Advanced Glossing — a language documentation format and its implementation with Shoebox

Sebastian Drude*

*Freie Universität Berlin and Museu Paraense Emílio Goeldi.

Abstract

This paper presents Advanced Glossing, a proposal for a general glossing format designed for language documentation, and a specific setup for the Shoebox-program that implements Advanced Glossing to a large extent.

Advanced Glossing (AG) goes beyond the traditional Interlinear Morphemic Translation, keeping syntactic and morphological information apart from each other in separate glossing tables. AG provides specific lines for different kinds of annotation – phonetic, phonological, orthographical, prosodic, categorial, structural, relational, and semantic, and it allows for gradual and successive, incomplete, and partial filling in case that some information may be irrelevant, unknown or uncertain.

The implementation of AG in Shoebox sets up several databases. Each documented text is represented as a file of syntactic glossings. The morphological glossings are kept in a separate database. As an additional feature interaction with lexical databases is possible. The implementation makes use of the interlinearizing automatism provided by Shoebox, thus obtaining the table format for the alignment of lines in cells, and for semi-automatic filling-in of information in glossing tables which has been extracted from databases.

1. Introduction

In recent years, the documentation of languages, especially of endangered languages, has received a growing interest within the linguistic community. Most researchers agree that the core of a language documentation should consist primarily of recorded and transcribed texts which should not only be translated but also annotated or *glossed* to be of use for a wide range of purposes.

A type of format widely used, especially in the context of typology or grammar writing within a functional framework, is *interlinear morphemic translations*, or *interlinear glossings (IG)* for short, first systematised by C. Lehmann (1982). However, although Lehmann proposes (1982:202) as the principal aim of that format "to make the grammatical structure [of a text in an unknown language] transparent", all it provides is a rendering of the "meaning or function" of the individual morphemes. This has proven useful when exemplifying established facts or illustrating a discussion, but it is by no means sufficient for the aims of language documentation. These aims include that it should be possible to write a grammar of the language or variety being documented, given a sufficient large number of texts that are completely documented.

During the first year of the Programme for the documentation of endangered languages (DOBES) funded by the Volkswagenstiftung, H. Lieb and the author of this contribution developed an extended glossing format, designed for purposes of language documentation, called Advanced Glossing. The primary aim of this paper is to present the most important ideas behind and features of Advanced Glossing (AG). First some relevant methodological and theory-related issues are considered. Then the basic ideas underlying AG will be presented, and a short characterisation of the details of the format will be given.

AG is a general format that does not stipulate details of a possible technical implementation. However, it is obvious that in the digital age any such format should be applicable by means of appropriate computer software. In the DOBES context, AG has been used as a frame of reference and basic orientation for the development of EUDICO (a multimedia tool for annotated language documentation) at the Max Planck Institute in Nijmegen. At a later stage, EUDICO will support AG. As a possible interim solution the author has set up a special complex configuration for the Shoebox-program, a tool designed for linguistic field work, especially for the creation of lexic(ographic)al databases and the interlinearisation of texts. The configuration allows for the partial or extensive documentation of texts within the AG format and also includes interaction with lexical databases. A presentation of this special configuration is the second basic aim of the present paper. The format of the different databases and their interaction, including semi-automatic filling-in of information and direct access to relevant database entries, will be shown.

Work on AG is not concluded yet. However, there is a first version available[1] which is fully usable in its present shape in actual language documentation. A second version of the description of the format will differ mostly in terms of explicitness, but the format itself will basically be the same, with some additions. This presentation is based on the version as yet available.

2. Glossing Formats, Theories, and Methodology

Some of the properties of Advanced Glossing that set it apart from traditional glossing formats follow directly from its different purpose: language documentation. If glossed texts are to serve as a basis for a complete language description, a proper place has to be made available for each different kind of information – phonetic, phonological, orthographical, prosodic, categorial, structural, relational, and semantic. Some of these information types occur twice, in syntax and morphology. It is for this reason that more than only one tier is needed.

