# Language Archive Newsletter

LAN Vol. 1, Nr. 2 April 2004 ISSN: 1573-4315 Contact: LAN@mpi.nl http://www.mpi.nl/LAN/
Editors: Hennie Brugman, Romuald Skiba (responsible), Peter Wittenburg

## Content:

New DoBeS Teams

Loretta O'Connor & Peter Kröfges - Lowland Chontal of Oaxaca 1
Dagmar Jung - Beaver knowledge systems: documentation of a Canadian First Nation language from a placenames' perspective 2
Frank Seifart, Nikolaus Himmelmann, Doris Fagua, Jürg Gasché and Edmundo Pereira - Documenting the Languages of the People of the Center, Especially Bora and Ocaina (North West Amazon) 2
Anna Margetts - Towards the documentation of Saliba/Logea an endangered language of Papua New Guinea 2

New developments

Herbert Baumann, Reiner Dirksmeyer, Peter Wittenburg - Long-Term Archiving 3
Daniel Broeder, Freddy Offenga - IMDI Metadata Set 3.0 3
Hennie Brugman – ELAN Releases 2.0.2 and 2.1 4
Romuald Skiba, Florian Wittenburg & Paul Trilsbeek - New DOBES web site: contents & functions. 4

News in brief

Andreas Claus - Access Management System 5
Jost Gippert - DoBeS conference and summer school 5
Asifa Majid – Data elicitation methods 6
Paul Trilsbeek - DOBES Training Course 6
Peter Wittenburg - Training Course in Lithuania 6
Peter Wittenburg - New Person at Archiving Team 6

## New DoBeS Teams

### Loretta O'Connor & Peter Kröfges - Lowland Chontal of Oaxaca

The Volkswagen Foundation has funded a three-year project to document Lowland Chontal of Oaxaca, an unclassified and highly endangered language spoken near the Pacific coast of southern Mexico. Principal investigators are linguist Loretta O'Connor (University of California, Santa Barbara, and Max Planck Institute for Psycholinguistics, Nijmegen) and anthropologist Peter Kröfges (State University of New York, Albany), with project director Prof. Dr. Ortwin Smailus (University of Hamburg).

The ethnic designation 'Chontal' derives from the Nahuatl term *Chontalli,* meaning 'stranger', which the Aztecs used to refer to various unfamiliar ethnic groups in ancient Mesoamerica. After the Spanish conquest, the Oaxaca Chontales were long perceived as barbaric cave-dwellers, a stereotype that hampered anthropological and linguistic scholarship. In many respects, the area still represents a land of strangers.

This project builds on data and analysis begun during the investigators' doctoral research. Primary linguistic results will include a series of thematically-based Chontal-Spanish dictionaries, a comprehensive grammatical description of the language, and an archive of digitized recordings and annotated texts. The anthropological component will focus on the documentation of local knowledge of landmarks, settlements and territorial boundaries, soil classification and agriculture, and sacred sites and religious practices. Members of the Chontal community will participate in all activities and will share in the results.

### Dagmar Jung - Beaver knowledge systems: documentation of a Canadian First Nation language from a placenames' perspective

Beaver is an endangered Northern Athabaskan language spoken in several communities in Canadian British Columbia and Alberta. The present number of speakers is estimated by our team as ca. 60-80 fluent speakers in British Columbia , and ca. 20-30 fluent speakers in Alberta. The documentation aspect of our study focuses on narratives of place, thereby creating a 'conceptual map' of the Beaver territory that has been defined by a traditional hunter-gatherer society. The heavy textual component provides the opportunity to collect 'rich' data, i.e. contextual data relating the significance of places to individuals and the overall community.

The core team members: Dagmar Jung (assistant professor at the University of Cologne, Linguistics) has a background of research and fieldwork on Southern Athabaskan languages, especially Jicarilla Apache. Julia Miller (PhD student with Sharon Hargus at the University of Washington in Seattle) started to do phonetic research on Beaver tone two years ago. Olga Müller (PhD student at the University of Cologne) works currently as a research assistant on a dictionary project of Tanacross Athabaskan. Patrick Moore (assistant professor of Anthropology at the University of British Columbia) has extensive experience documenting Kaska, one of the neighbouring Athabaskan languages.

