Digitizing and Annotating Texts and Field Recordings in the Awetí Project

Sebastian Drude

Freie Universität Berlin and Museu Paraense Emílio Goeldi

(Revised preliminary version, English has not been checked!)

1. Introduction

In what follows I will present the technical details of the workflow in the documentation of coherent speech ("texts") as it has been worked out in the Awetí Language Documentation Project (ALDP, or Awetí Project, for short).

The Awetí Project is one of several projects in the DOBES consortium (for the WEB-address see the references, below), sponsored by the Volkswagenstiftung, Germany. In many respects, the workflow in the ALDP may be taken as typical for DOBES projects. In some respects, however, the Awetí Project follows its own proposals, some of which are accepted by certain other projects, especially the other projects on Xingúan Languages (Trumai and Kuikuro; see the links at the DOBES homepage).
Fieldwork in the ALDP is carried out by the principal researcher, Sebastian Drude, who initiated research among the Awetí in 1998, together with his colleague Sabine Reiter (since 2001, first field stay in 2002). General matters, including methodological and theoretical matters, are under the responsibility of the applicant of the Project, H.-H. Lieb, and the principal researcher.

As the aim of this workshop is to arrive at recommendations of best practice, I shell focus especially on these points and dedicate some space to justify them. The main point here is the application of Advanced Glossing (AG), a general format for text documentation developed by members of the Awetí Team (Lieb+Drude 2000).
I presented the general ideas of AG and my implementation of AG with the Shoebox Program (see references) elsewhere (Drude 2002) and will not repeat most of the details here but focus on the single steps in the workflow, on one hand, and a more detailed argumentation that justifies this format, on the other.

2. Overview

Generally, the creation of a complete documentation of a text (a maximal "session": object language data with all annotations and metadata) in the Awetí Project includes the following steps, not all of which need to be done in this chronological order:

1. taping of speech events;
2. digitalization of tapings into computer-files (preliminarily in the field, later by the TIDEL - Team at MPI Nijmegen);
3. orthographical transcription of the text in the native language;
4. addition of a translation of the text into Portuguese, later English;
5. addition of unit-related morphological and syntactic annotation, look-up in a lexicon;
6. addition of complex structural (morphological and syntactic) and phonetic/phonological annotation;
7. creation of metadata;
8. conversion of annotation into the final format;
9. inclusion of the session into the corpus structure.

There are also questions of general character whose solution is presupposed by the single steps given above. These include:

10. the conventions for the organization and naming of the several files connected with a session.

All the steps but step (7) will be described in the following sections. As said above, focus will be on steps (5) and (6).

3. Raw data: taping and digitalization [steps (1) and (2)]

Since 2002, all audio and video tapings in the Awetí Project are done with digital devices (a SONY recording minidisc walkman for audio, and a SONY mini-dv camera for video, with, mainly, a Zennheiser MKE 300 directional stereo microphone). The minidisc-internal ATRAC format includes compression, which in principal means loss of quality. However, there are good reasons to believe that this does not affect the most common methods of linguistic analysis (cf. Wittenburg, 2001) .

Much linguistically relevant data (i.e., data to be further processed according to this presentation) is based on audio tapings only, especially "natural" conversations among native speakers and longer narratives. Some planned special sessions are taped with video and audio devices simultaneously, which probably will create problems of synchronization of the different signals later on. On the other hand, documentation of cultural events, usually not to be processed any further, is done mainly by video tapings only (these, of course, include an audio signal which, however, may be of inferior quality).

Audio and video tapings will be transformed into computer ("raw") data files. This process is often called "digitalization" – misleadingly so, as now the taping is done in digital format already and ideally no (intermediate) conversion into analogue signals is involved.
The creation of the final raw data files is done by the TIDEL group in the technical department of the Max Planck Institute in Nijmegen (see the link at the DOBES homepage, in the references). The digitalization process is done according to indications (especially with respect to start and end points of the single sessions) included in the metadata that are created for each session.

For the purpose of transcription and further processing right in the field, we produce a preliminar audio file from the relevant parts of the minidisk tapings. In the field, we use programs such as Sound Forge or Audacity (for both see references, below) to

create the file, usually without any further editing beyond signal strength adjustment. As the minidisk walkman permits analogue output only, the file is to be expected to be of inferior quality than one that was created by digital transformations only.
Hard-drive capacities are restricted in the field, so one could think of using formats that involve compression, such as the MP3 format, for the preliminar audio file. However, MP3 presents some acoustic distortion at the beginning of each playing, which is especially annoying when listening to small segments while transcribing the text. Therefore, we use the uncompressed Wave (Microsoft) format (.WAV), digitizing only the most needed parts of the audio tapings.

The annotations (including transcription and translation) are based on the segmentation of the provisional raw data file. The exact start and end points of this file will in many cases not mach those of the final file generated at the MPI, so a mismatch of segment borders is to be expected. I prepared two simple emacs-lisp scripts that allow to do the fine tuning of the segment borders in Transcriber and Shoebox files. The lisp code is distributed with this document (download here).

4. Field Work with native speakers: basic annotation (transcription and translation) [steps (3) and (4)]

The fluency of the researcher in Awetí, the native language studied, is still in intermediate resp. incipient stages. Therefore transcriptions and translations of the texts (based on raw linguistic data as described above) has to be done with the help of native speakers. (We suggest, however, that in any case, even when the researchers have good command of the language, all basic annotation should be checked by / with native speakers at some point.) We proceed as follows.

First we use the Transcriber program (see references) for segmenting the audio signal into segments that roughly correspond to sentences. For the time being, segmentation is done on somewhat impressionistic, mostly intonational criteria. This is certainly not the optimal way, more objective criteria are to be found and applied. This will be easier when the syntactic structure of the language is better understood. We aim at using sentences (full or elliptic), not phrases nor clauses nor intonational groups, as basic units for linguistic text annotation. (This is in accordance to the principles of Advanced Glossing, see below, sec. 5)

In Transcriber, adjustments of time segmentation are quite easily done. This includes division or merger of (adjacent) segments. Unfortunately, the same does not hold for Shoebox. However, it seems to be possible to write simple pearl or lisp scripts to automatize this, but this has not yet been undertaken in the Awetí Project. The same holds, to our knowledge, for ELAN, but again, it should be possible to implement such functionalities.

