Ulrike Mosel

Corpus building for under-researched languages - a practical guide.

to appear in
Firmin Ahoua, Dafydd Gibbon and Stavros Skopeteas.
Linguistic Fieldwork and Language Documentation
A Course Book on Foundational Skills

Corpus building for under-researched languages - a practical guide

Ulrike Mosel

Kiel University

Abstract:

Starting with definitions of basic corpus-linguistic concepts and a brief comparison of corpora of well researched languages and those of language documentation (LD) projects, this chapter provides a critical overview of corpus building methods for under-researched languages in LD projects and a guide to the collection of texts of different genres and registers with metadata and various kinds of annotations and eventually addresses the problem of how to evaluate language documentation corpora. The final section presents a commented list of further readings on aspects of corpus building, including the technology of recording and archiving, ethical questions and theoretical issues that are not dealt with in this chapter.

Keywords: Collaborative fieldwork, text collection, metadata, transcription, orthography, translation, morphological glossing.

1 Introduction

This article focuses on practical issues of corpus building in language documentation (LD) projects from a linguistic perspective. It first defines the basic concepts of text, corpus, annotation, metadata, genres, and registers in corpus linguistics (§2) and compares typical corpora of major languages with LD corpora with respect to their purpose, production, content, structure, and size (§3). The guidelines of how to build a LD corpus start with the collection of texts in collaboration with the speakers of the language (§4.1) and the various types of text found in LD corpora (§4.2) such as audio and video recordings, edited transcriptions and written texts without previous recordings, indigenous and non-indigenous genres, and elicited texts and wordlists.

For a better understanding of the corpus and its individual texts, the corpus should contain information, the so-called metadata, about the language and the corpus building process on the one hand, and the situational characteristics of each text on the other (§4.3). The individual texts of the corpus are identified short names (§4.4). The corpus should contain an introduction about its content and accessibility (§5.1) and can be organised into sub-corpora according to the genres, registers and topics of its texts (§5.2).

The texts are segmented into utterance units or clauses (§6.2). These segments are labelled, numbered (§6.3), and annotated, i.e. complemented by further information. In a LD corpus audio and video recordings need to be transcribed (§6.4), and the transcriptions as well as written texts of the corpus need to be translated (§6.5). In addition, annotations can represent a morphological analysis with glosses (§6.6), a syntactic analysis (§6.7), and any other kind of information like notes and keywords that is directly linked to the utterance units, words (§6.8 and §6.9). The exploitation and evaluation of corpora is discussed in the concluding remarks of this article (§7).

Technical, theoretical and ethical aspects of LD corpora are not dealt with, but references to relevant articles are given in the section on further readings with brief comments (§8.2). There you also find a list of handbooks relevant for language documentation (§8.1) as well as additional references to the particular issues dealt with in the preceding sections about the basic concepts of corpus linguistic, the ethnography of speaking, the collection of texts, the structure of the corpus, annotations and the use of corpora (§8.3 - §8.7).

2 Basic corpus-linguistic concepts

2.1 Text and corpus

In everyday language the word *text* refers to a piece of writing, whereas in corpus linguistics *text* is understood as any piece language usage in spoken, signed or written form (McEnery and Hardie 2012:252). Accordingly, not only stories and descriptions can be considered as texts in language documentation, but also the captions of pictures, greetings, or a native speaker's comments when he watches a video.

A *corpus* (pl. *corpora*) is a collection of any kinds of text in machine readable form which are annotated with various kinds of linguistic information. (McEnery and Hardie 2012:1-2; Kübler and Zinsmeister 2015:4-5). Spoken and signed texts are audio and video recorded and annotated by time-linked transcriptions.

2.2 Metadata

In general terms, metadata are data about data that allow the users to assess the suitability of the data for their research purposes. With respect to corpora, we can distinguish between two kinds of metadata: metadata about the corpus as a whole and metadata about the individual texts of the corpus. The first kind of metadata includes information about

- the purpose of the corpus
- the source of the of texts;
- the form in which the corpus data are presented;
- how the corpus can be accessed;
- what kind of items the corpus can be searched for and how.

The metadata of the text files inform the corpus user about the speakers or writers and the production circumstances (cf. §2.4). In professional archives the metadata are presented by standardised metadata vocabularies to facilitate systematic searches for texts of particular characteristics (see Lehmberg and Wörner 2008: 492-498; Thieberger and Berez 2012: 105-109).

2.3 Annotation

An annotated corpus is a corpus in which the primary data are enriched by further linguistic information such as the indication of parts of speech (Kübler and Zinsmeister 2015:4, 13). LD corpora are necessarily always annotated corpora because the audio and video recordings are transcribed, and all texts need to be translated into a language of wider communication (see §6).

2.4 Genre and register

Language usage varies not only regionally and socially across dialects and sociolects, but also within such varieties. This kind of variation is systematic as the selection of certain linguistic features depends on non-linguistic characteristics of the speech situation, see Table 1, which is a shortened and modified version of Table 2.1 in Biber and Conrad (2009:40).

We can distinguish between two perspectives of text classification: the distinction of genres and the distinction of registers. A genre is a type of text that has a more or less conventionalised structure which is indicated by certain formal features such as speech formulas at the beginning or the end. Compare the following two examples:

(1) *Dear Sir ..... With kind regards, yours sincerely John Smith*

(2) *Once upon a time ... lived happily ever after*

Obviously (1) and (2) represent the beginning and the end of the two English genres of letters and fairy tales.

Table 1: Situational characteristics of registers and genres

kind of metadata	examples
date	ideally the day of the recording, but if this is not clear, an approximate date must do
setting	private / public
participants	origin, age, gender, social status of speaker(s) and addressee(s) and their relationship; the name of the recording person;
mode of communication	speech, writing, signing
medium	audio or video recording, handwritten, printed
production circumstances	spontaneous, planned, edited version of a transcription
purpose	general purposes: narrate, describe, explain, entertain
topic	general topic domain, e.g. history, plants; material culture

Registers on the other hand are types of text which occur in certain speech situations and show certain phonological, grammatical and lexical features throughout any text of this type with a significantly higher frequency than other text types. In the German language of Radio or TV news you hardly hear any clitic pronoun, whereas they are frequent in casual conversations (3):

(3) German

written German:	*dann haben sie...*
spoken German in the news:	[dan habən zi:]
spoken German in casual conversations:	[dan ham=zə]
English translation:	‘then they have’

Texts that are ascribed to the same genre may belong to different registers depending on who speaks to whom in which situation. If, for example, fairy tales told by mothers to their children differ in phonological, grammatical and/or lexical features from fairy tales read by teachers to students, they would be ascribed to different registers.

Genres and registers are language and culture specific categories though they may be comparable across cultures if theses make use of conventionalised ways of speaking in comparable situations For further information on corpus linguistics and the ethnography of speaking see §8.3.

3 Some special features of language documentation (LD) corpora

LD corpora are corpora of under-researched languages and, consequently, quite different form corpora of so-called major languages with respect to their purpose, production, content, and size, as summarised in Table 2.

Table 2: Language corpora

	typical corpora of major languages	language documentation corpora
purpose	lexicography; research in theoretical and applied linguistics such as automatic translation and language teaching;	conversation of the cultural and linguistic heritage; educational materials; research on language and culture; (Mosel 2012b, Seifart 2011)
production	selection of previously existing printed and digitalised texts;	simultaneous linguistic research and corpus building;
languages	mostly monolingual;	at least bilingual;
compilers	team of professional native speakers;	a non-native speaker linguist with no or little experience in corpus building and a team of native speakers without any knowledge of linguistics and corpus building (see §4.1);
content	the content depends on the purpose of the corpus;	the content depends on the field situation and the purpose (see §4.2);
size	millions or billions of words.	much below half a million of words.

Most language documentation corpora, especially those archived by DOBES (Dokumentation Bedrohter Sprachen) archive and ELAR (The Endangered Languages Archive), are built with the multimedia annotation tool ELAN in combination with Toolbox (Field Linguist’s Toolbox) or FLex (FieldWorks Language Explorer). Guidelines for downloads and applications of these tools are found on the respective websites (see §11.2).

4 The collection of texts in language documentation projects

4.1 Collaborative fieldwork

In recent descriptions of fieldwork methods the authors emphasise the importance of collaborative language documentation. This implies that you are a guest in the speech community and that the people who invited you select the local research assistants and decide with them on the content, the format, and the accessibility of the texts that are recorded and processed in your project (see §8.1 on ethics and §8.4 on fieldwork). Since the quality of data collected for the corpus depends on your interaction with the local research assistants, you have to adapt to their various talents and train them in tasks they enjoy and can cope with. Even semi-speakers can be helpful.

If in the very beginning of your research you do not know the language, you ask local research assistants to teach you expressions and help you to record, transcribe and translate them into the contact language. These lists of expressions may serve as the first data for a description of the phonology and, if the language hasn't been written yet, as the basis for the development of a practical orthography (§6.4). At the same time they form the first texts of your corpus and require metadata (see §2.2; §4.3).

During the first recording sessions you can teach local research assistants to use the recording equipment so that they can make recordings without yourself being present. This will not only motivate them to actively participate in the project and take responsibility, but also result in recordings of a more natural way of speaking because in your presence speakers may use a more formal variety of speaking or even a kind of foreigner talk.

The transcriptions (§6.4) and translations (§6.5) of the recordings should be done with the help of native speakers during your fieldwork to make sure that you have understood everything correctly before you leave the fieldwork site. If the research assistants do the transcriptions by themselves, ask them to transcribe what they actually hear including false starts, hesitation phenomena like *ah, what's its name?* or repetitions, and not what they think would be the best way of saying things. If they want to edit the transcription, they should be encouraged to do so, but this edited version would then be a new text of a different register (§4.2.2).

If the research assistants are not able to write transcriptions or hesitate to do so for various reasons, you can ask them to listen to the recording utterance by utterance and slowly repeat what they hear. You can either immediately do the transcription or better record their repetitions and then do the transcription on the basis of these recordings, which can also become text files of the corpus with their own names and metadata. Later they could become a resource for research on the phonological differences between fast and slow speech.

In the beginning of your fieldwork you should be content with the recording of any kind of text you can get, but later I recommend to focus on a few genres and registers. The more genres and registers your corpus contains, the fewer texts you have of each genre or register so that it becomes difficult to identify linguistic features that are characteristic for a particular genre or register.

