2009 Fifth IEEE International Conference on e-Science

The SweDat Project and Swedia Database for Phonetic and Acoustic Research

Jonas Lindh and Anders Eriksson
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
Gothenburg, Sweden
jonas.lindh@ling.gu.se

The project described here may be seen as a continuation of an earlier project, SweDia 2000, aimed at transforming the database collected in that project to a full-fledged e-science database. The database consists of recordings of Swedish dialects from 107 locations in Sweden and Swedish speaking parts of Finland. The goal of the present project is to make the material searchable in a flexible and simple way to make it available to a much wider sector of the research community than is the case at present. The database will be accessible over the Internet via user-friendly interfaces specifically designed for this type of data. Other more specialized research interfaces will also be designed to facilitate phonetic acoustic research and orientation of the database.

Phonetics; Acoustics; Forensic Phonetics;Dialectology

I. INTRODUCTION

The database consists of recorded speech material from 107 Swedish dialects. The recordings were made as part of a research project carried out as a joint effort by the departments of linguistics at Umeå, Stockholm and Lund universities. The research project was funded by the Bank of Sweden Tercentenary Fund for the period 1998–2003. The full title of the project was The Phonetics and Phonology of the Swedish Dialects around the Year 2000. In the following it will be referred to as SweDia 2000 or simply SweDia.

To be able to perform empirically based linguistic research at present it is important to have access to large amounts of reliable data. The data should also be searchable in a user friendly way to facilitate answering research questions without any specialized knowledge of computation or data handling. On-line applications are suitable for some types of research, while other types are better used locally for reasons of security, speed etc. We are trying to provide tools and data for an as large user population as possible for the Swedia database as research questions span way beyond phonetic, phonological and acoustic questions; for example syntax [1] and biometric voice statistics [2]. In this paper we will describe some unique properties of the database, how the current project will continue the development of interfaces, functions and availability of the data.

II. COLLECTING THE DATA

Most of the recordings were made during the summer of 1999. The recording locations were evenly distributed over Sweden and the Swedish speaking parts of Finland, taking into account both geographical distribution and population density. For each location twelve speakers were recorded, representing two age groups – young adults aged 25–35 years of age and an older generation, 55–65 years of age, approximately equivalent to the parent generation of the younger speakers. Both age groups consisted of three male and three female speakers. In figure 1 on the next page the geographical distribution of all the recording places can be observed.

III. SOME UNIQUE PROPERTIES OF THE SWEDIA 2000 DATABASE

There are other dialect databases with a comparable number of speakers and dialects but the present database has certain properties, which as far as we are aware, do not exist in otherwise comparable databases.

Synchronic: All recordings were made within a narrow and very precisely defined time slice. That means that they represent the dialectal variation within the Swedish speaking community at a precisely defined moment in time.

Consistent: The recorded material has three parts that contain precisely defined speech material meant to represent three fundamental, phonological properties of Swedish – the quantity system, the accent system, and the phoneme inventory. This means that it is possible to analyze and compare speech material that is identical for all dialects.

Complete: In addition the database contains approximately 30 minutes of spontaneous speech per speaker, corresponding to 80–100.000 running words per dialect when transcribed. This part provides information about how the various phonological rules are implemented in everyday casual speech. But it may also be used for morphological and syntactic studies as mentioned in the introduction.

These different, well-controlled components make the database a unique source of information for answering research questions of the type mentioned in Section 5 in a systematic, varied and detailed way.