Transcriber is also useful for adding a transcription directly to each of the segments (sentences). We use an orthographical rather than a phonetic or phonological form for this first transcription. (An orthography for Awetí has been established during the last years and is beginning to be used by some speakers, and is applied in basic education.) Besides of being much quicker, this procedure avoids difficulties in entering phonetic symbols (depending on the configuration of the platform, Transcriber may present some difficulties here) and can be done by native speakers themselves.
Currently we are training some younger Awetí to use the computer and enter transcriptions with Transcriber. According to our experience, a well trained person can transcribe at least about two minutes in one hour of work, while one minute per hour is more common. While training the native speakers (this includes, in most cases, training to apply the Awetí orthography) we calculate as less as half a minute in one hour work.
Besides training of native speakers, doing the transcriptions offers many opportunities to discuss unclear or open points of the orthography which still may be submitted to some minor clarifications, additions or even changes.

After transcribing the text with the help of (and in future, by) native speakers, a translation is added. Transcriber is not designed to do this. We do all further processing using the Shoebox program (see references).

Conversion from Transcriber (which saves the transcriptions to XML files) to Shoebox (which uses plain text files with mark-up in SIL's "standard format"), and vice versa, can be most easily done using the Econv tool developed by the TIDEL team at MPI Nijmegen (see "MPI Tools" in the references). The resulting Shoebox databases are of a special Shoebox database type (database types are basically a mark-up schema, that is, a description of permitted field types in an entry, and their relationship among each other). This database type (econv.typ) can be expanded to include more lines with different annotations referring to the same segments of the raw data, principally lines with a rough Portuguese translation, a word-for-word Portuguese translation, and, later on, an English translation for each sentence.

The translations are obtained in different ways. In our case, the most common way is to print out the transcription in a big font and with much space between the lines, and add an translation between the lines with native speakers (some are able to do this on their own). The translations can than be typed into the correspondent Shoebox database. I distribute the Shoebox-export settings and the Word document style used for printing out the translation with this document (download here) .
Another way is to write a translation down on paper while the transcription is done. This slows the transcription process down and bears the difficulty of keeping track of the time codes. When several native speakers are involved, this may still be a valid method, also for purposes of training in Portuguese. Yet, especially when the researcher is working with only one informant and he himself is typing in the translation (using Transcriber), a good way is to tape the whole working session and ask the native speaker to give a translation of each sentence (this is almost automatically done in most cases). If loudspeakers are used while transcribing (and not just headphones), one can easily identify the current time position and type the translation later directly into Shoebox, listening to the taping of the transcribing process, possibly even after returning from the field.

Naturally, the process of translating gives access to a wealth of information, potentially rises many new questions and creates many opportunities for further elicitation, which should be catched as soon as possible in the field.

The result of these steps is a minimal text documentation as agreed among the DOBES projects – it's annotation includes a (time-linked) transcription (here of orthographical character) and a free translation (at least to Portuguese, but English translations are to be provided, too).

In order to understand our further processing of a sub-part of the corpus, I have to make a longer excursion in order to justify the need for the glossing format that is applied in this project. Therefore, the next section does not describe concrete steps in the workflow, but is of a general character.

5. The need for a complete annotation format: Advanced Glossing

Why Interlinear Morphemic Translations are not sufficient as an text annotation format

It seems to be unanimous that a transcription and a translation (at sentence or phrase level) alone are by far not sufficient annotation for the purpose of language documentation, especially in the case of endangered languages. One aim of a complete

documentation format should be to provide enough information to develop an in-depth description of the language even if there are no native speaker available. This means that it has to be possible to give information of several different linguistic types (at least those related to language structure as being described in a grammar), each in its proper place.

Our proposal for a complete text annotation format is Advanced Glossing (AG). It is meant as an elaborated extension of a well known glossing format, at least in functionalist/typologically oriented work: Interlinear Morphemic Translations (IMT), first systematized by Christian Lehmann (1982).

Some brand of IMT is possibly applied in many if not most current language documentation projects. (For a representative overview over the variation of displayed formats, see Bow, Hughes and Bird's (2003) contribution to this workshop.) However, I will argue that this format presents problems for this purpose, although it has proven useful for illustrative purposes, especially in descriptive grammars in a functionalist / typological framework.

For one thing, interlinear *morphemic* translations are by definition restricted to morphology. It's relations to phonetics, phonology, syntax, and semantics are, therefore, indirect at best. Several authors felt this restriction: As Bow et.al. (2003) show, there are expansions of the format so as to include glosses on the word level, in addition to or substituting annotation on the morph(eme) level. Still, in most cases the annotation is restricted to glosses that relate to isolated base forms – single words or single morphs.
Also in the EUROTYP project, the format was altered and expanded to ten or so lines with different information types (Bakker et al 1994). But there are still some problems with the format that cannot, it seems to us, be cured by adding some more lines. Besides the bias to morphology (or, more generally, base form related annotation), these problems include lack of inter-theoreticality and applicability to languages of different language types, and unclearness as to how to interpret the glosses in the case of grammatical / functional units.

There is no such thing as a theory-neutral language documentation, if documentation is to mean more than the mere recording of speech. Any annotation advances hypotheses, and virtually any hypothesis is formulated in terms of some presupposed theory. This means that there can be no theory-neutral glossing format either, but any such format should strive at inter-thoreticality, that is, to be usable by and compatible with all major models of linguistic theories.

