Essentials of Language Documentation

edited by

Jost Gippert
Nikolaus P. Himmelmann
Ulrike Mosel

Mouton de Gruyter
Berlin · New York

Chapter 1

Language documentation: What is it and what is it good for?

Nikolaus P. Himmelmann

Introduction

This chapter defines language documentation as a field of linguistic inquiry and practice in its own right which is primarily concerned with the compilation and preservation of linguistic primary data and interfaces between primary data and various types of analyses based on these data. Furthermore, it argues (in Section 2) that while language endangerment is a major reason for getting involved in language documentation, it is not the only one. Language documentations strengthen the empirical foundations of those branches of linguistics and related disciplines which heavily draw on data of little-known speech communities (e.g. linguistic typology, cognitive anthropology, etc.) in that they significantly improve accountability (verifiability) and economizing research resources.

The primary data which constitute the core of a language documentation include audio or video recordings of a communicative event (a narrative, a conversation, etc.), but also the notes taken in an elicitation session, or a genealogy written down by a literate native speaker. These primary data are compiled in a structured corpus and have to be made accessible by various types of annotations and commentary, here summarily referred to as the "apparatus". Sections 3 and 4 provide further discussion of the components and structure of language documentations. Section 5 concludes with a preview of the remaining chapters of this book.

1. What is a language documentation?

An initial, preliminary answer to this question is: **a language documentation is a lasting, multipurpose record of a language**. This answer, of course, is not quite satisfactory since it immediately raises the question of what we mean by "lasting", "multipurpose" and "record of a language". In the following, these constituents of the definition are taken up in reverse order, beginning with "record of a language".

At first sight, a further definition of "record of a language" may look like a bigger a problem than it actually is since it involves the highly complex and controversial issue of defining "a language". The main problem with defining "a language" consists in the fact that the word *language* refers to a number of different, though interrelated phenomena. The problems in defining it vary considerably, depending on which phenomenon is focused upon. That is, different problems surface when the task is to define *language* as opposed to *dialect*, or *language* as a field of scientific enquiry, or *language* as a cognitive faculty of humans, and so on. Unless we want to postpone working on language documentations until the probably never arriving day when all the conceptual problems of defining *language* in all of its different senses are resolved and a theoretically well-balanced delimitation of "a language" for the purposes of language documentations is possible, we need a pragmatic approach in dealing with this problem.

The basic tenet of such a pragmatic approach is implied by the qualifiers *multipurpose* and *lasting* in the definition above: The net should be cast as widely as possible. That is, a language documentation should strive to include as many and as varied records as practically feasible, covering all aspects of the set of interrelated phenomena commonly called *a language*. Ideally, then, a language documentation would cover all registers and varieties, social or local; it would contain evidence for language as a social practice as well as a cognitive faculty; it would include specimens of spoken and written language; and so on.

A language documentation broadly conceived along these lines could serve a large variety of different uses in, for example, language planning decisions, preparing educational materials, or analyzing a set of problems in syntactic theory. Users of such a multipurpose documentation would include the speech community itself, national and international agencies concerned with education and language planning, as well as researchers in various disciplines (linguistics, anthropology, oral history, etc.). In fact, the qualifier *lasting* adds a long-term perspective which goes beyond current issues and concerns. The goal is not a short-term record for a specific purpose or interest group, but a record for generations and user groups whose identity is still unknown and who may want to explore questions not yet raised at the time when the language documentation was compiled.

Obviously, this pragmatic explication of "lasting, multipurpose record of a language" rests on the assumption that it is possible and useful to compile a database for a very broadly defined subject matter ("a language") without being guided by a specific theoretical or practical problem in mind which could be resolved on the basis of this database. With regard to its use in scientific inquiries, the validity of this assumption is shown by the success of all those social and historical disciplines working with data not specifically produced for research purposes. Thus, for example, cave dwellers in the Stone Age did not discard shellfish, animal bones, fragments of tools, and the like within the cave with the purpose in mind of documenting their presence and aspects of their diet and culture. But archeologists today use this haphazardly discarded waste as the primary data for determining the length and type of human occupation found in a given location. Similarly, inscriptions on stones, bones, or clay tablets were not produced in order to provide a record of linguistic structures and practices, but they have successfully been used to explore the structural properties of languages such as Hittite or Sumerian, which had already been extinct for millennia before their modern linguistic analysis began.

However, it is also well known that historical remains and records tend to be deficient in some ways with regard to modern purposes. Stone inscriptions and other historic documents with linguistic content, for example, never provide a comprehensive record of the linguistic structures and practices in use in the community at the time when these documents were written. Thus, given that the Hittite records discovered to date mostly pertain to matters of government, law, trade, and religion, it remains unknown how Hittite adolescents chatted with each other or whether it was possible to have the verb in first position in subordinate clauses.[1]

The experience with historical remains and records thus is ambivalent: On the one hand, it clearly shows that they may serve as the database for exploring issues they were not intended for. On the other hand, they show that haphazardly compiled databases hardly ever contain all the information one needs to answer all the questions of current interest. Based on this observation, the basic idea of a language documentation as developed here can be stated as follows: The goal is to create a record of a language in the sense of a comprehensive corpus of primary data which leaves nothing to be desired by later generations wanting to explore whatever aspect of the language they are interested in (what exactly is meant by "primary data" here is further discussed in Section 3.1.1 below).

Put in this way, the task of compiling a language documentation is enormous, and there is no principled upper limit for it. Obviously, every specific documentation project will have to limit its scope and set specific targets. Guidelines and suggestions as to how to go about setting such limits and targets are further discussed below and in the remaining chapters of this book. But to begin with, the fundamental importance of taking a pragmatic stance in all matters of language documentation needs to be emphasized once again. There are major practical constraints on the usefulness of targets and delimitations for language documentations which are exclusively based on theoretical considerations regarding the nature of language and speech communities. In most if not all documentation settings, the range of items that can be documented will be determined to a significant degree by factors that are specific to the given setting, most importantly, the availability of speakers who are willing and able to participate in the documentation effort. In fact, recent experiences make it clear that encouraging native speakers to take an *active* part in determining the contents of a documentation significantly increases the productivity of a documentation project. Consequently, a theoretical framework for language documentation should provide room for the active participation of native speakers. While the input of native speakers and other factors specific to a given setting is not completely unpredictable, it clearly limits the level of detail of a general framework for language documentation which can be usefully explored in purely theoretical terms.

This assessment, however, should not be construed as denying the relevance of theorizing language documentations. Not everything in a documentation is fully determined by the specifics of a given documentation situation. Speakers and speech communities usually do not have a fully worked-out plan for what to document. Rather, the specifics of a documentation are usually established interactively by communities and research teams. On the part of the research team, this presupposes a theoretically grounded set of basic goals and targets one wants to achieve.

