed by data centres? | 12 | 21,1% |

*When producing their own resources, they used the following resources:*

- Self-produced tools (20)
  - Unix/Linux-based (3)
  - Perl-based (2)
  - .NET-based
  - alignment tools for parallel texts (4)
    - Stockholm TreeAligner
    - UPlug, preprocessing and alignment
  - C#-based
  - Database tools for lexicons
  - Dialogue tools
    - Collection of dialogues
    - Dialogue system development tools
  - Functional Morphology (FM), generation of lexicons
  - Grammatical Framework (GF), generation of speech recognition models
  - Lexicon Extraction
  - MS Visual Studio-based
  - MS Accessible-based
  - Python-based
  - Scania Checker, grammar checker
  - Web-based
  - WINGS, XML-based authoring tool
  - Word macros
- Off-the-shelf resources (15)
  - Annotation tools, manual or automatic (7)
    - part-of-speech taggers (3)
      - Connexor
    - syntax (3)
      - Annotate, treebank editor
      - vislcg, VISL constraint grammar parser/compiler
    - lemmatisation
    - morphology
      - lexc, two-level lexicon compiler
      - twolc, two-level morphology compiler
    - semantics
      - Salsa, frame-semantic editor
  - Computer-aided translation tools (4)
    - Trados Translator's Workbench (3)
    - Heartsome TMX Editor
    - Interverbum TermWeb for terminology
    - Wordfast, translation tool
    - Wordfinder lexicons
  - Audio and video tools, e.g. recording and annotation (3)
    - Multitool
    - Nuance
  - Adobe Framemaker
  - cat_text
  - TextWrangler, text editor
  - XCES
  - XFST, Xerox Finite State Tool

## *4.6 Validation of language resources*

This section deals with the validation of language resources concerning the production of guidelines for validation, the usage of standards, and benchmarks.

### 4.6.1 When producing language resources do you follow specific guidelines?

*Table 22. 29 or 51 % answered this question.*

| Do you follow specific guidelines? | Answers | Percent |
|---|---:|---:|
| Yes, we use internal specifications | 28 | 49,1% |
| Yes, we use external specifications, please specify: | 5 | 8,8% |

*The external specifications included the following:*

- Treebank guidelines:
  Penn Treebank
  German TIGER

- Guidelines for speech data:
  IPA
  Nuance, transcription
  SAMPA
  Speechdat
  Speecon

### 4.6.2 Do you follow specific standards?

*Table 23. 30 or 53 % answered this question.*

| Do you follow specific standards? | Answers | Percent |
|---|---:|---:|
| Yes | 20 | 35,1% |
| No | 10 | 17,5% |

*Standards specified were the following:*

- TEI (Text Encoding Initiative) (5)
  (X)CES ((XML) Corpus Encoding Standard) (4)
- XML (4)
- Internal standards (2)
- Translation standards (2)
  OLIF
  TBX
  TMX
- CA, transcription of conversations
- GTS
- IPA
- Multimodal coding
- PAROLE
- SAM
- SAMPA
- TIGER-XML, treebanks
- UTF-8

### 4.6.3 Do you use specific gold standards/benchmarks?

*Table 24. 24 or 42 % answered this question.*

| **Do you use specific gold standards/benchmarks?** | **Answers** | **Percent** |
|---|---:|---:|
| Yes | 6 | 10,5% |
| No | 18 | 31,6% |

*Standards/Benchmarks specified included the following:*

- Annotated corpora/treebanks (4)
    - SUC 2.0
    - Talbanken
    - Reference data for specific projects, e.g translations (2)
- Type control
- Speechdat
- Speecon

### 4.6.4 Are your LRs validated?

*Table 25. 30 or 53 % answered this question.*

| **Are your language resources validated?** | **Answers** | **Percent** |
|---|---:|---:|
| Yes | 18 | 31,6% |
| No | 12 | 21,1% |

Of the 18 that had validated resources, all had their resources validated in-house, and 3 had resources validated by independent/external organizations/experts.

# *5. Needs for language resources*

As was the case for existing resources, some subjects reported needs for resources not involving the languages we asked for, and those resources are not included in the summaries below.

## *5.1 Needs for written language resources*

Needs for lexical databases, terminological databases, semantic networks, corpora and grammars are given below.

**Needs for lexical databases**

*Table 26. 52 or 91 % answered this question.*

| **Needs for lexical databases** | **Answers** | **Percent** |
|---|---:|---:|
| Monolingual | 40 | 70,2% |
| Bilingual | 38 | 66,7% |
| Multilingual | 32 | 56,1% |

**Needs for terminological databases**

*Table 27. 40 or 70 % answered this question.*

| **Needs for terminological databases** | **Answers** | **Percent** |
|---|---:|---:|
| Monolingual | 28 | 49,1% |
| Bilingual | 29 | 50,9% |
| Multilingual | 23 | 40,4% |

**Needs for semantic networks**

*Table 28. 32 or 56 % answered this question.*

| **Needs for semantic networks** | **Answers** | **Percent** |
|---|---:|---:|
| Semantic networks (wordnets, thesauri, ontologies, etc.) | 32 | 56,1% |

*The following details were given (number of answers in parentheses):*

– Monolingual lexicons, term bases, semantic nets (4):
  Base lexicon, 10,000-100,000 entries (single- and multi-word units), general language, with information on frequencies, inflection, word formation, and lexical semantics (3)
  WordNet, 10,000-50,000 entries (2)

– Multilingual lexicons, term bases, semantic nets (9):
  Bi-/Multilingual lexicons (2)
  Swedish-English-German-French dictionaries with > 50,000 entries
  Swedish-Danish-Norwegian dictionaries with > 50,000 entries
  English-Swedish lexicons with semantic information (2)
  Multilingual concept-coded lexicon > 10,000 entries (most frequent from a balanced corpus)
  Automotive lexical databases
  Medical Swedish-English-German lexicons/term bases
  Technical lexicons/term bases

### Needs for Corpora

*Table 29. 51 or 89 % answered this question.*

| Needs for Corpora | Answers | Percent |
|---|---|---|
| Monolingual | 38 | 66,7% |
| Bilingual | 29 | 50,9% |
| Parallel | 29 | 50,9% |
| Enriched (annotated with e.g. tags, clusters) | 27 | 47,4% |
| Multilingual | 24 | 42,1% |
| Equivalent (non-parallel but same domain) | 20 | 35,1% |
| Translation memories | 20 | 35,1% |

*Detailed information included the following (number of answers in parentheses):*
Large, balanced (genre (facetted), (full) text/(transcribed) speech, synchronous/diachronous), monolingual corpus (10 million-1 billion words, 2 millions per genre) (12):

- Unannotated
- Annotated
  - Metadata (e.g. for splitting into subcorpora)
  - Linguistically annotated (18)
    - Part-of-speech tagged (5)
    - Baseforms
    - Named entities
    - Semantically annotated (e.g. FrameNet) (4)
    - Syntactically annotated (e.g. phrase structure and dependency relations) (5)
      - Treebank (2)
    - Discourse annotated (e.g. logical structure, utterance turns)

For the large balanced Swedish corpus, most subjects suggested a size of 100 million words, not all of which need to be equally much linguistically annotated, a minimum being (automatic) part-of-speech annotation. At least 10 million words of transcribed speech is needed (see details on speech resources), and about 10% need to be syntactically annotated.

- Specific monolingual corpora (7):
  - (Annotated) error corpus (2)
  - Parallel corpus of full texts and extracts/abstracts (larger, and more genres, than KTH eXtract) (2)
  - Compound corpus, with annotations on compound boundaries
  - (Annotated) humour corpus
  - Educational corpus (text, speech, teaching material and text books from various levels and disciplines)
- Sign language corpora (“grammatically” annotated)
- Multilingual (translation) corpora (11):
  - Domain-specific parallel Swedish-English corpora with semantic annotation (2)
  - Parallel corpora with EU texts (2)
  - Parallel Swedish-English(-German) corpora with medicine texts (2)
  - Parallel corpora for Swedish-English-Oriental languages > 2-3 million tokens each (300,000 sentence pairs)
  - Parallel corpora English-Swedish-German-French with > 10 million tokens each
  - Parallel corpora Swedish-English-German-French-Estonian/Finnish with syntactic and semantic annotations
  - Parallel corpora with automotive literature (manuals and functional descriptions)

In the DISC report, NLPLab made some comments on the size, content, and cost of a possible national translation corpus. To be of any use, the linguistically annotated corpus should span around 10 language pairs (both directions), 10 genres of about 500,000 source tokens each (approx. 15,000 source tokens per ouvre). A realistic cost estimation for such a corpus would be 1 million SEK per language pair (1.5-2 man years). As for corpora in general, copyright issues are the major obstacles for making corpora publicly available.

- "Minority" language resources (3):
  - Monolingual (balanced) corpora for (South/Swedish) Sami and Meänkieli (3)
  - Parallel corpora for (South/Swedish) Sami and Meänkieli (Swedish, Finnish, English) (2)
  - Bi-/Multilingual lexical and terminological resources for (South) Sami-Swedish and Meänkieli (2)

In this questionnaire, we did not focus on language resources for the other official ("minority") languages, only collecting contacts for a future survey, so we expect a greater need, once we start asking.

**Needs for grammars**

*Table 30. 39 or 68 % answered this question.*

| **Needs for grammars** | **Answers** | **Percent** |
|---|---:|---:|
| Rule-based | 29 | 50,9% |
| Language models | 32 | 56,1% |

### 5.1.2 Needs for genres

*Table 31. 46 or 81 % answered this question.*

| **Needs for genres** | **Answers** | **Percent** |
|---|---:|---:|
| News | 26 | 45,6% |
| Documentation | 25 | 43,9% |
| Balanced | 24 | 42,1% |
| E-mail | 23 | 40,4% |
| Reports | 21 | 36,8% |
| Chat | 17 | 29,8% |
| Other, please specify: | 22 | 38,6% |

*Other genres:*

- Texts from professional contexts (10)
  - Patient journals, health and medicine texts (2)
  - Research
  - Patent claims
  - Business communications (annual reports, press releases etc.) (2)
  - Finance documents, banking and insurance texts (2)
  - Judicial-technical texts
  - Decisions by authorities
- Fiction and simplified/abridged texts (7)
  - Fiction
  - Children's books
  - Easy-to-read texts (2)
  - Parallel texts: subtitles (2)
  - Parallel/Comparable texts: full texts vs. summaries
- Teaching and writing (5)
  - Text books, teaching material (3)

School texts, unedited (2)

- Transcribed spoken language (5)
    - Conversational speech in different varieties of Swedish
    - Human-computer dialogue
    - Human-human dialogue
    - Dialogue in various social settings (e.g. business transactions and call center calls)
- Informative texts (2)
    - Blog texts
    - Wiki and encyclopedia texts
    - User manuals
    - Tourist information
- Parallel/Comparable texts: native vs. non-native Swedish (2)
- Parallel/Comparable texts: texts on the same event, but from different perspectives
- Geographical data
- Historical corpora

### 5.1.3 Needs for annotation standards

*Table 32. 31 or 54 % answered this question.*

| Needs for annotation standards | Answers | Percent |
|---|---:|---:|
| XML | 26 | 45,6% |
| TEI | 7 | 12,3% |
| XCES | 7 | 12,3% |
| SGML | 4 | 7,0% |
| Other, please specify: | 7 | 12,3% |

A few subjects also commented that any standard would be appreciated.

*Other annotation standards needed were the following:*

- ISO TC37/SC4 (3)
    - Linguistic Annotation Framework (LAF)
    - Morphosyntactic Annotation Framework (MAF)
- Translation Memory eXchange format (TMX) (2)
- Term Base eXchange format (TBX)
- Göteborg Transcription Standard (GTS)
- Resource Description Framework (RDF)
- Simple Knowledge Organization System (SKOS)
- Web Ontology Language (OWL)
- EAGLES morphological standard

### 5.1.4 Needs for evaluation resources

*The following evaluation resources were needed:*

- Annotated balanced corpora, Swedish National Corpus, categorized corpora (3)
- Treebanks, test data and system for parser evaluation (3)
- Test data and system for part-of-speech tagger evaluation (2)
- Data, tools and measurements for semantic evaluation (word sense disambiguation, frames/roles, selectional restrictions etc.) (2)
- Summarization data (both abstracts/extracts, general/specific), tools and measures for evaluation of summarizers (2)
- Parallel aligned corpora, test data and system for translation evaluation (2)

- Parallel treebanks (with frame-semantic annotation)
- Annotated error corpora (analyzed and corrected), for evaluation of spell and grammar checkers
- Learner corpora (second language)
- Student texts (new academics)
- Data, tools and measurements for morphological analysis (inflection, derivation, compounding)
- Data, tools and measurements for named entity recognizers
- Data, tools and measurements for information retrieval (particularly with queries in a Swedish context, and other text types than news texts)

### 5.1.5 Needs for tools for processing written language

*Table 33. 41 or 72 % answered this question.*

| **Which tools do you need for processing written data?** | **Answers** | **Percent** |
|---|---|---|
| Morfological segmenter (stemmer, lemmatiser, compound analyser, etc.) | 25 | 43,9% |
| Sentence splitter | 23 | 40,4% |
| Part-of-speech tagger | 23 | 40,4% |
| Tokenizer | 22 | 38,6% |
| Clause splitter | 21 | 36,8% |
| Normalizer (upper;lower case, numeric expressions, etc.) | 20 | 35,1% |
| Parser | 20 | 35,1% |
| Formatter (character encoding, file format, etc.) | 18 | 31,6% |
| Lexical semantics analyzer (word sense disambiguation, etc.) | 18 | 31,6% |
| Named entity recognizer | 17 | 29,8% |
| Chunker | 15 | 26,3% |
| Text;Genre classifier | 14 | 24,6% |
| Word aligner | 14 | 24,6% |
| Optical character recognition | 13 | 22,8% |
| Formal semantics analyzer (reference resolution, etc.) | 13 | 22,8% |
| Term extractor | 12 | 21,1% |
| Sentence aligner | 12 | 21,1% |
| Generator | 10 | 17,5% |
| Discourse segmenter | 8 | 14,0% |
| Identifier of attitudinal expressions (attitudes, opinions, feelings, etc.) | 8 | 14,0% |
| Other, please specify: | 9 | 15,8% |

*The following other tools were also specified:*

- User friendly concordancing tools with statistics function, versatility of combined sorting options (cf. excel sheets) (2)
- Phrase linker
- Automatic summarizer
- Language recognizer
- Frame semantic annotator
- Annotation tools
- Visualization tools
- Tools for inductive learning
- Evaluation tools
- Translation memory (program)

*Details on the tools needed for written resources were given by 3 subjects:*

- Parser (functional dependency grammar) (2), with at least the same quality as Connexor's parser, but open source, and freely available, open-source grammars
- Wordform generator
- Basic, single-purpose, modules for tokenization, sentence segmentation and part-of-speech tagging, with standardized application programming interfaces for combining the modules

## *5.2 Needs for spoken language resources*

We received at most 16 answers to our questions regarding the needs for spoken language resources.