4.2 Types of text

4.2.1 Indigenous and non-indigenous genres

In all speech communities the most frequent use of language are spontaneous conversations in the family, at work and among friends and neighbours, but these are not only difficult to transcribe, they may also be considered as less suitable by the speakers for publication than folk-tales, personal narratives, and descriptions of traditional practices and customs.

Folk-tales are a good text type to start with because they have a conventionalised beginning and content so that speakers speak fluently without too many hesitation phenomena which would make the transcription and translation difficult. But ask your hosts to discuss what they regard as interesting for their language documentation. Are there other traditional genres such as poetry, oratory, proverbs, children's rhymes, riddles, or jokes they want to document? What about historical events; customs or the knowledge of plants and animals that they want future generations to remember? As the preceding questions reflect the standard European classification of genres, you should also ask the speakers if they have terms for different ways of speaking, and if they do, ask them for examples and explanations (see Senft 2010 for an outstanding example).

Personal narratives about past events and descriptions of customs, the natural environment, the material culture, and daily activities may not be an indigenous genre, but their content and form are a valuable resource for preserving the cultural memory of the community and for further linguistic or anthropological research. For example, it was only the descriptions of trees that showed us that Teop is not a subject-verb-object language as we first thought, but a verb-second language, because numerous clauses about trees and what they are used for started with a topical object referring to a tree or its parts rather than an agentive subject (Mosel 2014:151-153).

4.2.2 Edited texts

The transcriptions of audio recordings may not be suitable for educational purposes so that the authors or other members of the speech community want to edit them. Representing a distinct register, these edited versions form a special sub-corpus that may also be an interesting resource for linguistic research, because they show what people actually do when transforming more or less spontaneous speech into a written language.

When the Teop LD project started to edit folk-tales, the editors were asked to keep as closely as possible to the transcription of the original recording; only remove hesitation phenomena and speech errors, and only add words, phrases or sentences where they were absolutely necessary for the readers' comprehension. Nevertheless the speakers did many lexical and syntactic changes which show alternative ways of expressing the same content and thus provide a new type of data for research on the preferred lexical and grammatical features of spoken and written registers in comparable texts (Mosel 2012b, 2015).

To integrate edited texts into an ELAN corpus, you can first type them in Toolbox or FLex and then import these files into ELAN, or you ask a native speaker to read them, so that you can create ELAN files with the audio recordings. But note that the reader may more or less unconsciously change the text while reading it (see §6.8).

Another, but completely different kind of editing can be done when annotating an audio recording. When the transcription shows stuttering and self-corrections, you can add a separate annotation tier on which stuttering or false beginnings are removed and thus create a more reader-friendly version of the same text without changing the speaker's intentions.

4.2.3 Elicited texts

The recent developments of video technology and multimodal corpus building tools facilitate new methods of text collection as, for instance, the use of video clips as elicitation stimuli; see Hellwig 2019; Majid et al. 2007; Majid 2012, and the field manuals of the Max-Planck-Institute of Psycholinguistics (see §11.2). As elicitation in general, this method only provides the kind of data the researcher had in mind when preparing the elicitation and misses out on unexpected phenomena (see §4.2.5, §8.4). Furthermore, if the video clips are not produced in the speech community whose language is to be documented, they contain things and activities the local research assistants may not be familiar with. There are, for example, many ways of cutting up things; and if the video shows a way that is not practised in the speech community, how will you know what the speaker exactly means, when he uses one of the cutting words of his language? You first need to know how the native speakers cut up things and make pictures or videos of the event, then you ask them what this event is called, and also ask them, if they could explain in their own words what the word means. For a critique of elicitations by video-clips, see Haviland (2006:154) and Wierzbicka (2014:40-47), but such elicited texts can be a valuable resource if they are recorded to complement natural texts (cf. Hellwig 2006, Koptjevskaja-Tamm et al. 2016:442).

A further method of using video was practised by Margetts (2011) who filmed events in the community and later asked community members to comment the videos while watching them. This method can, of course, be applied for all sorts of community events or staged events, in which members of the speech community show how they perform certain activities. When filming an event such as butchering a pig is too complicated because you need to quickly change perspectives to capture different ways of cutting, a series of photographs will also do for elicitation, the lexical database and a PDF file or printed text, but they cannot be integrated into an ELAN file.

One and the same video or series of photographs can be used to elicit texts of distinct genres. For example, you can make a video or a series of photographs while people are butchering a pig. Then you show the video or the photographs to some people and either ask them to tell you what the people did while you were doing the video or the pictures, or to explain to you how pigs are butchered, so that you get narratives and procedural descriptions. Such texts are interesting for the grammatical comparison of narratives and procedural texts, because they are about the same kind of event and its sequence of actions (Mosel 2014:146-149).

Some linguists use children's picture books like the frog story (Mayer 1969) for the elicitation of narratives (Bowern 2008:116).While looking at the pictures, the speakers are asked to tell the story they see. But for LD projects this method cannot be recommended. When working with the Watam people of Papua New Guinea, Foley (2003) observes that the audio-recorded texts prompted by *Frog, where are you?* (Mayer 1969) showed different syntactic structures from traditional narratives, which probably resulted from the fact that retelling a story that is told by pictures is cognitively different from telling a story by memory. A further, stronger argument is that the elicited frog stories haven't anything to do with the speakers' language and culture and that their production is only motivated by the linguists' interests. The same holds true for the use of *The Pear Film* by Wallace Chafe which has been widely used for research on European and non-European languages (Chafe (ed.) 1980). For some critical remarks see Himmelmann (1998:187), for a website of *The Pear Film* see §11.2.

4.2.4 Written texts

Spoken texts are definitely preferred by linguists, but it may happen as in the Teop project that the local research assistants get tired of doing recordings, transcriptions and editorial work, while at the same time they gain confidence in writing stories or descriptions. As in the case of edited texts, such written texts can be integrated in the corpus as a special sub-corpus.

4.2.5 Elicited words, sentences and paradigms

The elicitation of words by using wordlists in English or any other language of wider communication and asking bilingual speakers to translate them is only suitable in the very beginning of a LD project when linguists and local research assistants start to analyse the phonology. As soon as the meanings of words come into play, the words need to be embedded into contexts, because usually both the word of the language of wider communication and the word given as its indigenous translation equivalent have more than one context-dependent meaning. Since, consequently, misinterpretations can hardly be avoided, non-translational elicitation methods have to be applied (see Mosel 2012a: 79-85 for an overview of elicitation methods and §8.4).

When translating texts, you frequently come across words that are new to you. Since a single context is not sufficient to fully understand the meaning of a word, you can prepare lists of new indigenous words and your research assistants can discuss their meanings with other native speakers and write or record typical examples. These examples represent a special genre. The lists of examples may form text files of what Ostler (2008) calls an "artificial" corpus. Artificial corpora or subcorpora "are artificial in the sense that the material has no social or cultural rationale for being collected. Such corpora can be generated by systematic elicitation ..." (Ostler 2008:459). For each speaker or writer you create a separate text file with metadata, because they may develop their individual styles of creating examples.

4.3 Metadata

In LD projects the metadata of the whole corpus gives information about

- the language, i.e. its genetic classification and geographical distribution;
- the dialect(s) and sociolect(s), genres and registers represented in the corpus;
- the approximate number of native speakers and their use of other languages;
- cultural and sociolinguistic characteristics of the speech community;
- the research team including local research assistants;
- the dates of the research project;
- the functions of the corpus within the research project;
- the methods of text collection;
- the number of speakers and writers;
- the recording tools and the software;
- the size of the corpus and its subcorpora;
- the script and orthography used in the corpus;
- the language or languages of the translations;
- the accessibility.

These metadata are included in the introduction of the corpus (see §5.1).

The metadata of individual texts concern their production circumstances, length, and special characteristics of their content, because the use of phonological, lexical and grammatical features varies across registers and genres. The most important characteristics are summarised in Table 1 in §2.4. But how the metadata of your corpus files have to be presented depends on the data management system of the archive where you want to deposit your corpus and the metadata entry tool the archive requires you to use to facilitate searches within and across corpora. Thieberger and Berez (2012:98-102) therefore recommend to plan for data management before you collect your first data.

When doing the recordings in the field, you, your research assistants, or the speakers themselves can start each recoding by saying the date, the place, the topic, and the names of the participants or, if they want to be anonymous, a nick name or number that allows you later to distinguish the participants. In my experience it can be difficult to get an ideal set of metadata because in the actual field situation asking personal questions may seem intrusive or impolite. In any case, you should note down the available metadata as soon as possible after the data have been collected, ideally, if you have electricity, by an electronic metadata tool (Thieberger and Berez 2012: 105-107). For further information see §8.4.