Furthermore, without theoretical grounding language documentation is in the danger of producing “data graveyards”, i.e. large heaps of data with little or no use to anyone. While language documentation is based on the idea that it is possible and useful to dissociate the compilation of linguistic primary data from any *particular* theoretical or practical project based on this data, language documentation is not a theory-free or anti-theoretical enterprise. Its theoretical concerns pertain to the methods used in recording, processing, and preserving linguistic primary data, as well as to the question how it can be ensured that primary data collections are indeed of use for a broad range of theoretical and applied purposes.

Among other things, documentation theory has to provide guidelines for determining targets in specific documentation projects. It also has to develop principled and intersubjective means for evaluating the quality of a given documentation regardless of the specific circumstances of its compilation. A further major concern pertains to the interface between primary data and analysis in a broad range of disciplines. Based on a detailed investigation and evaluation of basic analytical procedures in these disciplines, it has to be determined which type and format of primary data is required for a particular analytical procedure so that it can be ensured that the appropriate type of data is included in a comprehensive documentation.

The present book provides an introduction to basic practical and theoretical issues in language documentation. It presents specific suggestions for the structure and contents of language documentations as well as the methodologies to be used in compiling them. To begin with, it will be useful briefly to address the question of what language documentations are good for. That is, why is it a useful enterprise to create lasting, multipurpose records of a language?

2. What is a language documentation good for?

From a linguistic point of view, there are essentially three reasons for engaging in language documentation, all of them having to do with consolidating and enlarging the empirical basis of a number of disciplines, in particular those branches of linguistics and related disciplines which heavily draw on data of little-known speech communities (e.g. descriptive linguistics, linguistic typology, cognitive anthropology, etc.). These are language endangerment, the economy of research resources, and accountability.

Certainly the major reason why linguists have recently started to engage with the idea of multipurpose documentations is the fact that a substantial number of the languages still spoken today are threatened by extinction (see Grenoble and Whaley 1998; Hagège 2000; Crystal 2000; or Bradley and Bradley 2002 for further discussion and references regarding language endangerment). In the case of an extinct language, it is obviously impossible to check data with native speakers or to collect additional data sets. Creating lasting multipurpose documentations is thus seen as one major linguistic response to the challenge of the dramatically increased level of language endangerment observable in our times. In this regard, language documentations are not only seen as data repositories for scientific inquiries, but also as important resources for supporting language maintenance.

Creating language documentations which are properly archived and made easily accessible to interested researchers is also in the interest of research economy. If someone worked on a minority language in the Philippines 50 years ago and someone else wanted to continue this work now, it would obviously be most useful if this new project could build on the complete set of primary data collected at the time and not just on a grammar sketch and perhaps a few texts published by the earlier project. Similarly, even if a given project on a little-known language is geared towards a very specific purpose – say, the conceptualization of space – it is in the interest of research economy (and accountability) if this project were to feed *all* the primary data collected in the project work into an open archive and not to limit itself to publishing the analytical results plus possibly a small sample of primary data illustrating their basic materials.

While the set of primary data fed into an archive in these examples would surely fail to constitute a comprehensive record of a language, it could very well be of use for purposes other than the one motivating the original project (data from matching tasks developed to investigate the linguistic encoding of space, for example, are also quite useful for the analysis of intonation, for conversation analytic purposes, for grammatical analysis, and so on). More importantly, if it were common practice to feed complete sets of primary data into open archives (which do not necessarily have to form a physical unit), comprehensive documentations for quite a number of little-known languages could grow over time, which in turn would strengthen the empirical basis of all disciplines working on and with such languages and cultures. That is, while much of the discussion in this chapter and book is concerned with projects specifically targeted at creating substantial language documentations, the basic idea of creating lasting, multipurpose documentations which are openly archived is not necessarily tied to such projects. It is very well possible and desirable to create such documentations in a step-by-step fashion by compiling and integrating the primary data sets collected in a number of different projects over an extended period of time. In fact, it is highly likely that in most instances, really comprehensive documentations can only be created in this additive way.

Finally, establishing open archives for primary data is also in the interest of making analyses accountable. Many claims and analyses related to languages and speech communities for which no documentation is available remain unverifiable as long as substantial parts of the primary data on which the analyses are based remain inaccessible to further scrutiny. Accountability here is intended to include all kinds of practical checks and methodological tests with regard to the empirical basis of an analysis or theory, including replicability and falsifiability. The documentation format developed here encourages, and also provides practical guidelines for, the open and widely accessible archiving of *all* primary data collected for little-known languages, regardless of their vitality.[2]

3. A basic format for language documentations

This section presents a basic format for language documentations and then highlights some features which distinguish this format from related enterprises.

3.1. The basic format

3.1.1. Primary data

Continuing the argument developed in the preceding sections, it should be clear that a language documentation, conceived of as a lasting, multipurpose record of a language, should contain a large set of primary data which provide evidence for the language(s) used at a given time in a given community (in all of the different senses of "language"). Of major importance in this regard are specimens of **observable linguistic behavior**, i.e. examples of how the people actually communicate with each other. This includes all kinds of communicative activities in a speech community, from everyday small talk to elaborate rituals, from parents baby-talking to their newborn infants to political disputes between village elders.

It is impossible to record *all* communicative events in a given speech community, not only for obvious practical, but also for theoretical and ethical reasons. Most importantly, such a record would imply a totalitarian set-up with video cameras and microphones everywhere and the speakers unable to control what of their behavior is recorded and what not. A major theoretical problem pertains to the fact that there is no principled way for determining a temporal boundary for such a recording (all communicative events in one day? two weeks? one year? a century?).

Consequently, there is a need to sample the kinds of communicative events to be documented. Once again, we can distinguish between a pragmatic guideline and theoretically grounded targets. The pragmatic guideline simply says that one should record as many and as broad a range as possible of communicative events which commonly occur in the speech community.

The theoretically grounded sampling procedure will be determined to a significant degree by the purposes and goals of the particular project. The rather broad and unspecific goal of a lasting, multipurpose record of a language envisioned here implies that, as much as possible, a sufficiently large number of examples for every type of communicative event found in a given speech community is collected. This in turn raises the highly complex issue of how the typology of communicative events in a given speech community can be uncovered. Within sociolinguistics, the framework known as the *ethnography of communication* provides a starting point for dealing with this issue. Chapter 5 provides a brief introduction to major concepts relevant here. Chapter 8 lists a range of important topics and parameters.

Besides observable linguistic behavior, is there anything else that needs to be documented in order to provide for a lasting, multipurpose record of a language? Or can all relevant information be extracted from a comprehensive corpus of recordings of communicative events? One aspect of "a language" that is not, or at least not easily, accessible by analyzing observable linguistic behavior is the tacit knowledge speakers have about their language. This is also known as **metalinguistic knowledge** and refers to the ability of native speakers to provide interpretations and systematizations for linguistic units and events. For example, speakers know that a given word is a taboo word, that speech event X usually has to be followed by speech event Y, or that putting a given sequence of elements in a different order is awkward or simply impossible. Similarly, metalinguistic knowledge as understood here also includes all kinds of linguistically based taxonomies, such as kinship systems, folk taxonomies for plants, animals, musical instruments and styles, and other artifacts, expressions for numbers and measures, but also morphological paradigms.