### 5.2.1 Needs for speech types

Recorded speech is indispensable for speech research. The interest is moving from read and prompted speech to spontaneous, dialogue and multi-party speech although there still is most need for read speech.

*Table 34. 13 or 23 % answered this question.*

| Needs for speech types | Answers | Percent |
|---|---:|---:|
| Read speech | 9 | 69,2% |
| Dialogue speech | 7 | 53,8% |
| Multi-party speech | 4 | 30,8% |
| Spontaneous speech | 3 | 23,1% |
| Prompted speech | 3 | 23,1% |
| Other | 1 | 7,7% |

### 5.2.2 Needs for speakers

Speakers of both genders have the same need..

*Table 35. 11 or 19 % answered this question.*

| Needs for speakers | Answers | Percent |
|---|---:|---:|
| Male | 11 | 100,0% |
| Female | 11 | 100,0% |

The major demand is for adult speakers but some want child and adolescent speech. There is also a need of second language speakers of Swedish (with another L1 than Finnish). Good dialect coverage is furthermore mentioned.

### 5.2.3 Needs for bandwidths

Telephone speech is attractive for commercial companies, but also wide-band speech and radio speech is needed.

*Table 36. 9 or 16 % answered this question.*

| Needs for bandwidths | Answers | Percent |
|---|---:|---:|
| Telephone (fixed, mobile etc.) | 7 | 77,8% |
| Wide-band microphone | 4 | 44,4% |
| Broadcast news | 2 | 22,2% |
| Other | 1 | 11,1% |

### 5.2.4 Needs for databases/genres

This speech related question got the most answers of related speech, 16. Pronunciation lexica and language models demanded by many.

*Table 37. 16 or 28 % answered this question.*

| Needs for database genres/types | Answers | Percent |
|---|---:|---:|
| Pronunciation lexicon | 12 | 75,0% |
| Language model | 9 | 56,3% |
| Other | 2 | 12,5% |

### 5.2.5 Needs for annotation standards

XML is without any doubt the most popular annotation standard for speech.

*Table 38. 13 or 23 % answered this question.*

| Needs for annotation standards | Answers | Percent |
|---|---:|---:|
| XML | 10 | 76,9% |
| NIST SPHERE/LDC | 2 | 15,4% |
| SGML | 1 | 7,7% |
| SAM | 1 | 7,7% |
| Other | 3 | 23,1% |

Other formats mentioned were GTS, Göteborg Transcription Standard, and the W3C’s Resource Description Framework (RDF) and Web Ontology Language (OWL).

### 5.2.6 Needs for evaluation

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

### 5.2.7 Needs for tools for processing speech data

Quite many answered this question. The needs rather evenly spread across the tools asked for, but the interest in speaker recognition was low.

*Table 39. 14 or 25 % answered this question.*

| Which tools do you need for processing speech data? | Answers | Percent |
|---|---:|---:|
| Orthographic labeling of speech | 9 | 64,3% |
| Text-to-speech | 9 | 64,3% |
| Speech recording | 8 | 57,1% |
| Phonetic labeling of speech | 7 | 50,0% |
| Linguistic labeling of speech | 7 | 50,0% |
| Speech synthesis with augmented control | 7 | 50,0% |
| Checking of recording | 6 | 42,9% |
| Automatic speech analysis | 6 | 42,9% |
| Automatic phonetic segmentation | 5 | 35,7% |
| Speech recognition - a couple of thousand words | 5 | 35,7% |
| Pragmatic labeling of speech | 4 | 28,6% |
| Speech recognition - few words | 4 | 28,6% |
| Speech recognition - dictation | 4 | 28,6% |
| Speech response with prerecorded speech | 3 | 21,4% |
| Speaker recognition | 1 | 7,1% |
| Other | 0 | 0,0% |

### *5.3 Needs for multimodal language resources*

At most 12 answered questions on needs for multimodal language resources. This is not for from the 16 answers to the same question regarding speech resources and indicates a growing interest in this area considering the few answers to existing multimodal resources.

### 5.3.1 Needs for database genres/types

We got similar answers as when asking the same question regarding speech with domination for language models, but also an interest in pronunciation lexica.

*Table 40. 12 or 21 % answered this question.*

| Needs for database genres/types | Answers | Percent |
|---|---:|---:|
| Language model | 10 | 83,3% |
| Pronunciation lexicon | 6 | 50,0% |
| Other | 1 | 8,3% |

### 5.3.2 Needs for speakers

Also here we see no difference between genders. Thus the demand for multimodal speech is similar to that of the need for speech only recordings. Ages mentioned are from 20 to 50, but also younger and older are sought for. Second language speakers are also mentioned as well as a good dialect coverage of Swedish.

*Table 41. 8 or 14 % answered this question.*

| Needs for speakers | Answers | Percent |
|---|---:|---:|
| Male | 8 | 100,0% |
| Female | 8 | 100,0% |

### 5.3.3 Needs for annotation standards

The demand for multimodal speech standards is similar to those for speech only recordings. Also mentioned are GTS, Göteborg Transcription Standard, and the W3C’s Resource Description Framework (RDF), Web Ontology Language (OWL) and Ruby Annotation.

*Table 42. 7 or 12 % answered this question.*

| Needs for annotation standards | Answers | Percent |
|---|---:|---:|
| XML | 6 | 85,7% |
| SGML | 1 | 14,3% |
| SAM | 0 | 0,0% |
| Other | 3 | 42,9% |

### 5.3.4 Needs for evaluation

There was no answer to this question.

### 5.3.5 Needs for tools for processing multimodal language data

The need for tools was quite evenly distributed, as may be seen below. A specific demand for the annotation of face mimics and eye movements was mentioned.

*Table 43. 5 or 9 % answered this question.*

| **Which tools do you need for processing multimodal data?** | **Answers** | **Percent** |
|---|---|---|
| Orthographic labeling of speech | 5 | 100,0% |
| Checking of recording | 4 | 80,0% |
| Linguistic labeling of speech | 4 | 80,0% |
| Pragmatic labeling of speech | 4 | 80,0% |
| Speech recording | 3 | 60,0% |
| Phonetic labeling of speech | 3 | 60,0% |
| System for movem. measurements | 2 | 40,0% |
| Other | 2 | 40,0% |

## *5.4 Needs for other language resources*

*Table 44. 5 or 9 % answered this question.*

| **Do you need other language resources?** | **Answers** | **Percent** |
|---|---|---|
| Yes | 7 | 12,3% |
| No | 14 | 24,6% |

*The subjects who answered yes to the question on needs for other language resources needed the following resources:*

- FrameNet
- Resources for historical material
- Digital voices
- Linguistic descriptions for most European languages
- Alignment of transcriptions with recordings
- Synchronization of various recording sources, e.g. microphones and cameras
- Multimodal annotations
- Concept-coded language resources (natural languages, symbol languages, signs)

(We moved some specifications to detailed needs, instead, as they were not “other” resources as such.)

## *5.5 Plans to produce language resources within 2 - 5 years*

*Table 45. 35 or 61 % answered this question.*

| **Do you plan to produce any lang. resources within the next 2-5 years?** | **Answers** | **Percent** |
|---|---|---|
| Yes | 29 | 50,9% |
| No | 6 | 10,5% |

*Those who planned to produce language resources within 2 - 5 years, planned the following resources:*

- Treebank
- Multilingual resources (6)
    - Parallel treebanks (2)
        - with frame-semantic annotation
        - Swedish-English
    - Parallel corpora (non-West European languages)
    - Translation memories/databases (2)
    - Resources involving Sami

- Corpora (5)
    - Transcribed speech
    - Multimedial/multimodal
    - (Annotated) texts (3)
- Speech databases (6)
    - Voices
    - Spontaneous speech
    - Multimodal spontaneous speech
    - Dialogues
    - Speech styles (2)
        - Conversational Finland Swedish, Finnish second-language learner's Swedish
- Grammars (4)
    - HPSG
    - GF
    - Language models (2)
        - Part-of-speech tagging models
    - Morphological analyzers
- Lexicons (4)
    - Pronunciation
- Term databases (3)
- Concept-coded, ontology-based, multimodal language resource (including support for symbols and signs)
- Refinement/standardization of existing semantic resources
- Text resources for a minority language
- Writing guides
- Logs/Recorded utterances from services
- (Internal) language resources (3)

### *5.6 Re-evaluation of own resources and search for other language resources*

How often do you re-evaluate your Swedish LR needs and seek available LRs?

*Table 46. 44 or 77 % answered this question.*

| **How often do you re-evaluate your language resource needs?** | **Answers** | **Percent** |
|---|---:|---:|
| Monthly | 17 | 29,8% |
| Once per quarter | 7 | 12,3% |
| Once per semester | 3 | 5,3% |
| Once per year | 3 | 5,3% |
| Never | 3 | 5,3% |
| Once every 1-2 years | 1 | 1,8% |
| Other, please specify: | 12 | 21,1% |

Of the 12 subjects that answered other, 7 specified more often (4 almost daily), 1 more seldom, and 4 at irregular intervals (when needed).

## *5.7 Details about needs*

- Språkbanken/Litteraturbanken/Swedish National Corpus:
  Adding historical material (for linguistic research) and more genres of modern writing (for language technology research): 50-75 million SEK (corpus) + 40-50 million SEK (tools)
  Standardizing annotations/tools to a common format and level of annotations: 7 man years (3.5 full time staff over 2 years)
  Equipment: 1.5-2 million SEK (servers)

- Göteborg Spoken Language Corpus:
  Adding new recordings, digitizing, maintenance and development of corpus and tools: 5-10 man years (programmers/corpus development, 1-2 full time programmers/corpus developers over 5 years) and 2.5-5 man years (0.5-1 full time assistants over 5 years)
  Equipment: 50,000 SEK (storage), 1 man year (0.2 full time research developer over 5 years)
- The corpus Andraspråkets strukturutveckling (ASU, second language structural development):
  Digitizing: 70 man hours
  Aligning with transcriptions

- Pronunciation lexicon
- Tools for processing spoken data (2)
  Synthesizer, with the same quality as commercial systems (2)
  Speech recognizer, speaker independent, trainable (2)

  In several cases, some of the raw (unannotated) corpus resources for both texts and speech are available, and the need is rather time and funding for digitizing, annotating, and documenting the resources to make them more useful, and for collecting new resources to fill gaps in the corpora. Some estimations taken from the DISC report:

- Speech/Multimodal/Multimedia (annotated) databases (5)
  Spoken Swedish (2)
  Aligned data for training grapheme-to-phoneme-converter
  Dialogue/Multipart communication (4)
    General (4)
    Human-human
    Human-machine
    Call center calls

# 6 Acquisition of Swedish language resources

Of the 42 subjects that answered the question about the acquisition of Swedish LRs, 67% had acquired one or more Swedish language resources.

*Table 47. 42 or 74 % answered this question.*

| Have you ever acquired Swedish LRs? | Answers | Percent |
|---|---|---|
| Yes | 28 | 49,1% |
| No | 14 | 24,6% |

Of the 30 that had acquired Swedish language resources 15, or 50%, had acquired them from external vendors, while 20, or 67%, had acquired them from various sources as listed below.

*Table 48. 30 or 53 % answered this question.*

| From where did you acquire Swedish LRs? | Answers | Percent |
|---|---|---|
| External vendors | 15 | 26,3% |
| Other, please specify: | 20 | 35,1% |

The various sources and what was acquired from them are found below.

*Other, acquired from:*

- universities, (e.g. Språkbanken, KTH/Nada, Stockholm University, CSC in Finland),
- public organizations (e.g. Language Council of Sweden, EU's publication office, Lantmäteriet),
- researchers' web page (freely available corpora/lexicon/semantic resources, tools with open source code),
- Internet (e.g. translated texts)
- companies (e.g. translation memories)
- clients
- publishers

*Other, acquired resources:*

- monolingual corpora (Swedish Parole corpus, Stockholm EkonomiKorpus, Bonnier Product Classification corpus) containing fiction, newspaper text, older journals
- mono- or multilingual parallel corpora (Europarl, Acquis Communautaire), treebanks (Penn Treebank, LeMonde Treebank, NEGRA, TIGER, Talbanken)
- corpora containing transcribed speech with some multilingual information or dialects
- lexicon (pronunciation lexicon, lexicon containing synonyms, geographical names in Sweden for Sami, SAOL, Oxford thesaurus)
- semantic resources
- translation memories and lexicons, SAMPA transcription as well as tools such as modules for tagging and parsing, morphological analyzer (e.g. Swetwol)
- term- and grammar checkers
- tools for writing and translating display texts
- SDL Trados including Multiterm and Wordfinder dictionaries

Of those who acquired language resources 80% stated that the resources fulfilled their requirements. They also specified the standards they used.

*Table 49. 25 or 44 % answered this question.*

| **Did the acquired LRs fulfill your requirements?** | **Answers** | **Percent** |
|---|---|---|
| Yes | 20 | 35,1% |
| No | 5 | 8,8% |

*Specified standards for Yes answers:*

- ML
- RDF
- OWL
- SKOS
- ISO for thesauri
- TEI
- XCES
- PAROLE
- TMX (Trados' export-format)
- TBX
- TIGER-XML
- standard used by specific projects or product vendors (e.g. Lantmäteriet or the Nordic web lexicon format)

Concerning the question about portability, i.e., on whether standard interchange formats are useful, all of the subjects who answered this question (7/7) were positive. They also specified the formats they used, and they are listed below Table 50.

