Trumai Corpus: an Example of Presenting Multi-media Data in IMDI-Browser[1]

Raquel Guirardello-Damian,
Max-Planck Institute for Psycholinguistics, Nijmegen, Netherlands
Museu Paraense Emílio Goeldi, Pará, Brazil
raquelg@mpi.nl
Romuald Skiba
Max-Planck Institute for Psycholinguistics, Nijmegen, Netherlands
r.skiba@mpi.nl

Abstract

Trumai, a genetically isolated language spoken in Brazil (Xingu reserve), is an example of an endangered language. Although the Trumai population consists of more than 100 individuals, only 51 people speak the language. The oral traditions are progressively dying. Given the current scenario, the documentation of this language and its cultural aspects is of great importance. In the framework of the DoBeS program (Documentation of Endangered Languages), the project "Documentation of Trumai" has selected and organized a collection of Trumai texts, with a multi-media representation of the corpus. Several kinds of information and data types are being included in the archive of the language: texts with audio and video recordings; written texts from educational materials; drawings; photos; songs; annotations in different formats; lexicon; field notes; results from scientific studies of the language (sound system, sketch grammar, comparative studies with other Xinguan languages), etc. All materials are integrated into the IMDI-Browser, a specialized tool for presenting and searching for linguistic data. This paper explores the processing phases and the results of the Trumai project taking into consideration the issue of how to combine the needs and wishes of field linguistics (content and research aspects) and the needs of archiving (structure and workflow aspects) in a well-organized corpus.

1. The Trumai people and their language

The Trumai people live in the central area of Brazil, in an indigenous reserve named Xingu (which can be subdivided into two big areas, Upper and Lower Xingu). Several tribes, with different languages, live in the reserve. The four major stocks of Brazilian languages - Tupi, Arawak, Cariban, and Ge - are represented in the Xingu area. The Trumai language does not belong to any of these stocks nor to other small Brazilian linguistic families. Genetically speaking, Trumai is considered an isolate language.

The Trumai villages - Terra Preta, Boa Esperança, and Steinen - are located in the medial region of the Xingu reserve. However, in terms of tradition, the Trumai people belong to the Upper Xingu, a fascinating area that has attracted the interest of many researchers. The tribes of the Upper Xingu share a common cultural system, which is observed especially in the activities of surviving, in the diet, in the kinship system, in the mythology, and in the traditional ceremonies. The Trumai people were the last group to arrive to the Upper Xingu area; they came originally from another region, localized at the Southeast of the Upper Xingu. They migrated to the Xingu area in the first half of the XIX century, probably because of the attacks of another Brazilian tribe. The first contacts with the groups of the area were tense and generated some conflicts. However, over time the Trumai people became integrated to the new environment and assimilated several of its cultural patterns. In the 20th century, the group almost disappeared, because of wars and diseases. Epidemics of measles and flu decimated whole families. In 1952, according to sources (Galvão and Simões, 1966), there were only 18 Trumai individuals. The group eventually recovered thought marriages with people from other tribes and high number of births. However, these historical events had consequences for the future state of the Trumai language.

Although the Trumai population consists of more than 100 individuals, only 51 actually speak the language (Guirardello, 1999). The other members of the community speak one of the other indigenous languages of the region and/or Portuguese, the national language of Brazil.

The small number of speakers can be explained by historical facts: as already mentioned, the Trumai people almost disappeared because of the high number of deaths. In order to recompose the group, some individuals married people from other tribes. With these intertribal marriages, other languages of the Xingu region were introduced into the Trumai villages. Another strategy adopted by the Trumai people to survive was to move close to locations where they could get medical assistance. In the Xingu reserve, there are Indian posts maintained by the Brazilian government; these posts offer practical assistance to the groups of the reserve, such as health services provided by doctors or nurses. However, since the doctors and nurses are people from the city, the language spoken in the Indian post is primarily Portuguese. By living close to the Indian posts, the Trumai people became exposed to Portuguese and other aspects of the whitemen society, such as electricity, radios, and books. Some individuals learned to speak Portuguese, which then became a kind of *lingua franca* for couples that could not communicate in an indigenous language. As a result of that,

[1] This paper was written in full collaboration; the order of authorship is arbitrary.

the next generations of Trumai children grew up having at least three languages at home: Trumai, another Xinguan language, and Portuguese. For some time, Trumai was still the preferred language, because it is the language of the group and an important component of its ethnical identity. Eventually Portuguese started becoming more prominent, given its prestige in the area.