### Frank Seifart, Nikolaus Himmelmann, Doris Fagua, Jürg Gasché & Edmundo Pereira - Documenting the Languages of the People of the Center, Especially Bora and Ocaina (North West Amazon)

This project aims at documenting the endangered languages of the People of the Center, a culturally relatively uniform, but linguistically diverse group in the Peruvian part of the North West Amazon. Speaking seven mutually unintelligible languages, the People of the Center are characterized by some unique cultural practices, including completely memorized ritual discourses that may last up to three hours, repertoires of thousands of songs performed at festivals, as well as efficient systems of drum communication that build on the structures of the individual languages. Two of the seven languages, Bora and Ocaina, will be the subject of exemplary and comprehensive documentations, consisting of fully annotated video recordings of a representative sample of each major type of communicative event, including formal and informal discourses as well as drum communication. In addition, specimens for the moribund language Resígaro (three native speakers), will be included, as well as old audio recordings for Witoto which document types of ritual speech no longer practiced today. Taken together, this data set will be a representative documentation of the linguistic and cultural practices of the People of the Center as a whole.

### Anna Margetts - Towards the documentation of Saliba/Logea an endangered language of Papua New Guinea

Saliba and Logea are two closely related dialects spoken on neighboring islands in Milne Bay Province, Papua New Guinea. The estimated number of speakers  is 2,500. The dialects belong the Papuan Tip Cluster of the Western Oceanic language group.

Given that the community of speakers is traditionally small, Saliba/Logea must be considered highly endangered as English is encroaching on many aspects of daily life. While the degree of endangerment is serious, the documentation capacity is still very good. The Saliba and Logea people are continuing to lead a traditional life of fishing and subsistence farming and it is still possible to work with the last generation of old speakers who have essentially no knowledge of English, as well as with children who are growing up as monolingual speakers, at least in the first few years of their life.

The languages of the Papuan Tip Cluster are of special typological interest as they show features not found elsewhere in the Oceanic language group. Some of these features may be explained by early contact with Papuan languages.

The project aims at a multimodal documentation of the language in its cultural context. The main investigator will be Anna Margetts (Monash University) who wrote her Ph.D. thesis on aspects of Saliba grammar. The German host of the project will be Ulrike Mosel at the University of Kiel.

The team also includes John Hajek from Melbourne University who will work on phonetics and phonology, Rhys Gardner from the Auckland Museum working on ethnobotany and Andrew Margetts documenting the building and use of sailing canoes. The team is also seeking a German Ph.D. student in Linguistics to join in the documentation project.

# New developments

### Herbert Baumann, Reiner Dirksmeyer, Peter Wittenburg - Long-Term Archiving

It was reported frequently that two aspects are important to increase the chance of a survival of the bit-stream representations of the material we are storing about languages and music traditions that will be extinct soon. (1) The data has to be migrated frequently to guarantee that state-of-the-art storage media are used that are fully supported by hardware and software. (2) The data has to be copied and distributed to cope with all kinds of risks – even political ones - that could destroy the storage media used. The MPI team has finished its activity to have at least 5 copies of the DOBES data. Two copies are automatically created in the MPI storage system (RAID Disk Array and Tape Library). A third copy is stored on a standard PC system having a large RAID Disk Array that is within the control of the MPI team, but located in another building.

A fourth copy is transferred to the computer center of the Max-Planck-Society in Göttingen (GWDG) by using the RSYNC protocol provided for example in standard UNIX systems. The transfer is initiated by the GWDG, the protocol is efficient but lacks modern encryption capabilities. To achieve the full transmission speed of 5 MByte/sec five sessions are started in parallel. A fifth copy was generated in the mean time at the other computer center of the Max-Planck-Society in Munich (RZG). Here the well-known Andrew File System (AFS) is used as protocol. At the MPI an AFS client was installed that establishes connections with the AFS server in Munich, i.e., the transfer is initiated by the archivist. AFS makes use of state-of-the-art authentication and encryption. Also here several channels are opened in parallel to achieve the full 2.5 MByte/sec exchange speed.