The documentation of metalinguistic knowledge, while not involving principled theoretical or ethical problems, is also not a straightforward task because much of it is not directly accessible. To be sure, in some instances there are conventional speech events involving the display of metalinguistic knowledge, such as reciting a genealogy or lengthy mythological narratives which sketch a cognitive map of the landscape. In many societies, there are also a number of well established and much discussed topics where speakers engage in metalinguistic discussions regarding the differences between different varieties (in village X they say "da" but we say "de"; young people cannot pronounce our peculiar /k/-sound correctly anymore, etc.). Furthermore, transcripts prepared by native speakers without direct interference by a linguist often provide interesting evidence regarding morpheme, word, and sentence boundaries (see Chapters 3 and 10 for further discussion). But very often documenting metalinguistic knowledge will involve the use of a broad array of elicitation strategies, guided by current theories about different kinds of metalinguistic knowledge and their structure. One very important type of elicited evidence are monolingual definitions of word meanings provided by native speakers. See Chapters 3 and 6 for further discussion and exemplification.

The documentation of metalinguistic knowledge as understood here includes much of the basic information that is needed for writing descriptive grammars and dictionaries. In particular, it includes all kinds of elicited data regarding the grammaticality or acceptability of phonological or morphosyntactic structures and the meaning, use, and relatedness of lexical items. However, it should be clearly understood that documentation here means that the elicitation process itself is documented in its entirety, including the questions asked or the stimuli presented by the researcher as well as the reaction by the native speaker(s). That is, documentation pertains to the level of primary data which provide evidence for metalinguistic knowledge, i.e. what native speakers can actually articulate regarding their linguistic practices or their recordable reactions in experiments designed to probe metalinguistic knowledge.[3] A grammatical rule as stated in a grammar or an entry in a published dictionary are not primary data in this sense, even though some linguists may believe that they are part of a native speakers' (unconscious) metalinguistic knowledge. In this view, grammatical rules and dictionary entries are *analytical formats* for metalinguistic knowledge. Whether and to what extent these have a place in a language documentation is an issue we will take up in Section 4.2.

It is also worth noting that the documentation of observable linguistic behavior and metalinguistic knowledge are similar in that they basically consist of records of communicative events. In the case of observable linguistic behavior, the communicative event involves the interaction of native speakers among themselves, while in the case of metalinguistic knowledge it involves the interaction between native speakers and documenters. There is a superficial difference with regard to the preferred documentation format in that it is now standard practice to make (video) recordings of observable linguistic behavior, while for the elicitation of metalinguistic knowledge it is still more common simply to take written notes. In principle, (video-)recording would also be the better (i.e. more reliable and comprehensive) documentation format for elicited metalinguistic knowledge, but there may often be practical reasons to stay with paper and pencil (among other things, native speakers may be more comfortable to discuss metalinguistic knowledge without being constantly recorded). But, to repeat, regardless of the recording method, records of observable linguistic behavior and metalinguistic knowledge both contain primary data documenting linguistic interactions in which native speakers participate.

In the following, we will use the label *corpus of primary data* as a shorthand for *corpus of recordings of observable linguistic behavior and metalinguistic knowledge* for this component of a language documentation. Throughout this book it is assumed that this corpus is stored and made available in digital form.

To date, there is very little practical experience with regard to structuring and maintaining such digital corpora. Consequently, no widely-used and well-tested structure exists for them. Within the DoBeS program, it is a widespread practice to operate with two basic components in structuring primary data: records of individual communicative events and a lexical database (this obviously follows a widespread practice in linguistic fieldwork where apart from transcripts of recordings and fieldnotes the compilation of a lexical database is a standard procedure).

Records of individual communicative events are called **sessions** (alternative terms would be "document", "text", or "resource bundle"). In the manual for the IMDI Browser,[4] a session is defined as "a meaningful unit of analysis, usually [...] a piece of data having the same overall content, the same set of participants, and the same location and time, e.g., one elicitation session on topic X, or one folktale, or one 'matching game', or one conversation between several speakers." It could also be the recording of a two-day ceremony. Sessions are typically allocated to different sets defined according to parameters such as medium (written vs. spoken), genre (monologue, dialogue, historical, chatting, etc.), naturalness (spontaneous, staged, elicited, etc.), and so on. It is too early to tell whether some of the various corpus structures currently being used are preferable to others.

There are two reasons why a **lexical database** appears to be a useful format for organizing primary data. On the one hand, there is a need to bring together all the information available for a given item so that one can make sure that the meaning and formal properties of the item are well understood.[5] On the other hand, and perhaps more importantly, a list of lexical items is a very useful resource when working on the transcription and translation of recordings. One of the most widely used computational tools in descriptive linguistics, the program *Toolbox* (formerly *Shoebox*),[6] allows for the semi-automatic compilation of a lexical database when working through a transcript, and the existence of this program is certainly one reason why the compilation of a lexical database currently is almost an automatic procedure when working with recordings. However, as with all other aspects of organizing a digital corpus of primary data, it remains to be seen and tested further whether this is indeed a necessary and useful procedure.

3.1.2. Apparatus

Inasmuch as linguistic and metalinguistic interactions cover the range of basic interactional possibilities,[7] a documentation which contains a comprehensive set of primary data for both types of interactions is logically complete with regard to the level of primary data. However, it is well known that a large corpus of primary data is of little use unless it is presented in a format which ensures accessibility for parties other than the ones participating in its compilation. To be accessible to a broad range of users, including the speech community, the primary data need to be accompanied by information of various kinds, which – following philological tradition – could be called the **apparatus**. The precise extent and format of the apparatus is a matter of debate, with one exception: the uncontroversial need for **metadata**.

Metadata are required on two levels. First, the documentation as a whole needs metadata regarding the project(s) during which the data were compiled, including information on the project team(s), and the object of documentation (which variety? spoken where? number and type of records; etc.). Second, each session (= segment of primary data) has to be accompanied by information of the following kind:[8]

- a name of the session which uniquely identifies it within the overall corpus;
- when and where was the data recorded?;
- who is recorded and who else was present at the time?;
- who made the recording and what kind of recording equipment was used?;
- an indication of the quality of the data according to various parameters (recording environment and equipment, speaker competence, level of detail of further annotation);
- who is allowed to access the data contained in this session?;

- a brief characterization of the content of the session (what topic is being talked about? what kind of communicative event [narrative, conversation, song, etc.] is being documented?);
- links between different files which together constitute the session, e.g. a media file (audio or video) and a file containing a transcription, translation, and various types of commentary relevant for interpreting the recording contained in the media file (on which see further below).