The Trumai moved away from the Indian Posts in the 1980's, when they formed the Pato Magro village. However, Portuguese was already infiltrated in the group, and become more and more dominant. To make matters more complicated, the group started creating subdivisions because of internal problems. The main village was divided into three villages; in the 1990's, the group became more disperse, with families moving to other location in the reserve or to Canarana, one of the cities adjacent to the Xingu reserve. Because of this situation, the Trumai speakers, which were already few, started having less communicative interactions among themselves, since the contact of a Trumai speaker with some of his/her relatives became limited to sporadic visits. As a result, the new generation became less exposed to the language.

Nowadays, Trumai children and pre-teenagers are using Portuguese in their daily conversations, especially during their activities with older siblings or peers. These children are able to communicate with their parents because they still can understand Trumai and other Xinguan languages, and their parents are able to understand or speak Portuguese. It is very common to observe communicative interactions in which the parent speaks to the child in Trumai and the child replies in Portuguese. The young generation of Trumai individuals can still understand the language, but they themselves are not speaking it anymore. Without new speakers, Trumai is a language that is progressively dying. The oral tradition is also dying. The introduction of radio - and more recently, television - in the Trumai villages is having a serious impact in the old tradition of telling stories. Nowadays, it has become rather rare for people to get together to chat and tell stories at night. Only a few people still have good knowledge of the Trumai mythology.

As an attempt to revitalize the use of the language, education in Trumai started being developed in the 1990's. The schools in the villages are relatively new and are still in the process of organization, but they already offer classes in Trumai, taught by native teachers (before, there were no schools in the villages, only in the Indian posts, and only Portuguese was taught there). The bilingual education program is already showing good results, such as provoking in the students the desire to know more about their own culture. It is a nice initiative, but documentation of the language is also crucial. The Trumai speakers are in favor of documentation work, because they are aware of the importance of preserving their linguistic and cultural knowledge.

The organization of a multimedia archive is an essential step in the task of documenting the Trumai language. The current moment represents the right occasion for organizing such as archive, since the language is not lost yet - i.e., there are still people who can speak it well, therefore documentation is still possible - and since the Trumai community is already aware of the endangerment of the language. The speakers are willing to help preserve theirs linguistic traditions, especially the indigenous teachers of the Trumai villages, who would like to have an archive with information and data that they could use during the educational activities.

2. The Trumai Project

In order to conduct an extensive documentation of Trumai - preserving the knowledge that is under the risk of disappearing - a project is currently in development. This project participates of the program *Documentation of Endangered Languages* (Dokumentation Bedrohter Sprachen, DoBeS), sponsored by the Volkswagen Foundation.

The goal of the DoBeS program is the documentation of endangered languages and the cultural and social values associated with them. The type of documentation envisaged by the program is different from the conventional style of language description. It does not intend to produce a grammar and a dictionary of the language being documented, but rather to build a multimedia archive with texts and other kinds of data that can be used in the future for various purposes (i.e., the archive should have multifunctionality). Every project participating in the program should process and archive linguistic and cultural data. The work should alternate between field work (with the purpose of multimedia data collection) and the processing of the data with special computer programs. The data should be presented in an inter-theoretical format, that is, in a format that could be accessible for any linguist, independent of his/her theoretical orientation. It should also contain information that can make the materials useful for researchers of other disciplines. Besides the documentation of individual languages, the program also focus on the development of new methods of processing and archiving linguistic and cultural data.

The DoBeS program had a initial (pilot) phase - 2000/2001 - with the participation of some linguistic projects (documenting languages of various parts of the world) and one multimedia archive project (TIDEL). During this phase the technical, logistical, linguistic, and juridical framework of the program was discussed and defined. The processed data for each individual language will be stored in a common electronic archive, which will be based at the Max Planck Institute for Psycholinguistics in Nijmegen (the Netherlands). There are linguistic, technical, and other standards that should be followed by the projects participating in the program (for more information, cf. the website www.mpi.nl/DOBES). The DoBes program now is in its main phase, with more linguistic teams joining it.