Following a famous classification of linguistic theories set up by Hockett (1958), there are three basic models of linguistic description: The 'Item and Arrangement' and the 'Item and Process' models, on one hand, and the 'Word and Paradigm' model, on the other. The first two underlie American Structuralism and most theories that have developed from it, including typologist approaches to language description (most following the Item and Arrangement model) and most branches of the generativist theoretical reasoning (many following an Item and Process approach). These theories are so widespread that many a linguist may even ignore the existence of other models, namely of the Word and Paradigm model. During centuries, linguistic descriptions were based on theories belonging to this type, and there is a somewhat growing interest in modern theories that take the notion of paradigms as basic (some neostructuralist theories, especially in the European tradition of structuralism belong to this type).

So, a general glossing format should be compatible not only with theories based on an Item and Arrangement or Item and Process model, but also with Word and Paradigm theories. The IMT format, however, is clearly designed presupposing a 'Item and Arrangement' model of language theory (probably, it is compatible with some variants of the Item and Process model, too).

Restrictions of the underlying model type carry over to annotation formats based on them. The Item and Arrangement model works fine for most phenomena in agglutinating languages, but there are difficulties with synthetic word forms in inflecting languages, and with analytical word forms in arbitrary languages.

These difficulties are the main reason why I think we cannot recommend Interlinear Morpheme Translations, as they stand, as the principal device in providing grammatical information in text annotation in a language documentation. Let's discuss these via examples. Like Lehmann (1983), I will use Latin as the sample language, and I will even use variations on his examples.

Consider the following typical traditional example of an IMT:

> *time -o ne veni -a -t*
> fear **-1.SG NEG.VOL** come **-SBJV.PRES -3.SG**
> 'I am afraid he might come.'

For the purpose of language documentation, this annotation has to be considered as incomplete, as any customary account of Latin grammar describes verb paradigms that involve several categorizations for any (finite) verb form, at least:
person, (verbal) number, genus verbi, mode, and tense.
In this sentence, we have for instance an occurrence of the form *timeo*, a form that has to be characterized as belonging to the following categories: first person, (verbal) singular, indicative, active, present tense. (Other theories speak here of 'morphosyntactic features', a term that is frequently applied in many theories, but, as it turns out, has, to my knowledge, not been clearly defined anywhere.) Some of these categories can reasonably be related to the occurrence of the affix *-o*. (This way, *-o* would be analyzed as a heavily functionally loaded portmanteau-morpheme.) So the glossing of this form could be augmented as follows:

> *time -o*
> fear **-1.SG.IND.ACT**

Still, where do we indicate that *timeo* is a present tense form? We cannot ascribe this to the *-o* affix occurrence, as the same affix occurs in forms of other tenses, cf. *time -b -o* (future 1), or *tim -u -er -o* (future 2).

Lehmann gives two possibilities for such cases, which are explicitly both to be understood of notational variants of an analysis that includes the occurrence of a zero morpheme:

> *time -Ø -o* *time -o*
> fear **-PRES -1.SG.IND.ACT** fear**(PRES) -1.SG.IND.ACT**

As already holds for the case of portmanteau-morphemes, this analysis may be adequate or not: It is certainly not theory neutral, as many theories reject zero morphemes (traditional descriptions of Latin would ascribe the form to present tense to the occurrence of *time*, a present tense **form** of the **stem**). Even if different but valid interpretations of this glossing in terms of some other theories (which, e.g., avoid zero morphemes) can be found, we have still to conclude that there are problems of IMT with regard to theory neutrality, at least when it comes to synthetic word forms in inflecting languages.

The same holds even more in the case of analytic word forms. Consider the sentence:

> *monitus eram ut venirem*
> 'I had been asked to come.'

Here we have, again according to customary traditional descriptions, an occurrence of the analytic (or 'periphrastic') verb form *monitus eram*. This form could be ascribed to seven categories: first person, (verbal) singular, indicative, passive, pluperfect, (verbal) nominative, (verbal) masculine. So how would a interlinear morphemic glossing of this sentence look like, if it is to reflect this description?

Consider the following possible glossing for this sentence.

moni	*-t*	*-us*	*er*	*-a*	*-m*
ask	-PART.PF.PASS	-NOM.SG.M	PASS	-IND.PAST	-1.SG.ACT

ut	*veni*	*-re*	*-m*
[??]	come	-SBJV.PAST	-1.SG.ACT

Some of the categories that *monitus eram* belongs to (in other theories: some of it's "morphosyntactic features") appear directly under one morph occurrence (cf. first person, indicative, nominative, masculine), others even repeatedly (singular, passive). However, the categories hold for the whole form, and there is no easy way of recognizing that the two words (and hence, their glossings) belong together. There are proposals for using some type of brackets for these cases. But what when it comes to discontinuous forms? Consider the following possible variant of the above sentence:

> *monitus ut venirem eram.*

Here, bracketing becomes impossible if not special devices (such as indexing) are applied, which render the annotation not easily readable, at best, not to mention the difficulties for a digital implementation.

Even if these difficulties were overcome in some way, we still have a similar problem as with *timeo*, above – where do we indicate the tense, plu(squam)perfect? The whole form is to be categorized as pluperfect due to the combination of a participle perfect form with an auxiliary verb form in past tense. Each of these is indicated in the glossings, but there is no way to ascribe the category to the analytic verb form as a whole, because the IMT format is restricted to the glossing of morphs (or single words), and pluperfect should be seen, and is indeed seen in several theories, as a *syntactic* category (it can be shown that this holds for most "morphosyntactic features").

Other, similar problems exist, for instance in connection with so-called free morphemes (for instance, I did not give, by purpose, any proposal for a glossing of *ut*, above). The character (syntactic or morphological) of their glossings is often unclear. The list could be continued.

Again, these are factual questions that any theory has to cope with in its own terms; the point is that the IMT glossing format is not only far from being theory neutral (no glossing format can possibly be), but also not sufficiently inter-theoretical, because it is simply not compatible with a number of theories.