Note that the answers in Table 50 and Table 51 are compiled from various checkbox questions, some of them single, which makes it irrelevant to give the percentage that answered the question in the caption.

*Table 50. 11 answers to these questions.*

| **Are existing standard interchange formats useful?** | **Answers** | **Percent** |
|---|---|---|
| Yes | 7 | 12,3% |
| No, too many | 2 | 3,5% |
| No, too difficult to use | 1 | 1,8% |
| No, too (labour) resource demanding | 1 | 1,8% |

*Written formats specified:*

- RDF/OWL
- SKOS
- TEI
- TMX
- XCES
- XML
- TIGER-XML (very useful for treebank interchange)

*Spoken formats specified:*

- NIST/SPERE
- SAM

Only two people did not find standard interchange formats useful because they are too difficult to use and (labour) resource demanding.

Among the people who were not satisfied with the language resource, told that they had some problems with the recall of the resource, the documentation was missing or not complete, the resource needed some quality control, or the search tool provided did not work satisfactory or did not allow certain operations.

8 people out of 57 have never acquired any language resource. The reason for never acquiring any language resource is often that the quality of the data does not live up to the requirements (10/10). Data produced by others does not use standards or up-to-date standards so the resource has to be built again.

Other reasons for why people do not acquire language resource is that the data and/or the annotation of it might be ambiguous or erroneous without being validated, or the resource is only searchable instead of being available. According to 2 people, they have not acquired any language resource because of the price; it is cheaper to develop own ones from public sources.

The last, most obvious reason for why not acquiring language resources is that the resource needed is not available or non-existing (8/8), or possible end users do not know that the resource exists.

*Table 51. 41 answers to these questions.*

| What is the reason for not acquiring any Swedish LRs? | Answers | Percent |
|---|---|---|
| The LRs were not available or non-existing. | 8 | 14,0% |
| The available LRs were too expensive | 2 | 3,5% |
| The available data did not live up to your quality requirements | 10 | 17,5% |
| Lack of adaptability (smooth integration) | 3 | 5,3% |
| Lack of conformance (regarding format issues) | 1 | 1,8% |
| Lack of reusability | 1 | 1,8% |
| Lack of coverage | 2 | 3,5% |
| Lack of adequate information types | 3 | 5,3% |
| Lack of data quality (about the content) | 2 | 3,5% |

Clearly, one of the requirements on resources of today is that it should be standardized and freely available, the tools should have open source code, so they, when further developed, can be further distributed to others.

# 7 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 both industry, organizations and academia, who work with Swedish language resources and tools. We can conclude from the 71 answers that although many resources exist for Swedish, there is a need for freely available standardized resources and tools to be freely used by both industry and academia.

In the future, we do hope that an inventory will also be made for Finnish, Jiddisch, Meänkieli, Romani chib, and Sami, which are official languages in Sweden.

## References

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

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

E. Strangert. 2007. Vetenskapsrådets kartläggning av språkteknologiska databaser och framtida behov. (In Swedish.) URL http://sprakteknologi.se/dokument/disc-rapport-om-sprakteknologi.pdf

H. Strik, W. Daelemans, D. Binnenpoorte, J. Sturm, F. de Vriend, and C. Cucchiarini. Dutch HLT resources: From BLARK to priority lists. In Proceedings of ICSLP, Denver, pp. 1549-1552. 2002. URL http://lands.let.kun.nl/literature/strik.2002.2.pdf.

# Appendix A

## *Letter of invitation for participation*

Dear colleague,

Do you wish it would be easier to find Swedish language resources to use in applications or for educational purposes? You can contribute to this by answering a questionnaire.

The attached questionnaire 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.

Research and development of language technology systems needs an infrastructure of publicly available and standardized basic resources for the Swedish language. 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. 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 for each language separately. For Swedish, there are several resources, but it is unclear what type they are of, and to what degree they are available. Therefore, we 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. To make the Swedish BLARK as useful as possible, it is of great importance that everybody who works with Swedish language technology participate in the inventory process. In the future, an inventory will also be made for Finnish, Jiddisch, Meänkieli, Romani chib, and Sami, which are official languages in Sweden, so if you work with one of these languages, we would like you to fill in the first two sections in the questionnaire.

The work on surveying existing basic language resources and developing missing resources is carried out in three phases. In the first phase, we wish to find out what resources are needed, and get an overview of existing resources. As the next step, we will use the information gathered to define what types of resources should be part of the Swedish BLARK, describe the existing resources in uniform metadata, and point out what resources are missing. Lastly, missing language resources will be developed in the order of need.

For the survey of language resources and need, we would like to ask you to answer a questionnaire. You can do that on the Internet (http://www.speech.kth.se/prod/blark/blark.html) or in a text file downloadable from the same address. (You can send the text file with your answers to one of the contacts below.) There is a Swedish and an English version. We would like your answer as soon as possible, but before May 13, 2007, if your answers are to be included in the overview. During phase two, after the initial survey, we will contact you again with more specific questions about the resources.

The questionnaire mainly 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

If you have any questions, please, do not hesitate to contact us!

Thank you for your co-operation.

Sincerely,

Eva Forsbom (evafo@stp.lingfil.uu.se)
Beáta B. Megyesi (bea@stp.lingfil.uu.se)
Department of Linguistics and Philology, Uppsala University
Box 635, SE-751 26 Uppsala, SWEDEN

Kjell Elenius (kjell@speech.kth.se)
Rolf Carlson (rolf@speech.kth.se)
School of Computer Science and Communication
Department of Speech, Music and Hearing, KTH
Lindstedtsvägen 24, SE-100 44 Stockholm, SWEDEN

References:

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

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

E. Strangert. 2007. Vetenskapsrådets kartläggning av språkteknologiska databaser och framtida behov. (In Swedish.) URLhttp://sprakteknologi.se/dokument/disc-rapport-om-sprakteknologi.pdf

H. Strik, W. Daelemans, D. Binnenpoorte, J. Sturm, F. de Vriend, and C. Cucchiarini. Dutch HLT resources: From BLARK to priority lists. In Proceedings of ICSLP, Denver, pp. 1549-1552. 2002. URL http://lands.let.kun.nl/literature/strik.2002.2.pdf.

## *Questionnaire*

QUESTIONNAIRE ON SWEDISH LANGUAGE RESOURCES

0 GUIDELINES
Coding convention
() Tick only one alternative
[] Tick any number of alternatives
Answer in free text, or as specified in prompt

Terminology

* Organisation - company, university, institution, organisation, etc.
* Language resources (LRs) - corpora, lexicons, benchmarks, tools, etc. for language technology
* Evaluation resources - standardised test sets, metrics, benchmarks, etc.
* Tools for processing language data - tools for preprocessing text and sound (formatting, optical character recognition, sound capture, segmentation, normalisation, etc.), modules (part-of-speech taggers, parsers, text-to-speech converters), standards and tools for annotation, tools for searching and mining information from corpora

--- Beginning of questionnaire ---

1 PERSONAL CONTACT DATA

Name:
Position:
E-mail:
Web page:
Address, if not given on web page:

2 INFORMATION ABOUT THE ORGANISATION

2.1 Name of organisation:

2.2 Organisation contact data
Country of establishment, if not Sweden:
Web site:
Address, if not given on web site:

2.3 Type of organisation
() Company
() University
() Public organisation
() Other, please specify:

2.4 Number of employees
() Less than 10
() 10-49
() 50-99
() Over 100

2.5 Number of employees working with language resources (LRs):

2.6 Main activity
[] Software development
[] Language technology product vendor
[] Research
[] Teaching
[] Culture/Museum
[] Minority language organisation
[] Content provider
[] Interpreting/Translating/Localisation
[] Telecommunications
[] E-commerce
[] Banking/Insurance
[] Other, please specify:

2.7 Main language technology area
[] Language learning
[] Language resources
[] Speech technologies
[] Written technologies
[] Search and knowledge mining
[] Machine translation/Computer-assisted translation
[] Other, please specify:

2.8 Main language technology products/services:

2.9 Language coverage of language technology products/services
Monolingual
  [] Swedish
  [] Finnish
  [] Jiddisch
  [] Meänkieli
  [] Romani chib
  [] Sami
  [] Other, please specify:

Multilingual
  [] Swedish
  [] Finnish
  [] Jiddisch
  [] Meänkieli
  [] Romani chib
  [] Sami
  [] Other, please specify:

3 SWEDISH LANGUAGE RESOURCE NEEDS

What should a basic LR kit for Swedish contain to fulfil your needs? First, general questions are asked about your needs. Then, you can give as many details as possible about the resources, e.g. Language, size, format, genre, API, application usage.

3.1 Needs for written LRs
Needs for lexical databases

[] Monolingual
[] Bilingual
[] Multilingual

Needs for corpora

[] Monolingual
[] Bilingual
[] Multilingual
[] Parallel
[] Equivalent (non-parallel but same domain)
[] Enriched (annotated with e.g. tags, clusters)
[] Translation memories

Needs for terminological databases

[] Monolingual
[] Bilingual
[] Multilingual

Needs for grammars

[] Rule-based
[] Language models

Needs for semantic networks

[] Semantic networks (wordnets, thesauri, ontologies, etc.)

3.1.1 Needs for genres

[] News
[] Reports
[] Documentation
[] E-mail
[] Chat
[] Balanced
[] Other, please specify:

3.1.2 Needs for annotation standards

[] SGML
[] XML
[] TEI
[] XCES
[] Other, please specify:

3.1.3 If you need resources for evaluation, please specify:

3.1.4 Which tools do you need for processing written data?

[] Optical character recognition
[] Formatter (character encoding, file format, etc.)
[] Normaliser (upper/lower case, numeric expressions, etc.)
[] Tokeniser
[] Discourse segmenter
[] Sentence splitter
[] Clause splitter
[] Part-of-speech tagger
[] Morfological segmenter (stemmer, lemmatiser, compound analyser, etc.)
[] Named entity recogniser
[] Chunker
[] Parser

[] Generator
[] Lexical semantics analyser (word sense disambiguation, etc.)
[] Formal semantics analyser (reference resolution, etc.)
[] Term extractor
[] Identifier of attitudinal expressions (attitudes, opinions, feelings, etc.)
[] Text/Genre classifier
[] Sentence aligner
[] Word aligner
[] Other, please specify:

3.2 Needs for speech LRs

3.2.1 Needs for speech types
[] Read speech
[] Spontaneous speech
[] Prompted speech
[] Dialogue speech
[] Multi-party speech
[] Other, please specify:

3.2.2 Needs for speakers
[] Male
[] Female
[] Age:
[] Other (e.g. dialect or second-language speakers), please specify:

3.2.3 Needs for bandwidths
[] Telephone (fixed, mobile etc.)
[] Wide-band microphone
[] Broadcast news
[] Other, please specify:

3.2.4 Needs for database genres/types
[] Pronunciation lexicon
[] Language model
[] Other, please specify:

3.2.5 Needs for annotation standards
[] SGML
[] XML
[] SAM
[] NIST SPHERE/LDC
[] Other, please specify:

3.2.6 If you need resources for evaluation, please specify:

3.2.7 Which tools do you need for processing speech data?
[] Speech recording
[] Checking of recording
[] Orthographic labeling of speech
[] Phonetic labeling of speech
[] Linguistic labeling of speech

[] Pragmatic labeling of speech
[] Automatic speech analysis (formants, F0, etc.)
[] Automatic phonetic segmentation
[] Speech recognition - few words (voice call, etc.)
[] Speech recognition - a couple of thousand words (call center, etc.)
[] Speech recognition - dictation
[] Speaker recognition (verifying/identifying)
[] Speech response with prerecorded speech
[] Speech synthesis with augmented control (F0, emphasis, reductions, etc.)
[] Text-to-speech
[] Other, please specify:

3.3 Needs for multimodal LRs

3.3.1 Needs for database genres/types
[] Pronunciation lexicon
[] Language model
[] Other, please specify:

3.3.2 Needs for speakers
[] Male
[] Female
[] Age:
[] Other (e.g. dialect or second-language speakers), please specify:

3.3.3 Needs for annotation standards
[] SGML
[] XML
[] SAM
[] Other, please specify:

3.3.4 If you need resources for evaluation, please specify:

3.3.5 Which tools do you need for processing multimodal data?
[] Speech recording
[] Checking of recording
[] Orthographic labeling of speech
[] Phonetic labeling of speech
[] Linguistic labeling of speech
[] Pragmatic labeling of speech
[] System for movement measurements (Qualisys, etc.)
[] Other, please specify:

3.4 Do you need other LRs?
() Yes, please, specify:
() No

3.5 Do you plan to produce any LRs within the next 2-5 years?
() Yes, please, specify:
() No

3.6 How often do you re-evaluate your Swedish LR needs and seek available LRs?

[] Monthly<br>
[] Once per quarter
[] Once per semester
[] Once per year
[] Once every 1-2 years
[] Never
[] Other, please specify:

3.7 Details about your needs
Please add details about nature, size, etc. whenever appropriate and possible e.g. for a corpus of business documents, you may state it consists of 2 million words, Swedish-English dictionary, 50,000 entries, etc.:

4 INFORMATION ABOUT YOUR SWEDISH LANGUAGE RESOURCES

What resources do you have that could fit into a BLARK? First, general questions are asked about your existing resources. Then, you can give as many details as possible about them, e.g. language, size, format, genre, API, application usage.

4.1 Your written LRs
Your lexical databases
  [] Monolingual
  [] Bilingual
  [] Multilingual
Your corpora
  [] Monolingual
  [] Bilingual
  [] Multilingual
  [] Parallel
  [] Equivalent (non-parallel but same domain)
  [] Enriched (annotated with e.g. tags, clusters)
  [] Translation memories
Your terminological databases
  [] Monolingual
  [] Bilingual
  [] Multilingual
Your grammars
  [] Rule-based
  [] Language models
Your semantic networks
  [] Semantic networks (wordnets, thesauri, ontologies, etc.)