The metadata on both levels have two interrelated functions. On the one hand, they facilitate access to a documentation or a specific record within a documentation by providing key access information in a standardized format (what, where, when, etc.). In this function, they are similar to a catalogue in a library and we can thus speak of a *cataloguing* function.[9] On the other hand, they have an *organizational* function in that they define the structure of the corpus which, in particular in the case of documentations in digital format, in turn provides the basis for various procedures such as searching, copying, or filtering within a single documentation or across a set of documentations. Obviously, a metadata standard which targets the organizational function has to be richer and more elaborate than one which targets the cataloguing function. The former is actually a corpus management tool, which defines digital structures and supports various computational procedures, rather than just a standard for organizing a catalogue.

Currently there exist two metadata standards which in fact complement each other in that they target these different functions. The OLAC standard targets exclusively the cataloguing function and provides an easy and fast access to a large number of diverse repositories of primary data on a worldwide scale (in both digital and non-digital formats). The IMDI standard, which incorporates all the information included in the OLAC standard and hence is compatible with it, is actually a corpus management tool which primarily targets digitally archived language documentations. Further discussion of metadata concepts and standards is found in Chapters 4 and 13.

Apart from metadata, there is in most instances also a need for further information accompanying each recording as well as the documentation as a whole in order to make the corpus of primary data useful to users who do not know the language being documented. On the level of individual sessions, such additional information is called here an **annotation**.[10] Thus, in the case of audio or video recordings of communicative events, it is obviously useful to provide at least a transcription and a translation so that users not familiar with the language are able to understand what is going on in the recording.

However, the exact extent and format of the annotations that should be included in each session is a matter of debate. It is common to distinguish between minimal and more elaborate annotation schemes. A widely assumed minimal annotation scheme consists of just a transcription and a free translation which should accompany all, or at least a substantial number of, primary data segments. More elaborate annotation schemes include various levels of interlinear glossing, grammatical as well ethnographical commentary, and extensive cross-referencing between the various sessions and resources compiled in a given documentation. See further Chapters 8 and 9.

On the level of the overall documentation, information accompanying the primary data set other than metadata is, for lack of a well-established term, subsumed here under the heading **general access resources** (alternatively, it could also simply be called "annotation"). Such general (in the sense of: relevant for the documentation as a whole) access resources would include:

- a general introduction which provides background information on the speech community and language (language name(s), affiliation, major varieties, etc.), the fieldwork setting(s), the methods used in recording primary data, an overview of the contents, structure, and scope of the primary data corpus and its quality;
- brief sketches of major ethnographic and grammatical features being documented;
- an explication of the various conventions that are being used (orthography, glossing abbreviations, other abbreviations);
- indices for languages/varieties, key analytic concepts, etc.;
- links and references to other resources (books and articles previously published on the variety or community being documented; other projects relating to the community or its neighbors, etc.).

For further discussion of some aspects of relevance here, see Chapters 8 and 12.

Table 1 provides a schematic overview of the components of the language documentation format sketched in this section.

Table 1. Basic format of a language documentation

Primary data	Apparatus	
	Per session	For documentation as a whole
recordings/records of observable linguistic behavior and metalinguistic knowledge (possible basic formats: session and lexical database)	*Metadata* – time and location of recording – participants – recording team – recording equipment – content descriptors …	*Metadata* – location of documented community – project team(s) contributing to documentation – participants in documentation – acknowledgements …
	Annotations – transcription – translation – further linguistic and ethnographic glossing and commentary	*General access resources* – introduction – orthographical conventions – ethnographic sketch – sketch grammar – glossing conventions – indices – links to other resources …

3.2. What's new?

Language documentation in the way depicted in Table 1 is not a totally new enterprise. The compilation of annotated collections of written historical documents and culturally important speech events (legends, epic poems, and the like) was the major concern of philologists in the nineteenth century. Linguistic and anthropological fieldwork in the Boasian tradition has also always put major emphasis on the recording of speech events. Within linguistic anthropology, recording and interpreting oral literature is a major task. All of these traditions have had a major influence on documentary linguistics as developed in this book.

Nevertheless, the idea of a language documentation as sketched above is new for mainstream linguistics, and even compared to these earlier approaches, it is new with regard to the following important features:

- *Focus on primary data*: The main goal of a language documentation is to make primary data available for a broad group of users. Unlike in the philological tradition, there is no restriction to culturally or historically "important" documents, however such importance may be defined.
- *Explicit concern for accountability*: The focus on primary data implies that considerable care is given to the issue of making it possible to evaluate the quality of the data. This in turn implies that the field situation is made transparent and that all documents are accompanied by metadata which detail the recording circumstances as well as the further steps undertaken in processing a particular document.
- *Concern for long-term storage and preservation of primary data*: This involves two aspects. On the one hand, metadata are crucial for users of a documentation to locate and evaluate a given document, as just mentioned. On the other hand, long-term storage is essentially a matter of technology, and while compilers of language documentations do not have to be able to handle all the technology themselves, they need to have a basic understanding of the core issues involved so that they avoid basic mistakes in recording and processing primary data. Among other things, the quality of the recording is of utmost importance for long-term storage and hence needs explicit attention. See further Chapters 4, 13, and 14.
- *Work in interdisciplinary teams*: Work on a truly comprehensive language documentation needs expertise in a multitude of disciplines in addition to the basic linguistic expertise required in transcription and translation. Such disciplines include anthropology, ethnomusicology, oral history and literature, as well as all the major subdisciplines of linguistics (socio- and psycholinguistics, phonetics, discourse analysis, corpus linguistics, etc.). There are probably no individuals who are experts in all of these fields, and few who have acquired significant expertise in a substantial number of them. Hence, good documentation work usually requires a team of researchers with different backgrounds and areas of expertise.
- *Close cooperation with and direct involvement of speech community*: The documentation format sketched above strongly encourages the active involvement of (members of) the speech community in two ways. On

the one hand, as mentioned above, native speakers are among the main players in determining the overall targets and outcomes of a documentation project. On the other hand, a documentation project involves a significant number of activities which can be carried out with little or no academic training. For example, the recording of communicative events can be done by native speakers who know how to handle the recording equipment (which can be learned in very short time), and it is often preferable that they do such recordings on their own because they know where and when particular events happen, and their presence is frequently felt to be less obtrusive. Similarly, given some training and regular supervision, the recording of metalinguistic knowledge and also the transcription and translation of recordings can be carried out by native speakers all by themselves. See further Chapter 3.

3.3. Limitations

As with most other scientific enterprises, the language documentation format developed here is not without problems and limitations. Some of the theoretical and practical problems have already been mentioned in the preceding discussion, and it will suffice here to emphasize the fact that the documentation format in Table 1 is based on a number of hypotheses which may well be proven wrong or unworkable in practical terms (see further Section 4 below). In addition to theoretical and practical problems, there are also ethical problems and limitations which are related to the fact that even the most circumspectly planned documentation project has the potential to profoundly change the social structure of the society being documented. This may pertain to a number of different levels, only two of which are mentioned here (see Wilkins 1992, 2000; Himmelmann 1998; and Grinevald 2003: 60–62 for further discussion).