The project on the Trumai language started in December 2000, during the initial phase of the DoBeS program, and it will be developed for three more years. It has two major dimensions: (a) the documentation of the language, linked to cultural aspects; (b) research in historical anthropology, documenting elements of the history of the group.

In accordance to the guidelines of the DoBeS program, the target of the Trumai project is to organize a multimedia archive of the language, which will contain not only linguistic data (in the form of texts, word lists, field notes, etc), but also information about aspects of the cultural-social-historical context in which the language is embedded. The intention is to produce an archive that can have various possibilities of application, being a source of information for different kinds of people: (i) linguists, who could use the materials for typological studies, comparative work on Amerindian languages, studies on areal features of the languages of the Xingu reserve, analyses of the structure of the discourse in Trumai, and so on; (ii) anthropologists interested in the Upper Xingu tribes or Brazilian groups in general; (iii) historians trying to reconstruct the history of South-American tribes; (iv) other researchers (sociolinguists, researchers in the area of cognitive studies, etc); (v) general public, who could learn more about the peoples of South-America and their language and culture; (vi) the Trumai people themselves.

The documentation work has the support of the Trumai community, which has active participation in the organization of the archive, helping in the tasks of selecting topics and recording texts in the language. They also help the linguists in the transcription, analysis, and glossing of texts, and in the elaboration of comments about cultural facts. Not all materials will be available for consultation, because some have sensitive contents. The Trumai people will discuss the access to the materials, deciding which ones can be open to the general public, and which ones will have restricted access, being available only to researchers or only to the native speakers.

In the next sections we will see how the activities of the Trumai project are being conducted in the DoBeS program. Two aspects will be explored: the needs of the field linguists (in our case, the researchers of the Trumai project) and the needs of the archiver (i.e., the Tidel team and the DoBeS electronic archive that is being built).

3. The Integration of the Trumai Data into the Browsable Corpus

The main aims of data archiving from the linguistic point of view and from the point of view of an archiver seem to focus on different issues:

Main linguistic aims:

1. Preparing and organizing the data for further analysis;
2. Presenting the data to others;
3. Archiving the data for later use.

Main archive related aims:

1. Conserving the data, e.g. by digitizing and/or copying; data security issues;
2. Structuring the digitized data, e.g. by segmenting and defining hierarchies;
3. Storing technical and archiving information related to the data, e.g. data - base management;
4. Restoring data for the user (i.e. making the data available).

To deal with these issues, we have the Browsable Corpus. The Browsable Corpus is a concept for data management and archiving that satisfies both the needs of the linguist and the needs of the archiver. The central data unit of this concept is called a *session* (cf. Broeder et all. 2001). Understood as a linguistic unit, a *session* is a homogenous and comparable unit of analysis that has the same content, the same time and location, and the same set of participants (e.g. investigator and native speaker; interviewer and interviewee). Understood as a data unit, a *session* is a bundle of primary language data, additional information about those data (meta data), data analysis units (transcripts and other annotations) and technical information about links between the data (location and structure of data units). In the Browsable corpus all the information is stored in an xml file that follows the IMDI standard for metadata (ISLE Metadata Initiative, cf. Broeder et all. 2001).

The task of further processing the raw data can be divided into phases that are temporally or conceptually linked to each other. In the DoBeS program the routines for data processing are:

audio and video media processing;
meta data files (formats, conversions);
handling of other files (annotations, other media);
final organization of the archive (data structures, corpus info files).

Figure 1 shows the data processing workflow.