4.4 Naming of texts - file names

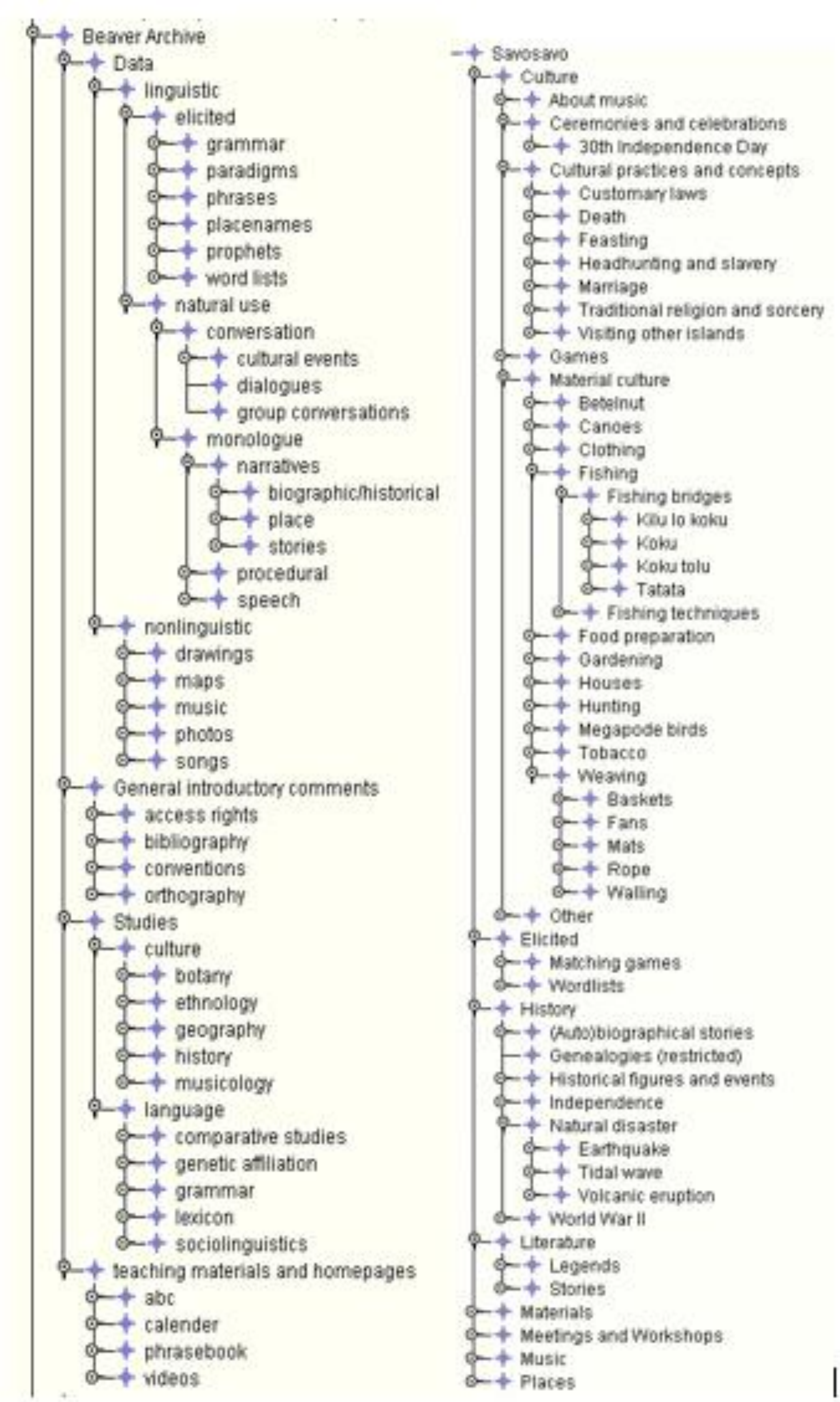
Each file of the corpus is given an individual name to distinguish it from the other files. Some archives require a particular format of file names. If there aren't any requirements, a useful practice is to give them a label that contains some metadata information about the medium, the speakers, the genre or register. Such information is essential if these annotations are quoted as examples in a grammar or dictionary. Reppen states:

> “File names that clearly relate to the content of the file allow users to sort and group files into subcategories or to create subcorpora more easily. Creating file names that include aspects of the texts that are relevant for analysis is helpful.” (Reppen 2010:33)

Nordhoff (2009:63), for example, uses abbreviations for the place of the text collection, the date of the recording and the genre, e.g. “nar” for narrative, “wrt” for written text, and “rec” for recipe. For further examples see §6.3.

5 The structure of corpora in language documentations

Fig. 1: The Beaver and the Savosavo documentations



5.1 The introduction to the corpus

Besides the metadata of the corpus as a whole (see §4.3), the introduction should as any kind of linguistic publication contain the following items:

- acknowledgements of the help the corpus compilers have received from other people and funding agencies;
- references to grammars, dictionaries, other books or articles that are in one way or the other related to the LD project in question.

The fact that the LD corpus represents an under-researched language also requires the inclusion of a sketch grammar (Mosel 2006b) that at least informs the user about

- the typological profile of the language;
- the orthography by presenting a table of the orthographic characters and the corresponding IPA symbols;
- a table that lists the abbreviations of glosses and explains their meanings.

5.2 The corpus and its subcorpora

The structure of the corpus depends on the technical format and the requirements of the archive (see §4.3, §7). If there are no requirements of how to organise the corpus as in the case of the DoBeS archive, the corpus compilers may divide their corpus into subcorpora. Some projects have their corpus divided into subcorpora according to the year of recording, but for users who are interested in particular aspects of the documented language and culture it is certainly more user-friendly when the texts are divided into subcorpora according to their contents.

Fig. 1 below shows the structure of the Beaver Archive and the Savosavo Documentation. Beaver is an Athabaskan language spoken in Canada, Savosavo a non-Austronesian language spoken on the Solomon Islands (see §11.2 for the websites). The two languages and cultures as well as the documentation projects are different, but both language corpora are hierarchically structured and contain some subcorpora of similar contents, see Table 3.

The elicited texts can be further divided into texts that are responses to questionnaires and stimulus based elicitations, whereas the non-elicited ones are divided according to topics, genres, and registers (see Table 1, §2.4).

Table 3: Similarities of the structure and content of the Beaver and the Savosavo corpus

Beaver	Savosavo
stories	legends, stories
biographic/ historical narratives	history
procedural texts, i.e. texts describing how something is done	texts about fishing, food preparation, weaving etc.
elicited wordlists	elicited wordlists

The division between elicited and natural texts should become a standard, while further divisions depend on the purpose and the size of the corpus and its subcorpora. The Beaver project, for example, separates monologues from conversations, whereas such a division in missing in the Savosavo project. But the Savosavo corpus structure is much more complex with respect to texts about the material culture.

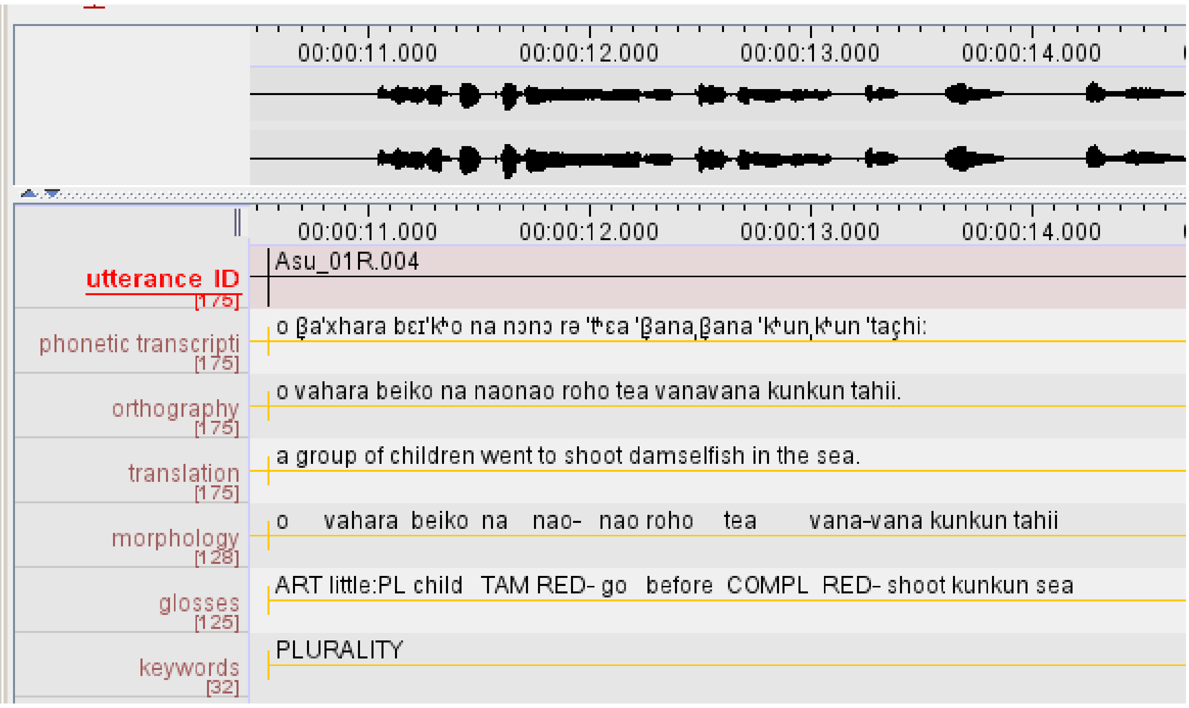
6 Annotations

6.1 Introduction

In a LD corpus in ELAN the annotations (see §2.3) are entered on tiers with each tier containing a particular kind of information as illustrated in Fig. 2, which shows an annotation unit of an audio-recorded Teop folk-tale.

The first tier called Utterance ID was created by segmenting the sound file into annotation units so that each unit of this tier is directly linked to a time interval of the media. Then the annotation units of this tier were automatically labelled by the name of the text and numbered. The subsequent tiers are associated with the Utterance ID tier so that their annotation units exactly match with the units of the Utterance ID tier.

Fig. 2: Annotation in ELAN



utterance ID [175] Asu_01R.004

phonetic transcripti [175] o β̞a'xhara bɛɪ'kʰo na nɔnɔ rə 'ʧɛa 'β̞anaˌβ̞ana 'kʰunˌkʰun 'taçhi:

orthography [175] o vahara beiko na naonao roho tea vanavana kunkun tahii.

translation [175] a group of children went to shoot damselfish in the sea.

morphology [128] o vahara beiko na nao- nao roho tea vana-vana kunkun tahii

glosses [125] ART little:PL child TAM RED- go before COMPL RED- shoot kunkun sea

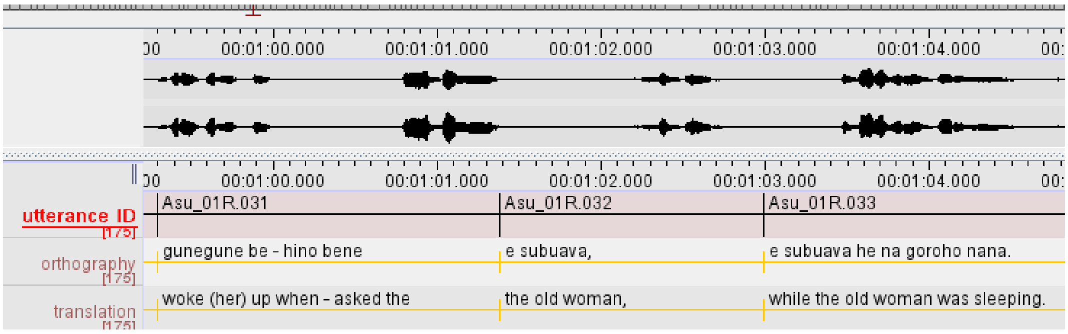
keywords [32] PLURALITY

The following sections do not describe how to use ELAN and Toolbox or FLex, as this information is given in the respective guidelines (see §11.2 for references), but focus on some practical issues that are not addressed in the guidelines of these tools. Due to my personal lack of experience with the annotation of gestures as documented in video recordings, this section only deals with the annotations of audio recordings (for the documentation of gestures see §8.6)

6.2 Segmenting the sound file

The first task of annotating the recording of a communicative event is to segment the flow of spoken language into intonational or grammatical units so that the transcription and translation can be aligned to the media file in a transparent way as shown in Fig. 2. Without a segmentation or with very large segments, it would be difficult to figure out how phonological, lexical or grammatical units of the sound file relate to the transcription and the translation. Furthermore, in corpus-based phonological, grammatical or lexical analyses you could only refer to particular data by indicating the time interval in terms of seconds and milliseconds, which would be very labour-intensive.