Both procedures guarantee that at regular intervals the changes in our DOBES archive are synchronized with the two computer centers. At both centers, GWDG and RZG, local strategies are applied to maintain several copies of all stored data in different buildings, i.e., all DOBES data is now stored in at least 7 different storage systems. The DOBES archivist sees it as an advantage that two different protocols are applied and that the two centers use different storage technologies. Currently, the Max-Planck-Society discusses at a high level what kind of guarantees can be given to the institutions for long-term storage support for their data, since many disciplines share this fundamental problem.

In the committees dealing with this question of long-term preservation it is consensus that guaranteeing the interpretation of the bit-streams is a task of the community and not the centers. Adherence to open standards and organizational, encoding and format coherence will be relevant criteria to determine the chance that the data will be migrated in time to state-of-the-art representation standards.

### Daniel Broeder, Freddy Offenga - IMDI Metadata Set 3.0

Based on the experiences and on a broad discussion process including field linguists, corpus linguists and language engineers, the IMDI set 3.0 [1] was designed as part of the INTERA and DOBES projects and is available as an XML-Schema. It was adapted to simplify the content description and the artificial distinction between collectors and other participants - probably influenced by Dublin Core - was removed.

Three major extensions were applied: First, it is now possible to describe written resources that are not annotations or descriptions. This was necessary, since most language collections contain written resources in the form of field notes, sketch grammars, phoneme descriptions and others more. Second, as a consequence of long discussions with participants of the MILE lexicon initiative, it is now possible to describe lexicons with a specialized set of descriptor elements. Third, it is now possible to define and add project-specific profiles. In the earlier version IMDI supported already the possibility of extensions at various levels in the form of user defined category–value pairs, i.e., the user was able to define a private category and associate values with it. This feature was used by individuals and also projects to include special descriptors, however, these descriptors were not fully supported by the IMDI tools. In the new version projects or sub-domains such as the Dutch Spoken Corpus respectively the Sign Language community can define a set of important categories and these are supported while editing or searching.

Therefore, IMDI exists of its core definitions that have to be stable to assure users that their work will be exploitable even after many years and of sub-community specific extensions, which nevertheless are result of discussion processes.

[1] Detailed description of the IMDI 3.0 metadata elements:
http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.4.pdf

[2] IMDI Web-site: http://www.mpi.nl/IMDI/

## Hennie Brugman – ELAN Releases 2.0.2 and 2.1

The version 2 is a major upgrade. Elan's viewer and media handling internals are completely re-engineered, as is the handling of user commands. The user interface is completely redesigned, including shortcut keys.

Main new features and changes:

- All viewers for one annotation document are now shown in one document window. The video panel can be detached into a second window. This can for example be useful to display MPEG-2 video on a separate monitor.
- Several new and/or revised viewers (for details see: http://www.mpi.nl/tools/elan/release-notes.html)
- 'Save As' is now supported.
- Time selections are now made or modified in a completely new way. Next to dragging and shift-click in the time line viewer or wave form panel Elan now has a special 'selection mode': all time navigation and playback buttons modify either the begin or the end of the selection when in selection mode.
- Two time-synchronized video panels are supported now. The user can specify the begin time for each of the two separately.
- Media files do not have to have the same name as the matching .eaf file anymore, and do not have to be in the same directory either. When files can not be found at the locations stored in the .eaf file, first the eaf file's directory is checked, then the user is prompted to specify a location.
- Elan's user interface can be localized on the fly. Currently supported languages are English and Dutch. It is now easy and straightforward to support other languages. Volunteers for translation of English user interface texts to some other language are welcome.
- Formats (for details see: http://www.mpi.nl/tools/elan/release-notes.html)
- Time accuracy: all times in all viewers are correct, and synchronized at all times. There is one annoying issue that can not be fixed on the short term: when playing a time selection, playback of the video continues a few frames after the end of the selection. How much depends on the computer or operating system running Elan. Right after this 'overshoot' the media time is set to the exact end time of the selection, resulting in a little jump in the video playback. Audio does NOT have this problem.
- Support for template documents to make reuse of tier setups easier.
- New Unicode input methods for Korean, Georgian and Turkish.
- Preferences are now stored between Elan working sessions. These are both preferences for Elan (like last used directories for eaf files, media files, shoebox type files) and preferences for individual documents (like media time, selection, active tier, etcetera).
- Even if media files for some eaf file are completely missing the document can still be opened for inspection and modification.
- A 'shift' mode is added to help alignment of imported data. Unlike the already existing 'bulldozer' mode gaps between annotations are maintained.

