LANGUAGE ARCHIVES NEWSLETTER

No. 5
May 2005

# In This Number:



*Language Archives Newsletter. www: http://www.mpi.nl/LAN/. Mail: LAN@mpi.nl. ISSN: 1573-4315.*
*Editors: David Nathan, Romuald Skiba, Marcus Uneson. Layout: Marcus Uneson.*

# Editorial

Dear Readers,

LAN is now into its second year. It has grown from a DoBeS-team newsletter presenting mainly technical information into a broader newsletter, covering a number of archives and a wider variety of topics. To mark this change, its name has been changed from Language Archive Newsletter to Language Archives Newsletter. The issue-numbering system has also been simplified.

This issue addresses topics ranging from the usage of documentations in support of local language programs to an introduction to the PARADISEC archive based in Australia; it introduces a new DoBeS team; it continues our recent call for equipment reviews; and it forwards a request from field linguists Umberto Ansaldo and Lisa Lim for assistance to communities affected by the recent Asian tsunami. And of course there is all the usual technical fare!

Best wishes,

*David Nathan, Romuald Skiba, Marcus Uneson*

# Endangered Languages

## Language Documentation and Development of Teaching Material

*Utta von Gleich*
Universität Hamburg

### Introduction

This article highlights both the usefulness and the limitations of language documentation (LD) as a source for preparing teaching material in endangered languages. In the wider context of language maintenance and revitalisation, using LD as a source provides opportunities for profitable interaction between field researchers, language communities and developers of materials. Although the main objective of the DoBeS project is to create multimedia documentation and standardized archiving of endangered languages, multimedia teaching materials are welcome secondary products that can benefit speech communities in their endeavours to revitalise their cultures and languages.

The major interest of endangered speech communities is education for mother tongue literacy. Literacy, in this sense, is more than the acquisition of purely technical skills. The following definition was developed at the UNESCO (2004) meeting for the assessment and evaluation of literacy worldwide:

> 'Literacy' is the ability to identify, understand, interpret, create, communicate and compute, using printed and written materials associated with varying contexts. Literacy involves a continuum of learning in enabling individuals to achieve his or her goals, develop his or her knowledge and potentials, and participate fully in community and wider society.

In this article, I use "literacy" as a socially relevant cultural activity (v. Gleich 2001). This is generally considered to be an essential, although not exclusive, aspect of language planning for language revitalisation (Fishman 1991, Hornberger 1997), and has become the challenge of the 21st century (López & Küper 2002, Chatry-Komarek 2003, Ouane 2003).

### The scope of a language documentation

A language documentation can deliver valuable inputs for language revitalisation and the production of teaching material if it is properly designed, produced, and archived to be a lasting multipurpose record of a language (Wittenburg 2003:122). Which types of LD are suitable for this purpose?

In 1996, Nikolaus Himmelmann defined a LD as greatly exceeding the scope of traditional language descriptions (Himmelmann 1996). A LD should:

- provide a multifunctional record of the linguistic behaviour and knowledge found in a given speech community;
- use several types of media: video, audio, images, text (transcription, description, analysis), metadata;
- include linguistic and metalinguistic annotation of the language;
- contain a wide range of observable linguistic behaviour, such as spoken language in a variety of styles and contexts, recorded material (audio, video, and/or text) with transcriptions, translations and annotations, relevant sociological and cultural information, dictionary, thesaurus, pedagogical materials and grammar; and
- be accessible to and helpful for a broad range of members of the respective language community.

I concentrate here on the development of teaching material for primary education. In most DoBeS projects, the associated speech community expects or claims support for revitalisation measures and researchers are willing to contribute, supporting the development

of an alphabet and orthography or participating in the development of materials. Such activities are recognized as legitimate products of documentation projects in the latest Volkswagenstiftung report (2004:7).

## Teaching and learning in indigenous languages

The first decision to be taken is how, to what degree, and for how long endangered languages can be used in education. Then we need a thorough consideration of the linguistic, multilingual, and pedagogical contexts in which multimedia language materials might be used for teaching endangered languages. Decisions have to take into account both the language community's participation in decision making and national education policies. Developing teaching materials in endangered languages is part of an active strategy in documentation and revitalisation. Teaching should encourage functional language usage and the modernisation of the language. Two major models of language education can be distinguished: monolingual, and bilingual/ multilingual.

In monolingual education, the endangered language is selected as mother tongue and medium of instruction in a variety of subjects. It allows revitalisation to take place through teaching at the primary level. Communities may opt for this type of education if (as might be the case for isolated speech communities) there is no need for every member to communicate with the dominant language groups. "Etnoeducación" in Latin America is an example.

Multilingual education models (often called bilingual education) have two socio-political objectives:

- *Homogenization*: social and cultural assimilation of linguistic minorities into the mainstream via transition programs. The endangered language is used as the language of instruction until students are sufficiently competent to transfer to learning in the dominant language(s).
- *Diversification*: maintenance of cultural and linguistic diversity by using endangered languages as language of instruction throughout schooling alongside the gradual introduction of official or co-official dominant languages (also known as language maintenance programs).

A thorough documentation could be a crucial resource for supporting education in indigenous languages as language of instruction in subject matters such as science, maths, modern terminology, native knowledge, and metalanguage for teaching languages and in general.