One important step in direction to inter-theoreticality is to make all tacit assumptions and theory-dependent concepts and conventions as explicit as possible, to allow followers of other theories to interpret the annotation in their own terms. However, unfortunately the attempts at formalizing IMT fall short in this respect, too. (This might be a consequence of the theories that underlie it.) When one wants to know what exactly the glosses in capitals stand for, one receives different and merely intuitively, if at all, understandable answers. Consulting Lehman (1982) or the EUROTYP guidelines (Bakker et al 1994), these glosses are explained as "grammatical category labels", but do they stand for categories? And if so, does "grammatical" mean "syntactic" or "morphological"? We also find in the same texts as explanations that they express the "meaning" of grammatical "elements" (i.e., morphs, or morphemes?), or indicate the "grammatical function" of these elements. So, after all, do they represent categorial, semantic, functional or what type of information?

Even worse, in many single language descriptions, the only explanation for these elements (naming "morphosyntactic features") is the long version for the abbreviations. (As long as the focus is on the abbreviations themselves, this might be appropriate, cf. Croft 2003 who does not make any ontological statement at all.) A general inspection of the abbreviation list for the EUROTYP project (see Lieb, Dwyer, Anderson 2001) demonstrated that the ca. 550 labels stand for linguistic entities that are usually interpreted as belonging to very different ontological types: almost the half stand for 'morphological' (or, as argued above, syntactic, at least word-paradigm-related) categories, a hundred or so for syntactical word classes, ca. 60 for syntactic relations, others for syntactic constituent categories, others for semantic roles, others designate word order properties, varieties, sentence types or other. Worse, about 75 or so are completely unclear or unspecific. It cannot be excluded that this hinders the future use of the EUROTYP annotated texts.

The point made here is that this vagueness has to be avoided in language documentations, if future linguists that might apply theories we are not even able to think of, are to make any use of the documentations.
The proposal of a general knowledge database or "ontology" such as GOLD (see Farrar and Langendoen 2003) may be a promising advance in the direction of avoidance of terminological ambigueness, but it runs the risk of providing just one more linguistic theory (theories consist to a large degree of definition of terms), or of including only the most popular theories at the present time or in a particular community. Compatibility of the proposed ontology with Word and Paradigm theories would have to be carefully checked. In any case, we need explicit explanations of terms, and the GOLD ontology may serve as a point of reference for any particular theory. (It is in a similar sense that Advanced Glossing is proposed as a maximal reference point for glossing formats.)

Basic features of Advanced Glossing and its application in the Awetí Project

In Lieb+Drude (2000), we made a proposal for a text annotation format, called Advanced Glossing. Advanced Glossing (AG) is designed for language documentation and is to be compatible with most if not all linguistic theories. AG is not a data model but rather a surface-oriented format. In principle, it may well be compatible with different general data structure models for interlinear texts, such as that proposed in Bow, Hughes and Bird (2003).
AG is not meant as an obligatory scheme that language documentation projects would have to follow, but as a maximal framework of reference, where every type of linguistic information we were able to think of has its proper place. Any concrete

text annotation can use a subset of the "tiers" we propose, or fill them in gradually.

The basic ideas of Advanced Glossing include strict separation of syntactic from morphological information, so for each sentence, on the one hand, and for each ('grammatical') word, on the other, there is to be a *glossing table*. Syntactic and morphological glossing tables are in many aspects analogous, but phonetic information (segmental, structural and intonational) is given only in syntactic glossing tables (which can, of course, correspond to one-word-utterances), nor do these include any information about the internal morphological make-up of the single words, which is given in the corresponding morphological glossing tables.

Each glossing table consist of 13 lines (in a revised version which is in preparation, two or three lines will be added). Some lines are holistic (such as an free translation of a sentence), some of which consist of lists (such as the lines for constituent structures which allow to deal with even discontinuous constituents). Most lines, however, consist of *cells*, each corresponding to one base units – the words, in the case of syntactic glossings, or the morphs, in the morphological case. Lines with cells include those for semantic information, which is separated from categorial information of two types: lexical (such as word classes, e.g. POS categories) and unit categories (such as cases, tenses etc., in the syntactic case). These categories are placed in two different lines that consist of cells, each cell containing a list of category names.
To each table, line or cell, there can be a comment that expresses doubts or gives other explanation. Cells or lines can be left empty, or by purpose, or due to lack of information. For further details, the reader is referred to the original proposal (Lieb+Drude 2000). The format stands as it is; however, a second version is in preparation that will include some few extensions and give more explanation on the format. Most of the discussion of the last subsection will be included the revised version.

In the Awetí Project, we apply AG as the frame of reference in all text annotation. This does not mean that for each text there will be complete syntactic glossings for all sentences and a morphological glossing for each word. On the contrary, complete glossings are planned for some 10 minutes or so of taped text only, for the process of filling in all types of information is very time consuming.
For a smaller sub - corpus (possibly some 2 hours of text), we plan to provide *basic* syntactic and morphological glossing tables, each table containing about eight partially or completely filled in lines, plus some derived additional lines (such as a Portuguese translation in addition to the English one) and comments.
The lions share of the corpus will consist of linguistically relevant sessions that have only the minimal annotation (i.e., only transcription and translation), or of culturally relevant data (without transcriptions) with only some accompanying comments. In a four - year documentation project, with the manpower available (including the native speakers, most of which are being trained during the project) we do even not expect to be able to provide basic annotation for most of the rich oral narrative threshold (mainly, myths)of the Awetí, some 70 to 100 hours, some 60 of which we hope to be able to at least tape and digitize.

As was said above, we apply the Shoebox program for annotating the texts and gathering information on lexical units (words and affixes). In Drude (2002), I described the quite complex Shoebox setting used to implement Advanced Glossing, and I refer the reader to this description. Each annotated text corresponds to a Shoebox database, where every record contains a syntactic glossing table together with its comments. In a second database (which will be split up for reasons of size), all morphological glossings are stored. The Shoebox field properties are set up in a way to reflect the ontological character of each glossing line. The interlinearization mechanism of Shoebox is used to implement the aligning into columns and hence, the cell structure of certain lines. Additionally, of course, support for interaction with the additional lexical databases (which Shoebox was originally designed for) is provided.