Fig. 3 Segmentation without the indication of pauses in the Teop corpus



Asu_01R.031 | Asu_01R.032 | Asu_01R.033

gunegune be - hino bene | e subuava, | e subuava he na goroho nana.

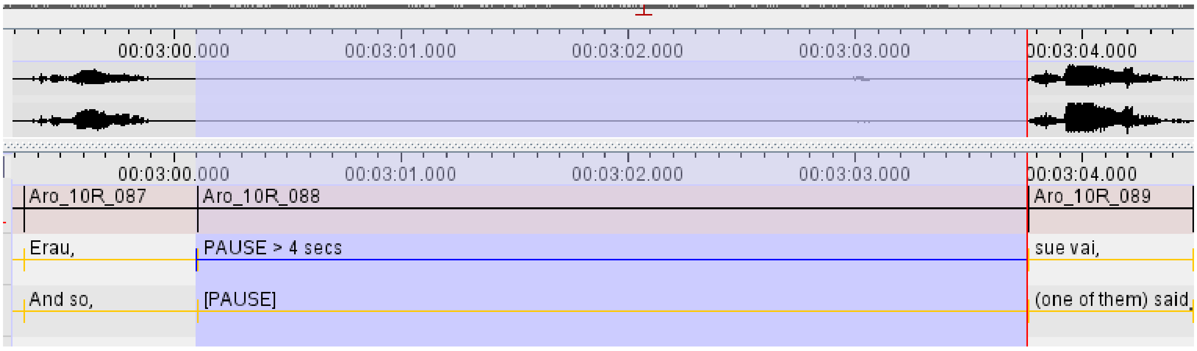
woke (her) up when - asked the | the old woman, | while the old woman was sleeping.

The intonational unit is, as Himmelmann (2006b:260) observes, “widely held to be the basic unit into which native speakers themselves chunk their utterances, i.e. it is seen as a unit of speech production which in some sense has a psychological reality for the speakers (as opposed to a purely analytic construct “invented`" by linguists).” Consequently, a segmentation of spoken texts by phonological boundaries that are marked by pauses, pitch and rhythm seems to be more adequate than a grammatical segmentation. However, “when transcribing spontaneous speech in a language one understands very well, there is a strong tendency for semantic and syntactic factors to interfere with one’s perception of prosodic boundaries” (Himmelmann 2006b:267).

This means that transcriptions done by native speakers in the field may be inconsistent and need to be checked and revised, if you aim at a consistent transcription of intonation units. But you also may consider the corrections of these interferences as too time consuming for the time being and leave them for later.

If your own research focuses on grammar, you probably prefer to segment the texts by grammatical criteria and have clauses or sentences as annotation units, but still the transcription should include false starts, pauses and hesitation phenomena, because these may be relevant for later research. With respect to pauses, the segmentation shown in Fig. 3 is not ideal; it would have been better to indicate the pauses and their lengths as shown in Fig. 4.

Fig. 4: Segmentation with the indication of a pause in the Teop corpus



Aro_10R_087 | Aro_10R_088 | Aro_10R_089

Erau, | PAUSE > 4 secs | sue vai,

And so, | [PAUSE] | (one of them) said,

6.3 Labelling of annotations

In Fig. 2 - Fig. 4 you see that each annotation segment (or unit) is labelled. The labels contain the name of the text (see §4.4) and the number the annotation unit has in this text. Thus Aro_10R.088 means ‘the 88th annotation of 10th recording of a spontaneously spoken text by Arovina’. Correspondingly, annotations of the edited version have the label Aro_10E plus their number, whereas the labels for written texts without a previous recording have W instead of E, e.g. Naph_02W_021-022 in (4). These labels and numbers are used when annotations are quoted in articles or books.

6.4 Transcriptions

Schultze-Berndt (2006:219-232) gives an excellent overview of the various types of transcription with references to further literature. In particular, she discusses the orthographical, phonemic, phonetic, and prosodic transcription, the transcription of paralinguistic and non-linguistic aspects of communicative events and, eventually, the transcription of multi-speaker and multilingual discourse. Therefore, a few remarks should suffice here.

With respect to the aim that the texts must be readable by community members (provided that at least some of them know how to read their language or a language of wider communication), the corpus should have a transcription tier in a script and orthography that is understood and accepted by them. If this is a non-Roman script, you add a transliteration or transcription in Roman script on a separate tier for other people who might be interested in this language documentation.

The orthography does not need to be officially standardised, but consistency is desirable, especially for texts that are edited for educational purposes. But if local research assistants occasionally use various spellings, it may be more efficient to accept spelling variants and avoid endless debates that are often of a more political than a linguistic nature. The development of a standard orthography which has to deal with numerous linguistic and non-linguistic factors (see §8.6), is too time consuming to be done in a general language documentation project with a duration of a few years. In the lexical database all spelling variants should become headwords (lemmas) of lexical entries with a cross-reference to the most frequently used variant.

Since the introduction to the corpus should contain a description of how the orthography is pronounced (see §5.1), the corpus ideally contains audio recordings of wordlists or short texts that are transcribed in the practical orthography and the IPA (International Phonetic Alphabet) as illustrated by the text segment in Fig. 2. The orthographical transcription of this text was done by a native speaker and then independently checked by two other native speakers, the IPA transcription by Radtke (2005), a master student of phonetics with a special training in transcriptions, who did not know Teop. In his IPA transcription he kept the word boundaries, i.e. the spaces between orthographical units, so that you can easily figure out that the intonational units marked by stress do not match with the orthographical units.

The orthographical transcription in Fig. 2 shows that the transcription done by a native speaker may not represent the actual phonetic form of a word. The phonetic form [rə], for example, is transcribed as *roho* which is the form of slow speech and written texts. That the transcriber renders [rə] by *roho* is not a mistake, but shows that he consciously or unconsciously perceives [rə] as the realisation of the word *roho* on the basis of his metalinguistic knowledge. For the notion of metalinguistic knowledge and its role in LDs see Himmelmann 1998 and Jung and Himmelmann 2011.

6.5 Translations

The choice of the target language or languages (i.e. the languages into which the indigenous language is translated) depends on the knowledge of the compilers and the prospective users of the corpus so that translations into more than one language could be desirable or even requested. The translation of a text must be exactly aligned to the annotation units of the transcription or written text to show the form-meaning correspondences on utterance, sentence or clause level. If the constructions of the source and the target language are very different, it may be helpful to supplement the free translation by a literal translation on a separate tier. In some cases this is necessary even for glossed texts. Consider the following quotation:

(4) Teop

o=re	*suguna*	*komana,*
3SG.PRON=CONSEC	arrive	indeed

‘until’
lit. ‘then it indeed arrives’

be=o	*iana*	*bero*	*va-antee*	*bono*	*rake*	*te=*	*ori*
COMPL=ART	fish	many	CAUS-enough	ART	wish	PREP=3PL.PRON	

‘they have as many fish as they want’
lit. ‘that the fish are many sufficing their wish.’ (Naph_02W_021-022)

The literal translation shows that there are hardly any correspondences between the Teop and the English syntactic units:

- the English conjunction ‘until’ corresponds to the first Teop clause *ore suguna komana*, lit. ‘then it indeed arrives’;
- the subject-predicate sequences ‘they have’ and ‘they want’ translate the object determiner phrase *bono rake teori* lit. ‘their wish’; and
- the English sequence ‘as many fish as’ corresponds to the Teop subject *iana* ‘fish’ and the predicative transitive serial verb construction *bero vaantee* ‘(are) many sufficing’.

A further, less complicated problem of translating texts of an under-researched language and culture is that you don’t know the translation equivalents of certain plants, animals or cultural items. But if you know what kind of higher order the plant or animal belongs to, you can simply form determinative compounds with the higher order term of the target language as the head and the indigenous term as the modifier as in (5).

(5) Teop

O	*kururu*	*na*	*vaapeha*	*mi*	*nana*	*bono*	*koverau.*
ART	kururu.bamboo	TAM	similar	with	TAM	ART	koverau.bamboo

‘The kururu bamboo is similar to the koverau bamboo.’ (Sii_40W_022)

The translation of culture specific terms may follow a similar strategy by using a suitable generic term as the head of a determinative compound and the indigenous term or a descriptive expression as a modifier, see Table 4.

Table 4: The translation of culture specific terms by determinative compounds

Teop word	English translation	dictionary definition
maurata	‘maurata girl’	girl who has her first menstruation and is going through a particular ritual which forces her to stay hidden in a particular place
upee	‘upee boy’	boy going through the initiation rites and wearing the upee hat
koopu	‘bamboo knife’	indigenous knife made of bamboo
rapisi	‘bush knife’	bush knife; machete

Note that the translation equivalent ‘bamboo knife’ would not be a suitable dictionary definition because ‘bamboo knife’ could be misinterpreted as ‘knife for cutting bamboo’.

6.6 Morphological segmentation and glossing

6.6.1 Introduction

The translation of an annotation unit only renders the meaning of the whole unit, but not the meanings of words and bound morphemes as illustrated by Fig. 2 above and the following German example (6).

(6) German

den Müttern der Kinder

‘to the children's mothers’

To show the form-meaning correspondences on word level, you automatically tokenise the transcription into words, i.e. the units that are separated by space in the transcription, on a separate tier so that your original transcription remains intact. Then you segment complex words into morphs and add an associated tier for the morphological glossing, see Fig. 2 and (7)

(7) German

den	*Müttern*	*der*	*Kinder*
den	Mütter-n	der	Kind-er
ART. PL.DAT	mother.PL-DAT.PL	ART.PL.GEN	child-PL

‘to the children’s mothers’

In general and as illustrated by the preceding example (7) and Fig. 2, there are two kinds of glosses: 1. the glosses of lexical items, e.g. *mother*, *child*, and labels that give information about grammatical properties, e.g. ART, PL, DAT, GEN.