4.1.1 Your genres
[] News
[] Reports
[] Documentation
[] E-mail
[] Chat
[] Balanced
[] Other, please specify:

4.1.2 Your annotation standards

[] SGML
[] XML
[] TEI
[] XCES
[] Other, please specify:

4.1.3 If you have resources for evaluation, please specify:

4.1.4 Which tools do you have for processing written data?
[] Optical character recognition
[] Formatter (character encoding, file format, etc.)
[] Normaliser (upper/lower case, numeric expressions, etc.)
[] Tokeniser
[] Discourse segmenter
[] Sentence splitter
[] Clause splitter
[] Part-of-speech tagger
[] Morfological segmenter (stemmer, lemmatiser, compound analyser, etc.)
[] Named entity recogniser
[] Chunker
[] Parser
[] Generator
[] Lexical semantics analyser (word sense disambiguation, etc.)
[] Formal semantics analyser (reference resolution, etc.)
[] Term extractor
[] Identifier of attitudinal expressions (attitudes, opinions, feelings, etc.)
[] Text/Genre classifier
[] Sentence aligner
[] Word aligner
[] Other, please specify:

4.2 Your speech LRs

4.2.1 Your speech types
[] Read speech
[] Spontaneous speech
[] Prompted speech
[] Dialogue speech
[] Multi-party speech
[] Other, please specify:

4.2.2 Your speakers
[] Male
[] Female
[] Age:
[] Other (e.g. dialect or second-language speakers), please specify:

4.2.3 Your bandwidths
[] Telephone (fixed, mobile etc.)
[] Wide-band microphone
[] Broadcast news
[] Other, please specify:

4.2.4 Your database genres/types
[] Pronunciation lexicon
[] Language model
[] Other, please specify:

4.2.5 Your annotation standards
[] SGML
[] XML
[] SAM
[] NIST SPHERE/LDC
[] Other, please specify:

4.2.6 If you have resources for evaluation, please specify:

4.2.7 Which tools do you have for processing speech data?
[] Speech recording
[] Checking of recording
[] Orthographic labeling of speech
[] Phonetic labeling of speech
[] Linguistic labeling of speech
[] Pragmatic labeling of speech
[] Automatic speech analysis (formants, F0, etc.)
[] Automatic phonetic segmentation
[] Speech recognition - few words (voice call, etc.)
[] Speech recognition - a couple of thousand words (call center, etc.)
[] Speech recognition - dictation
[] Speaker recognition (verifying/identifying)
[] Speech response with prerecorded speech
[] Speech synthesis with augmented control (F0, emphasis, reductions, etc.)
[] Text-to-speech
[] Other, please specify:

4.3 Your multimodal LRs

4.3.1 Your database genres/types
[] Pronunciation lexicon
[] Language model
[] Other, please specify:

4.3.2 Your speakers
[] Male
[] Female
[] Age:
[] Other (e.g. dialect or second-language speakers), please specify:

4.3.3 Your annotation standards
[] SGML
[] XML
[] SAM
[] Other, please specify:

4.3.4 If you have resources for evaluation, please specify:

4.3.5 Which tools do you have for processing multimodal data?
[] Speech recording
[] Checking of recording
[] Orthographic labeling of speech
[] Phonetic labeling of speech
[] Linguistic labeling of speech
[] Pragmatic labeling of speech
[] System for movement measurements (Qualisys, etc.)
[] Other, please specify:

4.4 Do you have other LRs?
() Yes, please, specify:
() No

4.5 Do you use LRs
[] produced internally?
[] produced by specific contracted vendors?
[] distributed by data centres?
4.6 Your tools
What kind of tools do you use to produce your LRs? Please list:

4.7 Your validation of LRs

4.7.1 When producing LRs, do you follow specific guidelines?
(If not go to question 4.8).
[] Yes, we use internal specifications
[] Yes, we use external specifications, please specify:

4.7.2 Do you follow specific standards?
() Yes, please specify:

() No
4.7.3 Do you use specific gold standards/benchmarks?
() Yes, please, specify:
() No

4.7.4 Are your LRs validated?
(ie. checked for compliance with given standard regarding format and content, and for complete, consistent, and correct mark-up, etc.)
() Yes,
[] by independent/external organisation/expert
[] in-house
() No

4.8 Details about your resources
Please add details about nature, size, etc. whenever appropriate and possible e.g. for a corpus of business documents, you may state it consists of 2 million words, Swedish-English dictionary, 50,000 entries, etc.:

5 ACQUISITION OF SWEDISH LANGUAGE RESOURCES

5.1 Have you ever acquired Swedish LRs?
() Yes (go to question 5.2)
() No (go to question 5.3)

5.2 (If you have ever acquired Swedish LRs)

5.2.1 From where did you acquire Swedish LRs?
[] External vendors
[] Others, please specify:

5.2.2 What kind of Swedish LRs do you get/buy? Please list:

5.2.3 When incorporating acquired LRs, have you ever benefited from existing standard interchange formats?
() Yes, please specify:
() No

5.2.4 Did the acquired LRs fulfill your requirements?
() Yes
() No, please specify:

5.3 (If you have never acquired Swedish LRs)

5.3.1 What is the reason for not acquiring any Swedish LRs?
[] The LRs were not available or non-existing, please specify which ones:
[] The available LRs were too expensive, please specify:
[] The available data did not live up to your quality requirements,
please specify below.
Portability
[] Lack of adaptability (smooth integration)
[] Lack of conformance (regarding format issues)
[] Lack of reusability
Functionality
[] Lack of coverage
[] Lack of adequate information types
[] Lack of data quality (about the content)
Other, please specify:

5.3.2 Portability
Are existing standard interchange formats useful?
[] Yes, please specify:
[] No, too many, please specify:
[] No, too difficult to use, please specify:
[] No, too (labour) resource demanding, please specify:

6 GENERAL COMMENTS
If you have general comments on the questions or the resources, you can give them here:

7 CONTACT COLLECTION

If you would like us to send this survey to other organisations involved with Swedish (Finnish,

Jiddisch, Meänkieli, Romani chib, Sami) basic language resources, please indicate the details below (duplicate as needed).
Name of organisation:
Contact person:
E-mail:
Website:
Address, if not given on web site:
Type of LRs available at that organisation:

--- End of questionnaire ---

Thank you for your participation!

References:

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

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

E. Strangert. 2007. Vetenskapsrådets kartläggning av språkteknologiska databaser och framtida behov. (In Swedish.) URL http://sprakteknologi.se/dokument/disc-rapport-om-sprakteknologi.pdf

H. Strik, W. Daelemans, D. Binnenpoorte, J. Sturm, F. de Vriend, and C. Cucchiarini. Dutch HLT resources: From BLARK to priority lists. In Proceedings of ICSLP, Denver, pp. 1549 1552. 2002. URL http://lands.let.kun.nl/literature/strik.2002.2.pdf.

# Appendix B: Swedish language technologies and tools

| Swedish language technologies and tools | Description | Provider |
|---|---|---|
| Masterin | Translation software for English, Swedish and Finnish | Master's Innovations Ab |
| Sprawk | Online application for improving vocabulary, comparing and translating of words by using semantic-centric language networks that represent links between words and meanings www.sprawk.com | Transmachina AB |
| CDORD | Help application for people with dyslexia, combining techniques from written language technology (corpus linguistics) and speech technology | Mikro Værkstedet a/s |
| Elixir | A tool for information filtering information retrieval, search, and cooperation | ASIMUS AB |
| Scania Checker | A language checker for technical writers (with Scania Swedish, a controlled language for technical documentation) | Scania CV AB |
| TermWeb | Terminology management system, consulting in database architecture and in-house terminology management; creation of company-specific terminology databases and glossaries; intranet or Internet access to terminology databases | Interverbum |
| Scania Checker | Controlled language checker | Scania CV AB |
| WordFinder | | Scania CV AB |
| Minspeak | Simple choice through speech synthesis or recorded speech | Rehabmodul AB |
| DART | Webb-based services NavigAbile vocabulary in symbol language (concept coding framework, SYMBERED)), http://www.dart-gbg.org/ | Sahlgrenska University Hospital |
| www-lemmie | A tool for accessing our corpora online | CSC - Scientific Computing Ltd. |
| VocabTool | A software-as-a-service platform for individual learning of, among others, language through the webb and/or cell phone | Vocab AB |
| Trados | Translation software, http://www.trados.com/en/ | SDL Trados Technologies |
| Convertus | Machine translation of academic course syllabi from Swedish to English | Department of Linguistics and Philology, Uppsala University and Convertus |
| A Swedish Systran Module | Scandinavian proof-reading tools | Department of Linguistics and Philology, Uppsala University |
| FASTY | Fast typing for disabled people | Department of Linguistics and Philology, Uppsala University |

| **Swedish language technologies and tools** | **Description** | **Provider** |
|---|---|---|
| PLUG | Parallel corpora in Linköping, Uppsala, and Gothenburg | Department of Linguistics and Philology, Uppsala University |
| MIA | Multra in work | Department of Linguistics and Philology, Uppsala University |
| KOMA | Corpus-based machine translation | Department of Linguistics and Philology, Uppsala University |
| MATS (Methods and Applications of a Translation System) | A Swedish machine translation module | Department of Linguistics and Philology, Uppsala University |
| GRAMEX | Automatic grammar extraction | Department of Linguistics and Philology, Uppsala University |

| Text processing technologies | Description | Provider |
|---|---|---|
| Grammar tools for Sami | Analyzers for North-Sami, Lule-sami and South Sami. | Trond Trosterud Universitetet i Tromsø |
| BaseModel | Perl package for baseform reduction (“lemmatizer”) or wordform generation, including a set of Swedish models derived from Stockholm-Umeå Corpus (SUC2.0). Statistically-based and trainable. Works for suffigating languages such as Swedish. http://stp.lingfil.uu.se/~evafo/resources/baseformmodels/ | Eva Forsbom, Dept. of Linguistics and Philology, Uppsala University |
| ConcApp | Free Windows concordancer http://www.edict.com.hk/PUB/concapp/ | (suggested by Tua Holm, Swedish Maritime Administration) |
| DAM-LR: Distributed Access Management of Language Resources | http://www.mpi.nl/dam-lr/lra-flyer | (suggested by Department of Linguistics and Phonetics, Lund University) |
| Dialog system modules | http://herd.ida.liu.se:8180/nlpfarm/index.jsp | Lars Ahrenberg, Natural Language Processing Laboratory, Department of Computer and Information Science, Linköping University |
| I*Link | Word-alignment system for sentence-aligned parallel texts http://www.ida.liu.se/~nlplab/ILink/ | Lars Ahrenberg, Natural Language Processing Laboratory, Department of Computer and Information Science, Linköping University |
| Granska | Grammar checker, Swedish | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Granska tagger | Part-of-speech tagger for Granska, Swedish | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Inflektor | Word inflector for Granska, Swedish | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| GTA (Granska Text Analyzer) | Analyzer for Granska, Swedish | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Compound splitter | Swedish http://www.csc.kth.se/tcs/humanlang/tools.html | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Stomp | Part-of-speech tagger, Swedish http://www.csc.kth.se/tcs/humanlang/tools.html | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Grim | Interactive language learning environment, Swedish http://www.csc.kth.se/tcs/humanlang/tools.html | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Missplel | Spelling error introducer http://www.csc.kth.se/tcs/humanlang/tools.html | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| JavaSDM | A Java package for Random Indexing, language independent http://www.csc.kth.se/tcs/humanlang/tools.html | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |

| Text processing technologies | Description | Provider |
|---|---|---|
| Infomat | Visualization tool for vector spaces, language independent http://www.csc.kth.se/tcs/humanlang/tools.html | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Language recognizer | | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Language classifier | | Trond Trosterud |
| MaltTagger | Probabilistic part-of-speech tagger http://w3.msi.vxu.se/users/jha/research/malttagger/ | Joakim Nivre, Växjö University, |
| MaltParser | Data-driven dependency parser http://w3.msi.vxu.se/~nivre/research/MaltParser.html | Joakim Nivre, Växjö University, |
| SPARK | A shallow parser developed for the analysis of the constituent structure of Swedish sentences http://stp.lingfil.uu.se/~bea/resources/spark/ | Beata Bandmann Megyesi, Dept. of Linguistics and Philology, Uppsala University |
| Nuance' tools | Parser, pronunciation generator (proprietory) | (suggested by Håkan Jonsson, Voice Provider Sweden AB) |
| Phrase aligner | | Mickel Grönroos, Master's Innovations Ab |
| Semantic tagging support tool | | Caroline Willners, Lund University |
| Stava | Spell checker, Swedish http://www.csc.kth.se/tcs/humanlang/tools.html | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Text summarizer | | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Tools for analyzing transcribed speech | A number of tools for coding, automatic processing, and multimodal transcription, such as Corpus Browser, Gorallt statistical measure,<br>Multitool transcription and coding tool. | Jens Allwood, Dept. of Linguistics, Göteborg University |
| Word predictor | | Daniel Ridings, Mikro Værkstedet a/s |

| **Spoken language technologies** | **Description** | **Provider** |
|---|---|---|
| Text-to-speech system | | Acapela Group |
| Voice-based phone applications | | Voice Provider Sweden AB |
| Speech synthesis and speech recognition | | Talboks- och punktskriftsbiblioteket |
| VocabTool | A software-as-a-service-platform for individual learning of e.g., language through the internet and cell phones | Vocab AB |
| Wavesurfer | An Open Source tool for sound visualization and manipulation | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| Snack | The Snack Sound Toolkit is designed to be used with a scripting language such as Tcl/Tk or Python. Using Snack you can create powerful multi-platform audio applications with just a few lines of code. | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| The NICO ANN tool-kit | The NICO Toolkit is an artificial neural network toolkit specifically designed and optimized for automatic speech recognition applications. | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| ESPS source code from the ESPS/waves+ package | The ESPS Toolkit has been licensed to the Centre for Speech Technology thanks to a generous donation from Microsoft and AT&T. An archive of source files is available for download. Only source code from the ESPS library is provided, no source code for waves. | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| MBROLA interface | http://www.ling.lu.se/persons/JohanF/php/mbrola.php | Dept. of Linguistics and Phonetics, Lund University |
| NLP components | text processing, pronunciation, and prosody implemented using Festival (www.festvox.org) framework http://www.ling.lu.se/persons/JohanF/php/festival.php | Dept. of Linguistics and Phonetics, Lund University |
| Text-to-speech conversion | Swedish (Festival based) | Dept. of Linguistics and Phonetics, Lund University |

