Proceedings LREC 2004
Lino, Xavier, Ferreira, Costa, Silva (eds.)
Vol IV, pp. 1377-1380

A Multi-Modal Documentation System for Warao

Stefanie Herrmann, Hartmut Keck, and Stephan Kepser

SFB 441, University of Tübingen
Nauklerstrasse 35, 72074 Tübingen, Germany
{herrmann,keck,kepser}@sfs.uni-tuebingen.de

Abstract

This contribution presents a multi-modal documentation system for the Amerindian culture and language of the Venezuelan Warao. The project has two main tasks: the documentation of language and culture of the Warao and the development of a documentation system. Our key aim is to allow the integration of different media and data types (photos, drawings, file cards, and audio data) in such a fashion that all of them would be first class objects. Furthermore the system has to be usable by ethnolinguistic field workers who are not trained computer scientists and has to guarantee integrability into larger documentation projects as well as data exchange. We therefore opted to use Microsoft Access as the underlying database system and XML as the annotation standard for corpora. One of the key features of the documentation system is an open concept space that allows the free definition of linguistic and cultural concepts and their relations and thus goes beyond the expression of hierarchical relations between concepts.

1. Introduction

As a consequence of the fact that literally hundreds of languages of the world will become extinct in the foreseeable future there is a growing insight that the documentation of endangered and minority languages is an important task in preserving the cultural heritage of mankind. Consequently, there exist now several documentation projects. Some of them also develop tools and provide infrastructure for this purpose.

Amongst the most important ones are DOBES[1] (*Dokumentation bedrohter Sprachen*) and EMELD[2] (Electronic Metastructure for Endangered Languages Data), both being larger projects for the documentation of several endangered languages. DOBES is funded by the *Volkswagen Stiftung*[3] and its data archive, which is supposed to cover sound material, video recordings, photos, and various textual annotations, will be housed by the MPI[4] for Psycholinguistics in Nijmegen (NL) (subproject TIDEL). There are currently 14 projects documenting languages all over the world predominantly in South America. While participation in DOBES is limited and subject to application and granting of funds, the EMELD archiving project is open to all researchers. In order to provide access to data on endangered languages in a useful form EMELD's main goal is to promote a consensus about aspects of archive infrastructure. The project is shared among five institutions (Wayne State University, Eastern Michigan University, The University of Arizona, The Linguistic Data Consortium and The Endangered Language Fund).

It is understood today that it is not enough to collect texts and vocabulary lists. A language has to be documented within its social and cultural context. In other words, the documentation should comprise the culture of which the language is but one important part. The documentation of language and culture naturally requires the collection of data on diverse media such as video, audio, (annotated) texts, lexica, research notes, diaries, and more. This holds especially true for ethnolinguistic research and minority languages, which are a stronghold of oral tradition. Therefore, any useful documentation system is inherently multi-modal. On the other hand media other than text are currently still put at a disadvantage regarding the availability of adequate tools and the possibility to integrate them into databases.

2. Aims of the documentation system

2.1 Warao

The multi-modal documentation system presented in this paper was developed for the Amerindian culture and language of the Warao. Spoken in the Orinoco delta of Venezuela by approximately 20000 people Warao is the second largest indigenous language in this South American country. Although not strictly endangered it clearly shares the common problems of minority languages such as limited usage in the educational system and official realms of national and regional life. At the same time Warao shows the richness in spoken genres predictable for an up to now predominantly oral culture. It is therefore of primary importance to document this language and in its genuine cultural context on a broad scale by use of various media such as photographs, audio recordings as well as written materials.

2.2 Tasks

The two main tasks of our project stem directly from these requirements. They consist at the same time of the documentation of language and culture of Warao and the development of a documentation system. The latter should not merely allow a digitalized storage of the data but also assist the researcher in her scientific analysis. Although not being a goal at the beginning of the project in 1999, the development of such a system became necessary as adequate documentation systems for the data and data types we needed to integrate were not available for most of the life time of the project. Nevertheless compatibility of our database modules remains one of our main concerns.

[1] DOBES: http://www.mpi.nl/DOBES/
[2] EMELD: http://saussure.linguistlist.org/cfdocs/emeld/
[3] *Volkswagen Stiftung*: http://volkswagen-stiftung.de/
[4] Max Planck Institute: http://www.mpi.nl/

3. Architecture of the system

3.1 Data sources and types

The data on Warao language and culture used as basis of our project were gathered in two fieldwork sessions from 1998 - 2003, one of which being a one year's stay, in a Warao village situated in the Western part of the Orinoco delta. Data from posterior sessions has been added gradually. The main data sources were

- audio recordings of interviews and narrations, ca. 30 hours
- photo series with 1665 photos
- drawings and figures
- text collections
- a lexicon of about 2000 file cards
- field diaries of about 1625 pages

Thus the key aim in the development of the documentation system is an integration of different media and data types in such a fashion that all of them would be first class objects. That is to say audio recordings and photos should be similarly accessible as textual data; a necessity still not provided for by standard applications. We would like to emphasize at this point that the challenge presented by this diversity of data was not something we volunteered for but a given requirement of ethnolinguistic field work. All of the existing data types had to be accounted for, none could have been discarded.