6.6.2 Grammatical glosses

Most linguists follow “The Leipzig Glossing Rules” (Bickel et al. 2004), which set standards for “the presentation of an example in research papers and books”. In particular they are concerned with the annotation of morpheme boundaries, abbreviations for functional words, e.g. ART ‘article’, and grammatical categories like DAT ‘dative’. Morpheme boundaries are indicated by a hyphen on the morphological and the glossing tier, e.g. *kind-er*; ‘child-PL’. But if a word form fuses the lexical item with a grammatical feature as in *Mütter* ‘mothers’, so that a segmentation is impossible on the transcription tier, the glosses are separated by a period, e.g. ‘mother.PL’ in example (7).

There aren't any two languages in which grammatical elements like articles or case markers exactly share all their semantic and functional properties so that you might think of giving some categories new names. But for mnemonic reasons, it is more user-friendly to use traditional names and briefly explain their language specific characteristics in the list of abbreviations in the introduction to the corpus.

The glosses can be more or less detailed, which depends on the type of language, the purpose of the documented texts, the stage of your grammatical research, the time planned for morphological annotations, and the grammatical and lexical material accompanying the corpus. Consequently, the glosses of LD texts may be different from the glosses in a specialised linguistic publication. Lehmann (2004:1844) lists 16 labels for glosses that are to be avoided, among them ART article’, ASP ‘aspect’ and TNS ‘tense’; instead you are supposed to use abbreviations like DEF ‘definite’, INDEF ‘indefinite’, SPEC ‘specific’, and NSPEC ‘non-specific’ for determiners and abbreviations for specific aspect and tense categories like IPFV for ‘imperfective aspect’ and PST for ‘past’. These recommendations, however, can only be followed after determination, aspect and tense categories have been analysed on the basis of the corpus. Therefore it is justified to use underspecified labels like ART as glosses in LD texts. But also in a linguistic publication it may be more user-friendly to use an underspecified abbreviation like ART as in the Teop examples (4) and (5) above, if a more specified abbreviation would not contribute to the argument that the example illustrates, but rather confuse the reader because of its complexity. In the determiner phrase *o iana* ‘the fish’ in (4) the article *o* marks specificity and plurality and also indicates that *iana* ‘fish’ belongs to the second noun class, but in our discussion about free and literal translations these details are irrelevant, see Table 5.

Table 5: Alternative glossing for *o iana* 'the fish' in example (4)

practical glossing		glossing according to Lehmann 2004	
o	*iana*	*o*	*iana*
ART	fish	SPEC.PL.CLASS2	fish
‘the fish’		‘the fish’	

If the grammar of the language has not been analysed yet, there may be morphemes the meaning of which is unclear. In this case you simply use the question mark as a gloss.

6.6.3 Glosses of content words

As for content words, Lehmann (2004:1841) recommends to have one and the same gloss for a lexeme irrespective its various context-dependent senses, because otherwise "it will confuse" the reader, "if Yucatec Maya *k'ìin* is once glossed 'sun' and the next time 'day'". Ignoring context dependent polysemy is also the fastest way for the half-automated glossing with Toolbox and FLex. But against Lehmann's approach it may be argued that it is less user-friendly for the reader who wants to see at a glance how the free translation matches with the glossed words of a clause or utterance unit. Furthermore, without a thorough analysis of all occurrences of a particular lexical form in the corpus, it may be difficult or even impossible to unequivocally decide which of the various translation equivalents of a lexical form, if any, would be the most appropriate one as a gloss in all contexts. We even do not know if the various translation equivalents of a particular lexical form represent distinct senses or if this lexical form is monosemous. It only may require different, context-dependent translations, because the target language - English, for example - lacks a suitable more general term. Evans warns:

> "Language particular studies of polysemy need to ensure that we are not dealing simply with monosemy, in the form of categories which are unitary from an emic viewpoint but which happen to involve more than one translation equivalent into English or some other metropolitan language of investigation." (Evans 2011:524)

The Teop word *toon* is a case in point. Table 6 shows five examples of *toon* with four distinct English translation equivalents: 'back, top, roof, surface'.

Table 6: The Teop noun *toon* with possessive attributes and their translation equivalents

Teop *toon* in different contexts	English translation
a toon ne sumeke	'the old man's back'
a toon na biroo	'the back of the lizard'
a toon na teevoro	'the top of the table'
a toon na inu	'the roof of the house'
a toon na tahii	'the surface of the see'

Do the four translation equivalents ‘back, top, roof and surface’ exactly match with four Teop senses? Do *toon* ‘back’ in ‘the old man’s back’ and ‘the back of the lizard’ represent a single sense, because both constructions denote a body part? Or do ‘the back of the lizard’, ‘the top of the table’, ‘the roof of the house’ and ‘the surface of the sea’ belong together, because in contrast to ‘back of a human being’ they denote the top of horizontally extended entities? Is there a basic or a general meaning of *toon* and a corresponding English word that could be used as a gloss instead of the four translation equivalents? As it seems impossible to spontaneously answer such questions when glossing a text (or reading such examples), I recommend to simply gloss content words by their translation equivalents and accept that the time-consuming semantic analysis has to be postponed. For a more detailed discussion of the difficulty to capture the meaning of lexical items in LD projects see Haviland 2006.

6.6.4 Summary: Two kinds of glossing

We need to distinguish between the glossing of examples in linguistic research publications and the glossing of texts in LD corpora. While the selection of the former type is based on a thorough analysis of lexical and grammatical data and is determined by the aim of the publication, the latter only reflects a preliminary analysis and serves as a tool for further corpus-linguistic research. But both of them have in common that the grammatical glossing should be consistent. Since without consistency searches result in inadequate data, it is recommended to have a list of glosses for grammatical categories to control the glossing. Otherwise one might, for instance, accidentally, gloss one and the same case marker first as LOC ‘locative’, but later as OBL ‘oblique’.

6.6.5 How many glossed texts does a LD corpus need?

Assuming that LD corpora can be submitted to a publisher for assessment and subsequent publication, Thieberger, Margetts, Morey and Musgrave argue from the perspective of linguists (Thieberger et al. 2015:15)

> “A substantial part of the corpus will be expected to be annotated with morphological breakdown and interlinear glosses. ... Corpora of transcribed and translated recordings without any further annotations relating to morphological breakdown and interlinear glossing can be submitted but will be considered to be of comparatively low quality ranking that it will be considerably more difficult for third parties to work with such corpora and given that they represent a lesser degree of linguistic analysis and annotation.”

Even on the assumption that LD corpora are exclusively exploited by linguists, the criterion that “a substantial part of the corpus” must be morphologically annotated is questionable. Firstly, it is unclear, how a “substantial part” is defined, and secondly, a corpus of an isolating language which is accompanied by a comprehensive dictionary does not need the same kind and amount of morphological annotations as a morphologically complex language without a dictionary. For further comments on the evaluation of LD corpora see §7.

6.7 Syntactic annotation in GRAID (Grammatical Relations and Animacy in Discourse)

An advanced glossing system called GRAID (Grammatical Relations and Animacy in Discourse) has been developed by Haig and Schnell (Haig and Schnell 2014) and is presently used in 11 small corpora of 4 Indoeuropean, 5 Austronesian, 1 Causasian and 1 Papuan language, see https://multicast.aspra.uni-bamberg.de/ .

GRAID facilitates quantitative morpho-syntactic analyses of

- the encoding of referential expressions with respect to their form, their grammatical and semantic categories, and their syntactic functions,
- the form of their predicates,
- the internal structure of NPs and predicates, and
- various types of dependent clauses and their syntactic functions.

With respect to referential expressions, GRAID distinguishes, for example, between

- NPs, independent, weak, cliticised and suffixed pronouns, and zeros, i.e. argument positions that are not filled by an overt referring expression,
- human, non-human, or anthropomorphised referents,
- 1st, 2nd, and 3rd person, and
- the syntactic functions of intransitive subject, transitive subject, and transitive object.

GRAID annotations are time consuming, but the annotator is rewarded by the possibility to answer a wide range of linguistically essential questions in a scientifically valid way for both language specific and cross-language typological and historical research (see Haig and Schnell 2016; Haig et al. 2011).

6.8 Notes or comments

A tier of notes allows to add various comments on particular annotations on the transcription, the translation, or any other kind of tier. For example, when you carefully listen to a text and realise that a sequence of words has been transcribed differently from what you hear, you can make a note like “we do not hear *vakokona*, but something like *vo koara,*” so that you won't change the transcription, but show that you and others have some doubts about its correctness.

If you have files of written texts that are read by a native speaker, it may happen that the reader makes some changes. In this case I recommend to make a note that shows the reader's change rather than changing the written text for two reasons. Firstly, you cannot change the original text without the writer’s consent, and secondly, you can later easily search for all the reader’s changes on the notes tier, if these notes have, for example, the form: “reader: ...”. The reader’s changes may be interesting for phonological or grammatical research.

As already mentioned in §6.4, the notes tier can also be used for literal translations. If they are consistently introduced by "lit.", you can later search for them and thus get a collection of examples for annotation units that show remarkable differences between the constructions of the source and the target language. For further examples see Table 7.

6.9 Keywords

The term keyword is used here in the sense of a word that represents an interesting research topic and leads to annotations that are difficult to find otherwise, because their content does not manifest itself in an easily searchable linguistic form on the transcription and/or the translation tier. A typical example are idiomatic phrases that denote emotions, see Table 7.

Table 7: Two Teop examples for the expression of emotions

Tier	example 1	example 2
transcription	*na koma hata ni nana*	*enaa na tii me nom o vuha ponis*
translation	‘is angry with’	‘I am worried.’
lit. translation	‘have a bad belly to’	‘I am with a heavy breath.’
morphology	*na koma hata ni nana*	*enaa na tii me nom o vuha ponis*
glossing	TAM belly bad APPL TAM	1SG TAM be with TAM ART breath heavy
keyword	emotion	emotion

Other examples are complex grammatical constructions like nominalisations, unmarked complement clauses, or serial verb constructions, or if you are interested in zero-anaphora, but don't want to apply the time consuming GRAID annotation, you may indicate zero anaphora simply by “zero” on the keywords tier.

