The DOBES Programme and its Contribution to Standardization and Revitalization

Peter Wittenburg, Ulrike Mosel*
MPI for Psycholinguistics
*University of Kiel

The DOBES programme to document endangered languages was started in 2000 and covers currently 21 documentation teams and one archiving team. The most important reasons for the documentation of endangered languages are the following two: (1) It is of greatest importance to preserve for future generations knowledge and information about those languages that will become extinct, since languages are a keystone of Human's intangible heritage. (2) Language material that is gathered and linguistically enriched during the process of documentation may be used to create material for language revitalization projects. The DOBES programme takes both reasons and intentions for the documentation of languages seriously.

During a pilot year intensive discussions about how to best fulfill both goals took place amongst the participants. The participants agreed upon a number of basic guidelines for language documentation projects:

- Language documentation should include the cultural background of the language in focus.
- The corpus should consist of a variety of text types and genres.
- Multimedia (sound and video) recordings form the basis of the documentation work. These recordings should be associated with an orthographic or phonemic transcription, a translation in one of the major languages of the world, and/or glossings in a local lingua franca and English.
- For some material a deep linguistic analysis should be provided such that later researchers will be able to reconstruct the (grammar of the) language.
- A topic-oriented lexicon should be provided as well as a sketch grammar, field notes, ethnologic descriptions etc.
- The language archive should be as open as possible; however, ethical and legal aspects have to be taken very seriously.

With respect to archiving the first priority was to look for standards and strategies to ensure accessing the archive via modern media such as the Internet. Moreover, the question of long-term survival of the material in our digital networks was one of the most important issues here. The following principles were worked out; they are now generally agreed upon amongst the major archives world-wide:

- Given the short life-time and the vulnerability of the current storage media, only continuous migration to new technologies and wide distribution of the data will ensure their survival.
- The archive content has to be based on stable and open standards such as MPEG and WAV for media streams, JPEG and TIFF for images and XML and UNICODE for textual material.
- With the long-term preservation goal in mind participants in language documentation projects were strongly advised to use good quality speech recording techniques.
- Relying on such standards should ensure seamless access to the archive for the years to come, however, no guarantee can be given for the seamless interpretability of the data after longer time spans (like, e.g., a hundred years or so).
- Well-organized corpora have a better chance to be easily accessed and maintained by future generations. Therefore all resources are described organized by metadata according to the IMDI standard.

The first teams that started their language documentation projects about 3 years ago are close to finishing their work. Although a more detailed look at all of their results can only be done in the second half of this year, it can be stated already now that these documentation projects very successfully reached their goals. Despite different fieldwork situations a convergence of methods and objectives can be observed. Given the fact that there were no documentation standards before, this convergence increased the confidence and the productivity of the language documentation teams.

Based on these data, the archivists have been setting up a well-organized and web-visible corpus consisting of more than 700 sessions that cover more than 1.500 hours of data recordings. State-of-the-art tools are available to the language documentation teams for annotating and analyzing their data. Moreover, conversion tools allow for converting all incoming data into standard formats. All material is automatically copied 5 times to ensure their security (2 copies remain in the archivist's building, one is kept at the campus, and two copies are stored in two large computer centers in Germany).

The DELAMAN network intends to develop technologies within the next 5 years that will allow for the world-wide distribution of the archived data on endangered languages and deal with the ethical aspects involved. Appropriate Data-Grid systems will be realized with the support by UNESCO and other such institutions.

Recently, a metadata-based framework was developed that allows archive managers and researchers to define access rights to the archived resources in a highly efficient way. It is hoped that this will encourage the researchers to make as much data openly available as possible. An Advisory Board was established to solve possibly complicated access rights problems.

Having established and consolidated this first phase of language documentation and archiving, the DOBES archivist has started discussing and implementing suitable ways of data access for different user groups, ranging from language community members to journalists and researchers. The presentation should ensure easy access to the archived materials for the members of the documented speech communities (via the web). Moreover, fully functional local data copies finally have to be transferred to local community centers. In addition we started working on transformation rules that allow researchers to create pedagogical course material from the XML data representations stored in the archive.

Language documentation with its careful linguistic analyses of the properties of a language and the culture of its speakers is seen as a necessary step to intensify language revitalization.

The proposed talk will present the documentation standards within the DOBES programme and illustrate the current state of the archive. It will discuss long-term data preservation strategies and ways of how to offer the archived data to different user-groups.

Our paper ***The DOBES Programme and its Contribution to Standardization and Revitalization*** is a contribution to the

Congress on Language Diversity, Sustainability and Peace

and there to the workshop B

Case Studies of Language Revitalization and Standardization

Abstract Length: 941 words

Author Information

Peter Wittenburg is technical director at the MPI for Psycholinguistics and also responsible for setting up the DOBES archive. He is also one of the initiators of the DELAMAN network that will bring together some major archives world-wide storing endangered languages and music material.
Ulrike Mosel is professor for linguistics at the university of Kiel and is responsible for the Teop documentation work within DOBES. Teop is a language spoken on the Solomon islands.
Both authors are members of the steering committee of the DOBES programme.

Contact Information

Peter Wittenburg
MPI for Psycholinguistics
Postbus 310
6500 AH Nijmegen
The Netherlands

Tel: +31-24-3521113
Fax: +31-24-3521213
email: peter.wittenburg@mpi.nl
web: http://www.mpi.nl
http://www.mpi.nl/DOBES