On a somewhat superficial level, there are usually a few, often not more than one or two native speakers who are very actively involved in the project work. Through their work in the project, their social and economic status may change in a way that otherwise may have been impossible. This in turn may lead to (usually minor) disturbances in the wider community, such as inciting the envy or anger of relatives and neighbors. It is also not unknown that affiliation with an externally funded and administered project is used as an instrument in political controversies and competitions within the speech community.

On a more profound level, in non-literate societies the documentation of historical, cultural, and religious knowledge generally introduces a new way for accessing such knowledge and thereby may change the whole psychosocial fabric of the society (Ong 1982). This is particularly true of societies where much of the social fabric depends on highly selective access to cultural and historical knowledge, transmission of such knowledge thus involving different levels of secrecy (see Brandt [1980, 1981] for a pertinent example). That is, in some instances a documentation project may contribute to the demise of the very linguistic and cultural practices it proposes to document. In these instances, it would appear to be preferable *not* to document, but rather to support language maintenance in other ways, if necessary and possible.

Note that in general, language documentation and language maintenance efforts are not opposed to each other but go hand in hand. That is, it is an integral part of the documentation framework elaborated in this book that it considers it an essential task of language documentation projects to support language maintenance efforts wherever such support is needed and welcomed by the community being documented. More specifically, the documentation should contain primary data which can be used in the creation of linguistic resources to support language maintenance, and the documentation team should plan to dedicate a part of its resources to "mobilizing" the data compiled in the project for maintenance purposes. Chapter 15 elaborates some of the issues involved here.

4. Alternative formats for language documentations

The format for language documentations sketched in the preceding section is certainly not the only possible format. In fact, within structural linguistics there is a well-established format for language documentations consisting primarily of a grammar and a dictionary. In this section, I will first briefly present some arguments as to why this well-established format is strictly speaking a format for language *description* and not for language *documentation* proper, and thus is not a viable alternative to the basic documentation format of Table 1. In Section 4.2, we will then turn to the question of whether it makes sense to integrate the grammar-dictionary format with the basic documentation format of Table 1 and thus make fully worked-out grammars and dictionaries essential components of language documentations.

It should be clearly understood that this section is merely intended to draw attention to this important topic at the core of documentation theory. It barely scratches the surface of the many complex issues involved here. For more discussion, see Labov (1975, 1996), Greenbaum (1984), Pawley (1985, 1986, 1993), Lehmann (1989, 2001, 2004b), Mosel (1987, 2006), Himmelmann (1996, 1998), Schütze (1996), Keller (2000), Ameka et al. (2006), among others.

4.1. The grammar-dictionary format

The grammar-dictionary format of language description targets the language system.[11] That is, it is based on the notion of a language as an abstract system of rules and oppositions which underlies the observable linguistic behavior. In this view, documenting a language essentially involves compiling a grammar (= set of rules for producing utterances) and a dictionary (= a list of conventional form-meaning pairings used in producing these utterances). To this core of the documentation, a number of texts are often added, either in the form of a text collection or in the appendix to the grammar, which have the function of extended examples for how the system works in context. These texts are usually taken from the corpus of primary data on which the system description is based, but they do not actually provide access to these primary data because they are edited in various ways. Providing direct access to the complete corpus of primary data is typically not part of this format.

The compilation of grammars and (to a lesser extent) dictionaries is a well-established practice in structural linguistics, with many fine specimens having been produced in the last century. But even the best structuralist grammars and dictionaries have been lacking with regard to the goal of presenting a lasting, multipurpose record of a language. Major problems with regard to this goal include the following points:

a. Many communicative practices found in a given speech community remain undocumented and unreconstructable. That is, provided with a grammar and a dictionary it is still impossible to know how the language is (or was) actually spoken. For example, it is impossible to derive from a grammar and a dictionary on how everyday conversational routines look like (how does one say "hello, good morning"?) or how one linguistically interacts when building a house or negotiating a marriage.

b. In line with the structuralist conception of the language system, grammars and dictionaries contain abstractions based on a variety of analytical procedures. With the data contained in grammars and dictionaries, most aspects of the analyses underlying the abstractions are not verifiable or replicable. There is no way of knowing whether fundamental mistakes have been made unless the primary data on which the analyses build are made available *in toto* as well.

c. Grammars usually only contain statements on grammatical topics which are known and reasonably well understood at the time of writing the grammar. Thus, for example, grammars written before the advent of modern syntactic theories generally do not contain any statements regarding control phenomena in complex sentences. Many topics of current concern such as information structure (topic, focus) or the syntax and semantics of adverbials have often been omitted from descriptive grammars due to the lack of an adequate descriptive framework. As pointed out in particular by Andrew Pawley (1985, 1993, and elsewhere), there is a large variety of linguistic structures often subsumed under the heading of *speech formulas* which do not really fit the structuralist idea of a clean divide between grammar and dictionary and thus more often than not are not adequately documented in these formats.

d. Grammars and (to a lesser extent) dictionaries provide little that is of direct use to non-linguists, including the speech community, educators, and researchers in other disciplines (history, anthropology, etc.).

These points of critique mostly pertain to the fact that structuralist language descriptions are reductionist with regard to the primary data on which they are based and do not provide access to them. Or, to put it in a slightly different and more general perspective, they document a language only in one of the many senses of "language", i.e. language as an abstract system of rules and oppositions. Inasmuch as structuralist language descriptions are intended to achieve just that, the above "critique" is, with the possible exception of point (b), not fair in that it targets goals for which these descriptions were not intended.[12]

In this regard, it should be emphasized that the above points in no way question the usefulness and relevance of descriptive grammars and dictionaries with regard to their main purpose, i.e. to provide a description and documentation of a language *system*. While there is always room for improvement (compare points (b) and (c) above), there is no doubt about the fact that grammars and dictionaries are essentially successful in delivering system descriptions. What is more, the above points also do not imply that grammars and dictionaries do not have a role to play in language documentations, as further discussed in the next section. The major thrust of the critical observations above is that a description of the language system as found in grammars and dictionaries by itself is not good enough as a lasting record of a language, even if accompanied by a text collection. And it is probably fair to say that the way primary data have been handled in the grammar-dictionary format is now widely seen as not adequate and thus in need of improvement.

From this assessment, however, it does not necessarily follow that the basic format of Table 1 is the only imaginable format for lasting, multipurpose records of a language. Instead, it may reasonably be asked, why not combine the strong sides of the two formats discussed so far and propose that language documentations consist of the combination of a large corpus of annotated primary data as well as a full descriptive grammar and a comprehensive dictionary? This is the question to be addressed in the next section.