| Multi-media technologies | Description | Provider |
|---|---|---|
| Wavesurfer | An Open Source tool for sound visualization and manipulation. May be used for parallel annotation of video and speech | Speech, Music and Hearing, School of Computer Science and Communication, KTH |

| Written language resources | Description | Provider |
|---|---|---|
| Corpus | Recordings of reading and writing activity online<br>(eyetracking, keystroke logging) | Dept. of Linguistics and Phonetics, Lund University |
| Umeå Keystroke logged writing corpus | A digital collection of young writers' keystroke log files of English as a Foreign Language and Swedish writings that are yet to be indexed into a database. The collection is held jointly with the Faculty of Teacher Education, Umeå University. | Umeå University, Department of philosophy and linguistics, General linguistics |
| A Finland Swedish text corpus (FISC) | 1995, 2.5 million words, containing an appendix on 80,000 words of transcribed, broadcasted conversations and interviews.<br>http://www.nord.helsinki.fi/fisc/press.html | Dept. of Nordic languages and Nordic literature, Helsinki University |
| Parallel corpus | 10 million words, English-Swedish, patents | Magnus Merkel, Fodina Language Technology AB |
| Parallel corpus | 10 million words, English-Swedish, software documentation | Magnus Merkel, Fodina Language Technology AB |
| Parallel corpus | 10 million words, English-Swedish, manufacturing documentation | Magnus Merkel, Fodina Language Technology AB |
| Parallel corpus | Swedish-German | Magnus Merkel, Fodina Language Technology AB |
| Parallel corpus | Around 1 million words, English-Swedish | Lars Ahrenberg, Natural Language Processing Laboratory, Department of Computer and Information Science, Linköping University |
| Parallel treebank | Around 100000 words, English-Swedish | Lars Ahrenberg, Natural Language Processing Laboratory, Department of Computer and Information Science, Linköping University |
| Semantic net | Swedish WordNet, with semantic relations such as synonymy, antonymy, hyponymy and meronymy; follows EuroWordNet's (EWN) principles; around 25000 concepts (synsets) and 34000 words (28000 nouns, 6000 verbs) general language (validated by SUC frequencies).<br>The concepts are connected to the EWN interlingual index (ILI). | Åke Viberg, Dept. of Linguistics and Philology, Uppsala University |
| Lexical database | FrameNet, Swedish, pilot study | Åke Viberg, Dept. of Linguistics and Philology, Uppsala University |
| Corpus | Learner data (Swedish), bilingual data, and native control data from five-year olds to adults, transcribed recorded speech mostly from 1980's and beginning of 1990's, with part-of-speech information | Åke Viberg, Dept. of Linguistics and Philology, Uppsala University |
| Parallel corpus | Around 200000 words, extracts from 10 Swedish novels and their translations into English, German, French, and Finnish | Åke Viberg, Dept. of Linguistics and Philology, Uppsala University |

| Written language resources | Description | Provider |
|---|---|---|
| Scania Corpus | Two collections of technical manuals from Scania CV AB.<br>Scania 95: around 220000 words each in Swedish, Dutch, English, Finnish, French, German, Italian, and Spanish<br>http://perdix.lingfil.uu.se/scania.html<br>Scania 98: around 1.2 million words each in Swedish, English, and German. | Anna Sågvall Hein, Dept. of Linguistics and Philology, Uppsala University |
| Scarrie Swedish newspaper corpus | 67 million tokens from Upsala Nya Tidning and Svenska Dagbladet (1995/1996).<br>Established by: Anna Sågvall Hein<br>Language: Swedish<br>Text type: newspaper text<br>Period: 1995-96<br>Usage: research, paper writing, undergraduate and postgraduate studies, including the Language Engineering Programme<br>SCARRIE SvD9596<br>Size: 47.4 miljoner ord<br>SCARRIE UNT9596<br>Size: 22.8 million tokens | Anna Sågvall Hein, Dept. of Linguistics and Philology, Uppsala University |
| UNT92 | From Upsala Nya Tidning.<br>Established by: Anna Sågvall Hein<br>Language: Swedish<br>Size: 6.5 million tokens<br>Text type: newspaper text<br>Period: 1992<br>Usage: research, paper writing, undergraduate and postgraduate studies, including the Language Engineering Programme | Anna Sågvall Hein, Dept. of Linguistics and Philology, Uppsala University |
| Scarrie Swedish Error Corpus Database | 9000 sentence fragments with errors and corrections, from the Scarrie corpus<br>http://www.lingfil.uu.se/ling/ecd/ (password protected)<br>ECD - Error Corpus Database<br>Established by: Anna Sågvall Hein<br>Language: Swedish<br>Size: about 9 000 sentence fragments<br>Text type: newspaper text<br>Period: 1992<br>Usage: research, paper writing, undergraduate and postgraduate studies, including the Language Engineering Programme | Anna Sågvall Hein, Dept. of Linguistics and Philology, Uppsala University |
| Swedish political texts | Swedish governmental inaugural speech in 5 languages: Swedish, English, French, German, and Spanish, since 1996, about 11,000 tokens. | Anna Sågvall Hein, Dept. of Linguistics and Philology, Uppsala University |
| Swedish Immigrant Newspaper corpus | Texts from Invandrartidningen in Swedish, Arabic, English, Finnish, Persian, Polish, Serbian/Croatian, Bosnian, and Spanish. | Anna Sågvall Hein, Dept. of Linguistics and Philology, Uppsala University |

| Written language resources | Description | Provider |
|---|---|---|
| Dictionary | 60000 lemmas and several thousand phrases, Swedish general language | Anna Sågvall Hein, Dept. of Linguistics and Philology, Uppsala University |
| Term base | 4000 Swedish terms on automotive maintenance and their English translations | Anna Sågvall Hein, Dept. of Linguistics and Philology, Uppsala University |
| SCANIA Swedish dictionary | SCANIA:<br>http://perdix.lingfil.uu.se/scania.html<br>Established by: Anna Sågvall Hein<br>Language: Swedish<br>Size: about 200,000 wordforms<br>Text type: language for specific purposes<br>Period: 1995-2005<br>Usage: research, paper writing, undergraduate and postgraduate studies, including the Language Engineering Programme | Anna Sågvall Hein, Dept. of Linguistics and Philology, Uppsala University |
| Parallel corpus | Turkisk-Svensk Korpus/Turkish-Swedish Corpus<br>Established by: Prof. Anna Sågvall Hein, Prof. Éva Á Csató Johanson, dr. Beáta Bandmann Megyesi, dr. Bengt Dahlqvist<br>Languages: Turkish-Swedish<br>Size: 118,000 tokens (Turkish), 144,000 tokens (Swedish).<br>Language type: fiction, non-fiction<br>Period: 20th and 21st centuries<br>Usage: research, paper writing, education in Turkish languages, including the programme on Oriental Studies.<br>Under construction | Department of Linguistics and Philology, Uppsala University |
| Parallel corpus | Swedish-Hindi<br>Under construction | Department of Linguistics and Philology, Uppsala University |
| Parallel copus | French-Swedish parallel corpus<br>Established by: Carina Andersson<br>Language: French-Swedish<br>Size: 2 milliontokens<br>Period: 20th century<br>Language type: novels<br>Usage: postgraduate studies, research | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| The Uppsala French TAP-Corpus | The Uppsala French TAP-Corpus ("Think-Aloud-Protocols")<br>Established by: Kerstin Jonasson<br>Language: French-Swedish<br>Size: about 50 pages<br>Texttyp: "think-aloud-protocols"<br>Period: established 1993-1995<br>Usage: research | Faculty of Languages (Anna Sågvall Hein), Uppsala University |

| Written language resources | Description | Provider |
|---|---|---|
| Elevers möte med skolans textvärldar: text books | Text book texts within the project "Elevers möte med skolans textvärldar"<br>Established by: Caroline Liberg, Agnes Edling, Jenny Folkeryd, Åsa af Geijerstam<br>Language: Swedish<br>Size: 27,953 tokens<br>Language type: fiction, discursive text book texts<br>Period: collected 1999-2003 (printed about 1980-2001)<br>Usage: research, postgraduate studies | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| Elevers möte med skolans textvärldar (texts written by students) | Texts written by students within the project "Elevers möte med skolans textvärldar"<br>Established by: Caroline Liberg, Åsa af Geijerstam, Agnes Edling, Jenny Folkeryd<br>Language: Swedish<br>Size: about 400 short texts<br>Language type: fiction, non-fiction<br>Period: collected 1999-2003<br>Usage: research, postgraduate studies | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| Facktexter under 1900-talet (Language for specific purposes during the 20th century) | Established by: Britt-Louise Gunnarsson<br>Language: Swedish<br>Size: The 20th century corpus contains 1,340 normal pages of 3,000 characters, in total (science 650 pages and popular science 690 pages).<br>Language type: 180 scanned scientific and popular science texts (90 in each genre) on (national) economy, (lung) medicine and (electrical) engineering (60 in each topic) from 1895-1905, 1935-1945, and 1975-1985.<br>Characteristics: From each period, 60 scanned texts on 2 topics: banking and credit system, and taxes (economy), lung diseases, and skin and veneral diseases (medicine), electrical engineering, and telcommunication (engineering). Each topic is represented by 60 texts, 5 from each topic, from each each respective genre and period. | Faculty of Languages (Anna Sågvall Hein), Uppsala University |

| Written language resources | Description | Provider |
|---|---|---|
| Fackspråkens framväxt (Development of language for specific purposes) | Established by: Britt-Louise Gunnarsson<br>Language: Swedish<br>Size: The 18th and 19th century corpus contain 1,590 normal pages of 3,000 characters (science 940 pages and popular science 650 pages).<br>Language type: 180 scanned scientific and popular science texts on (national) economy, medicine and engineering from the 18th century, 1800-1849, and 1850-1880.<br>Characteristics: From each period, 60 scanned texts on 2 topics: banking and credit system, and trade (economy), lung diseases, and skin and veneral diseases (medicine), electrical engineering, and telecommunication and mechanics (engineering). Each topic is represented by 60 texts, 5 from each topic, from each each respective genre and period. | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| Texter i europeiska skrivsamhällen (Texts in European writing societies) | Established by: Britt-Louise Gunnarsson<br>Language: Swedish<br>Language type: Collected and systematized Swedish, English, and German texts, written within the topic areas banking, engineering agency, department of occupational medicine, and department of history, in Sweden, Great Britain, and Germany.<br>Characteristics: The texts represent 4 genres: 1. directed to staff (staff magazine); 2. directed to owners, customers, and the public (press release, annual reports); 3. directed to (prospective) customers (tender, letter, booklet, expert opinion, research application, teaching material, project description, minutes, scientific article); 4. intended for image making (general presentation, advertisements). The number of genrer from each environment varies. | Faculty of Languages (Anna Sågvall Hein), Uppsala University |

| Written language resources | Description | Provider |
|---|---|---|
| Samnordisk runtextdatabas (Common Nordic rune text database) | Established under guidance of: Lennart Elmevik and Lena Peterson<br>Language: runskrift<br>Size: About 6,000 inscriptions.<br>Language type: Nordic rune texts, including texts found outside of the Nordic countries.<br>Characteristics: Texts are rendered in transliterated and normalised form, including translation to English. Metadata include information on time period, place of finding, excepted source, type of object, etc. Software for various searches.<br>Usage: research within several disciplines | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| ASU (Andraspråkets strukturutveckling) corpus | Recorded and transcribed conversations, written essays in Swedish, produced by young adult learners, comparable material in standard Swedish produced by native Swedes, 490000 words in total (415000 spoken, 75000 written).<br>Designed for longitudinal studies of second language development and comparisons of learner and native language production.<br>Spoken and written data are documented in parallel with fixed time intervals, transcribed and tagged, XML-based, with Java-based search tool, ITG, developed at Göteborg University, available via http://spraakbanken.gu.se | Section of Phonetics, Sign Language and General Linguistics, Dept. of Linguistics, Stockholm University |
| SSM (Svenska som målspråk) corpus | Learner corpus of essays (written 1973-1975) by students of Swedish for foreign students at Stockholm University (representing 10 L1), 112000 words, not balanced.<br>The corpus has 10 parts, with various course stages in each part.<br>XML-based, with Java-based search tool, ITG, developed at Göteborg University, available via http://spraakbanken.gu.se | Section of Phonetics, Sign Language and General Linguistics, Dept. of Linguistics, Stockholm University |
| SUC - The Stockholm Umeå Corpus | See Appendix C: Evaluation resources | Computational Linguistics section, Dept. of Linguistics, Stockholm University |
| SMULTRON: The Stockholm MULtilingual Treebank | 1000 sentences (50000 words) in English-German-Swedish, manually corrected, follows TIGER-XML and Corpus Documentation, Alignment and Annotation guidelines | Computational Linguistics section, Dept. of Linguistics, Stockholm University |