3.1. Audio and video media processing

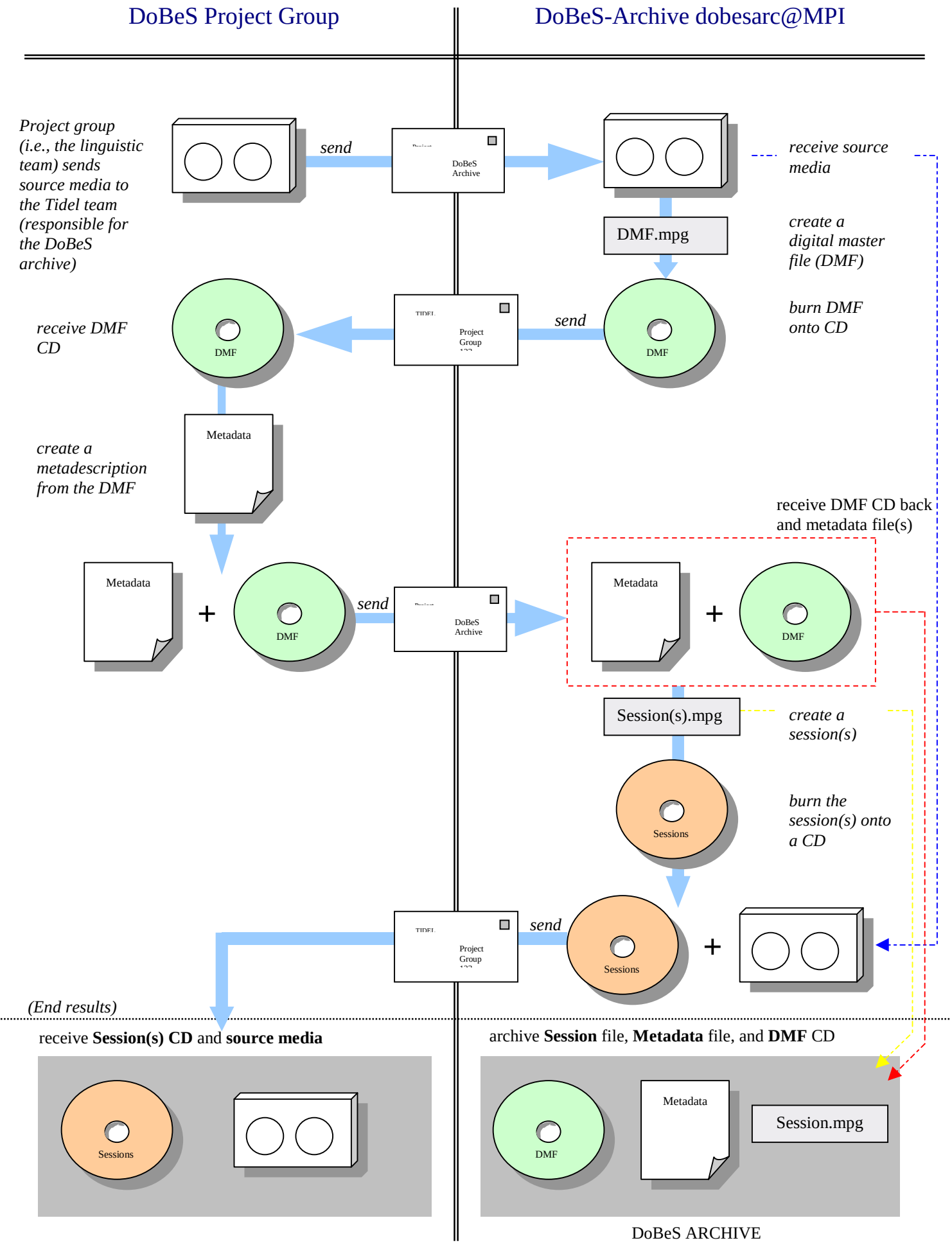
A master tape of good quality is the prerequisite for all further steps of data processing. The labeling of the recording media must be unambiguous. There are conventions for tape labeling. The aim is to ensure that all new segmentations of data can be traced back to the original master tape.

The first step of data processing consists of digitizing the tapes, i.e. creating a Digital Master File (DMF). The purpose of digitizing the data is: (i) to store and conserve the data in another physical form; (ii) to allow further data processing, i.e. segmentation.

In the case of the Trumai data, the original tapes were sent to the Tidel team, which digitized them. The resulting DMFs were burned onto CD-ROMs and sent to the Trumai project (cf. figure 1). The labeling conventions for the DMFs followed the general tape labeling conventions. The basis for the current Trumai corpus were 84 audio tapes (65 cassete tapes, 19 minidiscs) and 8 video tapes.

Master tapes can be split up into conceptual units (sessions) and a bigger conceptual unit can consist of several tapes. At the moment there are 142 media sessions in the Trumai archive. Segmenting raw data into sessions requires the definition of homogenous and comparable units of analysis that can be handled easily. These segments

Figure 1. Diagram of DoBeS Archive workflow (audio, video)



correspond later to the transcripts and other annotations. There are several possible ways of defining sessions: elicitation task (e.g. a questionnaire), text type (e.g. dialog), similar recording situation (e.g. a ritual).