So, semi-automatic filling in of certain information (lexical meanings, categorial information etc.) is possible: To fill out cells of syntactic glossing tables, information is retrieved from relevant morphological glossings, and for these, the look-up processes refer to the lexical databases. At the same time, one can 'jump' to a relevant lexical entry directly from the syntactic glossings.

With this document, I provide test and configuration files to exemplify the implementation of Advanced Glossing with Shoebox. (download here)

After this longer general section, we can proceed with the description of the concrete steps in the workflow of the Awetí Project, which eventually will make the last considerations somewhat more concrete.

6. Adding linguistic annotation [steps (5) and (6)]

The result of the previous steps (up to step (4), see section 4) is a Shoebox database with records corresponding to sentences, each record including an orthographical transcription, a translation to Portuguese (and English) and a time mark, that is, a unique reference to the starting point of a corresponding segment in the underlying audio file.

The database is still of the database type *econv.typ*, an extended version of the database type definition that comes with the Econv-tool, with additional definitions for the different translation fields. The next step is, then, conversion of these databases to databases of the type created for syntactic glossings in Advanced Glossing. Indeed, the conversion into the *AG-Syntax.typ* format is obligatory for the export process mentioned earlier used to print out the transcriptions if the translations are to be added later on on paper.
Again, I wrote a simple emacs-lisp script to automatize the conversion process, including conversion of provisionary work-arounds to mark nasal accents, and conversion of the field markers. I distribute this script with this document (download here). The result is exemplified by a data record as follows:

```
\ref      0002. 690
\per      mawalaja
\SXII     jatätsu jatä ozoporywyt:
\SIXp
\SXIIIn1 assim é que é nossa tradição:
\SXIIIe  it is like this that our tradition is:
\dt       02/May/2002
```

Each line represents a line in an syntactic glossing table. The transcriptions, for instance, are annotation in the syntactic glossing line XII (Shoebox field marker, in smaller fonts and starting with a backslash: **\SXII**), and the translations are different instances of data in line XIII (**\SXIIIe**, for English, and **\SXIIIn1**, for the first draft translation in the national language, Portuguese). In some cases, where no complete translation of the sentence could be obtained, or where certain word-related phenomena of further interest occurred, we make use of a further line with preliminary wordwise translations, **\SIXp**(here empty). The corresponding gloss line **\SIXn** will later be created in the interlinearization process, so we chose another name for not having this line

overwritten. The other fields are self-explanatory.

As said above, for most data this will be all annotation that is provided.

For as many as possible sessions, however, we will add further annotation. First, we add semi-automatically information extracted from morphological glossings, as shown below.

```
\ref      0002. 690
\per      mawalaja
\SXll     jatätsu jatä ozoporywyt:
\Sl       1      2     3           4
\SVl      jatä   tsu   jatä        ozoporywyt
\lx       jatä   tsu   jatä        porywyt
\SVll     dpron  pp    part        n
\SVlll    Unm_Nf Unm_Pf Unm_Pf      N_13 Unm_Ntense
\SlXn      este  como   é.que       costume
\SlXe      this  like   is.it.that  tradition
\CSlXn     este  como   é.que       nosso_costume
\CSlXe     this  like   is.it.that  our_tradition
\SlXp
\SXlln1   assim é que é nossa tradição:
\SXlle    it is like this that our tradition is:
\nts
\dt       02/May/2002
```

Orthographical words in line **\SXll** are split up into syntactic base forms (sometimes called 'grammatical words', including clitics), given in line **\SVl**; here, the clitical postposition *tsu* is separated from the governed pronoun *jatã*. To each word, a citation form is given in the **\lx** line (this is a new feature to be included into Advanced Glossing), cf. the citation form *porywyt* 'tradition, custom, culture', for the inflected form *ozoporywyt* 'our tradition'. Then we have word categories in line **\SVll** and form categories in line **\SVlll**. In this case, there is mostly only one relevant category for each word, but there may be several, as in the case of *ozoporywyt*, which belongs to the form categories "nominal first person plural exclusive" and to "unmarked for nominal tense". The numbers in line **\Sl** have been added by hand, but alignment to the columns is automatic.

In the next lines we have a gloss for the lexical meaning of the single words (function words would have entries of a different type here), in Portuguese (**\SlXn**) and in English (**\SlXe**). The next lines (**\CSlXn** and **\CSlXe**) are additions to Advanced Glossing, they contain a complete gloss, that is, a gloss that includes effects of inflection (here only relevant in the case of *ozoporywyt*).

Most of this information has been added semi-automatically, only in the case of *jatã*, which functions sometimes as a demonstrative pronoun and other times as a topicalization particle, we had to disambiguate. The information is obtained from the database of morphological glossings, of which I show the entry for *ozoporywyt*, the only morphologically complex word in the sample:

```
\MXll     ozoporywyt
\lx       porywyt
\MXllln   costume
\MXlle    tradition
\Ml       1        2
\MVl      ozo-     porywyt
\MVll     f:poss-  n
\MVlll    Aff-     Unm
\MlXn     13-      costume
\MlXn     13-      tradition
\nts
\SVll     n
\SVlll    n_13 Unm_Ntense
\CMXllln  nosso_costume
\CMXlle   our_tradition
\dt       02/May/2002
```

In the records of this database, we find not only fields that represent lines in a morphological glossing table according to Advanced Glossing (their field names start with **\M**), but also fields that serve basically for the interlinearization process as a source for the syntactic glossing tables (names of such fields start with **\S**, additionally **\lx**), besides the usual housekeeping fields such as the date-stamp field **\dt**. The (non-AG) word-glosses including effects of inflections are here stored in the fields with the labels **\CMXllln** and **\CMXlle**. In the beginning, the glossing is quite slow, as each morphological glossing for any word in any sentence has to be created. Later, when the most frequent words have been glossed morphologically, the annotation process goes much quicker. Still, it will admittedly never be as quick as pure morphemic interlinearization as done by the Shoebox standard setup, resulting in traditional IMTs. In our setup, more information has to be entered. Currently, for one minute of recorded text we calculate some 40 minutes of annotation / glossing / interlinearization work.