4.2. An extended format for language documentations

Assuming that the structuralist notion of a language as a system of rules and oppositions is a viable and useful notion of "a language", though not necessarily the only useful and viable one for documentary purposes, and assuming further that a descriptive grammar and a dictionary provide adequate representations of this system, it would seem to follow that a truly comprehensive language documentation does not simply consist of a large corpus of annotated primary data – as sketched in Section 3 – but instead should also include a comprehensive grammar and dictionary. Along the same lines, one may ask why the apparatus in Table 1 should only contain a sketch grammar and not a fully worked-out comprehensive grammar, thus replacing the format in Table 1 with the one in Table 2.[13]

Table 2. Extended format for a language documentation

Primary data	Apparatus	
	Per session	For documentation as a whole
recordings/records of observable linguistic behavior and metalinguistic knowledge	*Metadata* *Annotations* – transcription – translation – further linguistic and ethnographic glossing and commentary	*Metadata* *General access resources* – introduction – orthographical conventions – glossing conventions – indices – links to other resources . . . *Descriptive analysis* – ethnography – descriptive grammar – dictionary

The difference between the basic format for language documentations in Table 1 and the extended format depicted in Table 2 pertains to the addition of fully worked out descriptive analyses on various levels (as indicated by the shaded area in Table 2), replacing the corresponding sketch formats (sketch grammar, ethnographic sketch) under *general access resources* in the basic format. Whether this is in fact a fundamental difference or rather a gradual difference in emphasis, is a matter for further debate. In actual practice, the difference may not be as relevant as it may appear at first sight, as we will see at the end of this section. Still, in the interest of making clear what is involved here, it will be useful to highlight the differences between the two formats and to indicate some of the problems that are created by incorporating comprehensive descriptive formats in the extended documentary format. There are at least two types of such problems, one relating to theoretical issues, the other to research economy.

The theoretical problem pertains to the fact that it is not at all clear how exactly the descriptive grammar (or the ethnography or the dictionary)[14] should look that is to be regarded as an essential part of a language documentation. As is well known, for much studied languages such as English, Latin, Chinese, Arabic, Tagalog, Quechua, or Fijian, there exist not only different types of grammars (pedagogical, historical, descriptive) but also different descriptive grammars, each having its particular emphasis and way of presenting the structure of the language system. This simply reflects the fact that at least according to the current state of knowledge, there is not just exactly one descriptive grammar which correctly and comprehensively captures the system of a language. Instead, any given descriptive grammar is a more or less successful attempt to capture the system of a language (variety), rarely if ever comprehensive, and usually also including at least some contested, if not clearly wrong, analyses.

As a consequence of this state of affairs, the following problem arises with regard to the extended format for language documentations in Table 2. Either one has to specify a particular type of descriptive grammar as the one which is the most suitable one for the purposes of language documentations and thus is able to provide a reasonably precise definition of this part of a documentation. Alternatively, one allows for a multitude of descriptive grammars to be included in a documentation, thus declaring it a desirable goal to include a number of different analyses of the language system as part of the overall documentation of a language. The latter option clearly raises the issue of practical feasibility, which leads us to the second problem mentioned above, i.e. the essentially pragmatic problem of research economy.

Practical feasibility also is an issue if just one analysis of the grammatical system is assumed to be an essential part of a language documentation, for the following reason. It is a well-known fact that it is possible to base elaborate descriptive analyses exclusively on a corpus of texts (either texts written by native speakers or transcripts of communicative events) – and most good descriptive grammars are based to a large degree on a corpus of (mostly narrative) texts. A large corpus of texts in fact provides for the possibility of writing a number of interestingly different descriptive grammars, targeting different components of the language system and their interrelation. Consequently, one could argue that even if one accepts the claim that a comprehensive documentation should also document the language system, there is no need to include a fully worked-out descriptive grammar in a language documentation. The information needed to write such a grammar is already contained in the corpus and the resources needed to extract this information and to write it up in the conventional format of a descriptive grammar are not properly part of the documentation efforts. In this view, resources allocated to documentation should not be "wasted" on writing a grammar but are better spent on enlarging the corpus of primary data, the quantity or quality of annotations, or on the "mobilization" of the data (mobilization is further discussed in Chapter 15).

The major counterargument against this position would be the claim that actually producing a descriptive grammar is a necessary part of a language documentation because otherwise, essential aspects of the language system would be left undocumented. The evaluation of this claim rests on the question of whether there is some kind of important evidence for grammatical structure which, as a matter of principle, cannot be extracted from a sufficiently large and varied corpus of primary data as sketched in Section 3 above. As far as I am aware, there is especially one type of evidence of this kind, i.e. negative evidence. Obviously, illicit structures cannot be attested even in the largest and most comprehensive corpora.[15]

However, the lack of explicit negative evidence in a corpus of texts does not per se necessitate the inclusion of a descriptive grammar in a language documentation. On the one hand, with regard to the usual way of obtaining negative evidence (i.e. asking one or two speakers whether examples x, y, z are "okay"), it is doubtful whether this really makes a difference in quality compared to evidence provided by the fact that the structure in question is not attested in a large corpus. Elicited evidence is only superior here if it is very carefully elicited, paying adequate attention to the sample of speakers interviewed, potential biases in presenting the material, and the like. On the other hand, and more importantly, the basic documentation format of Table 1 does not only consist of a corpus of more or less natural communicative events but also of documents recording metalinguistic knowledge. Metalinguistic knowledge includes negative evidence for grammatical structuring, as already mentioned above.

Obviously, gathering negative evidence on grammatical matters presupposes that the researcher asks the right questions, which in turn presupposes grammatical analysis. In this regard, it bears emphasizing that documentation does not exclude analysis. Quite the opposite: analysis is essential. What the documentary approach implies, however, is that the analyses which are carried out while compiling a documentation do not necessarily have to be presented in the format of a descriptive grammar. Instead, analyses can (or should) be included in a documentation through (scattered) annotations on negative evidence, the inclusion of experiments generating important evidence for problems of grammatical or semantic analysis, and so on (see further Chapters 8 and 9).

The major reason for choosing a distributed grammatical annotation format instead of the established descriptive grammar format is one of time economy. The writing of a descriptive grammar involves to a substantial degree matters of formulation (among other things, the search for the most suitable terminology) and organization (for example, chapter structure or the choice of the best examples for a given regularity; see Mosel 2006 for further discussion and exemplification). These are very time consuming activities which in some instances may enhance the analysis of the language system, but in general do not contribute essential new information on it. Thus, with regard to the economy of research resources, it may be more productive to spend more time on expanding the corpus of primary data rather than to use it for writing a descriptive grammar.

In short, then, the difference between the basic and the extended formats as conceived of here is one between different formats or "styles" for the inclusion of analytical insights in a documentation. In the basic format, analyses are included in the form of scattered annotations and cross-references between sessions (and, of course, indirectly also by the fact that for topics for which little or no data can be found in the recordings of communicative events, elicited primary data are included). In the extended format, analyses are presented as such in full, i.e. as descriptive statements about the language system, usually accompanied by (links to) relevant examples.