The traditional IG format proposes one line, containing mainly morphological-semantic information and informa-

[1] See `http://www.mpi.nl/DOBES/applicants/Advanced-Glossing1.pdf` (Lieb and Drude, 2000).

tion of an unclear status. Consider example (1), adapted from Lehmann (1983:203):

(1) Or -e -mus!
pray -KONJ.PRS -1.PL
"Let us pray!"

In (1), "pray" is to render the lexical meaning of the stem *or*, but what, for example, is referred to by "1.PL" (short for "first person plural")? On page 200, such a part of an IG is characterised as "a configuration of symbols representing [the] meaning" of the morpheme named by "mus". On page 201, Lehmann speaks of "the meaning or function" of morphemes being rendered by IGs, but on page 221, such "labels taken from some grammatical metalanguage" are said to "represent the semantic or grammatical components", but they are consistently named "grammatical *category* labels". So do they actually refer to meaning, function, grammatical components, or categories?

Unfortunately, this vagueness is systematically present in glossings following the IG format. Often, no more explanation for "morpho-syntactic features" is given than a resolving of the abbreviations. This is partly due to the fact that the IG format is not as "theory-free" or "theory-neutral" as one may think (cf. Lehmann, 1983:199): the glossings make sense only if their interpretation in some framework of the Item-and-Arrangement or Item-and-Process model type is taken for granted. However, as already pointed out by Hockett (1958), there is a third model which he characterises as even "older and more respectable": the Word-and-Paradigm model.

There is no such thing as a theory-neutral documentation, if documentation means more than the mere recording of speech. Any annotation advances hypotheses, and virtually any hypothesis is formulated in terms of a presupposed theory. From this follows as an important condition for any general glossing format its *compatibility* with all major models of linguistic theory; not to be theory-neutral, but to be *inter-theoretical*. In particular, AG strives for *also* being usable with Word-and-Paradigm theories. (This sets AG apart from most of the current practice which is based on IG.) Consequently, a requirement for any documentation, especially for any glossing, is that underlying theoretical assumptions be made explicit and explained. This includes that, independently of any theory, the description language should be clearly interpretable; it is necessary to be able to distinguish between phonetic and phonological, morphological and syntactic, and between semantic and categorial, functional or relational information.

A last point has to be made with respect to methodology. In the case of an example that uses the IG format, the linguistic facts are established in the context. But in the case of glossings used in the documentation of languages, some information may not be known, or may be neglected systematically. This means that the documentation format must allow for the partial documentation of a text as well as for the gradual, systematic filling-in of gaps during the documentation process, and for the marking of missing or uncertain information. Gathering of information for the complete documentation of a text should, in principle, be possible in field conditions.

This does not mean at all that the researcher is bound by the format. A documentation format (and this incorporates a format for description) is not a research methodology or recipe. It is not meant to be an outline to be followed and filled in schematically.

After this brief explanation of the 'philosophy' behind AG, we can proceed to outline how the above requirements are put into practice.

3. Advanced Glossing: Basic Ideas

A first important feature of AG is that morphological information is strictly separated from syntactic information. For both levels, glossings are organised primarily in tables (glossing tables – GSs) that consist of several *lines*, one for each different type of information – phonological, semantic, categorial, relational and so forth.[2] The glossing of a text is primarily a sequence of syntactic glossing tables, one for each sentence.

The link between the two levels are the *syntactic base forms*. In syntax, they are taken as smallest building blocks that can be described, for instance, phonologically, semantically, or functionally or with respect to membership of syntactic categories. Several lines provide information that each applies to the same parts of a sentence which correspond to individual syntactic base forms or a certain number of these. Therefore, these lines are organised in *columns*, one for each syntactic base form, be it a particle ('free morpheme'), a clitic, or a form that could morphologically be analysed in stem or affix morphs. These baseforms I will henceforth call *words*.[3] In a syntactic glossing table (SGT), the morphological make-up of a given word is not accounted for.