For the users of your corpus (including yourself), it is important to have a list of the keywords in the introduction to the corpus.

7 Concluding remarks

The preceding sections describe the methods of corpus compilation in LD projects with respect to collaborative fieldwork, the selection of genres and registers, metadata, the structure of corpora and the annotation of texts, whereas the use and exploitation of LD corpora has only been mentioned in passing, because in contrast to the exploitation of corpora of major languages, that of LD corpora has not been been investigated yet (see §8.7). At best one can develop a set of criteria to assess the quality of corpora with respect to the archive where the corpus is deposited and the quality of the corpus. As for the first point, Thieberger et al. point out:

> “The repository must have a commitment to provide long term curation and access to the corpus, which includes creating a persistent identifier and a citation form for items within the corpus. The repository should provide access to metadata and a clear means for accessing primary data with clearly stated access conditions that may include restrictions.” (Thieberger et al. (2015:12)

The evaluation of the quality of the corpus includes introductory information (§5.1), the structure of the corpus (§5.2), the metadata (§4.4), the kind of texts (§4.2), and the consistency of annotations (§6), but the relationship between the size and content of a LD corpus on the one hand and its usability for different purposes has not been researched yet. Such a research would have to account for at least for the following user-oriented factors:

- the prospective users’ knowledge of the language or a genetically related language;
- the prospective users’ competence in using an electronic corpus;
- the prospective users’ motivation for using the corpus and the time they want to invest for their corpus-based work;
- the availability of other publications on the language, especially dictionaries;
- the content of the texts, i.e. the topics talked about; the variety of the speakers; the

registers and genres of the texts; and the kinds of specialised annotations besides the obligatory transcription and the translation;

- with respect to linguists, their current interests in the type of the corpus language with respect to its structure, genetic affiliation, geographical distribution, and sociolinguistic and cultural features;
- the access to texts of interest via the metadata and / or the structure of the corpus.

With respect to the second point, we should not forget that the corpus is meant to be accessible to native speakers (cf. §4.1, §8.2, §8.4), even if they do not have affordable internet access, suitable electronic devices, and the knowledge of how to use electronic corpora. Therefore I recommend to complement the digital corpus by PDF files and books that contain at least a good selection of texts in the practical orthography with a translation and possibly illustrations and photographs. The production of books and their storage in traditional libraries is also recommended in the case that the corpus is not deposited in an archive that guarantees long-term curation and accessibility.

8 Further readings

8.1 Introduction

This section recommends literature on subjects that are closely related to this chapter, but have not been dealt with or discussed in detail. During the last 20 years LD has become a linguistic discipline of its own right which manifests itself in several handbooks on endangered languages and linguistic fieldwork. The reader is recommended to browse through them to get an idea of the diversity of the field and its impact on corpus building in LD projects. Austin and Sallabank (eds.) 2011; Regh and Campbell (eds.) 2018; Chelliah and De Reuse 2011; Duranti (ed.) 2004; Gippert, Himmelmann and Mosel (eds.). 2006; Grenoble and Furbee 2010; Newman and Ratliff (eds.) 2001; Thieberger 2012.

Another important resource of articles on LD is the journal *Language Documentation and Conversation.* http://scholarspace.manoa.hawaii.edu/handle/10125/312

8.2 Theory, technology and ethics

Due to limits of space, this chapter does not deal with the theory, technology and ethics of language documentation.

Theoretical issues: Good 2010; Himmelmann 1998, 2006a, 2012; Mosel 2018; Woodbury

2003; Woodbury 2011.

Technology and recording techniques: Margetts and Margetts 2012; Rice and Thieberger 2018.

Ethical issues: Chelliah and De Reuse 2011: 139-159; Dwyer 2006, Good 2018; Macri 2010; Newman 2012; Rice 2010; Rice 2012, McCarty 2018.

8.3 Corpus linguistics and the ethnography of speaking

Handbooks: O'Keeffe and McCarthy (eds.).2010; Lüdeling and Kytö (2008); Biber and Reppen 2015.

Registers, genres and the ethnography of speaking: Agha 2001; Agha 2004; Bauman 2001; Biber 2010; Biber and Conrad 2009; Foley 1997:247-378; Franchetto 2006; Senft 2014: 120-123.

Metadata standards: Lehmberg and Wörner 2008.

Annotation: Gries and Berez 2017.

8.4 The collection of texts

Collaborative fieldwork: Chelliah and De Reuse 2011: 161-195; Dimmendal 2001; Dwyer 2010; Grinevald 2007; Mosel 2006a; Mosel 2012b; Sapién 2018. For a brief summary of the literature on academic exploitation, the communities' reactions and the researchers' responses see Tsunoda (2005:216-228).

Collection of texts: Mosel (2012a:85-88) summarises two valuable earlier works on text collection: Samarin (1967:55-68) and Rivierre 1992; Mosel 2018.

Elicitation: Chelliah 2001; Chelliah and De Reuse (2011: 229-231; 252-254; 357-412). Mithun (2001:34-48). The three authors survey the methods of lexical, phonological and grammatical elicitation in a critical way and warn about the shortcomings of translational and similar naive methods that are still practised. For the use of elicitation by stimili see Hellwig 2019 and her references.

Metadata: Austin (2006:93-94); Himmelmann (2006a: 11-14); Chelliah and De Reuse (2011:215-219) give some very useful advice for "record-keeping" during fieldwork.

Conathan 2011 and Thieberger and Jacobson 2010 explain the structures of archives and the necessity of metadata in view of long term preservation of LD data. Thieberger and Jacobson discuss two archives in detail, PARADISEC, the Pacific and Regional Archive for Digital Sources in Endangered Cultures in Melbourne, Australia, and the French “Archiving Project” of LACITO/CNRS, Laboratoire de Langues et Civilisations à Tradition Orale, called Pangloss (see § 11.2).

File names: Thieberger and Berez 2012:102-104.

8.5 The structure of the corpus

Trilsbeek and Wittenburg 2006:322-323;

8.6 Annotations

Gestures: Bressem 2014; Haviland 2004; Seyeddinipur 2012.

Transcription: Crowley 2007:137-141 gives some useful advice how to do transcriptions in the field with local research assistants; Mosel 2006a: 78-79; Schultze-Berndt 2006: 219-232.

Orthography: Cahill 2018. Lüpke 2011; Seifart 2006.

Translation: Mosel 2006:79-80; Schultze-Berndt 2006: 232-238.

Grammarical annotation: Schultze-Berndt 2006: 238-248; GRAID: Haig and Schnell 2016 show the application of GRAID in syntactic typology.

8.7 The use of corpora

For corpus based research of non-endangered languages see the table of contents of the handbooks on corpus linguistics: Biber and Reppen (eds.) 2015; Lüdeling and Kytö (eds.). 2008; O'Keeffe and Michael McCarthy (eds.). 2010. In documentary linguistics, corpora are used for writing grammars (Amber et al. 2018, Nordhoff 2012), whereas corpus-based dictionaries are usually supplemented by elicited data, see the introductions to the dictionaries published in *Dictionaria* edited by Haspelmath, Stiebels and Hartmann 2019.

Comparable analyses and overviews of how LD corpora are used for the research of linguistic characteristics, language acquisition, and language varieties, or the design of educational materials are still missing in documentary linguistics.

9 Exercises

Exercise 1

The purpose of this exercise is to motivate the reader to browse through the archives with a critical mind and become aware of what kind of information is desirable from a user's perspective and how it can be presented in a more or less user-friendly way.

Chose a corpus from the DoBeS, ELAR or pangloss archive (see §11.2 for the references) and answer the questions listed below. If you don't find any information, just answer "No information".

- When was the corpus compiled?
- Who was responsible for the compilation of the corpus?
- Who were the indigenous and non-indigenous research assistants?
- Are all texts freely accessible or are there any restrictions? If yes, what kind of restrictions?
- Who were the speakers and/or writers of the texts? Is there any information about their personal characteristics such as age, sex, place of birth, etc.?
- What kind of genres and registers are represented in the corpus? Note that in many LD corpora there is no clear distinction between genres and registers.
- What kind of topics are the texts about?
- What kind of information do you find about the language and its speakers in the introduction to the corpus? What kind of necessary information is missing?

Exercise 2

The purpose of Exercise 2 is to raise the awareness of the differences between a practical orthography and an IPA transcription and get some practice in making a table that shows how the practical orthography is pronounced.

- Compare the practical orthography and the IPA transcription in Fig. 2 and make a table of two columns in which the left column shows the letters and the right column the corresponding IPA symbols. Note that one letter may correspond to more than one IPA symbol, and vice versa,. one IPA symbol may correspond to more than one letter.
- What kind of phonetic features are not represented in the orthography?

10 Abbreviations

APPL	applicative particle
ART	article
CLASS2	second noun class
COMPL	complementiser
DAT	dative
GEN	genitive
PL	plural
PREP	preposition
PRON	pronoun

RED	reduplication
SPEC	specific
TAM	tense-aspect-mood particle

11 References

11.1 Books and articles

Agha, Asif. 2001. Register. In Alessandro Duranti (ed.). *Key Terms in Language and Culture*. Malden (USA), Oxford (UK), Carlton (Australia): Blackwell Publishing Ltd., pp. 212-215.

Agha, Asif. 2004. Registers of Language. In Alessandro Duranti (ed.). *A Companion to Linguistic Anthropology*. Malden (USA), Oxford (UK), Carlton (Australia): Blackwell Publishing Ltd., pp. 23-45.

Amber B. Camp, Lyle Campbell, Victoria Chen, Nala H. Lee, Matthew Lou-Magnuson, and Samantha Rarrick. 2018. Writing grammars of endangered languages. In Kenneth L. Rehg and Lyle Campbell (eds.). 2018. *Oxford Handbook of Endangered Languages*. Oxford: Oxford University Press, pp. 271-304.

Austin, Peter. 2006. Data and language documentation. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel (eds.). *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter, pp. 87-112.

Austin, Peter and Julia Sallabank (eds.). 2011. *The Cambridge Handbook of Endangered Languages*. Cambridge: Cambridge University Press.

Bauman, Richard. 2001. Genre. In Alessandro Duranti (ed.). *Key Terms in Language and Culture*. Oxford: Blackwell Publishing Ltd., pp. 79-82.