| Written language resources | Description | Provider |
|---|---|---|
| database of film subtitlesF | Swedish and Danish; >10 million entries; growing; commercial for building machine translation system; project-specific; use XML-standards for data interchange; regular maintenance | Computational Linguistics section, Dept. of Linguistics, Stockholm University |
| Tvärslå: Nordisk nätordbok | http://www.csc.kth.se/tcs/projects/netordbok | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Base vocabulary | 8220 Swedish baseforms (disambiguated for part-of-speech), derived from Stockholm-Umeå Corpus (SUC2.0), ranked by frequency and dispersion, contains wordforms, frequency, and morphosyntactic info (PAROLE tagset). http://stp.lingfil.uu.se/~evafo/resources/basevocpool/ | Eva Forsbom, Department of Linguistics and Philology, Uppsala University |
| Large, clean raw text corpora | Unsupervised acquisition | Chris Biemann, Universität Leipzig |
| The Language Bank of Finland | Contains text and speech<br>data in Finnish and other languages. | CSC - Scientific Computing Ltd. |
| Finnish Text Collection | | CSC - Scientific Computing Ltd. |
| Finnish-Swedish Text Collection | Around 34 million words in one hundred thousand documents, various genres http://www.csc.fi/english/research/software/fstc | CSC - Scientific Computing Ltd. |
| Finnish Broadcast Corpus | | CSC - Scientific Computing Ltd. |
| KOTUS Swedish-Finnish Parallel Corpus (KSFPC) | http://www.kotus.fi/ or 124 sentence aligned documents with approximately 130 thousand sentences (1.7 million words in Finnish, and 2.3 million words in Swedish) http://www.csc.fi/english/research/software/ksfpc | The Research Institute for the Languages of Finland (suggested by Mickel Grönroos Master's Innovations Ab*) or* CSC - Scientific Computing Ltd. |
| PAROLE corpus | Around 19 million words, Swedish http://www.csc.fi/english/research/software/parole-sv | CSC - Scientific Computing Ltd. |
| GSLC Göteborg Spoken Language Corpus | Around 1.5 million transcribed words ( see main entry under multimodal resources) | Jens Allwood, Dept. of Linguistics, Göteborg University |
| Learner corpus | 2 million words, Swedish | Jens Allwood, Dept. of Linguistics, Göteborg University |

| Written language resources | Description | Provider |
|---|---|---|
| Skrift hos barn och ungdomar, IKT och datorbaserat skrivstöd (Writing of Children and Adolescents in the Information Society) | Texts by children and adolescents, handwritten, written with a word processor or texted, e-mail, chat and diaries created over the Internet. The corpus contains over 600 school-related texts and 400 free-time texts, about 1/3 or 97,433 tokens transcribed and analyzed, using the transcription and coding format minCHAT (Codes for the Human Analysis of Transcripts) and the analysis tool CLAN (Computerized Language Analysis).<br>Video recordings of 14 students writing on computers. | Jens Allwood, Dept. of Linguistics, Göteborg University |
| Parallel corpora | (not for distribution) | Mickel Grönroos Master's Innovations Ab |
| Machine-readable dictionaries | Swedish-Finnish-Swedish and Swedish-English-Swedish<br>(not open source) | Mickel Grönroos Master's Innovations AB |
| Monolingual corpus | 4 million words (North Sami?) | Trond Trosterud |
| Parallel corpus | 400 000 words North Sami-Bokmål (Norwegian) | Trond Trosterud |
| KTH-text | 150 million words | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| CENTLEX | Pronunciation dictionary, 410000 entries, Swedish | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| Svenska språknämndens uttalsordbok (Swedish Language Council's Pronunciation Dictionary) | Pronunciation dictionary, 67000 entries, Swedish (printed?) | Rickard Domeij, Language Council of Sweden |
| Nyordsboken (Lexicon of new words) | New words (1980's to 1990's), 2000 entries, Swedish (printed?) | Rickard Domeij, Language Council of Sweden |
| Svenskt språkbruk (Swedish language usage) | Constructions and phrases, 85000 entries, Swedish (printed?) | Rickard Domeij, Language Council of Sweden |
| Svenska skrivregler (Swedish writing rules) | Rules and guidelines for writing in Swedish, more than 300 entries (printed?) | Rickard Domeij, Language Council of Sweden |
| Språklådan | Database with language questions and answers, more than 2000 entries, Swedish | Rickard Domeij, Language Council of Sweden |

| Written language resources | Description | Provider |
|---|---|---|
| LEXIN: Svenska ord | Available for research (Dept. of Swedish/Martin Gellerstam.<br>Distribution media: Internet. URL: http://scrooge.spraakdata.gu.se/lb/lexin/<br>No documentation.<br>Size: 28500 words.<br>Content: Swedish (printed) MRD, general language and terminology (Swedish Social Welfare), lexical unit is single word lemma and multiword unit, sources are corpus, MRDs, printed dictionaries. Manual compilation. MS Access DB, no validation.<br>Microstructure: level of representation (phonology/phonetics, morphology, syntax, semantics), orthography (spelling variants), morphology (inflection, derivation, compounding), syntax (valency information), semantics (senses), definitions, comments.<br>[More details in the ENABLER questionnaire] | ENABLER: Göteborgs University (Dimitrios Kokkinakis) |
| LEXIN | Dictionaries, mini=5000 entries, midi=8000 entries, large=17000 entries, maxi=28500 entries<br>online:<br>Swedish-English<br>Swedish-Albanian<br>Swedish-Arabic<br>Swedish-Bosnian<br>Swedish-Finnish<br>Swedish-Greek<br>Swedish-Croatian<br>Swedish-Kurdish (North Kurdish/Kurmancî)<br>Swedish-Russian<br>Swedish-Serbian (Latin letters)<br>Swedish-Serbian (Cyrillic letters)<br>Swedish-Somali<br>Swedish-Spanish (including Spanish American variants)<br>Swedish-Turkish<br>Swedish (including pronunciation and explanations)<br>Printed from electronic sources:<br>Svenska Ord - Swedish, with pronunciation and explanations (maxi)<br>BILDTEMAN - thematic illustrations, Swedish and English, more than 17000 pictures<br>Swedish-North Kurdish (mini)<br>Swedish-Persian (mini)<br>Swedish-Polish (mini)<br>Swedish-South Kurdish (mini)<br>Swedish-Somali (mini)<br>Swedish-Czech (mini)<br>Swedish-Turabdinian (mini)<br>Swedish-Vietnamese (mini)<br>Swedish-Tigrinya (midi) | Rickard Domeij, Language Council of Sweden |

| Written language resources | Description | Provider |
|---|---|---|
| | Swedish-Macedonian (large)<br>Swedish-Persian (large)<br>Swedish-Romanian (large)<br>Swedish-Turkish (large)<br>Swedish-English (maxi)<br>Swedish-Albanian (maxi)<br>Swedish-Arabic (maxi)<br>Swedish-Bosnian (maxi)<br>Swedish-Finnish (maxi)<br>Swedish-Greek (maxi)<br>Swedish-Croatian (maxi)<br>Swedish-Russian (maxi)<br>Swedish-Serbian (maxi)<br>Swedish-Spanish (maxi)<br>Swedish-Kurdish (North Kurdish/Kurmancî, maxi, in progress)<br>Swedish-Kurdish (South Kurdish, maxi, in progress)<br>Swedish-Somali (maxi, in progress)<br>In the future (not before 2009):<br>Swedish-Kalderash Romani<br>Swedish-Lovari Romani<br>Swedish-Thai (vague plans)<br>Swedish-Pashtu (vague plans)<br>Swedish-Aramaic (Assyrian, suggested)<br>Swedish-Chinese (suggested) | |
| Term bases | CD with Swedish-Finnish domain-specific dictionaries, around 23000 terms:<br>Ruotsalais-suomalainen koulusanasto - Svensk-finsk<br>skolordlista. 1985. (Education)<br>Ruotsalais-suomalainen työmarkkinasanasto - Svensk-finsk arbetsmarknadsordlista. 1989. (Labour market)<br>Ruotsalais-suomalainen sosiaalialan sanasto - Svensk-finsk socialordlista. 1992. (Socials)<br>Ruotsalais-suomalainen kasvien ja eläinten luettelo -<br>Svensk-finsk förteckning över växter och djur. 1999. (Flora and fauna)<br>Ruotsalais-suomalainen kirkollisen elämän sanasto -<br>Svensk-finsk kyrko- och församlingsordlista. 2001. (Church)<br>Ruotsalais-suomalainen pankkisanasto - Svensk-finsk bankordlista. 2001. (Banking)<br>Ruotsalais-suomalainen lääketieteen sanasto - Svensk-finsk medicinsk ordlista. 2004. (Medicine)<br>Ruotsalais-suomalainen urheilusanasto Svensk-finsk<br>idrottsordlista. 2004. (Sports) | Rickard Domeij, Language Council of Sweden |
| Pronunciation dictionary | 550000 entries | Kåre Sjölander, Christina Ericsson<br>Talboks- och punktskriftsbiblioteket |

| **Written language resources** | **Description** | **Provider** |
|---|---|---|
| Name lexicon | 23000 entries | Kåre Sjölander, Christina Ericsson Talboks- och punktskriftsbiblioteket |
| Corpus | 20 million words, raw text | Kåre Sjölander, Christina Ericsson Talboks- och punktskriftsbiblioteket |
| Corpus | Around 1 million words, easy-to-read texts and children's books | Katarina Mühlenbock DART, Sahlgrenska University Hospital |
| Dictionary | Extracted from a 40 million word news corpus (1998) | Katarina Mühlenbock DART, Sahlgrenska University Hospital |
| Onomastica | Pronunciation lexicon, 100000 Swedish names | Telia |
| Voice Provider | A large pronounciation lexicon with Swedish and Danish names (proprietory). | Voice Provider |
| A multilingual resource grammar | Grammatical Framework (GF), including Functional Morphology (FM) and Extract (lexicon extraction), 12 languages (Danish, English, Finnish, French, German, Italian, Norwegian, Russian, Spanish, Swedish), lexicons 500-15000 lemmas, http://www.cs.chalmers.se/~aarne/GF/lib/resource-1.0/doc/<br>Descriptions of basic grammar structure, including inflectional morphology and syntax (language-independent). Morphological descriptions of Arabic,<br>old Swedish, and Urdu. Morphological lexicons (3000-20000 lemmas) of several of these languages were created by using Extract and adapting existing resources.<br>Free software, open source. | Aarne Ranta, Chalmers, Department of Computer Science and Engineering |
| Translation memories | More than 500 customer-specific | Luca Vaccari, Viking |
| Translation memories | Many | Katherine Stuart, Katherine Stuart The Right Word |
| Glossaries | Small | Katherine Stuart, Katherine Stuart The Right Word |
| Resources in 20+ languages, mostly extracted from Wikipedia. Over 4 million words/phrases in about 2 million synsets. | | Nicholas Cottrell Transmachina AB |
| Dictionaries | About 20 Swedish & Swedish-English dictionaries, industry handbooks, Svenska Duden, Bevingade ord, etc. | Dan Lufkin Lufkin Colleagues |
| Corpora | Large, general and domain-specific | Joakim Cöster ASIMUS AB |
| Translation memories | Bilingual | Peter Cedermark Scania CV AB |
| Term database | Around 4000 entries | Peter Cedermark Scania CV AB |

| Written language resources | Description | Provider |
|---|---|---|
| Texts | The texts (regulations and general advice) published in<br>Swedish Maritime Administration Code of Statutes are free to use. Most other documents created within the authority are considered in the public domain. In principle, it is therefore free to use much of the authority's material for building a corpus. | Tua Holm, Maritime Safety Inspectorate |
| Translation memories | http://www.red.se | Irene Elmerot, red. (language consultant) |
| Word lists | http://www.red.se | Irene Elmerot, red. (language consultant) |
| Translation memories, corpus, word lists | Around 100 million words | Cecilia Falk, TransFalk AB |
| Language models | Swedish part-of-speech models, to be used with the Trigrams'n Tags tagger (Brants, 2000), trained on SUC and PAROLE tagset for text and transcribed speech<br>http://stp.lingfil.uu.se/~bea/resources/tnt/ | Beata Bandmann Megyesi, Dept. of Linguistics and Philology, Uppsala University |
| Language models | Swedish part-of-speech tagging models for various tagger tools (TnT, C&C, HunPos, Stomp, MBT), trained on Stockholm -Umeå Corpus (SUC2.0), and in turn, larger raw corpora (Scarrie, Parole, Europarl, and combinations of those).<br>Availability depends on the respective corpus license. | Eva Forsbom, Dept. of Linguistics and Philology, Uppsala University |
| Stava's word base | Some hundred thousand words, Swedish (spellchecking), based on SAOL 11 edition, with compound splitting information, coded for use in STAVA<br>http://www.csc.kth.se/tcs/humanlang/tools.html | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Lexicon | Information on part-of-speech and word statistics, Swedish, for use by Granska's tagger, inflector, and compound splitter<br>http://www.csc.kth.se/tcs/humanlang/tools.html | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| CrossCheck corpus | Learner corpus, Swedish, morphosyntactically annotated | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| KTH News Corpus | 13 million words, Swedish, hundreds of thousands of news articles downloaded from newspaper web pages | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |

| Written language resources | Description | Provider |
|---|---|---|
| Dictionaries within the Nordisk netordbog (Nordic Web Dictionary) | Collection of multilingual dictionaries in Swedish, Danish, Norwegian, Icelandic, Finnish, English, XML-based, includes the Lexin lexicons, searchable via Tvärslå http://ordbok.nada.kth.se | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Folkets synonymlexikon ('the people's synonym lexicon') | Around 70000 synonym pairs, Swedish, rated by Internet users 2005-2006, XML-based, free http://lexin.nada.kth.se/synlex.html | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| KTH Extract Corpus | A large collection of extract-based summaries of a few texts | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| QA corpus | 99 domain-central questions and answers for their respective news texts, for evaluation of question-answering system. | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Swedish treebank | Small | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Swedish compounds | Compound parts, with part-of-speech information for the parts, and texts with splitted compounds | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Swedish-Japanese lexicon | | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Thai-Swedish lexicon | | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| SALT: Språkbankens arkiv för länkade texter ("The Bank of Swedish Archive for Linked Texts") | Not available (will be available for research on completion) (Språkbanken/Lars Borin).<br>Distribution media: Internet (when available). URL: http://spraakbanken.gu.se.<br>Documentation in progress.<br>Size: Swedish originals (approx. 305000 words) and their translations into Dutch, English, French, German, Italian and Russian (approx. 1800000 words). Originals from each of these languages (approx. 1800000 words) and their translations into Swedish (approx. 1800000 words. A total of approx. 5700000 words.<br>Content: Parallel Swedish, Dutch, English, French, German, Italian and Russian (Unicode), general language (mainly composed of fiction published in 1960 and later, newspaper text), XML, structural annotation (text, paragraph, sentence), no linguistic annotation, partial formal validation.<br>[More details in the ENABLER questionnaire] | ENABLER: Göteborg University (Dimitrios Kokkinakis) |