But again, some of the information in the morphological glossing tables is obtained semi-automatically by looking up information in the relevant entries in the lexical databases. If an corresponding entry does not yet exist, it should be created, so the lexicon grows with the glossing of texts (the morphological glossings functioning as an intermediate level). In our case, for technical reasons, we had to set up three such lexical databases, one for affixes and one for simple words (and their stems, including particles, clitics and so-called 'free morphemes') – these two are used in the interlinearization processes. The third one is for complex (derived or composed) words. One can also jump directly from any word in a **\lx** field (in both, a morphological and a syntactic glossing) to the correspondent entry in the databases for simple or complex words. I show here an excerpt from the lexical entry for *porywyt*:

```
\lx  porywyt
\lc  [kaj]porywyt
```

```
\ps  n
\gn  costume
\ge  tradition
\dn   (um elemento da) cultura tradicional, os costumes antigos
\de   (an element of) traditional culture, the old costumes
\nq  obligatorily possessed?
\st  check
\simp  S?
\dat  12/Set/2001
\dt   02/May/2002
```

The structure of lexical entries is quite a complex matter that cannot be discussed here at any length. See, for instance, Wittenburg, Peters, Drude (2002) and Wittenburg (2001), for details. We use a (somewhat modified and extended) 'MDF'-structure in order to be convertible into print-out dictionaries via the Multiple Dictionary Formatter, provided with Shoebox.

As shown by Peter Austin (2002), Shoebox can (and should, I would recommend as best practice) be used to maintain and link several other databases relevant for any given language documentation project.
As an example, for reasons of consistency, all fields that contain abbreviations use controlled vocabulary (range sets, in Shoebox terminology), and for each abbreviation there has to be an entry in a further database. In such an entry one finds not just the long form, but an explicit description what type of entity is designated by the term. Eventually, this database can be related to other terminological systems such as the GOLD ontology mentioned above (Farrar and Langendoen 2003).

Confer the entry for **Unm_Ntense**, an abbreviation occurring in line **\SVIII** , in the morphological and syntactic glossings above.
The entry includes fields for: the abbreviation; the ontological domain (is it a syntactic or morphological category or feature, in case of a category: is it a lexical category (e.g., a word class, such as a part of speech) or a category of forms (such as a tense) etc.); the long name of the term (category), the theoretical status (is it part of the language system at the morphological / syntactic level, or is it a phonological or a semantic class, etc.); an explanation / description; a sample (here empty); and housekeeping fields.
Of course, the details of these depend on the linguistic theory underlying the annotation, which should be made as explicit as possible.

```
\abrv   Unm_NTense
\dom    Syntax:Form
\long   Unmarked for nominal tense
\type   SUO (Syntactic Unit Ordering)
\expl   It is still unclear if nominal forms with
     the suffixes _-(p)ut and _-(z)an_ should be treated
     as derived or as inflected.
     If inflected, the relevant categories could well
     be called Nominal Past and Nominal Future.
     Most other noun forms are, then, unmarked
     for nominal tense.
\sample
\create  11/Jun/2001
\dt    02/May/2002
```

As said above, for a smaller sub-corpus, the syntactic and corresponding morphological glossing tables will be further completed according to Advanced Glossing. I have shown a sample complete syntactic glossing table under Shoebox in Drude (2002) and need not to repeat this here. This work includes, especially, adding of phonetic and phonological information, and information on grammatical structure and relations.
We have not begun this phase for any substantial part of the raw data beyond for testing and demonstrating purposes. As for information on syntactic structures and relations, the correct filling in of the corresponding lines presupposes linguistic analysis which has, in most cases, not yet been concluded. For this and other reasons, we expect this phase to be very time consuming, and this is why only a very small part of the corpus will be completely documented this way.

We could, however, indicate, for instance, that words 1 and 2 in the syntactic glossing table, above, constitute a postpositional group. Analytical word forms and, in particular, discontinuous constituents should indeed be indicated immediately as they are identified. Many other details would have to be entered with question marks, indicating doubt, as is foreseen by the AG format.

7. Finishing a session: Conversion and Inclusion into the data corpus [steps (7) to (9)]

Currently, the Awetí Project does not use ELAN, for annotation, the annotation tool being developed at MPI Nijmegen. For us, the steps described above, using mainly Transcriber and Shoebox, are sufficiently efficient, as we do have no need for direct video annotation or entering complex UNICODE letters, functions that ELAN offers already while Transcriber and Shoebox do not (besides the disadvantages of Shoebox being a tool that is not free any more, and is, at least officially, not being developed any further). On the other hand, interaction between lexical databases and text data, semi-automatic filling in of annotation of different types ("interlinearization", see above) are features we would very much like to continue to use, and ELAN will need some time of development until being able to substitute Shoebox in these respects (if it ever will).

However, it is possible to convert the results of text annotation obtained in the way described above into the ELAN data format, using the same tool Econv that also converts Transcriber to Shoebox and vice versa. We have not been able yet to test the new version of ELAN, which should directly import Shoebox data.
It is obvious, though, that for purposes of presentation, ELAN is fundamental, as it allows synchronized playing of video images and audio data together with selected tiers of annotation. Our experiences with earlier versions have not yet given satisfactory results as to ease of operability for segmenting and entering annotation, but this should basically be solved with the new versions.

ELAN does already support the Advanced Glossing format to a large extent and will eventually allow the format (including comments etc.) in its entirety. In fact, when starting to develop ELAN, the Advanced Glossing proposal was used as a basic orientation for possible (maximal) linguistic needs of a language documentation tool.