Instead, for each such syntactic base form there may be a morphological glossing table. Morphological glossing tables (MGTs) are widely analogous to syntactic ones, they also consist of a number of lines that contain, for instance, phonological, semantic, categorial, structural or relational information. Most of these lines are also organised in colums, each column corresponding to a morphological base form, or *morph*. In both GTs, the intersection of a line and a column will henceforth be called a *cell*.

Not all lines are organised in cells, some provide global information which applies to the sentence (resp. the word, in the case of a MGTs) as a whole, such as constituency, grammatical relations or a rendering of the global meaning. For easy identification and re-use in other lines, especially in global lines (lines without cells), the columns of each table are numbered. This is achieved by a special line whose cells each contain a number. These numbers can be used, for instance, to refer to members of periphrastic word forms or even of discontinuous constituents if such entities are to be accounted for in a given linguistic approach.

Conforming the information type to be coded in a given cell or global line, these can be of different data types –

[2] This is consistent with other recent developments that build on the traditional IG format e.g. the format specified in the EUROTYP guidelines, (Bakker et al., 1994). However, these formats are still not designed for language documentation, and most of the points mentioned in the last section hold for these formats, too.

[3] Note that for purposes of documentation, or, more specifically, glossing, clitics are to count as words on their own.

some contain a single item, most often rendered by a string of letters or symbols, others can be lists of items. For instance, there will be at most one relevant lexical meaning for a given word in a SGT, but there may be several syntactic categories a given word form belongs to at the same time.

In addition to a (morphological or syntactic) GT organised in lines (and many lines, in cells), each glossing has a second part, a *comment*. The comment consists of (a) a global part that may contain relevant notes to the glossing table as a whole, and (b) a list of individual comments, each referring to a single cell of the glossing table or to one of its lines, be it global or divided in cells. For instance, if one is uncertain of the status of a putative syntactic base form (maybe what is seen as a clitic turns out to be a bound morph), this may be stated in the comment, in an entry referring to the whole line containing the numbers of the colums.

Each cell or line may be deliberately left *empty*, or may contain, for instance, question marks if the information is still missing but planned to be provided. Uncertain information may also be coded in combination with question marks, and an entry in the comment may then explain the nature of the doubt. Yet, we must not forget that also information not marked as uncertain is of hypothetical nature and may turn out to be factually wrong.

4. Glossing tables in detail

Not only are there parallel glossing tables for sentences (SGTs) and words (MTGs), these glossing tables (GTs) are also structured almost analogically. Therefore, they will be characterised together. In the case of a SGT, the term "*glossed unit*" refers to the whole sentence, in the case of a MGT, to a word (in the above defined sense). Analogously, a *base form* is a word, in the case of a SGT, or a morph, in the case of a MGT. Compare the accompanying sample glossing tables (tables 1 and 2).

The first nine lines of each GT are organised in columns (each line consisting of cells). The cells of the first line contain numbers that identify the columns for later reference and hereby record the order of the base forms of the glossed unit.

Despite the overall analogy, the next two lines differ in character between the syntactic and morphological GTs. MGTs account for abstract words that are used in utterances, but the utterances are always utterances of sentences.[4] It is the SGTs which document parts of speech events. Therefore, the phonetic shape is reflected only in SGTs. Line II in a SGT contains the segmental phonetic form of the whole sentence (including syllable breaks); line III the phonetic (sentential) intonation (in many cases the pitch contour will suffice, but other prosodic properties can be included here). In the case of lines II and III of a MGT, the *phonological* segmental shape and the phonological word intonation is given. In particular, in the case of tone languages, line M-III (i.e., line III of a MGT) is to represent the abstract tones (level pitches or glides).

[4]When eliciting word forms, an utterance of a single word form could be understood as an elliptic sentence where a part *"the word/form is:…"* has been omitted.

