On linguistic uses of language documentations

Nikolaus P. Himmelmann
Universität zu Köln

1. Introduction[1]

At the time of writing (mid to end 2009), first examples of modern multifunctional language documentations, in the sense of Himmelmann (2006), are becoming available (see in particular the DoBeS archive at www.dobes.nl). Typically, these documentations consist of a (largish) corpus of recordings of different kinds of communicative events (narratives, conversations, speeches, ceremonial exchanges, explanations of tools and cultural events, etc.), accompanied by introductory materials (remarks on the setting, grammar and orthography being used, corpus structure, documentation team and circumstances). Furthermore, many of the recordings are annotated to various degrees of detail. Typical annotation layers are a transcription in a practical (phonemic) orthography and a free translation into a national or international language (e.g. English, Spanish). Further annotation layers may include interlinear glosses, grammatical or ethnographic commentary, phonetic transcription, syntactic categories and structures, etc.

More often than not, the details and the consistency of the annotations for the recordings will vary widely across a given corpus and, of course, even more so among the different corpora found in an archive for language documentations such as the DoBeS archive. This heterogeneity implies that language documentations cannot be 'harvested' in a straightforward way by linguistic users interested in crosslinguistic generalisations and linguistic theorizing. Thus, a not uncommon reaction from these users when accessing language documentations with the expectation of finding a dearth of data relevant for their theoretical projects may go something like this: 'Well, obviously a lot of possibly interesting stuff, but not really useful for my purposes.'[2]

There are a number of reasons for this perceived lack of usefulness, including technical problems of online access, unhelpful data structures, lacking

1. Many thanks to Katharina Haude, Gabriele Schwiertz and Frank Seifart for most helpful discussion and suggestions.

2. Reactions from other disciplines may be similar, but the present squib will be limited to linguistics as the discipline the author is most familiar with.

guidance for accessing a given corpus, and, last but not least, wrong expectations with regard to what one can realistically expect to find in a documentation at this point in time. The present squib will be concerned exclusively with this last point and will attempt to show why these wrong expectations arise, and to contribute to a more realistic user approach to language documentations.

2. Whence wrong expectations

Wrong expectations regarding language documentations primarily arise because over the course of the 20th century linguists have become used to working solely with a single higher order data type, which is called *structural data* here and which is known as "facts" in generative jargon. With 'higher order' I am referring to the distinction of three major processing stages for linguistic data established in Himmelmann (in prep.) and summarized in Table 1.

Table 1. *Different types of data based on observable linguistic behaviour*

Raw data	*recording (audio/video)*
Methods for deriving	e.g. transcription, translation
Primary data	*transcript with translation*
Methods for deriving	distributional and frequency analysis, tagging, crosslinguistic comparison, "interpretation"
Structural data	*descriptive statement, dictionary entry, interlinear glosses, frequency data, typological database, treebank, implicational universal*

The example used to illustrate the three stages in Table 1 makes use of a very commonly used data type based on observable linguistic behaviour, i.e. an audio or video recording of a communicative event. The raw data processing stage pertains to the recording on an analogue or digital medium and all further physical processing, such as cutting, copying or converting the recording in between storage media (e.g. from audio tape to wav-file). Part of this stage is also the creation of a metadata set accompanying the recording, detailing its who-when-where-how and providing a brief indication of its content.

A first step in making recordings useable for further linguistic analysis and theorizing consists in preparing a transcription of the recording and a translation in case the recorded event involves a linguistic variety not well known to the researchers (the default case in current language documentation). Both transcription and translation require some amount of interpretation and are greatly facilitated by the direct input of native speakers. The processing stage reached in this way can be called *primary data* in the sense that this is the type of data

most commonly used for further linguistic analysis. That is, while the raw data (the original recording) may occasionally be consulted again, the transcript and translation often provide the only basis for further annotation and analysis.[3]