The usage of endangered languages in basic education can address communities' fear of losing their identity as a result of language loss, and can demonstrate that the language can serve as a language for literacy in competition with dominant languages. From a pedagogical perspective, the justification of a bilingual approach is quite simple: learning and teaching in a code common to both students and teachers guarantees better academic results (Cummins 2000).

## Producing multimedia teaching material using LD

A major factor in planning is whether, in the process of revitalisation, the language is to be taught as a mother tongue or as a second language. What kinds of courses will be delivered and who will they be for? There might be adult literacy courses combined with other training, such as basic education, formal primary teaching, or language and culture workshops. These options should be considered within the framework of the public education curriculum. It is advisable to consult the responsible education authorities, local, regional and central, and present your documentation project. Try to win their confidence and formal approval and decide together with them which kind of teaching materials can be developed. Also consider who can develop materials–for example, teachers and students may be able to contribute.

As a prerequisite to developing teaching material, researchers should check the contents of their LD to see what they can offer with regard to:

- an accepted alphabet, writing system, and orthography;
- grammars;
- knowledge of dialectal varieties;
- an analysis of spoken language, including different registers for children and adults;
- collections of oral literature and oral communication;
- culture specific vocabularies in natural sciences, social topics; and
- monolingual and bilingual dictionaries.

Possible outcomes that could be supported by a language documentation include:

- training and development of oral and written communication competences;
- a range of genres (from everyday communication skills to narratives, tales, reports, songs, music etc.);
- introduction to reading and writing in a standard orthography; and
- development of a body of literacy and literacy practice.

In adult education classes, students can benefit from participation in the production of media, which helps to overcome feelings of exclusion and disparagement of the language and culture. In Bolivia, for example, the production and use of radio-broadcast primary courses in the major indigenous languages by religious groups are proving useful due to the integration of multimedia

documentation materials such as songs, poems, tales, proverbs, or riddles.

CEFREC (Centro de Formación y Realización Cinematográfica), the video production training centre directed by Ivan Sanjinés (Grebe & v. Gleich 2001), also known as "Video indígena", has included ethnographic video in Latin American video production since 1980. Both this and recent television news programmes in indigenous languages have become valuable components of language documentation and contribute to the effort to preserve indigenous languages. Furthermore, I recommend learning from the experiences of the THOA (Taller de Historia Oral Andina) Workshop on Oral Andean History, an independent research group active in La Paz, Bolivia since the 1980s, which works directly with indigenous communities in the preservation and documentation of cultures, producing radio serials, theatre, reading material, etc. There still remains a great demand for documentation of the smaller Amazonian and Lowland Bolivian languages.

## Limitations

Despite the apparent advantages of multimedia teaching material, development of teaching materials in or for endangered languages requires an interdisciplinary team with skills in linguistics, anthropology, sociology, and pedagogy (Chatry-Komarek 1996), to work together with local teams of teachers, parents and others. Teams should work together with an independent research institute or an editorial network (e.g. APNET 2004). They should also cooperate with local education authorities to gain official recognition for the materials, and to improve prospects of being recognised, financed and distributed as part of the public education system.

The individual researcher can only contribute elements to the whole objective. If he/she has no teaching training or experience, it will be difficult to develop teaching materials without assistance. (For helpful advice, see Chatry-Komarek 1996 regarding materials developed for Quechua and Aymara in Peru, and Komarek (1998) for materials for Malgache in Madagascar and African languages in Ghana.)

Teaching one's own mother tongue without prior training is an even more difficult task for speakers of indigenous local languages, because the language may be undergoing standardization, and teaching materials may not be available. Therefore native speakers may need specialized training and follow-up.

There are two ways of assisting:

- The researcher can organize text collecting and text composing events in the community (both oral and written texts). Such events in Bolivia had more than 4000 primary school teachers participate (Proeibandes 2002) and generated a considerable amount of texts.
- The researcher can help teachers with grammar. Teachers will want to understand the structure of their language and how to integrate grammatical knowledge into teaching. Many will want to have access to a reference grammar for their classes, because they are expected to teach the language in a normative way.

As endangered language communities are in most cases poor and neglected by national education policies, archive data of the DoBeS type are not appropriate for immediate usage in primary education (an overview of factors to be considered in adapting raw language material for educational purposes appears in Skiba 2004). Communities may lack electricity, and no have no access to multimedia techniques or trained professionals (Volkswagenstiftung 2004:13). However, in many cases authentic language and culture documentation materials (video, film, and music) are powerful instruments in identity building. Cooperation with a supporting national academic institution may assist the development of teaching material. Funding for the additional costs of such collaboration must be included in the project budget.

Usually the interests of the speech community in the process of language preservation and revitalisation are not the same as those of the field linguist, but there are overlapping areas of interest that can be mutually useful, especially if the linguist has some professional training in the development of teaching material and is willing to cooperate with the community.

My own experience largely relates to larger language groups supported by nongovernment organizations and/ or government agencies in the process of a complex educational reforms, such as in Bolivia, Ecuador, Peru, Guatemala, and Honduras (v. Gleich 1987). This experience, over more than 20 years as consultant in bilingual intercultural education, teacher training, and production of teaching material for endangered Latin American indigenous languages, has shown me that language archives can provide effective means for the training of young local researchers and teachers, especially with an expansion of technological facilities in schools and adult education venues.