## Romuald Skiba, Florian Wittenburg & Paul Trilsbeek - New DOBES web site: contents & functions.

The new DOBES web site combines the information that was available on the old site with an adaptation of the DOBES-DEMO that was created for the VW-endorsed exposition "Science + fiction". The latter part is created in such a way that it is informative for the general public and not only for specialists.

The layout of the site allows for navigation in different ways. On the left side you find a traditional navigation panel for accessing different parts of the site quickly, e.g. using the Site Map. The main section on the right side starts off with a graphics based presentation that is intended for exploration rather than quick access. By slowly moving over the interactive dots, which are marked white, different topics of the site can be accessed.

The main page has three sections: Documentation, Endangerment and Languages. Each section is again subdivided in a number of topics:

*Languages*
Under this section you can find information on the following topics:

- projects: contains links to the websites of the individual DOBES projects
- data-types: gives an overview about the different sorts of data contained in the data base (arts & handcraft, religion & medicine, dance, music, everyday work, environment)
- field work locations: shows the worldwide location of the places where fieldwork for the DOBES project is done
- transcripts: contains several examples of modern and traditional transcriptions
- annotations: contains examples for grammatical analysis and translation of language samples

including direct links to the underlying media (videos)
- meta data: illustrates among other things how the IMDI Editor works (a tool for entering and organizing metadata)

*Endangerment*

Under the section Endangerment you can find some explanations about reasons of endangerment (e.g. death of speech communities, religious education, cultural dominance, industrialization, social reputation). Selected quotations from David Christal's book "Language death" are presented. The following submenu points are accessible:

- endangerment
- revitalization

The crucial points of revitalization are: shaping awareness and the creation of a positive image of the language, availibility of material on the internet and using of new technologies, culture specific teaching methods, teaching material (examples from DOBES are given), regional centers for language instruction and minority rights.

*Documentation*

Under the section Documentation you can find information on the following topics:

- goals (e.g. scientific analysis, archiving, material for teaching)
- stages shows the different stages that a prototypical piece of recorded data has to pass: recording - digitization - editing - metadescription - annotation - integration.
- tools (for some of the steps mentioned under "stages")
- archive (illustrates how the data are organized in the archive, what is done for the security of the data etc.).

At the moment the site is written in such a way that it works well with Internet Explore on Windows, with Windows Media Player to play the audio and video files. You can still view the site with other browsers and operating systems, but some things may not work. We will try to make the site more platform and browser independent in future versions.

We hope you will enjoy using the site!

http://www.mpi.nl/DOBES/

# News in brief

### Andreas Claus - Access Management System

We have released the first version of the Access Management System (AMS) for the Corpora housed by the Max-Planck-Institute in Nijmegen. The AMS can be accessed by clicking on the "set access rights" link at the URL http://corpus1.mpi.nl/BC/IMDI-corpora/

As of this release only the project coordinators have accounts. The coordinators (definers) can create groups and accounts. There are two kinds of accounts - users with read-permit and account managers (definers). Account managers can have the same rights as the project coordinators themselves: they can create accounts, groups and rules for the ARM.

Optionally the account managers can associate an acceptance declaration that pertains to the data in the archive. All users must agree to this declaration the first time they log in. The inclusion of the acceptance declaration is the first step towards a more elaborated AMS in the second version.

We also see the need that users should have the possibility to enter feedback to the results of the usage (e.g. references).

The resources are by default not accessible to everybody. It is possible that they can be made accessible to a certain group or to the world. Access can be defined for all video-, audio-, image-, info- and annotation-files which are linked to the metadata. You can define different rights for each of these types of data. By default only the metadata files are accessible to all.