The further analysis of primary data may then result in *structural data*, a term used here to cover a fairly broad and heterogeneous set of data types of various degrees of abstractness.[4] Thus, structural data include descriptive statements of the type *Language X has an accusative suffix –mu* and dictionary entries giving reasonably precise explications of the forms and functions of lexical items, which in turn are the prerequisite for a consistent and high-quality interlinear glossing of the texts compiled in the corpus. They also include frequency data (frequency of phonemes, words, affixes, constructions, collocations in a given (sub-)corpus) and syntactic annotations (segments of transcripts annotated for parts of speech, constituency, etc.). On a still higher level of abstraction, generalizing across descriptive statements for a sample of languages, there are the well-known implicational universals of the Greenbergian type (*If a language has a structural feature X, it tends to have – with overwhelmingly greater than chance frequency – also feature Y*) and other kinds of generalisations with a similar scope. These would also come under the broad rubric of structural data as conceived of here.

While most linguists probably would not expect to find the latter type of higher order generalizations in language documentations, they minimally appear to expect structural data on the lowest level of abstraction. This does not have to be a fully worked-out descriptive grammar (or dictionary, for that matter), but at least a level of grammatical and lexical annotation which allows one to derive, or check, descriptive or frequency statements more or less directly. A useful grammatical annotation in this sense would include, for example, a reasonably consistent indication of word and morpheme boundaries and a morpheme-by-morpheme interlinear translation in addition to the free translation given for larger units (clauses or intonation units). Such glossing is essential for frequency counts and for deriving or checking descriptive generalizations by, for example, being able to get a listing of all examples of the accusative suffix *–mu* using a general search option.

Note that according to the systematization of data types summarized in Table 1, transcription and free translation, on the one hand, and interlinear glossing and other types of lexico-grammatical annotation belong to two dif-

3. Of course there are also several kinds of analytical enterprises (e.g. phonetic or gestural analysis) where the recording is regularly accessed as well.
4. It is certainly possible and perhaps also useful to further systematize the very heterogeneous set of statements labeled *structural data* here, but this is not a topic that can be pursued further in this squib.

ferent stages, primary and structural, respectively. These two stages are distinguished by various conceptual and methodological differences which will not be further discussed here (cp. Himmelmann in prep. for details). The distinction is of relevance here, because, as mentioned in the introduction, annotation levels found in language documentations – at least at this point in time – vary widely and will often not meet even minimal expectations with regard to *structural* data. However, even these minimal expectations are misguided in that it is NOT a core function of language documentations to provide ready-to-use structural data. The main concern of language documentations is raw and primary data and their core functions include the task of making accessible the raw and primary data necessary for deriving structural linguistic data. In accessing language documentations, one cannot presuppose that these derivations have already been carried out but instead users should be prepared to do them on their own.

Consequently, language documentations make it necessary to change basic attitudes and procedures in the field regarding its empirical basis: Contrary to attitudes that have been dominating the field for much of the second half of the 20th century, generating valid and interesting structural data (aka "facts") is not an easy and straightforward process. Note that this insight is not confined to the community of linguists working on little-known languages. In fact, methods for obtaining valid structural data have always been a major concern in the subdisciplines of sociolinguistics (in particular in the work of Labov (cp., for example, Labov 1975, 1994) and in psycholinguistics. More recently, the need for such methods have also been acknowledged in more formal linguistic quarters, with Schütze (1996) providing the starting point for much current work on data structures (cp., for example, Keller 2000, Penke & Rosenbach 2004, Kepser & Reis 2005, Borsley 2005, Schlesewsky 2009 and other contributions to the "Forum" in vol 28.1 of *Zeitschrift für Sprachwissenschaft* (2009)).

