SweDat 2000 – A Swedish dialect research database

Anders Eriksson
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Gothenburg, Sweden
anders.eriksson@ling.gu.se

1. INTRODUCTION

Developing the *SweDat* database is an ongoing project and a continuation of an earlier project, *SweDia 2000* where the data were collected. In the present project, the aim is to organize the collected data into a full-fledged e-science database. The data consists of recorded speech material from 107 Swedish dialects. The recordings were made as a joint effort by the departments of linguistics at Umeå, Stockholm and Lund universities. The project was funded by the *Bank of Sweden Tercentenary Foundation* for the period 1998–2003.

2. DATA COLLECTION

The bulk of recordings were made during the summer of 1999. The recording locations were evenly distributed over Sweden and the Swedish speaking parts of Finland. Twelve speakers were recorded in each location. Informants were recruited from two age groups – young adults aged 25–35 years of age and an older generation, 55–70 years of age. An equal number of male and female speakers were recorded in each age group.

3. PROPERTIES OF THE DATABASE

The *SweDia 2000* database has some properties, which as far as we are aware, are not common in otherwise comparable databases.

Synchronicity: All recordings were made within a narrow and precisely defined time slice. They therefore represent the dialectal variation at a precisely defined moment in time.

Consistency: The material has three well controlled parts that represent three fundamental, phonological properties – the quantity system, the accent system, and the phoneme inventory. It is thus possible to analyze and compare speech material that is identical for all dialects.

Completeness: The recorded material also contains about 30 minutes of spontaneous speech per speaker. This gives us additional information about how observed phonological rules are realized in everyday speech. It may also be used for other types of study; for example studies of syntax and morphology (see below!).

4. LANGUAGE VARIATION DESCRIPTION FROM A SOMEWHAT DIFFERENT ANGLE

Traditionally, the driving forces behind language variation and change are considered to be geographical dispersion and isolation of groups of speakers as well as renewed contact as a result of migration. These factors are no doubt important, but if that were all there is, the observed variation is likely to be more chaotic than what we seem to observe. A basic tenet in the *SweDia* project is the belief that although there is certainly a random element involved in language change it is primarily rule governed. One way of approaching this question is to look for coherence, or clustering of phonological properties within the entire speech community rather than assuming any specific areal distributions.

Promising results along these lines have been obtained by approaching the description of regional distribution from an angle that does not assume any geographically based constraints at all. In three studies (Leinonen, 2010; Lundberg, 2005; Shaeffler, 2005) based on the *SweDia 2000* data, cluster analysis has been used as a means of creating dialect “areas” based only on acoustically grounded phonological properties. In those studies, geographical areas are defined by dialects whose properties cluster together. This approach could, in principle result in a very scattered picture with no obvious geographical coherence. This did not, however, turn out to be the case. On the contrary, dialects

grouped into geographical areas that in many cases closely resemble those suggested in traditional dialectology. If the clustering had been based on the same considerations as in the traditional analyses this would have to be seen as a rather trivial finding, but this is not the case at all. In all the above studies, cluster analyses were based solely on acoustic properties like formant frequencies (Leinonen; Lundberg) or segment durations (Shaeffler) never considered in traditional dialectology. The results in a study by Livijn (2010) on the articulation of coronals, using a similar approach but without using cluster analysis, point in the same direction. Moreover, there is considerable overlap between the areas resulting from these studies. This lends support for the assumption that dialectal change is rather strongly constrained by the compatibility of internal factors.

5. ADDITIONAL USES OF THE DATA

In addition to the research database, there is also a limited version of the database developed for educational purposes in university courses on dialectology, secondary schools and study groups of interested individuals. This database contains speech samples from all dialects represented by short sound files (30–50 seconds) from one speaker per category (age/gender) together with simplified phonetic like transcriptions and translations to standard Swedish. This database may be accessed over the Internet. At present the interface exists only in Swedish. There are no immediate plans to translate the interface.

A group of researcher at Lund university are using material from the database for studies of dialect syntax. They are part of a Nordic network of dialect syntax researchers (*ScanDiaSyn*). Studies of this kind were not envisaged when our data were collected, but we are pleased to see that the data can be fruitfully used also for such studies. To support their efforts we supply the *ScanDiaSyn* database hosted at the University of Oslo with data for their studies.

Although the data were collected for the primary purpose of studying language variation and change in the phonological domain, the usefulness is not necessarily limited to that area. As mentioned above, the data is now used also for the study of dialect syntax and another successful use of the data is as a reference database for automatic speaker recognition for forensic purposes. This has been described in Lindh and Eriksson (2009).

ACKNOWLEDGMENT

The present work on the research database is supported by a grant from the *Swedish Research Council* (grant # 825-2007-7432).

REFERENCES

Leinonen, Therese. (2010). *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects*. (Doctoral Dissertation), Groningen University, Groningen Dissertations in Linguistics, 83.

Lind, Jonas & Anders Eriksson. (2009). The SweDat Project and Swedia Database for Phonetic and Acoustic Research In *Proc. e-Science 2009, Oxford, UK*, 45–49.

Livijn, Peder. (2010). *En perceptuell och akustisk studie av svenskans koronaler i ett dialektperspektiv*. (Doctoral Dissertation). Department of Linguistics, Stockholm University.

Lundberg, Jan. (2005). *Classifying Dialects Using Cluster Analysis*. Master's Thesis in Computational Linguistics, Department of Linguistics, University of Gothenburg.

Schaeffler, Felix. (2005). *Phonological Quantity in Swedish Dialects*. (Doctoral Dissertation), Umeå University: PHONUM 10 – Reports in phonetics.

Links:

SweDat project presentation: http://swedat.ling.gu.se
The research (IMDI) database: http://corpus.sol.lu.se/ds/imdi_browser
The educational database: http://swedia.ling.gu.se
The ScanDiaSyn project home page: http://uit.no/scandiasyn/?Language=en
Reference list of Swedia 2000 publications: http://www.ling.gu.se/~anders/SWEDIA/publ_eng.html