Line IV of a GT is for representing the phonological shape of the occuring base forms (segmental and intonational). In fact, a cell in line S-IV corresponds to the concatenation of the cells in lines II and III of a corresponding MGT, and it may even be possible to fill it in (semi-)automatically, given a corresponding MGT. In the case of MGTs, information in lines II and IV, and in lines III and V, respectively, may greatly overlap, depending on the language structure and theoretical conception. In the sample table, only the syllable break points differ from the presentation of the glossed unit (word) as a whole (lines II and III) and the individual base forms (morphs, lines IV and V).

Line VI contains cells with orthographical names of the individual base forms. The concatenation of these may differ from the orthographical representation of the whole glossed unit (given in the global line XII). For instance, in a given orthography clitics may not constitute separate orthographical words.

Lines VII an VIII account for categorial information. At least for some approaches, categorial information may be of two different kinds. In syntax, we have word categories such as "Verb" "Masculine Noun", contrasting with word form categories such as "First Person" or "Nominative". In morphology analogous types of categories may be needed. The former (the *lexical* categories that concern whole lexical units including all their form variants) are given in line VII, the latter (*form* categories) in line VIII. If one did not differentiate between these two types of categories, e.g. in favor of 'morphosyntactic features', only line VII would be used.

Line IX most closely resembles the glossing line in IGs; it represents the meaning of each base form. In the case of content words (in SGTs) or content stems (in MGTs), a lexical meaning will be indicated. Other base forms may carry a 'grammatical meaning' (e.g. derivational affixes – in morphology– or, in syntax, function words). Still another type of 'semantic effect' may be relevant in the case of inflexional affix-morphs or auxiliary words. Here, names of *syntactic* categories are given that can be assigned to a corresponding syntactic unit 'based on' the occurence of the relevant base form. The conception of details will vary among different frameworks. It is important, however, that both, line VII and the categories indicated in line IX, in morphological as much as in syntactic GTs, are relevant if the glossing format is to account for complications such as periphrastic forms or categorial membership of forms that cannot be directly linked to the presence of a specific morph. AG does not prescribe to resort to null-morphemes or similar devices which are not acceptable in several approaches.

Lines X to XIII are global lines, they do not contain cells but are related to the glossed unit as a whole. Line X gives constituent structure information in a format that dispenses with bracketings or similar devices and uses the numbers of line I instead. A similar strategy is used to represent grammatical relations that hold between constituents.

Line XII renders the glossed unit in an established orthography. As said above, this may differ from information given in line VI. Finally, line XIII renders the meaning of

I	num	1	2	3
II	seg	[di.	ʔʊn̩ybɐ.zɪç.tˡɪç̩n̩	pʁo.blee.mə]
III	int	L	H_f M_f L H H H	H H_r L_f
IV	plb	/ˌdii/	/ˈʔun.ˌʔyy.bə.r-ˌziX.t-ˌliX̩ə.n/	/pro.ˈblee.mə/
V	pli	L	H L L H H H	H $\mathbf{H}_r$ L
VI	orb	*die*	*unübersichtlichen*	*probleme*
VII	lct	DefArt	Adj	Sub Neut
VIII	fct	Nom Pl Unm_G Str	Unm_C Pl Unm_G Wk	Nom Pl Unm_D
IX	gls	Nom Pl Def	'involved'	'problem'
X	str	1,3:Nf 2:Nf **1**,2,**3**:NGr		
XI	rel	mod:2 1,3		
XII	ort	Die unübersichtlichen Probleme.		
XIII	par	E: The involved problems.		

Figure 1: A sample syntactic glossing table.