Defining sessions contributes to the management of the whole corpus: after the preparation of several session media files, comparable and homogenous sessions can be bundled into groups. This is important for organizing the materials. When the linguists start structuring the multimedia archive (which, in the case of the Trumai project, is organized in the form of a tree structure), they group the linguistic/cultural materials into types and subtypes, so that the future users of the archive can navigate through them. Thus, the definition of sessions is a relevant step for the organization of the materials. From the technical point of view, the most important information about a session is the data source (DMF label) and the relevant time segments (beginning and end of a session file). From the linguistic point of view, the important information is the content of the session (genre; topic; participants; degree of interaction among the participants).

In the processing of the Trumai data, the information about beginning and end times of a session - together with other information (content, data collector, date and place of recording, etc) - was submitted to the archive. The data processing team then cut the segments (media data of a sessions) from the DMF and sent the new media files back to the project (cf. above figure 1). The time and source information is part of the metadata that belongs to each session.

3.2. Definition of sessions and metadata files

During the pilot phase of the DoBeS program the metadata set and the metadata formats changed. The DoBeS linguistic teams were partly involved in the process of seting up the IMDI-structure of metadata. The IMDI metadata set allows to store detailed and structured information about (among other things) the topic, genre and communicative task, and the age, social status or dialect of the speaker of the respective sessions. A very important function of the metadata is the integration and linking of all parts of a session: the original and the digital media unit, and the annotations and media files.

During the pilot phase there were two possibilities to enter and store metadata:

> Metadata description template: a MS Word template that helps to enter data in a structured form without any checkup procedure for the data; the result is a MS Word text file.
>
> IMDI-editor for metadata: a Java program that helps to enter data in a checked format (e.g. checked formats for date or age of participants); the result is an xml file.

Since in the pilot phase the IMDI-editor was still in development, the Trumai team was using the word template for entering metadata. They were later converted into the xml format by the Tidel team. About 140 sessions were defined for the Trumai project. The sessions are processed and now they are in the phase of being integrated into the data tree.

3.3. The Kinds of Data in the Trumai Archive

A variety of linguistic and non-linguistic materials have been selected to be included in the Trumai multimedia archive:

(i) Texts with audio and video recordings: in order to present a broad view of the language, texts of several genres were chosen, such as myths, historical narratives, explanations about cultural aspects of the Trumai society, procedural texts, conversations, descriptions of persons and objects, interviews, etc.
(ii) Annotations of the texts in different formats: the recordings in Trumai are presented with annotations, that is, transcription of the text, translation, comments, etc. The annotations are prepared with the programs Transcriber (Barras 1998), Shoebox (Buseman & Buseman, 1998), and ELAN (Brugman & Wittenburg, 2001).
(iii) Written texts from educational materials: these texts were produced in written format for educational purposes. Some of them are quite short, but they are included in the archive because they are real instances of the use of the language, presenting interesting data that could be used for studying the transition from orality to written tradition. The texts belong to several genres, and they will be presented into two formats: (a) as a PFD file, showing the text in the way it appears in the educational material. This is the version that is interesting for the native speakers themselves; (b) as a Shoebox or ELAN file, with annotations and glossing. This is the version that is interesting for linguists and researchers.
(iv) Data from elicitation: consisting of word lists, sentences in isolation, field notes, and data obtained with visual stimuli specially prepared for investigating specific constructions of the language. The word lists are organized by semantic field; some of them will be used in comparative studies between Trumai and other Brazilian languages.
(v) Drawings and photos: some texts are illustrated with photos or drawings. The archive also has a special node for presenting series of photos with captions, explaining aspects of the Trumai society or culture.
(vi) Traditional songs and instrumental music of the Trumai people.
(vii) The lexicon of the language: the lexicon is currently being organized in Shoebox. For each lexical entry, there is information about pronunciation, citation form, word class, meaning, semantic domains to which the word belong, scientific classification (in the case of terms for animals and plants), and examples of usage.
(viii) Results from scientific studies of the language, such as sound system, sketch grammar, studies about the history of the group, description of socio-cultural aspects, geographic information and maps, etc.
(ix) Comparative studies with other Xinguan languages: in the DoBeS program there are two other projects that are documenting Brazilian languages: Aweti and Kuikuro. The

project on Trumai is working in close collaboration with them, given that the three languages under documentation belong to the same geographical area, the Upper Xingu. The Trumai, Kuikuro, and Aweti languages are genetically unrelated, but their speech communities belong to the cultural system found in that region. If comparative studies are conducted, we may obtain a better understanding of their societies.