We start from the observation that there are in principle two different data formats existing across media bounds: corpus type data and database type data. Data that consist of a continuous stream rather than individual units are of corpus type and characteristically present temporal-sequential relations. Whereas data that have a distinguished atomic unit as a uniform element out of which the data are built are of database type. Consequently, the photos, the drawings, and the lexicon are of database type. And the audio recordings, the text collections, and the diary are of corpus type.

3.2 Choice of platform

The main considerations in the implementation of the documentation system were as follows. The system is used and has to be usable by ethnolinguistic field workers and not trained computer scientists. Therefore ease of use is of primary importance. The only experience one can assume these users to have is one of standard Microsoft office products. These products are not only the most widespread software in the world; they are by now already used by field workers. We therefore opted to use Microsoft Access as the underlying database system and did any additional programming in Microsoft Visual Basic and Sun Java. Data exchange with other researchers and integrability into larger documentation projects is also main concern. In particular, we plan to integrate our data collection into the DOBES project. Here again, it is the use of standard formats and standard software that warranties a seamless data exchange and integration into larger projects.

3.3 Storage of data

The photos, drawings, file cards, and part of the audio data are stored in Access databases. In the process of capturing the data in its corresponding Access database we tried to preserve the architecture the original field data was collected in. Thus for example lexicon file cards from the field were kept in a file card database especially designed for them. They were not just integrated in the lexicon we later using Shoebox.[5] As they were originally used by the ethnolinguist to document the process of language learning and acquisition of new words in the field they have a chronological dimension not represented in a standard lexicon. As simultaneous access to both of these lexicon type databases is desirable the system contains a tool (programmed in VB), which imports the Shoebox data into a temporary table so that it can be searched along with the existing Access file card database.

Thus transcriptions of audio material, such as interviews or answers to questionnaires[6] have been processed as interlinear texts in Shoebox, which is the most widely used tool for such annotations.

The largest portion of the corpora type data are the entries of 9 diaries. They are annotated in XML format. The according document grammar follows established standards loosely and is mainly focused on the particular needs and interests of the field worker doing the research. There is hardly a universal tag set for these types of corpora. The annotation of the diary is a skillful interpretation of the material annotated, and hence cannot follow some predefined standard tag set. We nevertheless attended to the preferences of the common XML search tools. Any common browser that works with cascading style sheets can be used for the visualization of the data. Details on the corresponding document type definition (diary.dtd), style sheet (diary.css) and an example diary file (diary.xml) are available at our website.[7]

3.2 Media modules

The documentation system for Warao consists of 6 modules (picture, audio, diary, file card, text, and conceptual space). This is one for each media type plus a so called "concept space" to be explained in the next section. Without going into the details of each module we would like to point out that obviously audio and visual data must be annotated with text layers (containing context information or transcription of speech) in order to make them accessible for operations such as searching.

Pictures can be included and displayed in an Access formula although with minor drawbacks. Departing from the observation that users often order and classify photographs, through theme related albums that normally comprise chronological information, the tool "Picture series" was programmed in VB. It allows the user to assign specific photographs to a series. The series is ordered (e.g., the documentation of a canoe building process) or left in free order (e.g., portraits of people). Contrary to albums in the material word, individual

[5] Shoebox is an application developed by SIL Int. Lately it is substituted for by Toolbox, which is built on Shoebox 5: http://sil.org/computing/shoebox/

[6] A sociolinguistic questionnaire of the MPI, Nijmegen (NL) was carried out in 2000. (Herrmann 2001)

[7] http://tcl.sfs.uni-tuebingen.de/a2/a2pubs.html/

photographs in a database can of course be assigned to and searched as part of various series at once. In this way the tool assists the user in her task of keeping track of and analyzing her visual data and has proven helpful in data exchange with other researchers.
Other than photographs audio files cannot be integrates directly into an Access database but must be referred to by ways of conventional hyperlinks. Tools that allow to display and edit audio data together with its multiple annotation layers are becoming available gradually.[8]
Textual material on the other hand is the media type for which annotation tools and techniques are the most widely available.