978-0-7695-3877-8/09 $26.00 © 2009 IEEE
DOI 10.1109/e-Science.2009.15

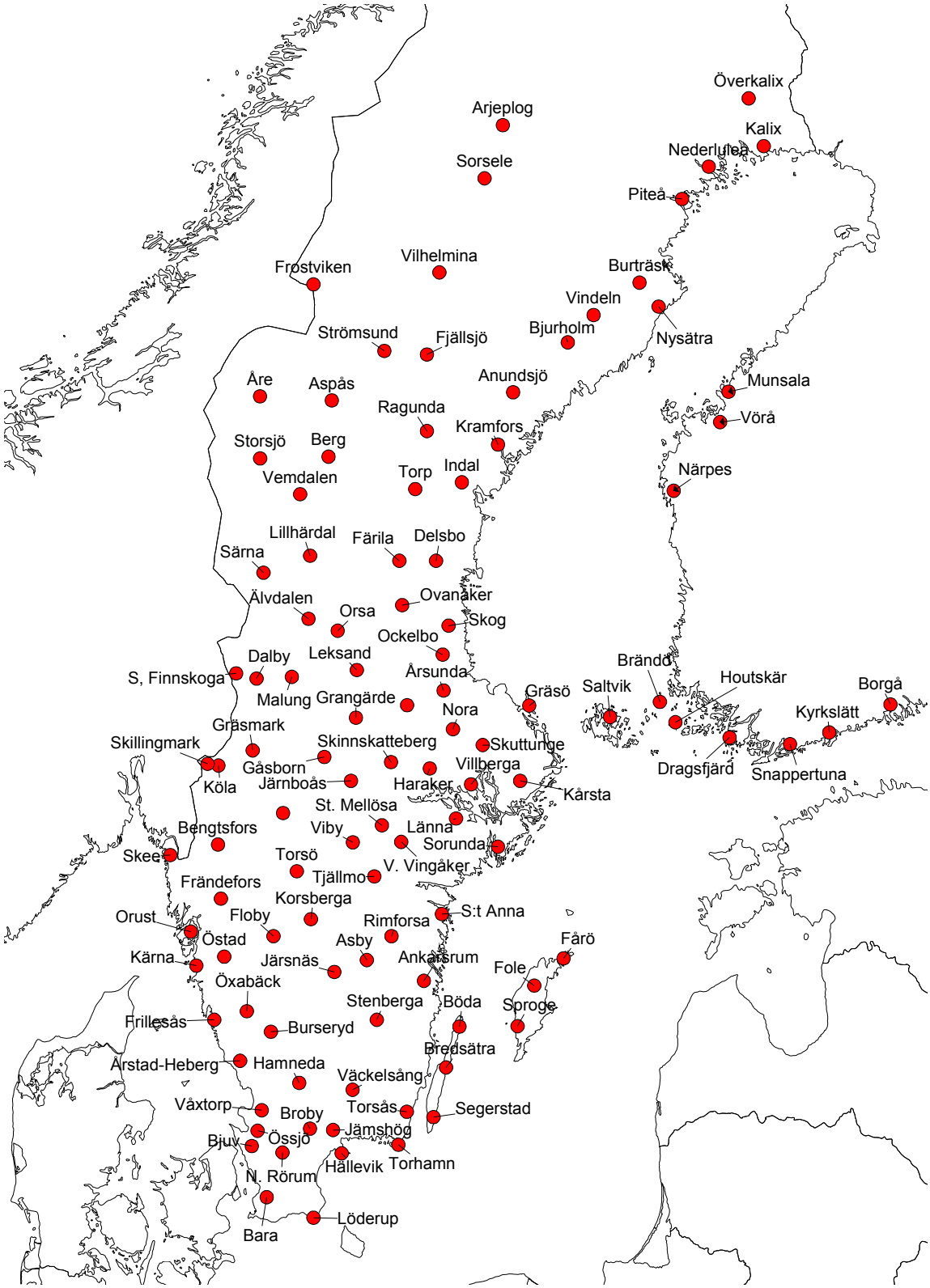

Figure 1. The geographical spread of all places where recordings took place.

IV. SOME SCIENTIFIC CONSIDERATIONS

There are several fundamental scientific questions that may be addressed in a fruitful way by an analysis based on data of the kind contained in the *SweDia* database. One such question which has played a central role in the *SweDia* project will be described in some detail.

In traditional dialectology, geographical dispersion and isolation of groups of speakers as well as renewed contact as a result of migration are considered the driving forces behind linguistic variation and change. There is no doubt that those factors are important, but is that all there is? If that were the case, then variation and change could take any direction and the end result would appear more or less chaotic. We know, however, that that is not the case so there must exist linguistic laws which govern development and change, constrain variation and make certain choices more likely than others. A basic tenet in the *SweDia* project is the belief those laws and constraints can be discovered, described and tested.

Interesting results have been obtained by approaching the description of regional distribution from an angle that does not assume any particular geographically based constraints at all. In three studies ([3], [4], [5]), cluster analysis has been used as a means of constructing dialect "areas" only based on individual acoustically grounded phonological properties. The resulting geographical areas are defined by dialects whose properties cluster together. In principle this could result in a very scattered picture with no obvious geographical coherence. But this does not turn out to be the case. On the contrary, dialects group in geographical areas that in many cases closely resemble those suggested in traditional dialectology. This would be a rather trivial finding if the clustering was based on the same considerations as the traditional analyses, but this is not the case at all. In both studies mentioned above, the cluster analyses are based on acoustic properties never considered in traditional classifications. This lends support for the assumption that dialectal variation is constrained by the compatibility of internal factors.

Dialectal variation as a function of age and gender of the speaker may also be studied using the *SweDia* database. We know from previous studies that these factors play a role, but the size of the *SweDia* database and the fact that the material is identically structured for all speakers represented in the database makes it possible to address this question with a greater degree of precision than has previously been possible.

Another important area of research is typologically oriented studies of sound systems and accent systems. Many examples will be found by looking at the *SweDia* publication list[1] but many more remain to be done.

V. PROJECT GOALS

In this section we will briefly describe and motivate the main goals of the current project named SweDat.