## Future prospects

In the near future, thanks to the possibilities opened up by advancing technology in linguistic documentation, we can develop authentic texts and teaching material using all the elements of a comprehensive multimedia language documentation: songs; music; video of everyday life, cultural and economic events; and local historical knowledge (Chatry-Komarek 2003, v. Gleich 2004).

Linguistics departments could not only follow the example of the Hans Rausing Endangered Languages Project at SOAS and include basic courses on the development of teaching materials and computer assisted learning in their documentation studies programs but also provide basic courses on bilingualism, child language acquisition and second language acquisition.

The reflections and recommendations in this paper are presented mainly in order to show the wide range of possibilities for field linguists to offer assistance in the production of teaching materials, while recognizing their own limitations and avoiding raising overambitious expectations among the speech community.

## References


APNET (Résau des Editeurs Africains) *Editer en Afrique*. Une collection de guides pratiques pour tous les intervenants de l'édition en langues nationales. http://www.apnet.org.

Chatry-Komarek, M. 1996. *Tailor-Made Textbooks*. A practical guide for the authors of textbooks for primary schools in developing countries. CODE Europe/Books Aid International, Oxford.

Chatry-Komarek, M. 2003. *Literacy at Stake*. Teaching and writing in African languages. Gamsberg. Macmillan Publishers, Windhoek. gmpubl@iafrica.com.na.

Cummins, J. 2000. *Language, Power and Pedagogy. Bilingual Children in the Crossfire*. Clevedon, Multilingual Matters.

Fishman, J. A. 1991. *Reversing Language Shift*. Clevedon, Multilingual Matters.

v. Gleich, U. 1987. *Latin American Approaches to Bilingual Bicultural Primary Education-Theory and Practice.* Education Report No. 35, Division 22, Education, Science and Universities Sports, GTZ, Eschborn.

v. Gleich, U. 2001. *Multilingualism and Multilingual Literacies in Latin American Educational Systems.* Arbeiten zur Mehrsprachigkeit Nr. 28. SFB, Universität Hamburg.

v. Gleich, U, 2004. New Quechua literacies in Bolivia. *International Journal of the Sociology of Language* 167 (2004), pp. 131–146.

Grebe, R. & v. Gleich, U. 2001. *Democratizar la palabra. Las lenguas indígenas en los medios de comunicación de Bolivia*. Goethe Institut, Universität Hamburg SFB 538. Universidad Católica Boliviana, La Paz.

Hornberger, N. (Ed.), 1997. *Indigenous Literacies in the Americas. Language Planning from the Bottom up.* Mouton de Gruyter, Berlin.

Himmelmann, N. 1996. Zum Aufbau von Sprachbeschreibungen. In: Linguistische Berichte 164 (1996), pp. 315–333.

Komarek, K. 1998. *Mother Tongue Education in Sub-Saharan Countries*, GTZ, Eschborn.

López, L. E. & Küper, W. 2002. *La educación intercultural en América Latina, balance y perspectiva*s. Informe Educativo No. 94 GTZ, Eschborn (English version available).

Ouane, A. (Ed.). 2003. *Towards a Multilingual Culture of Education*. Unesco Institute of Education, Hamburg.

PROEIBANDES 2002. *Informe final: Programa de capacitación docente en lectura y producción de textos en lenguas originarias: Aimara, Quechua y Guaraní*. Manuscript. http://www.proeibandes.org.

Skiba, R. 2004. Revitalisierung bedrohter Sprachen – Ein Ernstfall für die Sprachdidaktik. In: Hans Werner Hess (Ed.). *Didaktische Reflexionen "Berliner Didaktik" und Deutsch als Fremdsprache heute*. Arbeiten zur Angewandten Linguistik, Band 3, pp. 251–262. Staufenburg Verlag, Tübingen.

UNESCO. 2004. Literacy Assessment Workshop, Paris, June, 17–22 (Adama Ouane, UIE, p.c.).

Volkswagenstiftung. 2004. *Dokumentation bedrohter Sprachen. Eine Reise zu den weltweiten Projekten einer Förderinitiative*. Hannover.

Wittenburg, P. 2003. The DoBeS model of language documentation. In: Austin, P. (Ed.), *Language Documentation and Description*, Vol. 1, pp. 122–139. The Hans Rausing Endangered Languages Project, School of Oriental and African Studies, London. http://www.hrelp.org.


## New in DoBeS: Documentation of the Kola-Sámi Languages


*Jurij Kusmenko, Michael Rießler*
Nordeuropa-Institut, Humboldt-Universität zu Berlin


The aim of the Kola-Sámi Documentation Project (KSDP) is to provide a comprehensive linguistic and ethnographic documentation of the endangered Sámi languages Kildin, Skolt, Akkala, and Ter. The four languages are all spoken in the northwestern-most region of Russia (Murmansk Region – Murmanskaja oblast') on the Kola Peninsula. They belong to the group of eastern Sámi languages. Sámi itself is the westernmost branch of the Uralic language family.