Since the project is still in development, new materials may be incorporated, depending on the needs of the documentation work and the wishes of the Trumai community.

With the selection of materials, some organizational issues arise. First, there is the question of how to present them, that is, how the corpus should be structured (as already mentioned, the materials need some sort of internal organization, being grouped into types and subtypes). Then there is the question of how to integrate different types of information that belong to a same session. For example, a recorded text about a particular topic, such as "body paintings". Besides the audio or video file, there may be other kinds of data related to this session: the annotations of the text; drawings illustrating the body painting; photos of a person with the body painting, etc. The question is how to integrate the data, allowing the users of the archive to manage them easily.

The IMDI metadata is the key for that. As mentioned in section 3.2, one of the functions of the metadata is exactly the integration and linking of all parts of a session. The IMDI metadata allow to specify the kinds of data that are related to a particular session. The data can be defined as a *media*, *info*, or *annotation* file. The integration into the Browsable Corpus will be better explored in the following section (3.4). The structuring and accessing issues of the Trumai archive will be presented in session 3.5.

3.4. Processing of Other Data

As said above, the IMDI metadata set allows to define the various files that belong to a specific session. Two main types of files can be chosen for integrating the resource data: (i) media files; (ii) annotation files.

Usually a media file is a digitized audio or video file. However, it is also possible to define a PDF or a HTML file as a media resource file. For integrating photos, drawings, or written texts (or a combination of those), it is necessary to choose a file format that adequately presents the information. The Trumai project has been using both formats. The HTML and PDF files can be delivered ready to the DoBeS archive, or the archive can create them using the digitized photos and drawings. There are conventions for integrating pictures and texts into HTML files.

Annotation files have in most cases the pure text format (e.g. Shoebox files) or an xml format (e.g. ELAN or Transcriber files). However, there is another possibility: a PDF format file can also be used for annotating data. So far, this form of annotation has not been used in the Trumai archive.

In the Trumai project there is one basic type of annotation that is used for all texts. It consists of the Trumai text presented in orthographic representation, free translation into English and Portuguese (to allow both the Brazilian and the international public to have access to the information presented in the text), and notes about the content of the text or about linguistic facts observed in it. This kind of basic annotation is targeted to any kind of public, that is, linguists and non-linguists.

Some texts will have an extended annotation, presenting tiers with information that are of the interest of linguists, such as morpheme-by-morpheme glossing, phonological transcription, syntactic structure, etc. Below, we have the tiers for the extended annotation:

\trs	Text in orthographic representation (linguist's representation)
\nr	Number of the sentence in the text
\edu	Representation used in educational materials (i.e. native speaker's writing)
\tp	Translation in Portuguese
\te	Translation in English
\ntC	Notes about the content of the text (i.e. anthropological or cultural notes)
\div	**Division between General and Linguistic Information**
\m	Morpheme Breaks
\gp	Gloss in Portuguese
\ge	Gloss in English
\ps	Part of Speech
\pht	Phonetic Transcription
\syn	Syntactic Structures
\rel	Relational Information
\ntL	Notes about linguistic facts observed in the text

In the first part of the annotation schema, we have information that is interesting for a more general kind of public. The second part contains information that concerns linguists. The tier \trs presents the text in the representation used by the linguist; the tier \edu presents the same text, now in the way the Trumai speakers prefer to write in their educational activities. Although very similar, there are some differences between the representations in these two tiers. For example, the text in the tier \trs tries to be faithful to what the speaker is saying, showing hesitations or interruptions, while the text in the tier \edu is "filtered". There are also differences with regard to the representation of compounds and the placement of some clitics.