With regard to documentations of endangered varieties, the issue of useability for specific disciplinary interests (such as a typologist looking for structural data illustrating the structural diversity of natural languages) is further complicated by the fact that the raw and primary data compiled in them should be multifunctional, i.e. they should also be of use for other disciplines and outside academia proper. As different users have different needs and expectations, there is thus a high potential and probability for conflicts of interest as to which kinds of annotations should be included to make the data useful for as many different user groups as possible. In practice, this more often than not will lead to compromise solutions which put additional burdens on all user groups. At the risk of overstating the issue a bit and creating misunderstandings, one could thus perhaps say that language documentations of endangered varieties do not

provide *instantaneous use* options for any possible user group. That is, while the documentation may in fact contain the data relevant for the purpose at hand, it is not immediately available/accessible but needs the user's own work and input to get to it.

3. What realistically to expect: Examples from the Waima'a corpus

The introductory page for the corpus documenting Waima'a (an endangered Austronesian language from East Timor, cp. Belo et al. 2002-2006) provides the following information for outside users of the corpus (http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI77915%23):

> The corpus of materials deposited here contains a highly diverse set of naturally occurring communicative events, including rain invocation chants and mourning songs, taking cock fighting bets, everyday chatting while peeling peanuts, political discussion and, of course, folktales and personal and historical narratives. In addition, it contains a fair number of speech events elicited with the help of prompts such as the PEAR FILM, the FROG STORY, the SPACE GAMES and other prompting material developed at the MPI Nijmegen. While consisting mainly of Waima'a recordings, it also contains a number of segments in the local variant of Tetum, which is quite regularly used in, for example, political discussions.
>
> The quality of the annotation varies considerably across the annotated sessions. On the one extreme, there are a few sessions which have only been worked on by one or two team members. These may involve major inconsistencies in intonation-based segmentation, glossing mistakes, and translations which are difficult to parse and lack in coherence and cohesion across unit boundaries. On the other extreme, there are a few sessions which have been worked on repeatedly by three or four different team members in an attempt to weed out most inconsistencies and to provide for a coherent translation accessible to readers not familiar with language, culture and setting. Each annotation file contains information on the processing stage in its first record.
>
> The processing of the close to 4000 lexical entries in the toolbox lexicon file are at similarly divergent stages, some having been gone over repeatedly with two or more native speakers, others having been created on the spot in order to provide a quick gloss for a word encountered in a transcript with no attempt yet made to elucidate the full range of its meanings.
>
> In using this corpus, it should also be kept in mind that:
>
> 1) next to nothing was known about Waima'a before the documentation project started (a few descriptive materials have appeared since, see the separate document SETTINGSKETCH).
>
> 2) only East Timorese, none with any previous training in linguistics, worked full time or for substantial regular part-time periods on the documentation. The

professional linguists involved in the documentation supervised the work of the East Timorese team members and did the basic analysis needed to make a documentation possible (most importantly: orthography development), but so far had only limited time for actually working on annotations and analysis.

Let us try to spell out in a bit more detail what the statement "The quality of the annotation varies considerably across the annotated sessions" actually means. As probably many other language documentations, the Waima'a corpus includes segments of texts that have been annotated and edited to a degree that comes close to the kinds of annotated texts often included in the appendix to descriptive grammars, as in the following example:[5]

\ref Marcos_bubomara 009
\tx au na'i re'o dou
\gle 1sPOSS abdomen thirsty very
\tle I am very thirsty.

Evidently, this kind of annotation is only possible when the grammar of the language has been investigated in sufficient detail, which was – and still is in many respects – not the case for Waima'a. However, the segment shown in this example only involves relatively straightforward grammatical and lexical items for which the main functions and uses are not in doubt. Hence it comes fairly close to the standard found in the appendices of descriptive grammars. But note that even in the fairly short text where this segment comes from, and which is easy to understand and translate and which has repeatedly been checked by team members with linguistic training, there still are a few segments and elements which remain unclear, because the grammar has not yet been analysed in sufficient detail. Thus, for example, the following segment from the same text includes the postverbal clitic particle *lo* the function of which is not at all clear. Possibly, it has something to do with aspectual marking, hence the preliminary gloss ASP(ECT):

\ref Marcos_bubomara 013
\tx bu bubo mara ehe **lo**
\gle HON buttocks sharp say ASP
\tle the Pointed Back said:

5. Examples are copied directly from the toolbox files compiled in the corpus. Each line is preceded by an identifier: \ref = reference line, \tx = transcript of vernacular text in standard orthography, \gle = English interlinear gloss, \tle = free English translation, \tlt = free Tetum translation, \tlm = free Malay translation.