3.4 The concept space

One of the main tasks a field researcher is confronted with is to understand the central concepts of a language and culture and the relations amongst them. To aid this work there exist so called authority lists that are large, strictly hierarchically organized ontologies or classifications of keywords. We tried the "Outline of Cultural Materials" (OCM) as an example. OCM was first developed in the 1930s by George Murdock and his colleagues (Murdock et al., 1987). This numerological system for the categorization of cultural data is used in comparative anthropological research (in the "Human Relations Area Files" (HRAF))[9] and integrated in ethnolinguistic field work tools such as LinguaLinks and Shoebox.
Keywords that a researcher uses in order to classify and analyze her data such as "marriage" or "Warao marriage preferences" are abstract concepts. Ethnological research is aimed at the discovery of culture specific or "emic"[10] concepts that need not be found in our own cultures. Correspondingly languages can display phenomena not yet accounted for by traditional grammar concepts. Although authority lists can be adapted to the requirements of a specific culture and language to some degree they confront the researcher with a most fundamental problem. Their binary directional structure is inadequate as well for the description of the culture and language as it is for the way a field worker explores them. We therefore developed the notion of an open concept space which is not limited to the expression of hieratical binary relations but can also account for circular or reciprocal relations. A user can not only define the key concepts she regards adequate but, more importantly, she can define and use arbitrary relations between the concepts. Thus these relations need not be hierarchical or even binary. They can directly express, e.g., the provider, recipient, object and circumstances of a trade good exchange in a single relation by ways of hyper edges. Let us explain this in more detail:
Field data contains instances of abstract concepts as "marriage", e.g., the concrete marriage of two specific people agreed on and held under specific conditions. There can be established a directed binary relation between the keyword "marriage" and an instance such as a narration about this marriage. But as mentioned Warao "marriage" as concept and instance include an array of contextual factors like social situation, kinship system as well as group or personal preferences. This makes things far more complex and can not be represented in simple binary relations.
The mathematical model of the concept space is a (hyper-) graph. In this model, concepts are represented by nodes, and relations are edges between them. Binary edges link two concepts in a graph whereas a hyper edge is able to link a multitude of concepts in one relation.
One of the main components of the documentation system is a graphical user interface of the concept space, developed in Java. Its main goals are visualising and editing the concept space (see Figure 1). It is realised as a kind of tableau operating with two types of graphical objects, data objects on the one hand and relational objects on the other. Data objects, the nodes in the graph, can either be names for abstract concepts (taken from a keyword list) or concrete instances, i.e., records of the audio, picture or text data base. These nodes can be selected in the menu "concept" and added as graphical symbol to the concept space tableau. Their labels are chosen from a drop down list of available database records. Additionally nodes with arbitrary names can be inserted. This allows the definition of concepts that are not yet included in the database. Relational objects serve to establish links of various types between nodes. Relations can be defined in the menu "relation". They have four attributes: A name, a degree (binary or higher), a color (15 colors are available) and a directional type (directed or undirected). Once defined and selected from the menu they are graphically presented as lines of different colors that are drawn between the concepts in the concept space. Depending on whether they express directional relations or not they will end in a pointed arrow. For the graphical representation of hyper edges we use an "auxiliary node". The auxiliary node is not a database object in its own right but a mere graphical aid. Of course the concept space can be zoomed in and out and searched.
For displaying and manipulating the graphs of the concept space we use the library "yWorks", which has been developed at the University of Tübingen.[11] The concept space module is connected to Access via a JDBC interface and can be exported in the same way to every other SQL database.

3.5 Additional features

Besides the lexicon tool and the picture series tool already referred to we programmed two more tools in VB. Access does not enable search within all records of a subformula, one can only search within the main formula and the current record of the corresponding sub formula. In order to remedy this matter we programmed the tool "search title".
We also implemented a feature facilitating the relocation of links to audio-visual data files that are part of the modules "audio" and "picture". Relying on Access this would have to be done manually for each individual link, which of course is unacceptable for the user.

[8] ELAN: http://www.mpi.nl/tools/elan.html
and Transcriber: http://ldc.upenn.edu/mirror/Transcriber/
9 HRAF: http://sil.org/LinguaLinks/Anthropology/UsngThOtlnOfCltrlMtrlsOCM/TheHumanRelationsAreaFilesHRAF.htm
[10] The terms emic and etic were coined by Kenneth Pike (Pike [1954] 1967; Headland 1990).