I	num	1	2	3	4	5
II	seg	/ʔun.	ʔyy.bə.r-	ziX.t-	liX̩	ə.n/
III	int	H	H,L L	H,L	H,L	L
IV	plb	/ˈʔun./	/ˈʔyy.bə.r/	/ˈziX.t/	/ˌliX./	/ə.n/
V	pli	H	H,L L	H,L	H,L	L
VI	orb	*un*	*über*	*sicht*	*lich*	*en*
VII	lct	$Pref_i$	$PrepSt_j$	$SubSt_k$	$SubsSt_l$/$AdjSt_m$	$AdjFlex_n$
VIII	fct	–	–	–	–	–
IX	gls	not	'over'	'view'	suitable-for	Unm_C Pl Unm_G Wk
X	str	1:Af 2:Stf 3:Stf 4:Af 5:Af 2,**3**:Stf **2,3**,4:Stf 1,**2,3,4**:Stf **1,2,3,4**,5:StGr				
XI	rel	m-mod:2 3 m-qual:4 2,3 3 m-mod:1 2,3,4 m-qual:5 1,2,3,4 1,2,3,4				
XII	ort	unübersichtlichen				
XIII	par	E: 'involved'				

Figure 2: A sample morphological glossing table.

the glossed unit – a paraphrase of the sentence in another language in the case of SGTs, or a lexical meaning, in the case of a MGT (in the case of content words, the content of a cell in line S-IX is identical to the content of line M-XIII of a corresponding glossing table).

This short characterisation of AG is only to give a general idea of the format. For more details, the reader is referred to the available presentation of AG (Lieb and Drude, 2000). We now turn to the technical implementation of Advanced Glossing.

5. Implementation in Shoebox

The Shoebox program has been used in order to implement Advanced Glossing on the computer. In the DOBES context, eventually, Shoebox will be replaced by the newly designed EUDICO tool, but currently Shoebox seems to be still indispensable for many documentation projects. Since it has been developed over many years, it shows a number of features which are useful for the documentation of languages. One point is that Shoebox has the ability to organise the content of several lines by means of columns, thus providing a rudimentary table structure as required by AG. In addition it offers the possibility to use a combination of several data bases, either textual or lexical, and the information in these databases may be cross-referenced.

In Shoebox each text corresponds to a *database*, that is, a collection of *records*. Every record holds one syntactic glossing. So, in a first step, a *database type* for SGTs has been set up, providing Shoebox's *fields*, i.e., establishing the data types that may occur in a record of a database of this type. Some of the fields in a record are to function as lines in glossing tables, others for storing entries in the

comment to the table, and some for housekeeping data.

While each text may be stored as a separate database, all morphological glossing tables are collected in one single database of a second type. Due to the extensive analogy in the conception of morphological and syntactic GTs, the database types for MGTs and SGTs are almost identical, too. Under exceptional circumstances, one may even use morphological fields in SGTs and vice versa.

A new aspect not yet accounted for in the original AG proposal is the possibility of interaction between GTs and lexical databases. After all, Shoebox has been designed mainly as a lexicographic tool, that is, for constructing lexical databases. There is a comprehensive set-up for the organisation of lexical data which allows for databases to be converted to a file in Rich-Text-Format with the Multi-Dictionary Formatter (MDF) and thus to produce appealing hardcopies. Other ways of printing Shoebox's databases are conceivable. In particular, it seems to be a promising alternative to use the BiBTEX and TEX tools which are well-known for providing flexible and high-quality typesetting of databases. Nevertheless, as these alternatives do not exist yet (they may well turn out to be superfluous with the new tools being developed at the MPI in Nijmegen), it seemed advisable to stick as closely as possible to the MDF setup that comes with Shoebox.

For conceptional reasons, a somewhat modified version of the Shoebox database type designed to be used with the MDF has been set up. Also, for technical reasons, lexical data is stored in three separate databases (which all use the same database type): one for affixes, one for simple words (and, simultaneously, for their stems), and one for complex words. The technical point in question is that the databases do not merely co-exist, they may also refer to each other, in particular, for two two tasks: (a) semi-automatic filling-in of information ("*Interlinearisation*"), and (b) "*Jumping*", that is, looking up a relevant record in another database.