In actual practice, there will be many instances where this apparently clear difference will become blurred. For example, when the number and types of communicative events that can be recorded in a given community is severely limited, it may be more useful to work on full, and fully explicit, descriptions of aspects of the grammatical system not represented in the texts, rather than recording more texts of the same kind with the same speaker. Furthermore, on a much more mundane level, there are (individually widely diverging) limits as to the time and energy that can be productively spent on the not always thrilling routine work involved in documentation (filling in metadata, checking translations and glossing, etc.), and it would be a counterproductive and rather ill-conceived idea generally to restrict work with a speech community to "pure" documentation to the exclusion of all fully explicit (= publishable) analytic work. It is thus unlikely that linguists undertaking language documentations will stick to the basic format in its purest form and refrain from working on aspects of a fully explicit descriptive analyses while compiling the annotated corpus of primary data. It should, then, also not come as surprise that many researchers – including some of the contributors to this volume – tend to ignore the difference between the two formats and to remain implicit as to what exactly they have in mind when referring to grammatical analyses and dictionaries.

Most language documentations that have been compiled in recent years are actually hybrids with regard to the two formats. They tend to include many scattered analytical observations as well as substantial fully worked-out descriptive statements of some aspects of the language system (rarely comprehensive grammars). It remains to be seen whether this practice is actually viable in the long-term or whether there are clear advantages attached to adhering to either the basic or the extended format as discussed in this section.

5. The structure of this book

The following chapters provide in-depth discussions and suggestions for various issues arising when working on and with language documentations. While the authors have slightly different views of what a language documentation is (or should be) and clearly differ with regard to their major topics of interest and theoretical preferences, they share a major concern for the maintenance of linguistic diversity, including the quality, processing, and accessible preservation of linguistic primary data, which in some way or other all these chapters are about.

The focus of each chapter is on a topic which is rarely dealt with within descriptive linguistics (and mainstream linguistics in general), reflecting the fact that issues relating to the collection and processing of primary data have been widely neglected within the discipline until very recently. For each topic, both theoretical and practical issues are discussed, although the chapters differ quite significantly as to how much space they allot to either, in accordance with the topic being dealt with.

Apart from the present introduction, there are roughly four parts to this book which, however, are closely linked to, and overlap with, each other.

Chapters 2 to 4 deal with general (i.e. not specifically linguistic) ethical and practical issues which have to be considered and reconsidered from the earliest planning stage of a documentation project through to its completion. The guiding questions here are: How to interact with speech communities and individual speakers; and how to capture, store, and process relevant data. These issues are interrelated, in that data capture and processing is not just a technological issue, but also has to pay attention to sensitivities and interests of the speech community and the individual speakers contributing data. Chapter 3 includes suggestions for getting started with the actual linguistic documentation work in the field.

The next eight chapters (Chapters 5 to 12) pertain to the recording and processing of primary linguistic data from an anthropological and linguistic point of view. The first three of these chapters (Chapters 5 to 7, but also a considerable part of Chapter 8) are primarily concerned with the issue of how and what to document, given the goal of creating a lasting and multifunctional record of a language. Chapter 5 provides an introduction to a cultural and ethnographic understanding of language. This is essential for the success of a documentation project, not only with regard to the necessity of being able to identify the types of communicative events that should be recorded, but also for being able to successfully interact within a speech community which has a different set of norms of interaction. In the latter regard, Chapter 5 complements and expands Chapters 2 and 3.

Chapter 6 addresses the issue of how to access and represent metalinguistic knowledge, focusing primarily on lexical knowledge. Chapter 7 briefly discusses the kinds of data needed for prosodic analysis, while Chapter 8 reports on the demands of anthropologists for language documentations, which complements the discussion of this topic in Chapter 5.

Chapter 8 also addresses the issue of ethnographically relevant annotation and commentary and thus forms a group with the next four chapters (Chapters 9 to 12) all of which are concerned with the part of a documentation called "apparatus" in Table 1. That is, they deal with the processing of primary data necessary for them to become useful and accessible to a broad range of users. While Chapters 8 and 9 provide an overview of the basic structure and various practical aspects of ethnographic and linguistic annotation and commentary, respectively, the following two chapters address some more specific issues with regard to the written representation of recorded communicative events. Chapter 10 is concerned with one major aspect of transcription, namely, the need to segment the continuous flow of spoken language into smaller units, in particular words and intonations units. Issues relating to the development of a practical orthography which can be used for the written representation of the recordings, for educational materials, etc., and which is acceptable and accessible to the speech community are discussed in Chapter 11. The final chapter in this part of the book, Chapter 12, discusses the structure and format of the sketch grammar which is part of the overall apparatus of the documentation, intended to facilitate access to the primary data themselves as well as the grammatical information to be found in sessions and lexical database.

The last part of the book, consisting of the final three chapters, relates to the long-term perspectives of a documentation, in particular, archiving issues and its use in language maintenance. Apart from an obvious focus on technological issues, the main concern of Chapter 13 on "Archiving challenges" is a critical review of the different interests and goals of the three major groups involved in the archiving process: the donators (the people handing material to the archive), the archivists (the people running and maintaining the archive), and the users of archival sources. Chapter 14 takes up one particularly critical issue in long-term preservation, i.e. the changing standards in character and text structure encoding which very easily render digitally-stored information uninterpretable. Finally, Chapter 15 focuses on speech communities as potential users and argues that there is a need for elaborate and creative concepts for mobilizing primary data, i.e. creating language resources from archival data which are of interest and use to a given community.

There are a number of important topics which actually should also be dealt with in a book such as the present one but which unfortunately and for reasons beyond the control of the editors could not be included at this point. In particular, the following three topics are also of critical importance to language documentation (see the book's website for additional and up-to-date information on these and other topics).

- One major aspect of linguistic interactions which has to be attended to in documentations are so-called paralinguistic features, in particular gesture. The recent textbook on gesture by Kendon (2004) provides a thorough general introduction to this topic. See also Section 2.5 in Chapter 9 for a brief note on paralinguistics.
- There is no chapter on the basics of producing high-quality audio and video recordings. While this topic in part involves a lot of technological aspects which change rather rapidly and thus would in any event not have been included in this book, there is a need to be aware of what defines good recordings. In addition to the book's website, see the *Language Archiving Newsletter* and the DoBeS and ELDP websites for relevant pointers and links.
- Apart from the kind of mobilization of primary data for language maintenance purposes discussed in Chapter 15, there are also more traditional, but equally important contributions that a language documentation can make to language maintenance efforts. These include, in particular, the development of teaching materials in the documented variety. See von Gleich (2005) for a brief discussion and references.

The book is also heavily biased towards the more narrowly linguistic approaches to language. Documentary work that aims at a truly comprehensive record of a language also has to engage with ethnobotany, musicology, human geography, oral history, and so on. We hope that it will be possible before too long to compile a further introductory volume where the core issues and methodologies of these and related disciplines are presented from the point of view of enhancing language (and culture) documentations.