The annotation schema presented above is being prepared with various tools. Since the main annotating tool for the DoBeS archive (ELAN) was still in development during the pilot phase, the archive accepted other annotating formats (Transcriber, Shoebox) and offered converters for changing these annotations into ELAN (EAF) files. One of the tools was the ECONV program, which converts annotation data between Transcriber, Shoebox, and ELAN formats. The Trumai Project presents data in all mentioned formats. The ELAN tool allows the link of audio and video media files to the time codes of annotated media.

3.5. The Final Organization of the Archive: Integration into the IMDI Browser

The IMDI (ISLE Metadata Initiative) Browser is the main tool for accessing the data in the archive. It helps the user to browse and search in the DoBeS database, and to access the session data.

The IMDI Browser was developed at the Max Planck Institute for Psycholinguistics in Nijmegen (the Netherlands), with the purpose of facilitating the access to media, annotation, and metadata files (Broeder et all. 2001 and Brugman & Wittenburg 2001). The Browser displays a hierarchy of nodes that groups together sessions of a similar kind. It allows the linking and display of the metadata, the media file, and the annotations. It also allows direct access to them; that is, with the Browser one can: (i) read information about the kind of data that is contained in the corpora/multimedia archive; (ii) access the annotation and media files; (iii) initiate searches.

The user interface of the IMDI Browser contains several panels:

1. Browser Action Panel: one of its features is to allow searches on the metadata information;

2. Bookmark's Panel: in this panel, one can save shortcuts to specific parts of the corpus. The shorcuts allow quick access to the parts of the corpus or the sessions that one frequently works with. In this way, it is not necessary to navigate through the entire corpus hierarchy in order to access a particular node;

3. Info/Content Panel: through this panel, one can read information about the data that is contained in the corpus (descriptions) or read the content of metadata, annotation or info files;

4. Descriptions Panel: here, one can read a brief description of any item;

5. Meta Descriptions Tree Panel: allows one to see and access the data tree defined by each linguistic project.

Next, we have the tree structure proposed by the Trumai project, with its nodes and subnodes.

The materials selected for the multimedia archive (described in section 3.3) are organized into two main divisons:

TRUMAI DATA
TRUMAI STUDIES

The node Trumai Studies is further subdivided in aspects of the Language (genetic affiliations, sound system, sketch grammar, etc) and aspects of the People (ethnographic studies, sociocultural description, historical studies, etc). The main purpose of the node Trumai studies is to bring various kinds of information together, so that the user of the archive can have a better understanding about the Trumai group and its language. Materials from several researchers will be included here, with acknowledgements.

The node Trumai Data is the area of the archive that contains linguistic and cultural data. This node has internal divisions and subdivisions:

TRUMAI DATA:

NON-LINGUISTIC:	SONGS
	MUSIC
	IMAGES
	DRAWINGS
	BIBLIO-ICONOGRAPHIC
LINGUISTIC:	LEXICON
	NATURAL USE
	ELICITATION

The Non-linguistic node contains visual or audio materials that do not have linguistic data in Trumai (e.g. music with instruments). The node Images is for series of photos about cultural facts or videos with images only (e.g. a video showing visual aspects of a Trumai village). The node Drawings is for illustrations made by Trumai individuals, with captions explaining their contents. In the node Biblio-Iconographic, we intend to incorporate illustrations of cultural artifacts extracted from bibliographic sources, such as books or museum catalogues (with their permission).

As we can see, the node for Songs is also placed in the non-linguistic area. There is a reson for that. The Trumai people have various kinds of traditional songs. Some of them have lyrics that are in the Trumai language, but there are many other songs whose lyrics are not in Trumai. For example, the songs of the *Yamurikuma* ceremony, which were not created by the Trumai people (they learned them from other Upper Xingu tribes and incorporated them in their own tradition). Given this scenario, it seems to be more appropriate not to include songs among the linguistic materials of the language (i.e., the texts, which are in Trumai), but rather to have a separate category for them. In this way, we can have the whole collection of traditional songs of the group presented together.

With regard to the Linguistic data: in the archive there is a separation between data from Elicitation - where the speaker, following a request made by the researcher, produces specific pieces of information about the language - and Texts (Natural Use), where the speaker talks as in a natural situation of communication (the researcher is present during the speech event, but does not interfere verbally with it).

