The SweDat project: Making the SweDia 2000 database accessible to the scientific community.

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

The database consists of recorded speech material from 107 Swedish dialects. 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. Both age groups were represented by three male and three female speakers. The database has been used as a resource for many studies resulting in more than 70 publications.

The original database was stored and developed in formats compatible with the ESPS/Waves+ signal analysis environment. This is all very well and the environment contains powerful tools for all kinds of acoustical analyses. However, the complexity of the system and the fact that it is by and large a command-line oriented UNIX system required computer skills and experience by the user way beyond what many potential users possess. This problem was to some degree solved at the time by having access to competent programmers who could help individual researchers extract and process the data they needed for a given study. It is obvious, however, that such technical factors severely limit the number of researchers who may use the database for their research. The goal of the SweDat project is to address this problem in order to make the database available to a much wider sector of the research community than is presently the case. The database should also be accessible over the Internet via user-friendly interfaces specifically designed for this type of data.

In the course of this work many important issues have to be addressed, like choosing suitable sound file formats, annotation standards, metadata, etc. Another issue is interface language. Does it make sense to use any other language than Swedish for a database where all the data are in Swedish? How far should we go in preparing the data for certain types of pre-defined analyses? These and several other policy questions will be presented for discussion at the workshop.

A factor not to be forgotten in this context is the advantage of co-operating with other groups working with similar databases. In this spirit the SweDat project works closely together with people at the Text Laboratory at Oslo University which hosts a similar database. We have even developed our own little slogan for this type of co-operation – one plus one may be more than two – meaning that the joint work of two groups working together may be more than the sum of what each of them could have achieved on their own.