We deliver the data to the Archive at MPI organized in sessions, each session being described by a set of metadata. This metadata specification follow the IMDI layout (see IMDI 2003). It consists of a file that contains descriptions of the technical details of the recordings, of its content, and of the relevant files connected with the session. Some of the files are to be created

according to information in the metadata, as start and endpoints in audio and video files are used to cut session-relevant parts off the original digital master files.
To create the metadata, we use the *IMDI-Editor*, also developed by the MPI team. Although the tool is developing very quickly into a easy usable tool, entering of metadata continues to be a very time-consuming task, which binds much more of our manpower resources than we have planned at the beginning of the documentation project. However, if the data are to be accessible and searchable according to very different criteria, which is a necessary condition for any language documentation, the maximum of completeness for metadata is required.

The metadata, together with the relevant media and annotation files, is stored as part of the *corpus*, the central part of the language documentation, which will eventually be made available via the Internet, given the permission of access to the relevant pieces of data by the speech community and individual informants. (Some data will have to be given restricted access, or even be closed to the larger public at all. Mechanisms to protect the rights of the involved parties that still allow a maximum use of the archive are being developed, but this continues to be a very sensitive matter, often underestimated by language documentation projects. The situation may greatly vary in different countries or continents.)

The primary and secondary data of the Awetí Project in the MPI archive will be organized as follows:

1. Material FROM language / culture
(a) Linguistic Data
i. Non-elicited data
A. Monologue-like (myths, historical narratives, cultural explanations, procedural texts, descriptions etc.)
B. Dialogic (conversations, interviews etc.)
ii. Lexicon (including specific word lists)
iii. Elicitation (phrases and sentences)
(b) Non-linguistic Data (songs, instrumental music, photos, drawings made by speakers, iconographic material etc.)
2. Material ABOUT language / culture
(a) Language (sound system, orthography, closed classes, grammar sketch)
(b) People (ethnographic informantion, sociocultural informantion, historical informantion, geographic informantion (including maps), relations with other Xingú groups etc.)

If need be, any final branch of the above structure can be further subdivided. This holds especially for branch 1.a.i.A, which includes most of the linguistic data. Any terminal knot contains four directories for the different types of files: info files, metadata, annotation, media and other related files (such as scanned photos or graphics).

The metadata descriptions of each session must include or be accompanied by the information where to put the session-related files in the tree structure of the future browsable corpus. The corpus can then be searched, reorganized, and the individual sessions can be displayed, using the IMDI metadata browser, one further tool being developed by the MPI group.

8. The organization of files in the computer [step (10)]

This is an often neglected aspect of project work, or is taken for granted. However, when dealing with several hundreds of sessions, each with an audio and / or video media file and several annotation files in different formats, one has to spend some minutes and think of some conventions where to save the files and how to name them.

I will here give a brief overview over the solutions applied in the Awetí Project.

There is a naming convention for raw data files and the first version of the sessions based on those raw data established in the DOBES consortium during its pilot phase. One session based on a minidisk-taping in the Awetí Project has, for instance, the name `AWSDAM23Jun0201-S01`, where `AW` stands for the Awetí Project, `SD` for the recording researcher, Sebastian Drude, `AM` for Audio-Minidisk, `23Jun02` for the date of the recording, `01` indicates that this is the first (audio or video) medium started to be taped on that day, and `S01` that it is the first session based on this medium.
Several files, including the digital master files and the metadata files are based on this convention and differ mainly in the existence of the final part `-S01` and in the file name extension. `AWSDAM23Jun0201-S01.imdi`, for instance, is the metadata file that describes this session.

Other file names are created by the Awetí Team itself, the names may include parts of the conventions introduced above (such as recording date and tape number), but usually some element indicative of the content is added in order to recognize the file more easily. An example is `23Jun02-01-amanajup.sdb`, where `Amanajup` is the name of a speaker. `.sdb` is used as a file name extension for **S**hoebox-**D**ata**B**ases.

The project tree on the hard-disks of all project computers is organized in a way that facilitates their synchronization. It is organized as follows:

`X:\Aweti\` contains five folders:
`Papers+Results\` (all scientific products, such as this paper)
Contains folders such as: **`2003-07-EMELD`**
`MPI-Corpus\` (the final data tree, including the tree explained in the last section)
`Organizational\` (Budged, Reports, Research Permissions, Official Documents...)
Contains folders such as: **`VWS, FreeUniv, MuseuGoeldi, Internal, FUNAI`**
`Media\` (all audio and video files, images...)
`Data\` (all files related to annotation and processing)

The content of the **`Media`** and **`Data`** folders is further organized first by year, then by file type.
For instance,

`X:\Aweti\Media\2002\` contains the subfolders:
`Audio-Field\` (audio-files digitized in the field)
Contains files such as: **`23Jun02-01-amanajup.wav`**
`Video-Field\` (video-files digitized in the field – almost empty)
`Audio-DMF\` (Digital Master-Files digitalized by the MPI, each corresponding to one MD)
Contains files such as: **`AWSDAM23Jun0201.waf`**
`Video-DMF\` (Digital Master-Files digitalized by the MPI, each corresponding to one tape)
Contains files such as: **`AWSDAM23Jun0201.mpg`**

Audio-Sessions (ready cut session media files from the MPI)
Contains files such as: **amjp-biogr.wav**
Video-Sessions (ready cut session media files from the MPI)
Contains folders such as: **amjp-biogr.mpg**
Photos (Digital Photos or scanned material etc.)