Even though the focus is on linguistic approaches to language, it should be clearly understood that even for this domain the ability to engage in language documentation projects cannot be gained by mastering only the topics and techniques presented here. Ideally, training in language documentation includes a training in the basics of a broad range of linguistic subdisciplines and neighboring disciplines. Training in descriptive and anthropological linguistics is indispensable.

The latter two topics are not dealt with here because good textbooks for them are readily available. As for descriptive linguistics, the classic textbooks by Hockett (1958) and Gleason (1961) still provide an excellent introduction which, however, should be complemented by typologically grounded surveys of major categories and structures as, for example, in the second edition of Shopen's *Language Typology and Syntactic Description* or in Kroeger (2005). As for anthropological linguistics, Duranti (1997) introduces the most important concepts and issues, which could be complemented with the more in-depth discussion of the ethnography of communication by Saville-Troike (2003). Finally, the contributions in Newman and Ratliff (2001) combine descriptive and ethnolinguistic topics and insights and complement the discussion of linguistic fieldwork in Chapters 2 and 3 of this volume.

In conclusion, it may be worthwhile to emphasize the fact that documentary linguistics is an emerging field where many things are still in flux. Most importantly perhaps, large multimedia corpora on lesser-known languages are very new and largely unexplored entities. It is very well possible that new techniques for working with such corpora will emerge before too long, requiring major adjustments to the format for language documentations discussed in this chapter and book. But rather than a shortcoming, this should be seen as one of the exciting aspects of language documentation. Apart from being a useful introduction to language documentation, providing theoretical grounding as well practical advice, this book should make it clear that language documentation is an important, engaging and rewarding enterprise with many repercussions for linguistics and other language-related disciplines and projects.

Acknowledgements

I wish to thank my co-editors and Eva Schultze-Berndt for critical discussion of many of the issues touched upon here, as well as helpful comments on the draft version of this chapter.

Notes

1. With regard to the latter point, compare the following quote from Luraghi (1990: 128 FN1) which nicely illustrates the problems arising when data types are missing in a given corpus: "As to the position of the verb, the most important difference [between main and subordinate clauses, NPH] lies in the absence of VSO sentences in subordinate clauses. It can of course be objected that this may be due simply to the shortage of sources, since VSO sentences are on the whole very infrequent. However, in the light of comparative data from other Indo-European languages, this objection could perhaps be rejected ..."
2. The major limitation here are restrictions on access to recordings imposed by speakers or communities which, of course, should be observed.
3. "Experiment" here is to be taken in a broad sense, including, for example, the testing of the acceptability of invented examples.
4. IMDI = ISLE Metadata Initiative. The manual can be downloaded at http://www.mpi.nl/IMDI/tools.
5. Note that this does not necessarily imply that all the information for a lexical item has to be gathered in a single location (i.e. an entry in the database), as it is currently done by most researchers. Alternatively, the lexical database could consist simply of links to all the sessions where the item in question occurs. This could include a session where the item is elicited as part of the elicitation of a word list or semantic field, a session where the item has been recorded in a list of items or a carrier phrase in order to document characteristic sound patterns, and a session where it occurs as part of a procedural text.
6. Please refer to the appendix for further information on this program.
7. Note that *linguistic interaction* here includes interactions with native speakers of other varieties inasmuch as they are a common occurrence in the speech community which is being documented.
8. The following list takes an audio or video recording as its main example. Of course, the same type of metadata is needed for primary data gathered in a different way such as written fieldnotes or photos.
9. Note that the term *cataloguing* is used here in a somewhat broader sense than in Chapter 4 where it is used to refer to one particular subtype of metadata.
10. Strictly speaking, "annotations" could also be called metadata since the term "metadata" in general refers to all kinds of data about data. However, within the

context of language documentations it is useful to distinguish between different types of metadata (in this broad sense), and it is now a widely-used practice to use the term "metadata" in the context of language documentations exclusively for data types which have a cataloguing or organizational function and to use "annotation" (or "commentary") for other types of information accompanying segments of primary data.

11. The structuralist idea of language as an abstract system has been articulated in a variety of oppositions including the well-known Sassurean distinction of *langue* vs. *langage* vs. *parole* and the Chomskyan distinction of *competence* vs. *performance*. For the present argument, the details of how the abstract language system is conceived of do not matter and thus are ignored.
12. With regard to falsifiability (point (b)), not providing access to the primary data is indeed a major problem for the scientific status of these descriptions. However, the basic assumption here appears to have been that whoever wanted to replicate and possibly falsify a descriptive analysis on the basis of material other than the one made available in examples and texts could compile their own set of primary data. This assumption is no longer viable in the case of endangered languages and, as already pointed out in Section 2, it is hence not by chance that a close connection exists between language endangerment and the recent increased concern for the preservation of primary data in linguistics and related disciplines.
13. The part called "descriptive analysis" in the rightmost column could also be added in other ways to the overall format, for example as an additional column of its own, on a par with "primary data" and "apparatus". While there are theoretical issues associated with these alternative overall organizations, these do not play a role for the argument in this section and hence can be safely ignored.
14. Essentially the same points made here and in the following with regard to descriptive grammars could also be made with regard to conventional dictionaries and ethnographic monographs (see Chapter 6 for a brief discussion of different types of dictionaries, which is also relevant here). Including these two other main analytical formats in the discussion would, however, unnecessarily complicate the exposition. Hence, dictionaries and ethnographies are not further discussed in this section. The choice of descriptive grammars as the main example is simply due to the fact that it is the format the author is most familiar with.
15. Very occasionally, though, especially in the interaction between parents and children, unacceptable or highly marked structures might be attested in admonishments of the form: Don't say X, say Y.

Chapter 2

Ethics and practicalities of cooperative fieldwork and analysis

Arienne M. Dwyer

Introduction

This chapter examines central ethical, legal, and practical responsibilities of linguists and ethnographers in fieldwork-based projects. These issues span all research phases, from planning to fieldwork to dissemination. We focus on the process of language documentation, beginning with a discussion of common ethical questions associated with fieldwork: When is documentation appropriate in a particular community, and who benefits from it? Which power structures are involved, both in and out of the field? Section 1 explores key concepts of participant relations, rights, and responsibilities in fieldwork in the context of ethical decision-making. It introduces a set of guiding principles and examines some potential pitfalls. Section 2 discusses the legal rights issues of data ownership (intellectual property rights and copyright) and data access. Such information aids planning before fieldwork and especially the archiving phase.

Sections 3 and 4 cover the more concrete practical aspects of the fieldwork situation: developing a relationship with a speech community and organizing and running a project. We survey what may be termed "the five Cs" critical to planning and executing a project: criteria (for choosing a field site), contacts, cold calls, community, and compensation. Finally, since even the best-planned projects encounter logistical and interpersonal challenges, we present several generic case studies and some possible methods of resolving such disputes.

Such ethical and logistical planning is essential to successful community-centered knowledge mobilization, from which documentation products useful for both academics and community members are produced in an environment of reciprocity. It is the linguist's responsibility to focus on *process* (Rice 2005: 9)[1] as much as the end goals.