A. Making the database searchable

In order to make the speech data searchable, the audio files have been tagged (a kind of time aligned annotation). The tagging consists of text files containing various relevant labels and information about where in the sound file the corresponding feature may be found. There are several analysis programs available which allow the audio files to be displayed (and listened to) with the tag files time aligned with the audio signal in a separate window. We have chosen a sound format and a tag file format that is compatible with the two most widely used (freeware) analysis programs (*Praat* and *WaveSurfer*) [6] [7].

B. Tagging the Parts of the Controlled Speech Material not yet Tagged

It goes without saying that gender and age are crucially important factors in dialect research. As was detailed above, the database covers variation with respect to these factors. A problem that remains to be solved before these factors may be included in searches at all levels, however, is to complete the tagging of the entire material. The resources available for the project within which the data were collected did not permit tagging of the entire material. As a compromise it was decided that it would be more useful to have all types of material tagged for at least one category of speaker than parts of the material for all speakers. The category we chose for complete tagging was the older generation of male speakers. In addition, as much of the remaining material as could be treated using available economic resources was also tagged. Approximately 85 % of the work has been now completed. This kind of work can only to a limited extent be performed using automatic methods and is therefore very time consuming. However, the current project described here has the goal to complete as much as possible of the tagging and hopefully complete the database.

C. Tagging the spontaneous speech material

The spontaneous speech material has not previously been tagged at all and it was never the intention in the *SweDia* project to do so. The material was primarily meant to be used as an extra resource from which additional material could be retrieved. It would, of course, be a useful resource if the spontaneous speech material had also been tagged to make it searchable, but at the time of the original SweDia project, this would have required an inordinate amount of work and would have increased the costs way beyond available resources. In recent years, however, several methods have been developed for automatic or semi-automatic alignment of a text transcribed from a sound file and the corresponding sound file. We have tested one such program and found that under favourable conditions the alignment of text and sound can be quite good [7]. We have access to orthographic transcriptions of about 50 speakers representing 15 different dialects. These transcriptions were made by transcribers engaged in a small pilot study of dialect syntax done in Umeå and by people working for the *Scandinavian Dialect Syntax* project who have been given access to *SweDia* recordings. What we intend to do within the present SweDat project is to use an automatic aligner to tag those files and see how far we can get. Our own resources will not primarily be spent on transcription work but with developing automatic or semi-automatic alignment routines so that, hopefully, spontaneous speech material from more dialects may be incorporated as orthographic transcriptions become available from various other project groups and individual researchers. Would it not be possible to use the same

[1] http://www.ling.gu.se/~anders/SWEDIA/publ_sv.html

method to semi-automatically transcribe the controlled material one might ask? And the answer is that we are already doing that. But for this type of material it is essential for many types of analyses that the alignment between the tags and the corresponding sound is very precise. So although the semi-automatic methods save quite a bit of time, all the files must nevertheless be checked very carefully by human transcribers to make sure that the precision is at the required level. If one accepts that the tagging of the spontaneous material will primarily be used to search the material for occurrences of certain words or phrases, then the same degree of precision is not required. The alignment must nevertheless be checked manually. The results of this checking will also provide us with valuable information useful for fine tuning the automatic methods.

VI. Developing easy-to-use user interfaces

When the work on setting up the database began towards the end of 1999, the only suitable programs were developed for a UNIX environment. They were fine, worked well and were easy to use for people with a reasonable UNIX experience. They were expensive, however, and so were the machines. But at the time and for the purpose it was a very satisfactory solution. Several search and analysis tools were developed for use on UNIX platforms. Today the situation is radically different. Two of the programs most widely used for acoustic analysis are freeware programs. They cannot be used to run the tools that were developed, but they both accept the sound file formats and the time aligned tag files. So listening to the original recordings and using the time aligned tag-files is possible without any further work. At least one of the programs, *Praat*, provides an environment which would permit the development of all the tools needed to do the work that was previously done on UNIX machines. But the tool scripts and programs must be developed from scratch, using the same ideas but implemented in a different way. This is, however, only the first step. The programs developed for the UNIX environment were largely command line based, which is fine for those familiar with such an environment but an insurmountable obstacle for many of the users we now have in mind. In addition to porting the tools to the *Praat* environment, we must therefore also develop user friendly, intuitive, graphical interfaces which can be understood and used by any moderately computer literate user.

Examples of tools already existing in the "old" system are search tools by which the user can search for all occurrences of a certain word uttered by a given speaker category, say older women, and a freely chosen set of recording locations. As a result of the search, the requested words are copied from the respective sound files and concatenated in a result file together with the corresponding labels from the tag files. The result can then be listened to or saved for further analyses.