Analogously,

X:\Aweti\Data\2002 contains the subfolders:
Transcriber (Transcriber-generated XML-files such as:)
23Jun02-01-amanajup.trs
Shoebox-texts (Shoebox databases for texts as discussed above, such as:
23Jun02-01-amanajup.txt, the Econv-generated Shoebox-file, and
23Jun02-01-amanajup.sdb the database in AG-Syntax format)
Shoebox-lists (For transcribed wordlists and similar)
Print-out-versions (For RTF-files used for elicitation and translation, such as:)
23Jun02-01-amanajup.doc
Shoebox-lexical-databanks
(for morphological glossing tables and year-specific lexical databases, e.g.:)
23Jun02-01-amanajup-MGT.sdb
Metadata-work (For metadata files that have still to be revised or completed, such as:)
AWSDAM23Jun0201-S01.imdi
Metadata-final
(For metadata files as sent to the MPI, named as the above.
Later replaced by renamed files sent back from the MPI such as:)
amjp-biogr.imdi

Besides the year-specific folders below **Data,** there is a folder **General** which includes, among other, the three general lexical Shoebox databases.

In the folder **Organization/Internal,** there are such files as Excell sheets for all media, sessions, annotations etc. for each year.

With these remarks I finish my presentation of the work done in the Awetí Project. I hope some of our solutions can count as "best practice", or at least as a starting point for a fruitful discussion to arrive at these.

References

Audacity (computer program)
http://audacity.sourceforge.net/

Austin 2002:
Peter K. Austin. 2002.
Developing Interactive Knowledgebases
for Australian Aboriginal Languages - Malyangapa.
Paper presented at the Workshop on Australian Aboriginal Languages,
University of Melbourne, March 2002.
http://www.linguistics.unimelb.edu.au/contact/staff/peter/Malyangapa.pdf

Bakker et al 1994:
Dik Bakker, Oesten Dahl, Christian Lehmann, and Anna Siewierska. 1994.
Eurotyp guidelines. Technical report.
Fondation Europeenne de la Science, Strassbourg. (EUROTYP Working Papers).

Bow, Hughes and Bird (2003):
Cathy Bow, Baden Hughes and Steven Bird. 2003.
Towards a General Model of Interlinear Text.
Paper to be presented at the third EMELD Conference on
Digitizing & Annotating Texts & Field Recordings.
LSA Institute, Michigan State University, July 11th -13th.
http://emeld.org/workshop/2003/bowbadenbird-paper.html

Croft 2003:
William Croft. 2003.
Abbreviations and symbols for interlinear morpheme translation.
In: Typology and universals, 2nd edition.
http://lings.ln.man.ac.uk/Info/staff/WAC/Papers/TypAbbrev.pdf

DOBES:
http://www.mpi.nl/DOBES
http://www.mpi.nl/DOBES/teams/Aweti/Aweti.html
http://www.mpi.nl/DOBES/teams/Kuikuro/Kuikuro.html
http://www.mpi.nl/DOBES/teams/Trumai/Trumai.html
http://www.mpi.nl/DOBES/teams/TIDEL/TIDEL.html

Drude 2002:
Sebastian Drude. 2002.
Advanced Glossing - a language documentation format
and its implementation with Shoebox.
Talk at the LREC-Workshop in May 2002, Las Palmas.
http://www.mpi.nl/DOBES/meetings/lrec2002/lrecWorkshop.pdf

Farrar and Langendoen (2003):
Scott Farrar and D. Terence Langendoen. 2003.
Markup and the GOLD Ontology.
Paper to be presented at the third EMELD Conference on
Digitizing & Annotating Texts & Field Recordings.
LSA Institute, Michigan State University, July 11th -13th.
http://emeld.org/workshop/2003/paper-terry.html

Hockett 1958:
Charles F. Hockett. 1958.
Two models of grammatical description.
Word, 10:210–234.

IMDI 2003:
ISLE Metadata Initiative. 2003.
Metadata Elements for Session Descriptions.
Draft Proposal Version 3.0.3, July 2003.
http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.3.pdf
On ISLE see also:
http://www.ldc.upenn.edu/sb/isle.html

Lehmann 1982:
Christian Lehmann. 1982.
Directions for interlinear morphemic translations.
Folia Linguistica (Acta Societatis Linguisticae Europaeae), XVI:199-224.

Lieb, Dwyer, Anderson 2001:
Hans-Heinrich Lieb, Arienne M. Dwyer, Gregory D. Anderson. 2001.
Approaches to Morphosyntactic Annotation
Unpublished DOBES internal Working Paper.
http://www.linguistlist.org/~workshop/markup/DOBES-markup.html

Lieb+Drude 2000:
Hans-Heinrich Lieb and Sebastian Drude. 2000.
Advanced Glossing: A language documentation format.
DOBES internal Working Paper.
(1st version; a second, more explicative version is in preparation).
http://www.mpi.nl/DOBES/applicants/Advanced-Glossing1.pdf.

MPI Tools:
ELAN, Econv, the IMDI editor and IMDI browser, among others,
are developed by the TIDEL Team at the Max Planck Institute in Nijmegen.
They are freely available at:
http://www.mpi.nl/tools

Shoebox (computer program):
www.sil.org/computing/shoebox/

Sound Forge XP (computer program):
http://www.sonicfoundry.com/Products/showproduct.asp?PID=668

Transcriber (computer program):
http://www.etca.fr/CTA/gip/Projets/Transcriber/

Wittenburg 2001:
Peter Wittenburg. 2001.
Survey of lexical structures.
(Internal working paper in the DOBES Consortium. Cf. Wittenburg et.al. 2002)

Wittenburg 2002:
Peter Wittenburg. 2002.
Effects of Compression on Linguistically Relevant Speech Analysis Parameters.
(Internal document presented to the DOBES teams, among others.)

Wittenburg, Peters, Drude 2002:
Peter Wittenburg, Wim Peters, Sebastian Drude. 2002.
On lexical structures in language engineering and field linguistics.
Talk at the LREC-Conference in May 2002, Las Palmas.
http://www.mpi.nl/DOBES/meetings/lrec2002/lrecWorkshop.pdf

Program	Readings	Participants	Instructions for Participants	Workshop Homepage
Registration	Local Arrangements	Emeld 2001	Emeld 2002	Emeld Homepage