| Written language resources | Description | Provider |
|---|---|---|
| Modern written Swedish text corpora | Available for research (Språkbanken/Lars Borin). Distribution media: Internet. URL: http://spraakbanken.gu.se. Documentation in Swedish.<br>Size: Approx. 65 million words.<br>Content: (modern) Swedish (ISO-8859-1), general language, mySQL DB, text files, XML, various formats, structural annotation (text, paragraph, sentence), no linguistic annotation, no validation.<br>[More details in the ENABLER questionnaire] | ENABLER: Göteborg University (Dimitrios Kokkinakis) |
| Historical/diachronic written Swedish text corpora | Available for research (Språkbanken/Lars Borin). Distribution media: Internet. URL: http://spraakbanken.gu.se. Documentation in Swedish.<br>Size: Approx. 9 million words.<br>Content: (historical/diachronic) Swedish (ISO-8859-1), general language, mySQL DB, text files, various formats, structural annotation (text, paragraph, sentence), no linguistic annotation, no validation.<br>[More details in the ENABLER questionnaire] | ENABLER: Göteborg University (Dimitrios Kokkinakis) |
| POS-tagged modern written Swedish text corpora | Available for research (Språkbanken/Lars Borin). Distribution media: Internet. URL: http://spraakbanken.gu.se. Documentation in Swedish.<br>Size: Approx. 20 million words.<br>Content: (modern) Swedish (ISO-8859-1), general language, mySQL DB, text files, structural annotation (text, sentence), linguistic annotation (morphosyntactic), no validation.<br>[More details in the ENABLER questionnaire] | ENABLER: Göteborg University (Dimitrios Kokkinakis) |
| SynTag, Swedish treebank | Available for research (Språkbanken/Lars Borin). Distribution media: Internet. URL: http://spraakbanken.gu.se. Documentation in Swedish.<br>Size: Approx. 0.1 million words.<br>Content: (modern) Swedish (ISO-8859-1), newspaper texts, text files, proprietary format, structural annotation (text, paragraph, sentence), linguistic annotation (morphosyntactic, syntactic (dependency) parsing), no validation.<br>[More details in the ENABLER questionnaire] | ENABLER: Göteborg University (Dimitrios Kokkinakis) |

| Written language resources | Description | Provider |
|---|---|---|
| SAOL (Swedish Academy Glossary) | Available for research (Dept. of Swedish/Martin Gellerstam.<br>Distribution media: Internet/CD.<br>URL: http://scrooge.spraakdata.gu.se/saol/.<br>No documentation.<br>Size: 120000 words (12th edition, 1998).<br>Content: Swedish (printed) MRD, general language, lexical unit is single word lemma, sources are corpus, MRDs.<br>Manual/Semi-automatic compilation.<br>Content validation.<br>Microstructure: level of representation (morphology), orthography (spelling variants, syllabification), morphology (inflection, derivation, compounding), definitions, comments, usage examples.<br>[More details in the ENABLER questionnaire] | ENABLER: Göteborg University (Dimitrios Kokkinakis) |
| The Swedish PAROLE-SIMPLE lexicon | Available for research(will be available for research on completion) (Dept. of Swedish/Maria Toporowska Gronostaj).<br>Distribution media: Internet.<br>Documentation<br>Size: 20000 morphological units, 30000 syntactic units, 8000 semantic units.<br>Content: Swedish computational lexicon (English metalanguage), general language, lexical unit is single word lemma, sources are corpus, MRDs. Semi-automatic compilation (automatic extraction from MRDs, manual annotation).<br>SGML/text, partial validation.<br>Microstructure: level of representation (morphology, syntax, semantics), morphology (inflection), syntax (complements, functions of complements, morphosyntactic restrictions), semantics (senses, ontological typing, argument structure), definitions, usage examples.<br>[More details in the ENABLER questionnaire] | ENABLER: Göteborg University (Dimitrios Kokkinakis) |

| Written language resources | Description | Provider |
|---|---|---|
| GLDB: Göteborgs lexikaliska databas ("Gothenburg Lexical Database"). | Available for commercial research and use (Göteborgs universitet/Jerker Järborg). Distribution media: CD/Internet. Documentation.<br>Size: Modern Swedish (1950-). Some 60000 expressional entities (lemmas); some 70000 content entities (senses), comprising also some 25000 subsenses. (total size as a text available on request).<br>Content: Swedish MRD, general language, domain specific language, lexical unit is single word lemma or multiword unit, sources are corpus, printed dictionaries, and informants. Manual compilation. INGRES database (deliverable XML), partial content validation.<br>Microstructure: level of representation (phonology/phonetics, morphology, syntax, semantics, some pragmatics and phraseology), ortography (spelling variants), etymology, morphology (inflection, derivation, compounding), syntax (simply valency information; many special constructions and morphosyntactic restrictions), semantics (senses with descriptions of relations between main (core) sense and subsense(s) when applicable, ontological typing where motivated and relevant, some semantic characterization of argument structure), definitions, comments, usage examples.<br>[More details in the ENABLER questionnaire] | ENABLER: Göteborg University (Dimitrios Kokkinakis) |
| Lexical data in Språkbanken | Searchable over Internet.<br>Söderwall (old Swedish, 23,000 entries)<br>Söderwall, supplement (old Swedish, 20,000 entries)<br>Schlyter (old Swedish, law vocabulary, 16,000 entries)<br>SAOB (Swedish Academy Dictionary, 51,000 entries)<br>LEXIN: Svenska ord (Swedish base vocabulary, 20,000 entries, see details above)<br>Svenskt associationslexikon (Swedish Associative Thesaurus, 72,000 entries)<br>TERMIN (bilingual social terminology, Swedish-immigrant languages, 4,400 entries)<br>AVENTINUS (multilingual drug-related terminology, 15,000 entries)<br>In total: 221,400 entries<br>http://spraakbanken.gu.se/ | Lars Borin, Språkbanken (the Swedish Language Bank), Göteborg University |
| Swedish corpora in Språkbanken | Searchable over Internet.<br>Press 65 (news text, 990,989 tokens) | Lars Borin, Språkbanken (the Swedish Language Bank, Göteborg University, |

| Written language resources | Description | Provider |
|---|---|---|
| | Press 76 (news text, 1,156,958 tokens)<br>DN 1987 (news text, 4,132,784 tokens)<br>Press 95 (news text, 6,769,649 tokens)<br>Press 96 (news text, 5,755,168 tokens)<br>Press 97 (news text, 11,900,570 tokens)<br>Press 98 (news text, 9,239,336 tokens)<br>SVD 00 (news text, 13,131,043 tokens)<br>GP 01 (news text, 15,257,883 tokens)<br>GP 02 (news text, 18,434,005 tokens)<br>GP 03 (news text, 16,663,701 tokens)<br>GP 04 (news text, 19,406,813 tokens)<br>Stockholm Umeå Corpus (see separate entry, 1,166,590 tokens)<br>SYNTAG (see separate entry, about 100,000 tokens)<br>Talbanken/Bruksprosa (see separate entry, about 87,000 tokens)<br>PAROLE corpus (PoS-tagged mixed published written Swedish, about 19,000,000 tokens)<br>SVANTE (written learner corpus, 204,398 tokens)<br>ASU (see separate entry, about 730 000 tokens)<br>Forskning & Framsteg (popular science, 669,893 tokens)<br>Older Swedish novels (late 1880s and early 1900s, 3,702,748 tokens)<br>Bonniersromaner I (modern novels, 1976/77, 5,626,348 tokens)<br>Bonniersromaner II (modern novels, 1980/81, 3,715,690 tokens)<br>Strindberg's letters (1,223,288 tokens)<br>Strindberg's novels and plays, 2,461,426 tokens)<br>Svenska dagbladets årsbok 1923-1958 (yearbook of news text, about 1,500,000 tokens)<br>Psalmboken (1937, hymns, 111,304 tokens)<br>Svensk författningssamling 1978-81 (The Swedish Code of Statutes, 612,688 tokens)<br>Bellman's collected works (about 360,000 tokens)<br>Riksdagens snabbprotokoll 1978-79 (parliament proceedings, 4,420,767 tokens)<br>Källtext (old Swedish, 1,096,244 tokens)<br>Manuductio (1651, on poetry, 28,202 tokens)<br>MEDLEX corpus (medicine, about 10,000,000 tokens)<br>Litteraturbanken (fiction, about 1,500,000 tokens)<br>SAOB (Swedish Academy Dictionary, as text, 28,375,720 tokens)<br>In total: about 209 935 801 tokens<br>http://spraakbanken.gu.se/ | |

| Spoken language resources | Description | Provider |
|---|---|---|
| Archiving of (existing materials of) sound, video and transcriptions of conversations (NorDiga) | A project to be commenced in August 2007 | Dept. of Nordic Languages and Nordic Literature, Helsinki University |
| Diphone database | Swedish, MBROLA | Dept. of Linguistics and Phonetics, Lund University |
| Letter-to-sound rules | Swedish<br>http://www.ling.lu.se/persons/JohanF/php/ltsr.php | Dept. of Linguistics and Phonetics, Lund University |
| Swedia 2000 | The Phonetics and Phonology of Swedish Dialects around the year 2000. A central goal for this project, was to develop a digital database for research on Swedish dialects. The database was created on a national basis, in cooperation between the phonetic departments at Umeå, Stockholm and Lund universities and intended as a resource for research both within and outside the project. The corpus includes speech samples from 12 speakers (including younger and older men and women) from 107 locations across the Swedish-speaking areas of Sweden and Finland, altogether more than 1200 hours of collected speech. The database material of which a reasonable part has been annotated, has been used by researchers from several different areas. | Dept. of philosophy and linguistics, General linguistics, Umeå University,<br>Dept. of Linguistics and Phonetics, Lund University,<br>Dept. of Linguistics, Göteborg University |
| IRIS (Invandrarröster i Sverige) | Recordings of immigrant's speech | Section of Phonetics, Sign Language and General Linguistics, Dept. of Linguistics, Stockholm University |
| Speech database | Pathological speech (after glossectomy) | Section of Phonetics, Sign Language and General Linguistics, Dept. of Linguistics, Stockholm University |
| VaKoS, Variation in Consonant Clusters in Swedish | A partially online database. The controlled labeled material is accessible by all researchers. The spontaneous material is yet to be labeled and published. This database was created within a project of the same name | Umeå University, Department of philosophy and linguistics, General linguistics |
| UC3 Corpus, Umeå Child Consonant Cluster Corpus | A corpus of 22 children’s productions of a set of control words. Each child was followed over a period of twelve months. This corpus is digital, yet not generally accessible. Only certain aspects are marked-up. This database was created as part of Fredrik Karlsson’s PhD research; the findings of which are published in Karlsson (2006) The Acquisition of Contrast: A longitudinal investigation of Initial s+plosive cluster development in Swedish Children. | Umeå University, Department of philosophy and linguistics, General linguistics |

| Spoken language resources | Description | Provider |
|---|---|---|
| UDID, Umeå Disguise and Imitation Database | This is a digital and partially marked-up database contain voice imitations and attempts at voice disguise by 20 speakers. Work based on the database has been conducted together with Robert Rodman of North Carolina State University. This database is being constructed within Erik Eriksson's PhD project that is a part of the external project Imitated Voices awarded to Umeå University. | Umeå University, Department of philosophy and linguistics, General linguistics |
| The Archive of Accented Swedish | The analogue database on which Robert Bannert's (1990) book På väg mot svenskt uttal is based. This database contains a rich set of data on accented Swedish. | Umeå University, Department of philosophy and linguistics, General linguistics |
| SpeechDat, fixed telephone | Automatic speech recognition, ASR, for the fixed telephone network. 5000 speakers | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| SpeechDat, mobile telephone | ASR for mobile phones: in the office, pavement, vehicle or public place, 1000 speakers | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| SpeeCon | ASR in the home, office, outdoors or car, 4 mics, 550 adult and 50 child speakers | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| Rafael | Speech Recognition, 1000 speakers | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| Gandalf | Speaker verification, 86 speakers | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| PER | Speaker verification, 52 speakers | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| Waxholm | Dialogue system, 68 speakers | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| August | Dialogue system, 2650 speakers | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| Adapt | Dialogue system, 57 speakers | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| PF-Star | ASR for children, 198 children, dialogue system | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| Higgins | Human dialogue system, speech and some video, 16 speakers | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| CHIL KTH Connector | Dialogue system, 36 speakers | Speech, Music and Hearing, School of Computer Science and Communication, KTH |

| **Spoken language resources** | **Description** | **Provider** |
|---|---|---|
| Voice Provider, KTH | 20000 calls (60000 utterances) from voice controlled telephone services | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| Voice Provider | Dialog/spontaneous speech from voice controlled telephone services (recordings from some million calls). | Voice Provider |
| Intervjuer (formella och informella | | Department of Linguistics and Philology, Uppsala University |
| Voice database | 16 hours, phonematically annotated | Kåre Sjölander, Christina Ericsson<br>Talboks- och punktskriftsbiblioteket |
| The Swedish Map Task Corpus | The Swedish Map Task Corpus<br>Established by: Pétur Helgason<br>Language: Swedish<br>Size: 70 minutes, about 8,000 tokens<br>Language type: conversation<br>Period: contemporary<br>Usage: 3 postgraduate study programs, C- and D-level papers | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| Svenska stadsmål 1. Eskilstuna (Swedish city language 1: Eskilstuna) | Established by: Bengt Nordberg<br>Language: Swedish<br>Size: 53 hours, transcriptions<br>Language type: Conversational interviews<br>Characteristics: 83 native Eskilstuna inhabitants, differenciated socially, by age and sex, are interviewed on everyday topics, personal memories and local history. Some conversations without an interviewer are also included.<br>Period: 1967-68 | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| ÅbEsk - Kontinuitet och förändring i nutida talspråk. Återbesök i Eskilstuna (Continuity and change in contemporary spoken language. Revisiting Eskilstuna) | Established under guidance of Bengt Nordberg<br>Language: Swedish<br>Size: 78 hours, partly transcribed<br>Language type: Conversational interviews<br>Characteristics: The corpus is collected for a project following up the Eskilstuna study from 1967. The native informants have the same social distribution and the recordings the same stylistic qualities as in the previous study (see Svenska stadsmål 1). 13 speakers are the same as in the previous study, while the other 72 speakers are new.<br>Usage: research | Faculty of Languages (Anna Sågvall Hein), Uppsala University |