The Kola-Sámi languages are spoken in one historically, geographically, and administratively coherent territory. Each of the original speech communities once formed a historically based cultural and linguistic-sociological unity. Today, however, most of the speakers of these four Sámi languages live together in ethnically and linguistically mixed communities. They are all subject to the same political and economic influences and their languages face serious decline and endangerment (even though on different levels). This makes it reasonable to involve all four Kola-Sámi languages together in the proposed documentation project.

Today not more than 40% of the around 1800 Sámi living in Russia speak and understand a Sámi language fluently. Most of those who do are elder speakers. The younger generation has either a very limited knowledge of the Sámi languages, or does not know them at all. The absence of social motives for language use and the absence of a language environment in which the language is spoken by everyone all the time are

weakening the knowledge of the languages. Taking into account the age of the active speakers, the end of the Sámi speech communities in Russia is probably not far away.

The recording of language data will be undertaken during several field trips to the Kola Peninsula. We plan to work with Kildin speakers in both rural and urban environments. The principal methods to be used are extensive interviews, group discussions and questioning of local experts. Recordings made during periods of participant-observation will complete the documentation.

The processing of the recorded language data will be carried out mainly in Germany. The work in Russia will centre on transcription and translation of the recorded data.

The project will be carried out at the Nordeuropa-Institut at the Humboldt University of Berlin. The team includes Sámi, Russian and German linguists.

It is a primary goal of the project that the broadest possible variety of spoken language data from the Kola-Sámi languages should be systematically recorded, transcribed and translated. The recordings will be provided with rich linguistic and ethnographic annotations. The documentation is expected to reflect active and passive native-speaker competence, situational and social structuring and the geographical distribution of the Kola-Sámi language according to their current patterns of use.

# Archiving

## The PARADISEC Archive

*Nick Thieberger*
University of Melbourne

PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures) is an archive established in 2003 to preserve the field recordings of endangered languages and musics made by (mainly) Australian researchers outside Australia (mostly in the Pacific and Asian regions); and to make these recordings accessible to researchers and originating communities. In our first years of operation we have concentrated on salvaging orphaned academic collections that are themselves endangered through lack of a safe repository when researchers retire or pass away. We intend to extend the collection to include digital textual data such as dictionaries, grammars, text collections and so on.

As of 1 April 2005, we have 828 hours of audio digitised. The archive is held in mass storage systems in Canberra and Sydney, with considerable assistance from the Australian Partnership for Advanced Computing. The current collection represents data in 254 languages from 39 countries, mainly, but not exclusively, in the Pacific region. Important audio collections that have now been digitised to international standards (the audio standard used by PARADISEC is the Broadcast Wave Format or BWF, a WAV file at 96khz/24bit stereo, with an encapsulated metadata header including basic textual information about the audio file) include the following:

- The Dutton collection: 266 hours of recordings covering 63 different languages, and some music, made between 1958 and 1995 by the ANU (Australian National University) linguist Tom Dutton. The material is mostly of Papua New Guinea languages but also includes recordings from the Solomon Islands, Australia and the USA.
- The Wurm collection: 115 hours, still in process, and covering many different languages in the Pacific and Papua New Guinea. This series includes recordings made by speakers themselves, with transcripts and translations, sent to the late ANU linguist Stephen Wurm (1922–2001).
- The Capell collection: 10 hours, still in process. PARADISEC is beginning to digitise the original non-Australian field tapes and related field notes of Oceanic materials recorded by the late Arthur Capell (University of Sydney) from as early as the 1940s.

We also have been encouraging current researchers to deposit their recordings as soon as they return from fieldwork. This allows us to create citable archival versions with persistent identification that all of their subsequent work can be referenced to.

OAI/OLAC conformant metadata for our collection can be searched via the OLAC and LinguistList gateways noted below. Later in 2005 we will be providing a geographic and text-based search facility with password-protected access to the digital files in the collection. For copyright and intellectual property reasons, access is limited to depositors and those authorised by them.

PARADISEC is keen to establish relationships with local cultural centres in its region of interest, and we have already begun cooperation with the Institute of Papua New Guinea Studies (Port Moresby, Papua New Guinea), the Vanuatu Kaljoral Senta (Port Vila, Vanuatu), and local researchers in New Caledonia and Rapa Nui. We have assisted local cultural bodies by providing copies of relevant parts of our collection in an appropriate format, and helping with digital preservation of selected parts of their existing collections. In 2005 we are also trialling a variety of digital field recording technologies.

As well as assisting with general advice on digital

preservation, we run training workshops on digital documentation, and are co-founders of the Resource Network for Linguistic Diversity (RNLD), an alliance which provides information and advice via a mailing list and FAQ pages. PARADISEC cooperates with international digital archives through OLAC and DELAMAN (the Digital Endangered Languages and Musics Archives Network)

### Links

PARADISEC: http://paradisec.org.au

RNLD: http://www.linguistics.unimelb.edu.au/RNLD.html

OLAC: http://www.language-archives.org

DELAMAN: http://delaman.org

OLAC Gateway:
http://www.language-archives.org/tools/search/?archive=paradisec.org.au

Linguist List Gateway:
http://cf.linguistlist.org/cfdocs/new-website/LL-WorkingDirs/olac/olac-search-advanced.cfm

## The Language Archive at the MPI: Contents, Tools, and Technologies

*Peter Wittenburg, Romuald Skiba, Paul Trilsbeek*
MPI, Nijmegen

The language resource archive at the MPI is not a classical archive that is only concerned with long-term data preservation, nor is it a strictly digital library that only gives access to the data. It is a modern digital archive that combines safeguarding of the stored material such that future generations can access it and providing immediate access to those who are currently interested in the material. The basic technological requirements are driven by long-term considerations, i.e. representational aspects. Presentation aspects are seen as something that will be derived from the standards-based representation (Wittenburg et al. 2004).