[11] yWorks: http://yWorks.de/

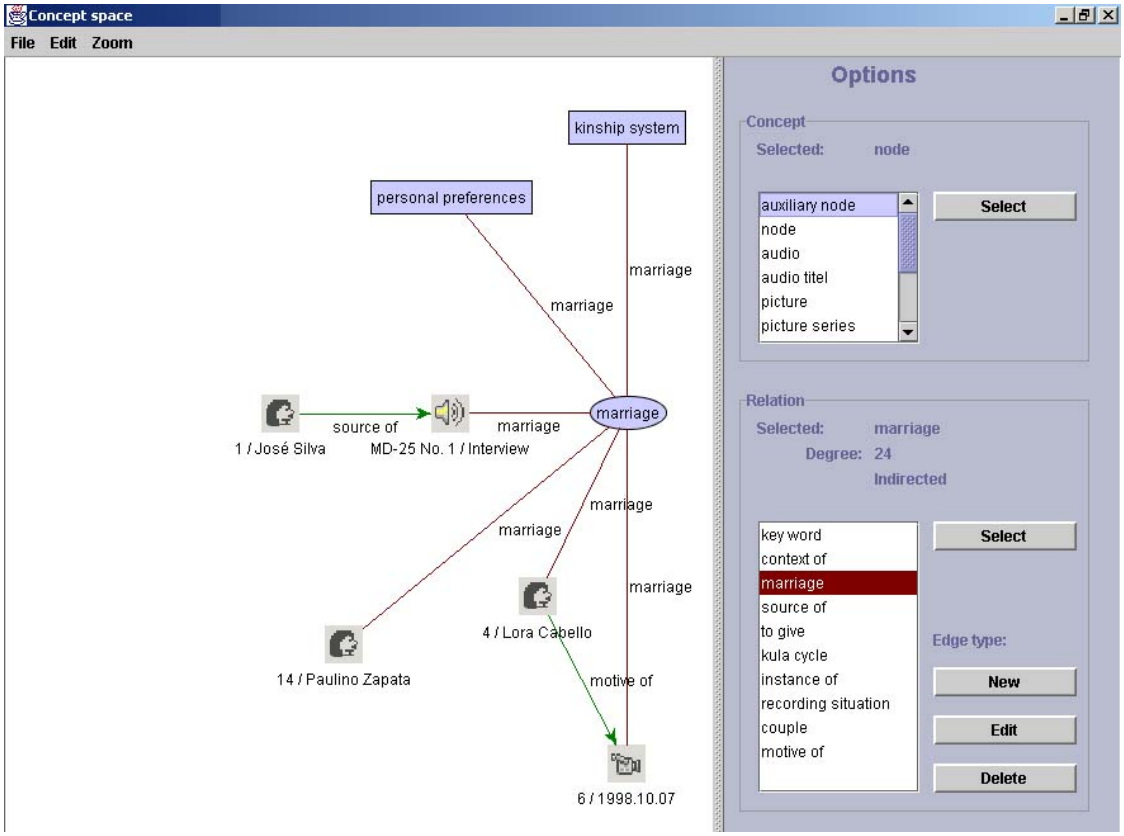

Figure 1: the concept space

4. Conclusion

In this paper, we have presented a multi-modal documentation system for the Amerindian language and culture of the Warao. The choice of Warao is a prototypical one, the system can be used to document other languages and cultures as well. In order to accommodate the various data types and sources used in ethno-linguistic field work, the system contains databases for photos, drawings, file cards, and audio data as well as XML-annotated corpora for diaries and narratives.

The documentation system houses the data collected during two extended fieldwork sessions at a Warao village comprising ca. 30 h audio recordings of interviews and narrations, photo series with 1665 photos, drawings and figures, text collections, a lexicon of about 2000 file cards and field diaries of about 1600 pages.

One of the outstanding features of the documentation system is the concept space. It supports the researcher in exploring the relevant concepts of a culture and their relationships by providing a tableau on which the researcher may define arbitrary concepts and link them and concrete instances of them from the databases using arbitrary relations.

An issue not addressed here are tools for (multi-level) annotation of streaming media. We do not question their necessity. But their development is beyond the limited resources of the current project. There are several such tools under development, for example the one from the MPI as part of the TIDEL project, to which the interested reader is referred.

5. References

Headland, T. N. (1990) Emics and etics: the insider/ outsider debate. Newbury Park, Calif.: Sage Publ.

Herrmann, S. (2001). Warao Demonstratives. (In print. Available at: http://tcl.sfs.uni-tuebingen.de/a2/ warao3.pdf)

Keck, H. (2004) (unpublished). Dokumentationssystem ethno-linguistischer Feldforschungsdaten. Master thesis at the Eberhard Karls Universität Tübingen.

Murdock, G. P., C. S. Ford, A. E. Hudson, R. Kennedy, L. W. Simmons, & J. W. M. Whiting. (1987). Outline of Cultural Materials. New Haven, Connecticut: Human Relations Area Files.

Pike, K (1967) 2nd edition (1954). Language in relation to a unified theory of the structure of human behavior. The Hague: Houton. (First edition in three volumes 1954, 1955, 1960.)