Obviously, anyone who wanted to make generalizations involving this particle *lo* would have to do a more detailed analysis before actually being able to do so. Which in this case is reasonably straightforward in so far as it will be easy to locate many further examples of this particle by doing a combined search on *lo* and ASP.[6]

Note that even in those instances where a reasonably comprehensive descriptive analysis of the grammatical structure of the language being documented has already been achieved and the meaning of lexical items has been investigated in sufficient detail, it does not necessarily follow that all texts/communicative events included in a documentation of the variety at hand will be completely annotated in the way exemplified by segment MARCOS_BUBOMARA009, because annotating and editing recordings to this degree of detail is a very labour-intensive enterprise. Consequently, there is a trade-off between the amount of (raw and primary) data to be found in a given language documentation and the degree and quality of its annotation: the more specimens are annotated in detail the smaller the overall size of the documentation will tend to be. And conversely, the more specimens are included, the sparser the annotation will tend to be.

In fact, documentations more often than not will include specimens without any linguistic annotation, i.e. a recording with metadata detailing the circumstances of recording and some indication of its content, but no transcription, translation or interlinear glossing. There is a great variety of such recordings, for example, ones with very little spoken content or ones where the linguistic interaction makes use of a better known contact language, but also ones where the recording involves (almost) exclusive use of the endangered variety and thus is not directly accessible to any outsider.

One reason to include this last type of recordings despite the fact that it is probably of no, or very little, direct practical use for anyone not mastering the language is the speculation that it will be possible in the future to transcribe and translate such a recording even without direct input by a native speakers on the basis of what can be learned about the language on the basis of better annotated specimens in the collection and any analyses already published for the variety in question. Obviously, this would involve traditional philological techniques as they have been successfully employed in reconstructing and interpreting written texts of times long gone that are only available in corrupt manuscripts or inscriptions.

Returning to the annotation standards and practices found in the Waima'a corpus, this corpus also contains recordings which are only accompanied by a

6. The grammatical sketch accompanying the corpus of recordings explains the glosses and indicates the extent to which the analysis of the formative in question has already been carried out.

transcription and a free translation, i.e. only by primary data, which constitute the minimal annotation standard in the DoBeS program,[7] as in this segment from another recording:

```
\ref Sabino_carpent 12
\tx  aku de loo harewai see
\tlt ha'u la halo natar
\tlm saya tidak kerja sawa
```

Apart from the general annotation level, this example also illustrates another, and perhaps not untypical, feature of language documentations, which makes for a further obstacle to their direct and immediate use by linguists: Free translations are not necessarily in English, but may involve other languages which are reasonably well-known to a sizable group of non-specialists. In this example, the free translations are in Tetum (\tlt), one of the two official languages of East Timor, and in Malay (\tlm). In English, the example means "I do not have (lit. make) a paddy field" (from a story where a carpenter lists his means of livelihood, which do not include a paddy field). As the Waima'a documentation also includes a brief grammatical sketch and a lexical database, it is not difficult, but admittedly somewhat cumbersome, to come up with the appropriate interlinear morpheme glossing and English translation, as in this edited version:

```
\ref Sabino_carpent 12
\tx    aku  de   loo  harewai      see
\gle   1s   NEG  make paddy field  one
\tlt ha'u la halo natar
\tlm saya tidak kerja sawa
\tle I do not have a paddy field
```