| Spoken language resources | Description | Provider |
|---|---|---|
| Språkanvändning och språkmiljöer i staden och på landet (Language use and environments in urban and rural areas) | Established by: Bengt Nordberg<br>Language: Swedish<br>Size: 29 hours<br>Language type: Informal interviews<br>Characteristics: 20 informants 30-50 years, 50% living in an urban environment - 50% in a rural environment (Mid-Sweden, southern Norrland), 50% men - 50% women, 50% workers - 50% white collar workers. The interviews concern the participants communication environment, contact nets and language activities, and any perceived changes and attitudes in connection with them. | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| Stad och omland: Urbaniseringen speglad i språket (City and countryside: urbanization mirrored in language) | Established by: Mats Thelander<br>Language: Swedish<br>Size: 52 hours<br>Language type: Semi-structured telephone interviews<br>Characteristics: 120 informants from northern Västerbotten and Eskilstuna municipality, respectively. The informants (20-80 years) were randomly selected from 7 geographically defined populations:: born and living in rural Västerbotten (12 informants), born and living in Skellefteå city (12), born in rural Västerbotten but living in Skellefteå city (24), born and living in rural Eskilstuna municipality (12), born and living in Eskilstuna city (12), born in rural Eskilstuna municipality but living in Eskilstuna city (24), and born in rural Västerbotten but living in Eskilstuna city (24). Each interview lasts about 30 minutes, is more or less conversational, and mostly concerns the topics moving and language. The interviewer is a 35-year-old male, spekaing standard Swedish.<br>Period: 1978-79 | Faculty of Languages (Anna Sågvall Hein), Uppsala University |

| Spoken language resources | Description | Provider |
|---|---|---|
| Språk, roll och sociala relationer - Burträskundersökningen (Language, role and social relations - the Burträsk study) | Established by: Mats Thelander<br>Language: Swedish<br>Size: 29 hours group conversation, 3.5 hours interviews.<br>Language type: group conversation<br>Characteristics: 14 situationally varied group conversations with 4 participants in each group - all 56 speakers (14-64 years) from former Burträsk municipality. Each recording lasts about 2 hours, and in mostly leisurely conversational style. During the second half of the conversation an external researcher i s present. The topic is Rural habitat in transformation (reflections on Burträsk in view of the upcoming municipality merging). 9 participants were later recorded in more formal interviews.<br>Period: 1973-75 | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| Samtal, åldrande och identitet 1: identitetsskapande strategier i äldresamtal (Conversation, aging and identity 1: identity-establishing strategies in conversation among elderly women) | Established under guidance of Bengt Nordberg<br>Language: Swedish<br>Size: dialogue 27 hours, group conversation 14 hours, partly transcribed<br>Language type: arranged dialogues med middle-aged and retired women, informal group conversation among elderly women (coffee drinking, card playing, etc).<br>Characteristics: In the dialogues, 40 women in total take part. The purpose is to get acquainted, over and within generations. In the group conversations, most participants are already acquainted with each other.<br>Period: 1997-2000<br>Usage: research | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| Ungdomars samtalsstil (Adolescent conversational style) | Established by: Bengt Nordberg<br>Language: Swedish<br>Size: 2 hours, transcriptions<br>Language type: Free conversation between close friends<br>Characteristics: Girls and boys, 12-16 years, in single-sex groups. Topics include personal interests, everyday events, gossip. The participants come from Uppsala.<br>Period: 1984-88 | Faculty of Languages (Anna Sågvall Hein), Uppsala University |

| **Spoken language resources** | **Description** | **Provider** |
|---|---|---|
| Barnets språkliga identifikation - BSI (Linguistic identification of children) | Established by: Bengt Nordberg<br>Language: Swedish<br>Size: 78 hours, partly transcribed<br>Language type: Recorded situations: reading aloud, communication games, card playing and discussion.<br>Characteristics: Situationally varied recordings of 85 native students in Eskilstuna during 3 consecutive years, grades 1-3, 4-6, and 7-9. The students are evenly distributed as regards sex, but socially differentiated.<br>Period: 1977-79 | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| Talutveckling (Speech development) | Established by: Birgitta Garme<br>Language: Swedish<br>Size: About 170 conversations<br>Language type: Conversations with students of varying age, from grade 2 to upper secondary school, working with various conversational tasks. Mainly group conversations without a teacher, but the corpus also includes discussions led by a teacher.<br>Period: 1988-92 | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| Interaktionen vid seminarier (Interaction at seminars) | Established under guidance of Britt-Louise Gunnarsson<br>Language: Swedish<br>Size: 20 PhD student seminars, transcriptions<br>Language type: PhD student seminars from 3 departments at Uppsala University: 1 humanistic, 1 social sciences, 1 natural sciences.<br>Period: 1992-95 | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| FORTIS - Sverigefinnars två språk - språkbruk och attityder hos två generationer (The two languages of Sweden Finns - language use and attitudes in two generations) | Established under guidance of Bengt Nordberg<br>Language: Finnish-Swedish<br>Size: 54 hours, transcriptions<br>Language type: formal and informal interviews, group conversations with self-recruited participants<br>Characteristics: Situationally varied recordings with 2 generations Sweden Finns (16-19 and 35-55 years, children and parents, male and female) living in a Stockholm suburb. Free topics in the group conversations. Topics in the informal interviews include everyday life, migration history of the informants, and the situation as Sweden Finns. The formal interviews are authentically faithful job and radio interviews. | Faculty of Languages (Anna Sågvall Hein), Uppsala University |

| Spoken language resources | Description | Provider |
|---|---|---|
| GIC - Samtal i nödsituation. Telekommunikationen vid en giftinformationscentral (Conversation in emergency situations. Telecommunication at the Swedish Poisons Information Centre) | Established under guidance of Bengt Nordberg<br>Language: Swedish<br>Size: 17 hours, transcriptions<br>Language type: The corpus contains 377 authentic telephone calls from private persons all over Sweden to the Swedish Poisons Information Centre regarding acute or presumed poisonings.<br>Characteristics: The advising staff includes 13 pharmacists/advisors. | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| Kommunikationen vid en larmcentral (Communication at an alarm call centre) | Established by: Bengt Nordberg<br>Language: Swedish<br>Size: 8 hours, transcriptions<br>Language type: Authentic recordings of telephone and radio communication at a local alarm call centre between callers (requesting help), operators and relief staff. The calls concern acute illness, accidents, fire, ambulance requests, stoppages, etc.<br>Period: 1986 | Faculty of Languages (Anna Sågvall Hein), Uppsala University |
| Texter i europeiska skrivsamhällen (Texts in European writing societies) | Established under guidance of Britt-Louise Gunnarsson<br>Language: Swedish<br>Size: 70 interviews, transcriptions<br>Language type: 70 interviews with persons responsible for the writing of various texts in a bank, an engineering agency, a department of history, and a department of occupational medicine, in 3 countries: Sweden, German, and Great Britain.<br>Period: 1994-96 | Faculty of Languages (Anna Sågvall Hein), Uppsala University |

| **Multimodal language resources** | **Description** | **Provider** |
|---|---|---|
| GSLC, Göteborg Spoken Language Corpus | Several sub corpora:<br>1) Kernel Corpus - adult Swedish L1: Recorded (about 60% video) and transcribed speech from some 30 social activity types, collected during 30 years, transcribed with GTS (Göteborg Transcription Standard) and MSO (Modified Standard Orthography) for Swedish. In total around 2,000 hours of speech (400 hours digitized).<br>Activity types:<br>Discussion<br>Retelling Of Article<br>Interview<br>Task-Oriented Dialogue<br>Informal Conversation<br>Role Play<br>Trade Fair<br>Arranged Discussions<br>Formal Meeting<br>Consultation<br>Shop<br>Dinner<br>Market<br>Auction<br>Factory Conversation<br>Party<br>Games & Play<br>Phone<br>Travel Agency<br>Court<br>Church<br>Lecture<br>Hotel<br>Therapy<br>Bus Driver-Passenger<br>2) Adult language learners of Swedish: recorded audio/video, transcribed<br>3) A number of small sub corpora in various languages, including 90 hours recordings of Finnish spoken in Sweden.<br>4) A number of small sub corpora of pathological speech, etc.<br>http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3 | Jens Allwood, Department of linguistics, General linguistics, Göteborg University, |
| child language databases | Digital, audio, video, babble and early speech (up to 30 months) | Section of Phonetics, Sign Language and General Linguistics, Dept. of Linguistics, Stockholm University |
| printed sign language dictionary | Swedish sign language dictionary, with transcriptions and photos | Section of Phonetics, Sign Language and General Linguistics, Dept. of Linguistics, Stockholm University |

| Multimodal language resources | Description | Provider |
|---|---|---|
| interactive sign language database | Digital version of Swedish sign language dictionary, 2968 entries, with information from the printed dictionary and video sequences, animated images, general language.<br>Searchable by Swedish word, hand shape, ID in printed dictionary, and domain. Viewable in isolation, or in an example sentence.<br>http://www.ling.su.se/tsp | Section of Phonetics, Sign Language and General Linguistics, Dept. of Linguistics, Stockholm University |
| domain-specific sign language databases | Digital Swedish sign language dictionaries, 11500 entries in total, with transcriptions, photos, and video sequences.<br>Sports (1056)<br>Bridge (370)<br>Math (469)<br>Religion<br>Linguistics<br>Geographical names (500)<br>Personal names (3150)<br>Everyday signs (1000)<br>Old and regional signs (1600)<br>http://www.ling.su.se/tsp | Section of Phonetics, Sign Language and General Linguistics, Dept. of Linguistics, Stockholm University |
| PF-star, emotion | Multimodal + Qualisys, video, audio, 2 speakers | Speech, Music and Hearing, School of Computer Science and Communication, KTH |
| Corpora | Swedish (and Thai) longitudinal child language corpora - approximately half a million running words each plus extensive video linkage | Dept. of Linguistics and Phonetics, Lund University |
| Läkarens och lekmannens begreppsvärldar (Doctors' and laymen's conceptual worlds) | Established by: Ulla Melander Marttala<br>Language: Swedish<br>Size: 95 interviews, 15 doctor-patient conversations<br>Language type: A semantic and conversation-analytical study of conceptions and concepts regarding rheumatic diseases. Two parts: 95 semantic in-depth interviews, audio recordings; 15 doctor-patient conversations with interviews and video displays, audio and video recordings.<br>Period: 1986-89 | Faculty of Languages (Anna Sågvall Hein), Uppsala University |

# Appendix C: Evaluation resources

| Evaluation resources (written) | Description | Provider |
|---|---|---|
| Analyzer and corpus texts for North Sami | http://giellatekno.uit.no/english.html | Trond Trosterud, University of Tromsø |
| MaltEval | Software for evaluation of dependency-based syntactic analysis, http://w3.msi.vxu.se/~nivre/research/MaltEval.html | Joakim Nivre, Växjö University, |
| AutoEval | Evaluation tool, language independent | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| MT Quality Evaluation Toolbox | Prototype for evaluation of translation quality and meta-evaluation of evaluation measures, http://stp.lingfil.uu.se/~evafo/software/MTQualEvalTool/ | Eva Forsbom, Department of Linguistics and Philology, Uppsala University |
| Stockholm-Umeå Corpus (SUC) | A balanced corpus of one million Swedish words. Manually annotated with information about part of speech, inflectional morphological features, and base form. Version 2.0 is also annotated with functionally interpreted structures, e.g. information on paragraphs, quotes, abbreviations, and named entities.<br>For linguistic research and training and evaluating language technology systems; SGML; XML; TIGER-XML standard; using Corpus Documentation and Annotation guidelines, http://www.ling.su.se/DaLi/suc/index.htm | Martin Volk, Department of Linguistics (Computational Linguistics), Stockholm University |
| Test data for baseform reduction and wordform generation | 164,000 pairs from DSSO 'The large Swedish lexicon', with PAROLE part-of-speech tags, semi-automatically extracted, http://stp.lingfil.uu.se/~evafo/resources/baseformmodels/ | Eva Forsbom, Department of Linguistics and Philology, Uppsala University |
| KTH eXtract Corpus | Text summarization test data. A small corpus of manually made extracts. | Viggo Kann, Human Language Technology group, School of Computer Science and Communication, KTH |
| Translation test data | | Lars Ahrenberg, Natural Language Processing Laboratory, Department of Computer and Information Science, Linköping University |
| Talbanken | Treebank from the seventies, in a modern format, http://w3.msi.vxu.se/~nivre/research/Talbanken05.html | Joakim Nivre, Växjö University |

| Evaluation resources (written) | Description | Provider |
|---|---|---|
| Swedish-SENSEVAL | Annotated corpus instances with semantic tags and associated lexicon for the instances used. Available for research (Språkdata/Jerker Järborg).<br>Distribution media: Internet.<br>URL: http://svenska.gu.se/~svedk/SENSEVAL/senseval.html.<br>Documentation in Swedish/English.<br>Size: Corpus instances for 20 nouns, 15 verbs and 5 adjectives (8716 training instances 1525 test instances) and detailed lexical descriptions for these 40 words.<br>Content: Swedish (ISO-8859-1), general language (mainly composed of fiction published in 1960 and later, newspaper text), XML, structural annotation, instance-id etc., no linguistic annotation, full formal validation.<br>[More details in the ENABLER questionnaire] | ENABLER: Göteborg University (Dimitrios Kokkinakis) |