### History

The MPI Language Archive in its current version has existed since 2000; however, the original archive is older than that. The local-access intranet-based language data samples that have existed since the early nineties were hybrid: partly HTML-based structures and partly UNIX-based file structures. The first corpus accessible via the internet was the "ESF Second Language Acquisition Corpus", which was published in 1996 and was subsequently added to the IMDI metadata corpora. The 2000 version of the archive was entirely IMDI-metadata based (see Wittenburg & Broeder 2002). All data in the archive is described by IMDI-session metadata, i.e. the primary language data (recordings and/or annotations) is described as a bundle of metadata information and also as part of the archive tree structure (IMDI-corpus metadata). The metadata information is freely accessible. One can browse through the metadata tree (Browsable Corpus) and finally reach all the available information on the primary data. Access to the primary data can be restricted by the owner of the data (the researcher, group or institution).

### Contents

Currently the MPI corpus contains more then 20,000 sessions (the entire IMDI-based corpus contains twice as much, Broeder 2004:12). There are data from the Max Planck Institute for Psycholinguistics, from the DoBeS program (Documentation of Endangered Languages, links: DoBeS), ECHO (European Cultural Heritage Online), and other projects. Child, sign, and second language acquisition data, are represented, as well as data for many endangered languages. The contents are described in information files at the entry point of the corpora (links: IMDI-corpora).

### Data formats

*Recording formats*
The sources of the data are recordings on many different media, representing the long history of technological development of recording devices. For video, this ranges from celluloid film to analog and digital video formats such as U-matic, VHS, Hi-8 and DV. Audio material ranges from recordings made on reel-to-reel tapes to recordings made with the latest flash-memory recorders. The MPI stores these recordings in the archive as high-quality digital audio and video files, making conversions when necessary.

*Digital data formats and codecs*
In order to preserve digital data for future use, the file formats and codecs that are used should be explicitly documented and preferably conform to open standards (Wittenburg et al. 2004:3).

For the MPI, there are different rules with respect to the permitted file formats and codecs for the different corpora. For example, for the DoBeS archive, there are more restrictions than for some of the MPI-internal corpora, due to the fact that certain agreements have been made within the DoBeS program to maximize the probability of long-term data preservation. Currently, the archive contains data in the following data formats and codecs:

- Audio: WAV files of linear 16-bit PCM audio with a sampling frequency of 44.1 kHz (CD quality) or 48 kHz (DAT quality). Also lower quality recordings, e.g. made with a minidisk recorder, are converted to this format in order to have a more consistent archive and provide a greater chance of data survival (even though

the conversion doesn't add any extra quality). As well as the WAV files, MP3 files are created as a web presentation format.

- Video: MPG files containing MPEG1- or MPEG2-encoded video are the main formats. Even though MPEG video is compressed and reduces the quality of the original recording, it is currently the only feasible encoding for the MPI archive. It will still take a number of years before digital storage technology will allow us to store uncompressed video. MPEG4 video is created as a web presentation format.
- Images: JPEG files containing JPEG compressed data are still the standard for digital photography at this point and are therefore accepted as an archive format. However, digital photography is moving towards uncompressed formats, which will be the preferred formats in the future.

*Metadata formats*
Metadata descriptions are in accordance with the IMDI standard, a scheme-based XML format (Wittenburg & Broeder 2002). While the IMDI standard contains a large set of fixed metadata description elements, it also allows for user extensions.

"Info" files can be seen as another kind of metadata, created to provide a more detailed description of the content of a specific point in the corpus. The structure of these files is open, but the file formats are restricted to HTML, plain text, and PDF.

*Annotation formats*
Annotations are typically documents that refer to the content of media files. These documents conform to various standards and formats. The primary format of the archive is the EAF (ELAN Annotation Format, Brugman & Russel 2004), which is XML-schema based. The ELAN annotation tool (see below) creates annotations in EAF format. Another format is Shoebox/Toolbox (links: Shoebox:31). Shoebox data can be imported and manipulated using ELAN. The CHAT format (MacWhinney 1995) is also supported by ELAN. Some less-represented annotation data are in PDF and MediaTagger format. The recommended character encoding is Unicode (UTF-8). We are trying to convert all annotation formats to the EAF format, i.e. XML-based with Unicode UTF-8.

*Lexicon formats*
Lexicon data related to EAF annotations are in LMF format (supported by LEXUS, see Wittenburg & Kemps-Snijders 2005; see also section "Accessing Archival Content" below). Shoebox and CHAT-lexicon files are accepted, but conversion to LMF format is recommended. Character encoding for lexica is again Unicode (UTF-8).