The practice of including free translations in Malay *and* Tetum, despite the fact that the latter is not widely known by non-specialists, is due to practical and political considerations. In the Waima'a project, the corpus has been mostly compiled and processed by Maurício Belo, a native speaker of Waima'a who knows both Tetum and Malay well, but no other national or international languages. Hence, his translations had to be in either language and, as it turned out that in difficult cases it is helpful to be able to access both translations, he

7. Cp. The *Information for Applicants 67* of the DobeS program of the Vokswagen Foundation, version December 2009, p3 "[the] minimal annotation standard [consists of] a transcription, a translation, and explanatory comments on the speech situation of the recording and its content".

was asked to provide free translations in both languages. Note that Tetum is structurally quite similar to Waima'a. Hence, the Tetum translation is quite often a word-by-word translation, while the Malay translations tend to be freer, focusing on the meaning. Metaphoric expressions, for example, are usually rendered literally in Tetum, but in terms of their meaning in Malay.

Politically, from the point of view of the speech community and the government, it would not have been suitable to provide translations in Malay only, as this was the language of the oppressor of 25 years. In fact, both in glossing and translation an attempt was made to add also the second official language of East Timor, Portuguese, at least for some texts. This was done with the possibility in mind that such texts, with further editing and correction, could then also be used for Portuguese classes in schools.

Finally, the Waima'a corpus contains many specimens which, to quote again from the cautionary notes on the introductory page of this corpus reproduced in full above, "have only been worked on by one or two team members. These may involve major inconsistencies in intonation-based segmentation, glossing mistakes, and translations which are difficult to parse and lack in coherence and cohesion across unit boundaries." This is illustrated by the following segment from a recounting of the pear story, reproduced from a version which had not been checked in detail by a team member with linguistic training:[8]

```
\ref pear_santina 091
\tx  inke ne lara  ne  lara  ne
\gle where 3s obtain FOC obtain FOC
\tle So he picked one and

\ref pear_santina 092
\tx  huo  teme oro  bote   oli se  ke
\gle lift all  just basket big one DEM
\tle he just took a whole big basket

\ref pear_santina 093
\tx  tii   thau  la
\gle arrive place LOC
\tle placed it on
```

8. Tetum and Malay glosses and translations have been omitted, spelling mistakes and punctuation of the English translation are unchanged. The English glosses and translation, which were prepared by a Timorese team member not familiar with Waima'a, are based on the Malay ones, which were prepared by a native speaker of Waima'a. The version quoted from here was part of the uploaded corpus in between 2006 and 2009, but was replaced by the further amended version quoted from below in late 2009.

\ref pear_santina 094
\tx thau bete la
\gle place nearly LOC
\tle placed it near

\ref pear_santina 095
\tx ah:::

\ref pear_santina 096
\tx biskalita nini ke ne buni n'ai lo laku-tuo ana ita de buni ne
\gle bicycle POSS DEM 3s see go.up ASP old man DIM DIST EMPH see 3s
\tle When the boy went back to his bike, the old man didn't see him.

\ref pear_santina 097
\tx n'ai hile biskalita ne thau hite la
\gle go up again bicycle 3s place PTL LOC
\tle Upon riding his bike he plced the fruits on top of

\ref pear_santina 098
\tx biskalita oo ke
\gle bicycle above DEM
\tle the pushbike.

\ref pear_santina 099
\tx ke ne uko ruo
\gle DEM 3s run and
\tle Then he took it away.

As already mentioned above, the grammar of Waima'a has not yet been analysed in sufficient detail, so quite a few aspects of this segment cannot be correctly analysed and glossed at this point. This pertains in particular to most function words such as *ke* DEM(ONSTRATIVE), *ana* DIM(INUTIVE), or *hite* P(AR)T(IC)LE. As explained in the grammar sketch accompanying the corpus of recordings, glosses for these items are very preliminary and have mostly been added to make it easier to locate further attestations of the same formative. Reasonably clear, however, is the fact that *la* (glossed LOC(ATIVE)) is, among other things, a kind of very general locative preposition, as well illustrated by its uses in this segment.