Biber. Douglas. 2010. What can a corpus tell us about registers and genres? In Anne O'Keeffe and Michael McCarthy (eds.). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, pp. 241-254.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, Douglas and Randi Reppen (eds.). 2015. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press.

Bickel, Balthasar, Bernard Comrie, and Martin Haspelmath. 2004. *The Leipzig Glossing Rules. Conventions for Interlinear Morpheme by Morpheme Glosses.*Last change May 31, 2015 Leipzig: Max Planck Institute for Evolutionary Anthropology. https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf (accessed 19/09/2019).

Bowern, Claire. 2008. *Linguistic Fieldwork. A Practical Guide.* Basingstoke, New York: Palgrave Macmillan.

Bressem, Jana. 2014. Transcription systems for gestures, speech, prosody, postures, gaze. In Cornelia Müller, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill, and Sedinha Teßendorf (eds.). *Body-Language-Communication: An International Handbook on Multimodality in Human Interaction*. (Handbooks of Linguistics and Communication Science 38.1) Berlin, Boston: De Gruyter: Mouton, pp. 1037-1058.

Chafe, Wallace (ed.). 1980. *The Pear Stories. Cognitive, Cultural, and Linguistic Aspects of Narrative Production.* Norwood: Ablex.

Cahill, Michael. 2018. Orthography design and implementation for endangered languages. In Kenneth L. Rehg and Lyle Campbell (eds.). 2018. *Oxford Handbook of Endangered Languages***.** Oxford: Oxford University Press, pp. 327-346.

Chelliah, Shobhana L. 2001. The role of text collection and elicitation in linguistic fieldwork. In Paul Newman and Martha Ratliff (eds.). 2001. *Linguistic Fieldwork.* Cambridge: Cambridge University Press, pp. 152-165.

Chelliah, Shobhana L. and Willem J. De Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork.* Dordrecht, Heidelberg, London, New York: Springer.

Conathan, Lisa. 2011. Archiving and language documentation. In Peter Austin and Julia Sallabank (eds.). *The Cambridge Handbook of Endangered Languages*. Cambridge: Cambridge University Press, pp. 235-254.

Crowley, Terry. 2007. *Field Linguistics. A Beginner's Guide.* Edited and preapred for publication by Nick Thieberger. Oxford: Oxford University Press.

Dimmendaal, Gerrit. 2001. Places and people. In Paul Newman and Martha Ratliff (eds.). *Linguistic Fieldwork.*. Cambridge: Cambridge University Press, pp. 55-75.

Duranti, Alessandro (ed.). 2004. *A Companion to Linguistic Anthropology.* Malden (USA), Oxford (UK), Carlton (Australia): Blackwell Publishing Ltd.

Dwyer, Arienne. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel (eds.) *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter, pp. 31-66.

Dwyer, Arienne. 2010. Models of successful collaboration. In Lenore A. Grenoble and N. Louanna Furbee (eds.). *Language Documentation. Practices and Values.* Amsterdam and Philadelphia: John Benjamins Publishing Company, pp.193 - 212.

Evans, Nicholas. 2011. Semantic typology. In Jae Jung Song (ed.). *The Oxford Handbook of Linguistic Typology.* Oxford: Oxford University Press, pp. 504-533.

Foley, William A. 1997. *Anthropological Linguistics*. Oxford: Blackwell Publishers.

Foley, William A. 2003. Genre, register and language documentation in literate and preliterate communities. In Peter K. Austin (ed.) *Language Documentation and Description,* Vol. 1. London: Hans Rausing Endangered Languages Project, Department of Linguistics, School of Oriental and African Studies, pp. 85-98.

Franchetto, Bruna. 2006. Ethnography in language documentation. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel (eds.) *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter, pp. 183-211.

Gippert, Jost, Nikolaus Himmelmann and Ulrike Mosel (eds.). 2006. *Essentials of Language Documentation.* Berlin, New York: Mouton de Gruyter.

Good, Jeff. 2010. Valuing technology. Finding the linguist's place in a new technological universe. In Lenore A. Grenoble and N. Louanna Furbee (eds.). *Language Documentation. Practices and Values.* Amsterdam and Philadelphia: John Benjamins Publishing Company, pp. 111-131.

Good, Jeff. 2018. Ethics in language documentation and revitalization. In Kenneth L. Rehg and Lyle Campbell (eds.). *Oxford Handbook of Endangered Languages***.** Oxford: Oxford University Press.

Grenoble, Lenore A. and N. Louanna Furbee (eds.). 2010. *Language Documentation. Practices and Values.* Amsterdam and Philadelphia: John Benjamins Publishing Company.

Gries, Stefan Th. and Andrea Berez. 2017. Linguistic annotation in/for corpus linguistics. In N. Ide and J. Pustejovsky (eds.). *Handbook of Linguistic Annotation*. Dordrecht: Springer Science+Business Media, pp. 379-409. DOI 10.1007/978-94-024-0881-2_15

Grinevald, Colette. 2007. Linguistic fieldwork among speakers of endangered languages. In Osahito Miyaoka, Osamu Sakiyama, and Michael E. Krauss (eds.). *The Vanishing Languages of the Pacific Rim.* Oxford: Oxford University Press, pp. 35-76.

Haig, Geoffrey and Stefan Schnell. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse) Introduction and guidelines for annotators*. Manual Version 7.0 https://www.uni-bamberg.de/fileadmin/aspra/Publications/GRAID7.0_manual.pdf (accessed 20/09/2019)

Haig, Geoffrey and Stefan Schnell. 2016. 'The discourse basis of ergativity revisited', *Language* 98.3, pp. 591-618. https://muse.jhu.edu/article/628202/pdf (accessed 20/09/2019).

Haig, Geoffrey and Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts.* https://multicast.aspra.uni-bamberg.de/ (accessed 20/09/2019)

Haig, Geoffrey, Stefan Schnell and Claudia Wegener (2011). Comparing corpora from endangered languages: Explorations in language typology based on original texts. In Geoffrey L.J. Haig, Nicole Nau, Stefan Schnell, Claudia Wegener (eds.). *Documenting Endangered Languages. Achievements and Perspectives.* Berlin, Boston: De Gruyter Mouton, pp. 55-86.

Haviland, John B. 2004. Gesture. In Alessandro Duranti (ed.). *A Companion to Linguistic Anthropology*. Malden (USA), Oxford (UK), Carlton (Australia): Blackwell Publishing Ltd., pp. 197-221.

Haviland, John B. 2006. Documenting lexical knowledge. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel (eds.). *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter, pp. 129-162.

Hellwig, Birgit 2006. Field semantics and grammar-writing: Stimuli-based techniques and the study of locative verbs. In Felix K. Ameka, Alan Dench and Nicholas Evans (eds.). *Catching Language. The Standing Challenge of Grammar Writing.* Berlin, New York: Mouton de Gruyter, pp. 321-358.

Hellwig, Birgit. 2019. Linguistic diversity, language documentation and psycholinguistics: The role of stimuli. In Aimée Lahaussois and Marine Vuillermet (eds.). *Methodological Tools for Linguistic Description and Typology*. Language Documentation and Conservation Special Publication No. 16 (2019), pp. 5-30 https://scholarspace.manoa.hawaii.edu/handle/10125/24855

Himmelmann, Nikolaus 1998. Documentary and descriptive linguistics. In *Linguistics* 36, pp. 161-195.

Himmelmann, Nikolaus. 2006a. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel (eds.). *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter, pp. 1-30.

Himmelmann, Nikolaus. 2006b. Prosody in language documentation. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel (eds.). *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter, pp. 163-185.

Himmelmann, Nikolaus 2012. Linguistic data types and the interface between language documentation and description, *Language Documentation and Conversation*, Vol. 6, pp. 187-207.

Jung, Dagmar and Nikolaus Himmelmann. 2011. Retelling data: Working on transcription. In Geoffrey L.J. Haig, Nicole Nau, Stefan Schnell, Claudia Wegener (eds.). *Documenting Endangered Languages. Achievements and Perspectives.* Berlin, Boston: De Gruyter

Mouton, pp. 201-220.

Koptjevskaja-Tamm, Maria, Ekaterina Rakhilina and Martine Vanhove. 2016. The semantics of lexical typology. In Nick Riemer (ed.). *The Routledge Handbook of Semantics.* Abingdon, Oxon and New York: Routledge, pp. 434-454.

Kübler, Sandra and Heike Zinsmeister. 2015. *Corpus Linguistics and Linguistically Annotated Corpora*. London, New Delhi, New York, Sydney: Bloomsbury.

Lehmberg, Timm and Kai Wörner 2008. Annotation standards. In Anke Lüdeling and Merja Kytö (eds.). *Corpus linguistics*. *An International Handbook.* Berlin, New York: Walter de Gruyter, pp. 484-501.

Lehmann, Christian. 2004. Interlinear morphemic glossing. In Geert Booij, Christian Lehmann, Joachim Mugdan, Stavros Skopeteas in collaboration with Wolfgang Kesselheim (eds.). *Morphologie Morphology. Ein internationales Handbuch zur Flexion und Wort-Formation. An International Handbook on Inflection and Word-Formation.* Berlin/ New York: Walter de Gruyter, pp. 1834-1857.

Lüdeling, Anke and Merja Kytö (eds.). 2008. *Corpus Linguistics. An International Handbook.* Handbücher der Sprach- und Kommunikationswissenschaft Bd. 29. Berlin, New York: Walter de Gruyter.

Lüpke, Friederike. 2011. Orthography development. In Peter Austin and Julia Sallabank (eds.) *The Cambridge Handbook of Endangered Languages*. Cambridge: Cambridge University Press, pp. 312-336.

Macri, Martha. 2010. Language documentation. In Lenore A. Grenoble and N. Louanna Furbee (eds.). *Language Documentation. Practices and Values.* Amsterdam and Philadelphia: Benjamins Publishing Company, pp. 37-47.

Majid, Asifa. 2012. A guide to stimulus-based elicitation for semantic categories. In Nick Thieberger (ed.). *The Oxford Handbook of Linguistic Fieldwork*. Oxford: Oxford University Press, pp. 54-71.

Majid, Asifa, Melissa Bowerman, Miriam van Staden and James S. Boster. 2007. The semantic categories of cutting and breaking events: A crosslinguistic perspective. In *Cognitive Linguistics* 18-2, pp. 133-152.