## The architecture of the MPI archive

The archive distinguishes four groups of people involved in digital archiving: (1) donators; (2) corpus managers; (3) system managers; and (4) users. Donators and users should not see the physical storage structure; they should operate in virtual domains defined by linguistic terminology and should have the flexibility to create their own personalized virtual archival domain. Systems managers must ensure that the data can reliably meet the criteria for long-term storage: accessibility and protection against unauthorized access. Corpus managers create and maintain a canonical organization of all material, map virtual and physical data organization, ensure that the archive is consistent, and mediate between the different groups involved.

To meet these requirements the archive uses the IMDI metadata framework to organize and maintain all material, and also to offer the user a browsable and searchable virtual archive. The virtual IMDI domain is the core of all donator, user, and management activities. Underlying this is the physical structure maintained by the system managers. This principal distinction allows system managers to migrate and copy data without affecting users. To meet the requirements of long-term storage the archive has several layers of redundancy. In the center is a Hierarchical Storage Management system (HSM) running on two parallel multi-processor SUN servers. These servers are connected to a two-layer RAID system – a faster one for small files and another, slower, for large media files – and a tape library. Via fast European network connections, complete copies of the data are dynamically created at two large computer centres at the Max Planck Society and the MPI for Evolutionary Anthropology. The Max Planck Society has committed to being responsible for a portion of the data (e.g. DoBeS data) for a period of 50 years. An additional group, the developers, ensure that there are methods for uploading resources into the archive, for maintaining the archive, for monitoring its consistency, and for accessing its content at different layers. These layers are described below.

## Archive technologies and tools

*Digitization, capturing, and transcoding tools*
Capturing and transcoding of audio and video is performed using various hardware and software. Information about current procedures is available via the Web (links: A4 guides).

*Metadata tools*
The IMDI infrastructure is the core of the archive's architecture. For the user it is like a browsable and searchable catalogue for discovering resources; for the archive manager it is the basis of maintenance operations. The developers provide a professional editor that allows the creation of valid IMDI files for various

resource types; a native IMDI browser operating in the domain of linked XML-based metadata descriptions; and a server program for browsing the IMDI domain as if they were HTML pages. All IMDI tools are freely accessible via the Web (links: IMDI-tools).

*Annotation tools*
The main annotation tool for creating the standard EAF archiving format is ELAN. ELAN allows the creation of media annotation files in an open format. ELAN can import and export other accepted formats, such as CHAT and Shoebox files. (As stated above, the ultimate goal of the archive is to unify all formats to XML-based formats. For the moment we also accept and archive annotations in their original format.)

*Accessing archival content*
There are various methods of accessing the archival content. The browsers allow the downloading of individual or groups of resources once they are identified by their metadata. Download allows for local operations on resources, such as playing audio or video, viewing annotations, or using tools such as ELAN or LEXUS on the resources found. Currently, a web-based exploration framework is being developed.

*Uploading and management*
Until now the corpus managers had to upload and check the data manually. Metadata descriptions were also checked and integrated, and media resources were digitized and linked with the metadata descriptions. With an increasing number of teams contributing to the archive, this process proved to be no longer efficient and so LAMUS (Language Archive Management and Upload System) was developed. It offers a web-based interface so that donators as well as managers can use a standardized interface to prepare and check the material and then upload it.

*Access management*
The archive is bound by a number of rules restricting access. The archivist has the right to store the data, but copyright is held by the donators and consultants. Therefore, the researcher is seen as the central figure in determining access to the material. The archivist assumes that the researcher has obtained the consent of the language community and/or consultants for making the resource available to others. Potential users are required to accept a Code of Conduct developed for the DoBeS program and to declare their intended use before they are granted access to the resources. To support this, the Access Management System has been developed to deal with access management issues electronically. In addition, it allows the delegation of the right to grant access to others, such as the responsible researchers. It is efficient insofar as it allows the granting of access rights to groups of people with a single command from a particular node in the linked IMDI domain.

## Future aspects

The DoBeS archive, hosted at MPI, intends to intensify its cooperation with other archives within the DELAMAN initiative (links: DELAMAN). It will focus on integrating its collection with others to make it more useful to users and will seek to improve the chances of data survival by distributing copies of its data worldwide.

## References


Broeder, D. 2004. 40,000 IMDI sessions. *Language Archive Newsletter* 4, p. 12.

Brugman, H. & Russel, A. 2004. Annotating multi-media/ multi-modal resources with ELAN. In: F. Ferreira, R. Costa, R. Silva, C. Pereira, F. Carvalho, M. Lopes, M. Catarino, & S. Barros (Eds.). *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2004, Lisbon), X., pp. 2065–2068. European Language Resources Association (ELRA), Paris. CD-ROM version available.

MacWhinney, B. 1995. *The CHILDES Project: Tools for Analyzing Talk*. 2nd Edition. Erlbaum, Hillsdale, NJ.

Wittenburg, P. & Kemps-Snijders, M. 2005. LEXUS – a new lexicon tool. *Language Archive Newsletter* 5, p. 10.