The access rights are hierarchically organised. A change at a higher point in the corpus structure will be handed down to the 'child records'.

### Jost Gippert - DoBeS conference and summer school

The Volkswagen Foundation has confirmed the funding of both the conference on "A World of Many Voices" (Frankfurt, Sep. 4-5th, 2004) and the summer school on Language documentation (Frankfurt, Sep. 1-11th, 2004).

The ten-day summer school is intended to introduce promising students (max. 50 persons) of linguistics and adjacent disciplines (ethnology, anthropology, African Studies, Asian Studies, etc.) into the aims, objectives and methods of fieldwork with a view to the documentation of endangered languages. The participants will be taught and trained by members of the DoBeS programme and other internationally renowned specialists. The teaching will be undertaken in form of lectures, lecture tutorials, and seminars; the application of fieldwork methods will be trained in fieldwork tutorials. Please note that the deadline for applications is May 15th, 2004.

More details under:
http://titus.fkidg1.uni-frankfurt.de/curric/dobes/ssch2cir.htm

### Asifa Majid – Data elicitation methods

The Language & Cognition group of the Max-Planck Institute for Psycholinguistics is involved in language documentation, i.e., describing previously under-described languages; linguistic typology, i.e., establishing how similar and different languages are from one another; and investigating the relationship between language and thought. To this end, the group maintains about a dozen fieldsites around the world at any one time in which research can be conducted in a sustained way, using a full range of anthropological, linguistic and psychological methods.

In order to conduct comparative research, the Language & Cognition group publishes a *field manual* annually. The field manual consists of a series of tasks to help researchers in different fieldsites to collect data in a standardised way. The tasks belong to one of the core projects of the group, such as Space, Event Representation, or Multimodal Interaction. Each task addresses a specific research question about language documentation, linguistic typology or the relationship between language and thought. For further details please contact the Language & Cognition group, and see the website http://www.mpi.nl/DOBES/INFOpages/Fieldmanuals-LAC/index.html

### Paul Trilsbeek - DOBES Training Course 2004

A new training course is scheduled for the second week of may (10 to 14 May). This course is devoted to very practical matters as they are relevant to keep the documentation work within DOBES at a maximum level of coherence. Therefore, it is dedicated to participants that primarily come from existing and new DOBES teams. In contrast to the DOBES summer-school that is directed to a broader scope of topics relevant for the documentation work and to interested young people, the coming training course has to cover topics such as the concrete agreements within DOBES and the necessary workflow aspects as well. We invite everyone to comment on the suggested schedule (see http://www.mpi.nl/DOBES/training/training2004program.pdf).

### Peter Wittenburg - Training Course in Lithuania

Due to his experience gathered within the DOBES and ECHO projects Peter Wittenburg was invited by the UNESCO to carry out a 5 days workshop about "Digital Archiving" for the major cultural heritage institutions in Vilnius (Lithuania) together with a colleague from Lund University. The members of the various institutions participated with great enthusiasm and the workshop mutated to an interactive seminar about ongoing developments. The program was modified almost every day to fit with the expectations of the participants as closely as possible. The major topics were metadata, metadata interoperability, archiving standards, container models, architectures for long-term preservation of digital data, the difference between presentation and representation formats and management issues. The final agenda of the workshop that took place at the Lithuanian Folklore Center is available under http://www.ling.lu.se/projects/echo/contributors/events/vilnius_course.html

Most of the presentations were developed online on flip charts, however, and are now owned by the Folkcenter in Vilnius that took care of an excellent and creative environment and atmosphere.

### Peter Wittenburg - New Person at the Archiving Team

The MPI team in DOBES realized that more conversions will have to be carried out to come to a coherent archive of language resources. In particular in the area of textual material, people are obviously using different tools and mixing various character sets. All this material has to be converted to proper XML and in the case of annotations to EAF. To better cope with these needs Paul Trilsbeek was integrated into the team. He will take care of technical archive matters and will interact with DOBES members about these aspects. Paul has a musical background and will also become active in the ethnomusicology working group.

Send contributions for the next
issue to: LAN@mpi.nl
before June 31, 2004