A common difficulty in glossing function words is illustrated by the form *ne* which, on the one hand, is clearly a third singular personal pronoun (glossed 3s), but which also seems to have some clause structuring function glossed as FOC(US), although it is quite clear that not all its uses mark focus in a narrow

sense. As the first line (091) shows, it is not always easy to decide which of these two quite different functions are involved. Thus, *ne lara ne* could also be analysed as involving two pronouns "he took it/him/her". In fact, as a close checking of the original recording reveals, this segment actually does not involve a single coherent construction, but rather a repeated attempt at rephrasing a state of affairs (as shown in the amended version further below where all three *ne* in this segment are analysed as (subject) pronouns).

Other problems in interpreting this segment relate to slips in glossing and the well-known difficulties of providing useful translations for segments of spoken language, which can only be done when paying close attention to the actual recording. In addition to line091, which was just discussed, these problems are well illustrated by line 096, repeated here for convenience:

\ref pear_santina 096
\tx biskalita nini ke ne buni n'ai lo laku-tuo ana ita de buni ne
\gle bicycle POSS DEM 3s see go.up ASP old man DIM DIST EMPH see 3s
\tle When the boy went back to his bike, the old man didn't see him.

The particle *de* toward the end of this unit is glossed as EMPH(ATIC) but context and translation make it clear that it should rather have been glossed as NEG(ATIVE). The slip leading to this mistake is quite common when programs such as TOOLBOX are used for semi-automatic interlinear glossing. Whenever the program happens upon *de*, it offers both EMPH and NEG as a possible gloss. Wrong choices easily occur and usually are only noticed when the glossing is carefully checked again.

A perhaps more serious problem pertains to the first part of the unit for which the glosses and the translation do not match very well. Clearly, there are no words in the vernacular unit which would correspond to *boy* and *went back* in the translation. Conversely, the vernacular line contains the word *bun(i)* 'see' twice, but the translation only once. Furthermore, it is unclear how the Waima'a word *n'ai* 'go.up' is rendered in the translation. There may be two very different sources for these divergences. On the one hand, the translation may be rendering a constructional meaning that is not obvious from the literal translation of the Waima'a morphemes making up this unit, taking the liberty of making explicit elements which are only implicit in the Waima'a original. Note that subject and object NPs are freely omissible in Waima'a, but are usually required in producing grammatical English sentences.

On the other hand, the translation may be misleading in that it does not properly take into account the previous context and the fact that this segment represents spontaneous speech which does not follow the model of a written,

edited narrative. And this, in fact, is the case, as a closer inspection of the previous units makes clear. In these units it is recounted that the boy takes a basket of pears and puts it onto something, the speaker obviously having troubles in expressing exactly where he puts it, as she repeats the phrase *thau la* 'place on' twice (in 093 and 094), then hesitates (095) before she finally names *biskalita nini ke* 'his bicycle' at the beginning of 096. Thus, the initial words of this unit belong syntactically and semantically to previous units, although prosodically the preceding boundaries are heavier than the one separating *biskalita nini ke* from the remainder of 096.

Consequently, *ne bun(i) n'ai lo* introduces the beginning of a new (subevent) and the literal meaning 'he looks up' makes good sense in the context at hand: The boy stealing the pears is looking up towards the old man to check whether he is watching him. The idea of 'checking' is nowhere explicitly mentioned in the Waima'a original, so this remains an inference which in the following emulated translation is put into parentheses: 'looks up to the old man (to make sure that he) isn't watching'.

Finally, the final *ne* in this unit belongs not only syntactically but also prosodically to the following unit, as an inspection of the audio files makes clear, where one also hears a final *ko* of unclear meaning and function.