Wittenburg, P. & Broeder, D. 2002. Management of language resources with metadata. In: L. Romary, C. Galinski, N. Ide, & K.-S. Choi (Eds.), *Proceedings of the Third international Conference on Language Resources and Evaluation* (LREC 2002). Workshop on international standards of terminology and language resources management, pp. 49–53. European Language Resources Association (ELRA), Paris.

Wittenburg, P., Skiba, R. & Trilsbeek, P. 2004. Technology and tools for language documentation. *Language Archive Newsletter* 4, p. 12.


## Links

A4 guides: http://www.mpi.nl/corpus/a4guides/

DELAMAN: http://delaman.org

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

IMDI-corpora: http://corpus1.mpi.nl/BC/IMDI-corpora/

IMDI-tools: http://www.mpi.nl/tools/

Shoebox: *The Linguist's Shoebox: Tutorial and User's Guide.* SIL International. http://www.sil.org/computing/ shoebox/ShTUG.pdf

**Suggestions and contributions welcomed at:**
**LAN@mpi.nl**
**Next deadline for copy:**
**August 1, 2005**

# Technical Section

## Sony HiMD Recorders

*Gerd Klaas*
MPI, Nijmegen

Sony has released a new line of MiniDisc recorders that allows the recording of both compressed (ATRAC) MD format and high-quality linear PCM on inexpensive 1GB discs (approximately 8 euros). While ATRAC compression filters out important components of the sound signal and therefore cannot be recommended for archive material, high-quality linear PCM (16bit, 48 kHz) is a direct conversion of the analogue sound signal, making the new HiMD recorders very interesting for field research purposes. However, when testing the device and capturing sound files on a notebook, we found that Digital Rights Management encryption rendered the recorded WAV files unreadable on a notebook, and thus of little use.

Since then, we have found that capturing and decryption is possible, but not particularly convenient. It requires several steps:

1. Connection of the recorder via a USB port to a PC. Any files other than audio (photos or other data) will map directly as files in a folder in the current directory structure.
2. Retrieval of audio files by running Sony's "SonicStage version 2.3" software, which can be downloaded from Sony's web site. WAV files will first be captured to a temporary file (note: care must be taken during capturing, because the original file on the HiMD disc can become corrupted or even be deleted).
3. Conversion of the resulting OPENMG format files using the WAV conversion tool, which can also be downloaded from Sony. This tool allows the decrypted files to be saved as normal WAV files.

Since this procedure may take up to 2½ times the actual recorded time when performed on a normal notebook, we cannot yet recommend use of these HiMD recorders when direct digital transfer is required.

**LAN back issues available at:**
**http://www.mpi.nl/LAN**

## A4 Guides – Technical and Organisational Factsheets

*Paul Trilsbeek*
MPI, Nijmegen

The technical group at the MPI in Nijmegen has started creating guides on technical and organisational aspects of the MPI archives. The aim is to create short one- or two-page items that each fit on one sheet of A4 paper. They will cover a wide range of topics, some explaining step-by-step how to perform tasks using certain software, others explaining more about the structure of the archive, and yet others explaining technical phenomena. Some of them have already been completed:

- DV video processing using Adobe Premier Elements;
- encoding MPEG using TMPGEnc;
- encoding MPEG using Mainconcept encoder;
- cutting MPEG1 video using MyFlix;
- audio compression techniques; and
- DoBeS format and encoding agreements.

Many more are planned for the next few months, including:

- power management in the field;
- storage management in the field;
- tape labeling and time marking;
- archive organisation; and
- creating IMDI metadata.

The guides are available on the Web:

http://www.mpi.nl/corpus/a4guides

# News in Brief

## LEXUS – A New Lexicon Tool

*Peter Wittenburg, Marc Kemps-Snijders*
MPI, Nijmegen

The first version of LEXUS, a flexible web-based lexicon framework, has been completed. The underlying model is based on LMF (Lexical Markup Framework), an ISO TC37/SC4 effort to arrive at a flexible lexicon model. The

idea to create such a tool emerged from DoBeS fieldwork and NLP lexica that use different structures and linguistic concepts or terminology. LMF is a structural framework that allows lexica to be combined and their building blocks to be reused and/or restructured. Since it is strongly linked with domain ontologies such as the ISO Data Category Registry, both GOLD ontology and Shoebox MDF registry interoperability will be offered.

LEXUS can be used via the Web or locally on a notebook, although a simple installation procedure has still to be worked out. Users may create and modify structures, add lexical content manually, and import Shoebox, CHAT and XML files. This version allows the user to carry out searches–even across several lexica. Although it is capable of exporting to Shoebox, limitations of the Shoebox format itself restrict this function.

LEXUS comes with a flexible user interface for defining both onscreen and printout formats. Further, it provides a basic merging function so that it is possible to merge two lexicon versions created by field linguists who have worked independently on the same lexicon.

LEXUS will be used by the DoBeS archive as a tool to access lexica via the Web and to create XML representations in the archive (the XML-based structure representation is one of the main principles for DoBes; see the article on the MPI language archive by Wittenburg et al. in this issue). The MPI team plans to continue its development, in particular to increase the merge functionality, to improve cross-lexica operations using ontologies, and to develop interaction with the ELAN and ANNEX annotation frameworks (ANNEX can be regarded as a web-accessible version of ELAN). LEXUS is available via the Web at http://www.mpi.nl/lexus (or http://lux13:8080/lexiconapp if accessed within MPI Nijmegen). A manual is currently in production. Both LEXUS and ANNEX will be demonstrated at the upcoming DoBeS conference in May (see http://www.mpi.nl/DOBES/workshop/workshop.html).