Interlinearisation has the convenient side-effect of arranging data in several consecutive lines in columns. This provides a table structure as required by AG. When interlinearising sentences, different occurrences of one and the same word will refer repeatedly to a single morphological glossing (MG) for that word. Therefore, it turned out to be practical to store certain syntactic information relevant to a given word in the MG-record for that word. In the case of syncretism or other types of polysemy or poly-functionality of words, several MGTs are needed (independently of the hybrid character of the records).

So far as semi-automatic filling-in of information is concerned, data needed for the SGTs is searched for in the database for MGTs and, in order to parse and interlinearise words in the latter, information is looked up in the lexical databases for simple stems (simple words) and affixes. The solid arrows in figure 3 symbolise the search relations.

Second, appropriate jumping had to be set up. This means, when looking at a certain item in a (morphological or syntactic) GT, one would like to compare the relevant lexical entry or, in the case of a word in a SGT, the relevant MGT. The relevant dependencies are symbolised in figure 3 by dotted arrows. (More could have been and occasionally will be set up.)

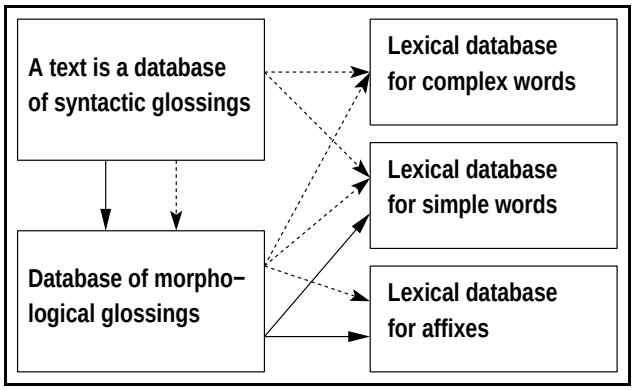

Figure 3: Databases and their relations in Shoebox.

After this general characterisation of the implementation of AG in Shoebox, some details of the work flow and some features of the setup will be given.

6. Work flow and implementation details

The general work flow of a text documentation in the Awetí Language Documentation Project (one of the DOBES projects) is as follows.

A text is recorded (usually in audio and video) and digitalised. First the Transcriber tool is employed to segment the text (only audio) in sentences ('time-linking'), and each sentence is transcribed orthographically. The transcription (with its time-linking) is converted into the Shoebox format by means of Econv, a conversion tool developed at the Max Planck Institute in Nijmegen.

In Shoebox, the translations are added with the help of native speakers (usually, there is first a word-wise translation to the national or contact language, which in this case is Portuguese, then a free translation into English). The result is a minimal documentation as agreed among the projects in the DOBES pilot phase. The transcription of a sentence is interpreted as its orthographic representation in a SGT (AG line S-XII, in Shoebox: a field `\SXII`, where "S" stands for "syntactic"). The free translation into English is AG line S-XIII.[5]

For a smaller sub-corpus, a more complete documentation is aimed at, as far as permitted by the current knowledge of Awetí. For this purpose, the result is interlinearised, and during this process, missing MG-records and lexical data base entries are created. If the relevant records in the database for MGs have been provided for and filled in correctly, interlinearisation leads to a SGT with several types of information which have been added semi-automatically: lines S-VI (orthographical words), S-VII (word categories), S-VIII (word form categories), S-IX (word glosses). Also, one more line `\lx` has been added. It contains names of citation forms that are used in order to 'jump to' (look up) the corresponding entries in the lexical databases.

If need for disambiguation arises, Shoebox will present the different possibilities from which the correct one may

[5]The entries in most cells in line `\SXI` are based on the word-for-word translations obtained with the help of the native speakers, with some complications for function words.

be chosen. The remaining information which is needed for an exhaustive documentation of the sentence according to the AG scheme has to be filled in by hand.