Margetts, Anna. 2011. Filming with native speaker commentary. In In Geoffrey L.J. Haig, Nicole Nau, Stefan Schnell, Claudia Wegener (eds.). *Documenting Endangered Languages. Achievements and Perspectives.* Berlin, Boston: De Gruyter Mouton, pp. 321-336.

Margetts, Anna and Andrew Margetts. 2012. *Audio and video recording techniques for linguistic research.* In Nicholas Thieberger (ed.). *The Oxford Handbook of Linguistic*

Fieldwork. Oxford: Oxford University Press, pp. 13-53.

Mayer, Mercer. 1969. *Frog Where are You?* https://archive.org/details/frogwhereareyou0000maye_q9x4. (accessed 20/09/2019).

McCarty, Teresa L. 2018. Indigenous language rights—Miner's canary or mariner's tern? In Kenneth L. Rehg and Lyle Campbell (eds.). 2018. *Oxford Handbook of Endangered Languages***.** Oxford: Oxford University Press, pp. 82-104.

McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics.* Cambridge: Cambridge University Press.

Mithun, Marianne. 2001. Who shapes the record: the speaker and the linguist. In Paul Newman and Martha Ratliff (eds.). 2001. *Linguistic Fieldwork.*. Cambridge: Cambridge University Press, pp. 34-54.

Mosel, Ulrike. 2006a. Fieldwork and community language work. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel (eds.). *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter, pp. 67-85.

Mosel, Ulrike. 2006b. Sketch grammar. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel (eds.). *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter, pp. 301-309.

Mosel, Ulrike. 2011. Lexicography in endangered language communities. In Peter Austin and Julia Sallabank (eds.). *The Cambridge Handbook of Endangered Languages*. Cambridge: Cambridge University Press, pp. 337-353.

Mosel, Ulrike. 2012. Morphosyntactic analysis in the field - a guide to the guides. In Nicholas Thieberger (ed.). *The Oxford Handbook of Linguistic Fieldwork*. Oxford: Oxford University Press, pp. 72-89.

Mosel, Ulrike. 2014. Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language. In Toshihide Nakayama and Keren Rice (eds.). *The Art and Practice of Grammar Writing*, Language Documentation and Conservation Special Publication No. 8 (July 2014), pp. 135-157.

Mosel, Ulrike. 2015. Putting oral narratives into writing - experiences from a language documentation project in Bougainville, Papua New Guinea. In Bernard Comrie and Lucia Golluscio (eds). *Language Contact and Documentation. Contacto lingüístico y documentación*. Berlin, Munich, Boston: De Gruyter Mouton, pp. 321-342.

Mosel, Ulrike. 2018. Corpus compilation and exploitation in language documentation projects. In Kenneth L. Rehg and Lyle Campbell (eds.). 2018. *Oxford Handbook of Endangered Languages***.** Oxford: Oxford University Press, pp. 248-270.

Newman, Paul. 2012. Copyright and other legal concerns. In Nicholas Thieberger (ed.). *The Oxford Handbook of Linguistic Fieldwork*. Oxford: Oxford University Press, pp. 430-456.

Newman, Paul and Martha Ratliff. 2001 (eds.). *Linguistic Fieldwork.* Cambridge: Cambridge University Press,

Nordhoff, Sebastian. 2009. *A Grammar of Upcountry Sri Lanka Malay*. Utrecht: LOT Publications.

Nordhoff, Sebastian (ed.). 2012. *Electronic Grammaticography*. Language Documentation and Conservation Special Publication No. 4, https://scholarspace.manoa.hawaii.edu/handle/10125/24244 (accessed 20/09/2019).

O'Kaeffe, Anne and Michael McCarthy (eds.). 2010. *The Routledge Handbook of Corpus Linguistics.* Abingdon: Routledge.

Ostler, Nicholas. 2008. Corpora of less studied languages. In Anke Lüdeling and Merja Kytö (eds.). *Corpus Linguistics*. *An International Handbook*. Berlin, New York: Walter de Gruyter, pp. 457-483.

Radtke, Alexander. 2005. Explorative Studie zur phonetischen Realisierung des Teop auf perzeptorischer Basis. https://hdl.handle.net/1839/00-0000-0000-0001-3F64-D (accessed 24/09/2019.

Rehg, Kenneth L. and Lyle Campbell (eds.). 2018. *Oxford Handbook of Endangered Languages***.** Oxford: Oxford University Press.

Reppen, Randi. 2010. Building a corpus: what are the key consideration? In Anne O'Keeffe and Michael McCarthy (eds.). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge. pp. 31-37.

Rice, Karen. 2010. The linguist's responsibilities to community speakers: Community-based research. In Lenore A. Grenoble and N. Louanna Furbee (eds.). *Language Documentation. Practices and Values.* Amsterdam and Philadelphia: John Benjamins Publishing Company, pp. 25-36.

Rice, Karen. 2012. Ethical issues in linguistic fieldwork. In Nicholas Thieberger (ed.). *The Oxford Handbook of Linguistic Fieldwork*. Oxford: Oxford University Press, pp. 407-429.

Rice, Karen and Nick Thieberger. 2018. Tools and technology for language documentation and revitalization. In Kenneth L. Rehg and Lyle Campbell (eds.). 2018. *Oxford Handbook of Endangered Languages*. Oxford: Oxford University Press, pp. 225-247.

Rivierre, Jean Claude. 1992. Text collection. In Luc Bouquiaux and Jacqueline Thomas (eds.). *Studying and describing Unwritten languages.* Dallas, Tex.: SIL, 56-63.

Samarin, William J.. 1967. *Field linguistics: A guide in Linguistic Field Work.* New York: Holt, Rinehart and Winston.

Sapién, Racquel-María. 2018. Design and implementation of collaborative language documentation projects. In Kenneth L. Rehg and Lyle Campbell (eds.). 2018. *Oxford Handbook of Endangered Languages*. Oxford: Oxford University Press, pp. 203-224.

Schultze-Berndt, Eva. 2006. Linguistic annotation. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel (eds.). *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter, 163-181.

Seifart, Frank. 2006. Orthography development. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel (eds.) *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter, pp. 275-299.

Seifart, Frank. 2011. Competing motivations for documenting endangered languages. In Geoffrey L.J. Haig, Nicole Nau, Stefan Schnell, Claudia Wegener (eds.). *Documenting Endangered Languages. Achievements and Perspectives.* Berlin, Boston: De Gruyter Mouton, pp. 17-32.

Senft, Gunter. 2010. *The Trobriand Islanders' Way of Speaking.* Berlin, New York: De Gruyter Mouton.

Senft, Gunter. 2014. *Understanding Pragmatics.* Routledge: London and New York.

Seyeddinipur, Mandana. 2012. Reasons for documenting gestures and suggestions for how to go about it. In Nicholas Thieberger (ed.). *The Oxford Handbook of Linguistic Fieldwork*. Oxford: Oxford University Press, pp. 147-165.

Thieberger, Nicholas (ed.). 2012. *The Oxford Handbook of Linguistic Fieldwork*. Oxford: Oxford University Press.

Thieberger, Nicholas and Andrea Berez. 2012. Linguistic data management. In Nicholas Thieberger (ed.). *The Oxford Handbook of Linguistic Fieldwork*. Oxford: Oxford University Press, pp. 90-118.

Thieberger Nicholas and Michel Jacobson. 2010. Sharing data in small and endangered languages. Cataloguing and metadata, formats, and encodings. In Lenore A. Grenoble and N. Louanna Furbee (eds.). *Language Documentation. Practices and Values.* Amsterdam and Philadelphia: Benjamins Publishing Company, pp. 147-158.

Thieberger, Nick, Anna Margetts, Stephen Morey and Simon Musgrave. 2015. Assessing annotated corpora as research output. *Australian Journal of Linguistics*, Vol. 36, 1-21.

Trilsbeek, Paul an Peter Wittenburg. 2006. Archiving challenges. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel (eds.) *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter, pp. 275-299.

Tsunoda, Tasaku. 2005. *Language Endangerment and Language Revitalization.* Berlin, New York: Mouton de Gruyter.

Wierzbicka, Anna. 2014. *Imprisoned in English. The Hazards of English as a Default Language*. Oxford, New York: Oxford University Press.

Woodbury, Anthony. 2003. 'Defining documentary linguistics'. In Peter K. Austin (ed.) *Language Documentation and Description*, Vol. 1, London: Hans Rausing Endangered Languages Project, Department of Linguistics, School of Oriental and African Studies, 35-51.

Woodbury, Anthony. 2011. Language documentation. In Peter Austin and Julia Sallabank (eds.). *The Handbook of Endangered Languages*. Cambridge: Cambridge University Press, pp. 337-353.

11.2 References to websites of archives and tools

Beaver language documentation http://dobes.mpi.nl/projects/beaver/ (accessed 20/09/2019)

Dictionaria 2019. edited by Martin Haspelmath, Barbara Stiebels, and Iren Hartmann. https://dictionaria.clld.org/ (accessed 20/09/2019)

Dokumentation Bedrohter Sprachen Archive (DoBeS-Archive) http://dobes.mpi.nl/projects/ (accessed 20/09/2019)

Endangered Languages Archive (ELAR) https://elar.soas.ac.uk/ (accessed 20/09/2019)

ELAN https://tla.mpi.nl/tools/tla-tools/elan/ (accessed 20/09/2019)

Field Linguist's Toolbox https://software.sil.org/toolbox/ (accessed 20/09/2019)

FieldWorks Language Explorer (FLex) https://software.sil.org/fieldworks/ (accessed 20/09/2019)

L&C Field Manuals and Stimulus Materials, Language and Cognition Department, Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands http://fieldmanuals.mpi.nl/ (accessed 20/09/2019)

Language Documentation and Conversation. http://scholarspace.manoa.hawaii.edu/handle/10125/312 (accessed 20.09.2019)

Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) http://www.paradisec.org.au/ (accessed 20.09.2019)

Pangloss http://lacito.vjf.cnrs.fr/pangloss/corpus/index.html (accessed 19/09/2019)

Savosavo http://dobes.mpi.nl/projects/savosavo/ (accessed 19/09/2019)

Teop http://dobes.mpi.nl/projects/teop/ (accessed 19/09/2019).

The Pear Film http://pearstories.org/ (accessed 20.09.2019).