There are many further smaller and larger infelicities in the transcription, glossing and translation of this segment which will not be commented on in detail. The interested reader can easily find them by comparing the version above with the following, emulated one. Note that many of the changes introduced here are based on checking the original recording for prosodic cues as to the proper segmentation, and hence the interpretation, of the units contained in this segment.

```
\ref pear_santina 091
\tx  inke  ne: lara ne lara (0.1) ne:
\gle which 3s  take 3s take      3s
\tle So he: takes he takes, he:

\ref pear_santina 092
\tx  huo  teme oro  bote   ol  see ke
\gle lift all  just basket big one DEM
\tle lifts a whole big basket,

\ref pear_santina 093
\tx  tii    thau  la
\gle arrive place LOC
\tle places (it) on ahm
```



```
\ref pear_santina 094
\tx  thau  bete   la
\gle place nearly LOC
\tle places (it) close to ahm

\ref pear_santina 095
\tx ah:::

\ref pear_santina 096
\tx  biskalita nini ke  ne bun n'ai  lo  laktuo  ana ita  de  buni <ko>
\gle bicycle POSS DEM 3s see go.up ASP old man DIM DIST NEG see
\tle to his bike, looks up to the old man (to make sure that he) isn't
      watching,

\ref pear_santina 097
\tx  ne (0.3) n'ai  hile  biskalta ne thau  hite la
\gle 3s       go up again bicycle  3s place PTL  LOC
\tle he: picks up his bike, he places (the basket) on

\ref pear_santina 098
\tx  biskalita oo    ke
\gle bicycle   above DEM
\tle the bike,

\ref pear_santina 099
\tx  ke  ne  uko ruo
\gle DEM 3s  run with
\tle then he runs off with (it).
```

The need to check the original recording in order to be able to properly interpret a given segment as a prerequisite for using it as a datum for a descriptive or typological analysis (e.g. of clause structure or serial verb constructions in Waima'a) clearly is rather cumbersome and not something linguists used to work with descriptive grammars are inclined to do. It is certainly not the intention of the present squib to argue that this is the best one can hope for in language documentation. Quite the opposite: Ideally, of course, all recordings compiled in a language documentation should be properly glossed and annotated according to the standards found in descriptive grammars.

However, it is not realistic to expect that this can be done in the time frames for which documentation projects may receive funding (currently, in between 2-4 years). During such a project, there is a constant need to decide between whether to spend more time on getting more recordings with (often only rough) transcriptions and translations or on properly analysing and annotating the ma-

terials already gathered. Decisions in this regard will always depend on the specific circumstances of a project, and no general rule or advice can be formulated here.

Realistically speaking, then, the emulation of the quality of linguistic annotations found in the first upload of a (largish) documentation corpus is a matter of many years, if not decades. Chances for actually achieving higher consistency and greater comprehensiveness in linguistic annotation significantly increase the more hands join in this task. This means that one needs to extend the group of annotators well beyond the team who has originally compiled the corpus. Currently, there are no well established procedures and codes of conduct for such an enterprise. But it is clearly one of the most important directions that need to be further explored in documentary linguistics.

Note that recommending that language documentations be made available only when a consistently high level of linguistic annotation has been achieved is not a viable alternative. Among the many reasons against such a recommendation, the one perhaps most pertinent here is that this would significantly raise the chances that a documentation is *never* made available.

4. Conclusion

Clearly, the lack of consistency in annotation practices in a language documentation is a major obstacle to its immediate useability for linguists interested in typological generalisations and linguistic theorizing. But this does not mean that the whole enterprise is misguided. Rather, two things are needed now: First, preconceptions in the field as to what kinds of data can be expected and usefully exploited have to change so as to include the kind of raw and primary data characteristic of language documentations, thus leading to realistic expectations of prospective users of language documentations. Second, we need efficient ways for emulating annotation levels in language documentations, both in terms of accepted and respected academic practices and in the way of automatized tools helping in the process.