The texts are subdivided into two types: texts in which there is basically one main speaker (the listener does not interact at all, or talks very little), and texts where there are two or more speakers interacting and alternating turns.

NATURAL USE:
"MONOLOGICAL" (i.e. sessions that are non/semi-interactive)
"DIALOGICAL"(i.e. interactive sessions)

It is important to note here that during the narration of myths, it may happen sometimes that a person of the audience poses a question to the narrator, and the monologue turns into a dialogue for a while, with

interaction and exchange of turns. Thus, the division among the various kinds of texts is not so strict, but rather used for practical reasons (it is for the question of how to organize the annotation. The annotation of a session with one main speaker is not so complicated; a session with two or more speakers interacting requires a little different organization). It is also the case that some Trumai myths may contain "non-linguistic" materials, such as songs (for example, when a character of the myth starts singing).

The text are then further subdivided according to genre:

"MONOLOGICAL":
- EXPLANATION
- NARRATIVE: Mythical, Historical, Personal
- PROCEDURAL
- etc.

"DIALOGICAL":
- CONVERSATION
- INTERVIEW
- CONSULTATION
- etc.

Finally, the Trumai multimedia archive contain *info* files in certain nodes. These files - which are in the HTML or PDF format - provide information with the purpose of helping the user of the archive to understand its structure and to be able to search for materials. As examples of info files, we can mention:

Text types: with information about the various types of texts and how they were subdivided;
Data types: providing information about kinds of data from elicitation;
Annotation: explaining how the texts are annotated (tiers, types of grammatical information, etc);
Ortho: presenting the orthographic conventions used in the texts;
Tags: presenting the linguistic tags and other conventions used in the annotations of texts.

All nodes of the tree will have a short description to help the user to browse through the data structure. The DoBeS archive has the possibility to separate access to different kind of data: acces to the metadata should be free; access to the media, annotation, and info files can be protected.

As a final remark, it should be mentioned here that both the linguistic teams and archiving team of the DoBeS program are aware that data security and ethical issues are closely connected. The legal and ethical rights of the endangered-language community will always be given priority.

Acknowledgments

We would like to present our acknowledgments to the Volkswagen Foundation (Germany) for funding to the Trumai project; to the Fundação Nacional do Índio (Brazil), for providing assistance to the researchers during their visits and stays at the Trumai villages in the Xingu reserve; to the Max Planck Institute for Psycholinguistics (Holland) and Museu P. Emílio Goeldi (Brazil), for administrative and other kinds of support; and to the Trumai speakers, for their support of the documentation work.

References

Barras, Claude et. al (1998). Transcriber: a tool for segmenting, labeling and transcribing speech. France: DGA/Ministery of Defense, Linguistic Data Consortium (LDC).

Broeder, D., Offenga, F., Willems D., & Wittenburg, P. (2001). The IMDI Metadata set, its Tools and accessible Linguistic Databases". In Proceedings of the IRCS Workshop on Linguistic Databases (pp. 48--55). Philadelphia: University of Pennsylvania. [http://www.ldc.upenn.edu/annotation/database/papers/Broeder_etal/32.3.broeder.pdf]

Brugman, H. & Wittenburg, P. (2001). The Application of annotation models for the construction of databases and tools – an overview and analysis of the MPI work since 1994. In Proceedings of the IRCS Workshop on Linguistic Databases (pp. 65--73). Philadelphia: University of Pennsylvania. [http://www.ldc.upenn.edu/annotation/database/papers/Brugman_Wittenburg/20.2.brugman.pdf]

Brugman, H. & Wittenburg, P. (2001). Eudico Linguistic Annotator (ELAN), version 1.1. Nijmegen, The Netherlands: Max Planck Institut for Psycholinguistics.

Buseman, A. & K. Buseman (1998). The Linguist's Shoebox, version 4.0. Waxhaw, North Carolina: Summer Institute of Linguistics.

Galvão, E. & M. Simões (1966). Mudança e sobrevivência no Alto Xingu, Brasil Central. Revista de Antropologia, 14, 37--52.

Guirardello, R. (1999). A Reference Grammar of Trumai. Ph.D. Thesis. Linguistics Department, Rice University. Houston, Texas, U.S.A.

Monod Becquelin, A. (1975). La practique linguistique des Indiens Trumai. Paris: Selaf.