Cartography is also an important and much used method for exploring dialect data for regional distribution of various properties. There exist several studies using *SweDia* material which have applied this approach. Three have been mentioned here – an ongoing Doctoral thesis in Groningen, Netherlands, a completed Doctoral dissertation from Umeå and a Master's thesis from Gothenburg. They used cluster analysis or multidimensional scaling to map the results of acoustic analyses on a geographical map. We intend to make some form of automatic mapping of various types of analysis results an integral part of the user interface we plan to develop.

VII. Already in use and ongoing development

So far, some implementations and development has been done and some are still in a developing stage.

A. *Web Based Tools*

When it comes to web based applications we have developed cooperation with the Text Laboratory in Oslo, a unit at the Department of Linguistics and Scandinavian Studies, University of Oslo. They are mainly involved in research and development of new resources such as software and tools for user interfaces, searching, databases, text research, linguistic research and corpora. One of the tools under development is Glossa [9], a web based search tool for researchers in linguistics. The tool can be used for multi-lingual corpora and provides highly sophisticated ways to search annotated corpora and provide you with statistics. The annotated corpora can also include audio and video that can be shown integrated in the browser. We have provided them with parts of our database along with annotations in a suitable format. Some conversion tools between formats had to be developed within our project, however. The opportunity to be part of this joint effort is highly rewarding for our purposes.

B. *Locally Based Tools*

For acoustic and phonetic research purposes it is sometimes better to work with the database in a locally based environment. The database in its current format needs at least around 200 GB of disk space. The tools for acoustic research will be ported to the open source environment Praat. A very suitable plug-in function is provided for developing interfaces with a scripting environment. In this way the database can be provided with a suitable Swedia plug-in with functions for acoustically displaying figures for groups of data or just to search and extract desirable parts.

C. *Forensic Phonetic Biometric Research*

A research group at the department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, is currently investigating the use of automatic methods in speaker identification. To increase reliability, using Bayesian statistical methods, it is important to have access to databases with data representative for the population in question. We intend to use the data in the SweDia database which contains data from about 1300 speakers varying in dialect, speaker age and speaker sex as part of such a population database. For those purposes yet another open source environment for biometric research and development, Mistral/Alize [10], is used integrated with our other tools. So far a pilot study of parameter statistics has been done [11] and several so called universal background models (UBMs) have been created [2] [12].

VIII. Future development of the present project

The aim is now to complete the database with described tools and research environment and make it

available for as many areas of research as possible. As described above several important initiatives and first steps have been taken. However, the ongoing developments are only first steps and a lot of work remains. The result of our efforts will hopefully make the database a very useful resource in our own discipline, but we also hope to contribute to other areas when it comes to handling databases, developing tools and encourage interesting interdisciplinary collaborations.

ACKNOWLEDGMENT

This research is supported by a research grant from the Swedish Research Council (grant # 825-2007-7432).

REFERENCES

[1] *http://uit.no/scandiasyn?Language=en*

[2] J. Lindh, "Pick a Voice among Wolves, Goats and Lambs." Proceedings of IAFPA2009, Cambridge, UK, August 2009.

[3] F. Schaeffler, "Phonological Quantity in Swedish Dialects." (Doctoral Dissertation) Umeå University: PHONUM 10 – Reports in phonetics, 2005.

[4] Lundberg, Jan, "Classifying Dialects Using Cluster Analysis." Master's Thesis in Computational Linguistics, Department of Linguistics, University of Gothenburg.

[5] T. Leinonen, "Classifying Swedish Dialects based on Vowel Pronunciation." Workshop of Production, Perception, Attitude. Leuven, Netherlands, 2009.

[6] P. Boersma, "Praat, a system for doing phonetics by computer." Glot International 5:9/10, pp. 341-345, 2001.

[7] K. Sjölander and J. Beskow, "Wavesurfer – An Open Source Speech Tool." In Proc. ICSLP, volume 4, pp. 464 – 467, 2000.

[8] K. Sjölander, "An HMM-based system for automatic segmentation and alignment of speech." in Proc of Fonetik, Umeå University, PHONUM 9, pp. 93-96, 2003.

[9] L. Nygaard, "Glossa – The Corpus Explorer, Version.0.9" *http://www.hf.uio.no/tekstlab/glossa.html*, 2009

[10] J-F. Bonastre, F. Wils. and S. Meigner "ALIZE, a free toolkit for speaker recognition", in Proceedings of ICASSP, 2005, pp. 737–740.

[11] J. Lindh, "Preliminary Descriptive F0-statistics for Young Male Speakers", Papers from FONETIK 2006, Working Papers, 52, Department of Linguistics and Phonetics, Lund University, 89-92, 2006.

[12] J. Lindh, "A first step towards a text-independent speaker verification Praat plug-in using Mistral/Alize tools", The XXII[nd] Swedish Phonetics Conference, Department of Linguistics, Stockholm University, 2009.