Abbreviations used in glossing

3s	Third person singular	FOC	focus
ASP	aspect	HON	honorific marker
DEI	deictic	LOC	locative
DEM	demonstrative	NEG	negation
DIM	diminutive	POSS	possessive
DIST	distal	PTL	particle
EMPH	emphatic		

References

Belo, Maurício C.A, John Bowden, John Hajek, Nikolaus P. Himmelman & Alexandre V. Tilman, 2002-2006, *DoBeS Waima'a Documentation*, DoBeS Archive MPI Nijmegen, http://www.mpi.nl/DOBES/

Borsley, Robert D. (ed.), 2005, *Data in Theoretical Linguistics*, Special issue of *Lingua* 115: 1475-1665

Himmelmann, Nikolaus P., 2006, "Language documentation: What is it and what is it good for?", in: J. Gippert, N. P. Himmelmann & U. Mosel (eds), *Essentials of language documentation*, Berlin: Mouton de Gruyter, 1-30

Himmelmann, Nikolaus P., in prep, "Linguistic data types and the interface between language documentation and description", revised version of plenary lecture at *1. International Conference on Language Documentation and Conservation*, Honolulu, March 2009

Keller, Frank, 2000, *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*, PhD Thesis, University of Edinburgh

Kepser, Stephan & Marga Reis (eds), 2005, *Evidence in Linguistics*, Berlin: Mouton de Gruyter

Labov, William, 1975, *What is a Linguistic Fact?*, Lisse: de Ridder

Labov, William, 1994, *Principles of Linguistic Change*, vol. 1, Oxford: Blackwell

Penke, Martina & Anette Rosenbach, 2004, "What counts as linguistic evidence", *Studies in Language* 28:480-526

Schlesewsky, Matthias 2009. „Linguistische Daten aus experimentellen Umgebungen: Eine multiexperimentelle und multimodale Perspektive." *Zeitschrift für Sprachwissenschaft* 28: 169–178

Schütze, Carson T., 1996, *The empirical base of linguistics*, Chicago: The University of Chicago Press.

SELECTED PAPERS

From the

International Conference on Language Documentation and Tradition

with a special interest in the Kalasha of the Hindu Kush valleys, Himalayas

7 – 9 November 2008

Edited by
Carol Everhard & Elizabeth Mela-Athanasopoulou

School of English, Department of Theoretical & Applied Linguistics
Aristotle University of Thessaloniki, Greece

Thessaloniki 2011

ISBN 978-960-243-688-2

Layout - Printing
UNIVERSITY STUDIO PRESS S.A.
Publishers of Academic Books and Journals

32, Armenopoulou str., 546 35 Thessaloniki, GREECE
Tel. +30 2310 208731, + 30 2310 209837, Fax +30 2310 216647
E-mail: info@universitystudiopress.gr
www.universitystudiopress.gr

STOA TOU BIBLIOU – 5, Pesmazoglou str., 105 64 Athens, GREECE
Tel. & Fax +30 210 3211097

Table of Contents

Preface 7

Acknowledgements 9

Part I. Plenary papers

Elena Bashir
Kalasha: Past, present, and possible futures 13

Nikolaus Himmelmann
On linguistic uses of language documentation 37

Jan Heegård Petersen
Prescriptivity in fieldwork data – examples from Kalasha 53

Part II. Linguistic papers

Anvita Abbi
Documenting Great Andamanese: Challenges and Solutions for a Dying Language 65

Sergio Baldi
A Brief Sketch of Arabic Influence on Dagbani 75

Greg Cooper
History is being written: Documenting, revitalizing and developing the Kalasha language 85

Chloé Darmon
Collecting data in Xamtanga: A case of two-way transfer inside the Ethiopian linguistic area 89

Pierpaolo di Carlo
Two clues of a former hindu Kush linguistic area 101

Ekaterina Gruzdeva
Archival data and modern language documentation 115

V. Koui, I. Mouhika, G. Migdalia, Th. Papadopoulou
Documenting the Dialect of Smyrna – Comparison with Standard Greek 129