## ELAN 2.4 Now Available

*Hennie Brugman, Han Sloetjes*
MPI, Nijmegen

We are happy to announce release 2.4 of ELAN. We have added a substantial number of new features and improvements. We feel that the core functionality of ELAN is now ready, although we still have a long list of ideas and requests and we will continue its support, maintenance, and development. All new features and details may be found in the release notes and manual, but here is a list of the primary ones:

- The maximum number of coupled video players has increased from 2 to 4.
- Media formats other than MPEG and wav are supported.
- Linked media files of an ELAN document can be added, removed or changed at any time.
- It is possible to switch between Annotation Mode and Media Synchronization Mode even if annotations already exist. Offsets for audio players can also be specified.
- It is possible to time-shift all annotations in a document.
- Annotation start and end times can be changed by Alt-dragging the ends of the annotation segments in the time line viewer.
- "Open" vocabularies are supported.
- Segmentation mode: it is possible to make a fast initial time segmentation on a given tier.
- Filter tier: annotations on a source tier can be copied to a (symbolically associated) target tier. Filter value strings can be specified. These strings are filtered out of the annotation values. This feature allows filtering out embedded encoding or copying a tier's annotations.
- Localization: ELAN's user interface can be switched to German.
- ELAN is available for Linux. Note that media playback behavior is not optimal. Use ELAN on Linux only when exact media alignment is of lesser importance.
- Import and export:
  - The visible columns of the Grid Viewer can be exported to tab-delimited text from the viewer's popup menu. This is also the case when the Grid Viewer is in multi tier mode. The option is also available in the 'search results' table.
  - Export as Traditional transcript generates a text export of annotation values on selected tiers, ordered in time.
  - Any document opened in ELAN can now be exported to CHAT format.
  - Transcriber import is added.
  - Export to Shoebox/Toolbox format is revised.

ELAN can be found at http://www.mpi.nl/tools.

# Announcements

## PARADISEC Workshop on Linguistic Data Management

*Nick Thieberger*
University of Melbourne

PARADISEC will offer a one-day training workshop on linguistic data management on September 27th 2005 as part of the Australian Linguistic Society Conference, to be held at Monash University's Melbourne City campus.

The workshop will cover the following topics:

- archiving data before analysis rather than some time after;
- creating persistent identification and citability of data;
- creating metadata and managing small-medium scale linguistic data collections;
- transcribing digital media with time alignment of transcripts;
- interlinearising multilingual data with Shoebox;
- managing a lexical database as the basis for dictionary production, using Shoebox; and
- conversion of data using regular expressions.

Web: http://paradisec.org.au/

## DGfS Workshop on Language Archives Bielefeld, February 2006

The upcoming annual conference of the German Linguistics Society (DGfS, Deutsche Gesellschaft für Sprachwissenschaft) "Language Documentation and Description" (Universität Bielefeld, February 22–24 2006) will offer a workshop on "Language Resource Archives – Standards, Creation, and Access". Details can be found at http://www.spectrum.uni-bielefeld.de/DGfS/.

## Call for Help to Sri Lanka Malay Tsunami Victims

As some of you may know, Umberto Ansaldo and Lisa Lim have been doing research on Sri Lanka Malay for the last couple of years. Two of the communities they have been working with very closely are Hambantota and Kirinda, both small localities in the south east of the island which have been almost completely devastated by the tsunami. Hambantota, as you may have seen in news reports, was the second worst hit in the country, with most of the town destroyed and many thousands killed. Kirinda was a small fishing village further east on the south coast; it was almost entirely destroyed though so far the casualty count is relatively low.

These communities have been harder to reach from Colombo, and getting less attention than, for example, the north-east Tamil Tiger regions. As general relief is not reaching these communities in these more remote areas, a special Relief Fund has been set up by one of the main Sri Lanka Malay associations (COSLAM) to send aid directly to them: in the first phase, to help the victims, and in a next phase to contribute to rehabilitation and reconstruction.

Lisa and Umberto are in personal contact with the organisation and believe this the best and most direct way to help these communities; the process is transparent and accountable. At the moment COSLAM have managed to raise just under US$4,000 there. Lisa and Umberto have been collecting money from friends and family and will be either bringing the funds over to them or sending them very soon. If you would like to contribute to this, or if you would like more details, you can either e-mail Umberto at U.Ansaldo@uva.nl or phone Lisa at +65-96621807. Thank you.

LANGUAGE ARCHIVES NEWSLETTER

www: http://www.mpi.nl/LAN/
Mail: LAN@mpi.nl
ISSN: 1573-4315

Editors: David Nathan (ELAR, SOAS, London)
Romuald Skiba (MPI, Nijmegen)
Marcus Uneson (Lund University)
Layout: Marcus Uneson

**Contributions welcomed at:** **LAN@mpi.nl**

**Last submission date for the next issue:** **August 1, 2005**