So, in order to document all linguistic aspects foreseen in AG, the remaining lines in the GT are added, some of which share the cell structure. This concerns the numbers in line S-I and phonological words (line S-IV). The phonetic lines S-II and S-III are by their very nature global lines, although in the AG proposal they share the cell structure. In Shoebox, they are not to be broken up into cells. The same holds for the lines S-X and S-XI (syntactic constituents and relations).

As AG allows for incomplete and partial successive filling-in, lines considered to be irrelevant for a specific purpose may be left empty, and the content of specific cells may be marked as uncertain if linguistic knowledge with regard to the language does not yet allow a complete description.

Comments on specific cells or lines may be added (for instance in the case of uncertainty). The Shoebox solution is to create additional fields which refer to lines in the GT, for instance, a field with a *marker* `\SIXc` which contains a comment referring to line S-IX. In the case of comments on different individual cells, one field for each such comment is created, the comment beginning with the number of the relevant column.[6]

A complete SGT in Shoebox is shown in figure 4. Most of the features mentioned above are illustrated in figure 4.

When interlinearising a SGT, Shoebox looks into the MGT-database and finds the matching entries (if already created) which justify to split up the orthographical words into orthographical representations of individual syntactic base forms. Usually, there will be a one-to-one correspondence, but observe the case of *jatãtsu* in the SGT (in field `\SXII`) which corresponds to two words in the above defined sense, *jatã* and *tsu*, in field `\SVI`. Right-clicking on a word in field `\SVI` will carry out a "jump" to the record for the corresponding MG.

After parsing the text and producing the parsed line, additional information for each phonological word is filled in semi-automatically by usage of the same relevant records in the MG-database. First, the field `\lx` is filled with the names of lexical words whose forms occur in the sentence. Right-clicking on such a name will cause a "jump" to the corresponding entry in a lexical database (either for simple or complex words).

Then, the part of speech, the relevant syntactic word form categories and a gloss are added accordingly, recurring to information stored together with the MGTs (these records are thus, for technical reasons, hybrid with respect to the strict separation of morphological and syntactic information). If a word was a form of different lexical words at the same time (in the case of polysemy or homophony), or if it could be assigned different syntactic categories (e.g. due to syncretism), several entries in the MG database would be needed, and Shoebox would, again, ask for disambiguation when filling in.

[6]Note that in Shoebox several fields of the same data type can be repeated as often as required.

Interlinearisation can be done also in a record for a MG. In this case, we start from the orthographical representation of the whole word (field `\MXII`). The process is almost analogous to the syntactic case, the word is "parsed" (i.e., split up) into morphs. In order to do this semi-automatically, Shoebox accesses information stored in the lexical databases for affixes and simple words. This concerns the allomorph of a stem or affix occuring in the word (field `\MVI`), to which the canonical name of the morphological lexical unit in question is added in an additional `\lx` field (not yet foreseen in the current version of AG).

The information in fields `\MVII`, `\MVIII` and `\MIX` is also filled in by recurring to data in the corresponding lexical records in the databases for simple stems (which is the same as for simple words) and for affixes. Again, 'jumping' from a name of a morpheme to the relevant entry in these databases is possible.

The three Shoebox database types and a sample Shoebox-'project' with test-files are available to anybody and any language documentation project that would like to test and apply Advanced Glossing and its Shoebox setup. Please look at the MPI-web-site where you also can find AG (Lieb and Drude, 2000).

7. Acknowledgements

The Awetí Language Documentation Project is included in the research program Dokumentation Bedrohter Sprachen (Documentation of Endangered Languages, DOBES) funded by the Volkswagenstiftung. I thank H.-H. Lieb for fruitful discussions that contributed to this paper. H.-H. Lieb is also the principal developer of Advanced Glossing, while the Shoebox implementation is my own work. Thanks also to P. Wittenburg and the TIDEL group at the Max Planck Institute in Nijmegen for technical support and general discussion. Sabine Reiter kindly helped to improve my English.

8. References

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

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

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

Hans-Heinrich Lieb and Sebastian Drude. 2000. Advanced glossing: A language documentation format (1st version). http://www.mpi.nl/DOBES/applicants/Advanced-Glossing1.pdf.

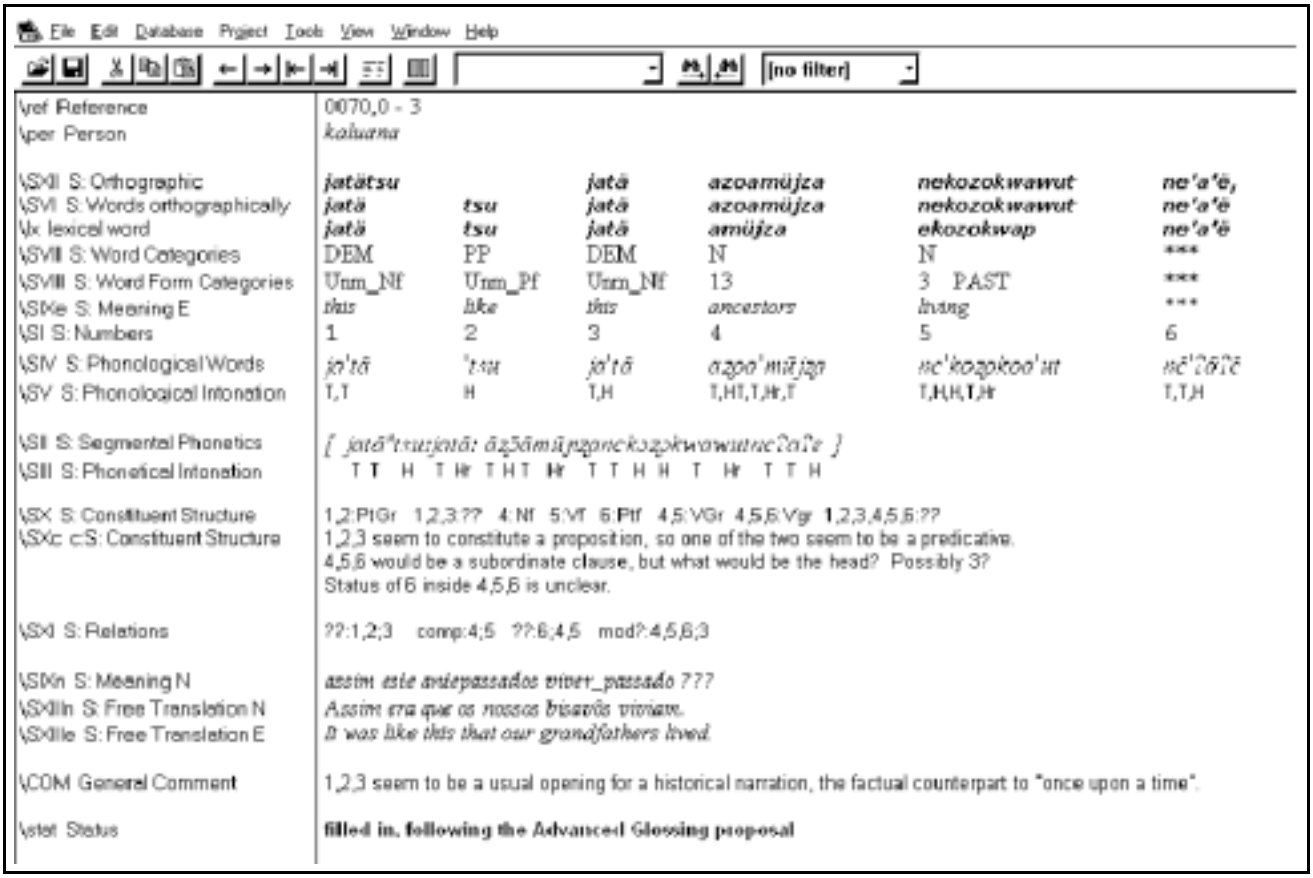

Figure 4: A sample syntactic glossing record in Shoebox.