9

Language Documentation & Conservation Special Publication No. 8 (July 2014):
The Art and Practice of Grammar Writing,
ed. by Toshihide Nakayama and Keren Rice, pp. 135-157
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/4589

Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language

Ulrike Mosel
Universität Kiel

Drawing on her experiences with writing a grammar in the course of the Teop language documentation project, the author explores how corpus linguistic methods can be employed for the analysis and description of a previously undescribed language. After giving a short introduction into the creation of a digital corpus and complex corpus search methods, the chapter focuses on the importance of creating a diversified corpus. It demonstrates that different text varieties such as spoken and written legends, procedural texts and descriptions of objects show different preferences for certain ways of expression and thus represent valuable resources for various grammatical phenomena. Accordingly, a grammar which is based on texts should account for this variation by incorporating a detailed description of the corpus, giving references and metadata for each example and providing information on the kind of contexts particular grammatical features are usually associated with.

1. INTRODUCTION. Most publications on linguistic field methods emphasize that a collection of recorded, transcribed and analyzed texts is the most important source for the grammatical description of a previously undescribed language (see Bright 2007:16, Chelliah 2001, Crowley 2007:121, Dixon 2010:321 among many others). But only two older field manuals (Samarin 1967:55-68, Rivierre 1992:56-63) and the recently published *Handbook of descriptive linguistic fieldwork* (Chelliah & De Reuse 2011:422-44) give some information on what constitutes a good corpus for grammaticographers and how the texts that are typically collected during fieldwork can be classified. The crucial questions, however—how an annotated corpus of texts is created and what kind of grammatical information can be gained from different text varieties—have been neglected in descriptive and typological linguistics.

Therefore, I would like to open the discussion on this topic by making a few suggestions of how the writers of grammars of previously undescribed languages can build up a diversified text corpus, and illustrate this corpus linguistic approach by examples from my own research on Teop. Teop is an Oceanic Meso-Melanesian language of the North-West Solomonic linkage (Lynch et al. 2002:101-102), spoken by approximately 6000 people in the Autonomous Region of Bougainville, Papua New Guinea. Our project was one of the first language documentation projects funded by the Dokumentation Bedrohter Sprachen program of the Volkswagen Foundation (Mosel et al. 2007), but besides this documentation I continuously worked on a Teop Reference Grammar. At the same time I learned to use the language documentation tool ELAN (see §2) and became interested in modern corpus linguistics, which completely changed my way of writing a grammar compared to the methods we employed when writing the *Samoan Reference Grammar* (Mosel & Hovdhaugen 1992).

ISBN: 978-0-9856211-4-8

In the following, I will briefly explain some corpus linguistic methods of grammatical analysis and grammar writing in §2, then in §3 discuss how a corpus can be compiled that meets both the wishes of the speech community and the interests of the grammaticographer, and in the following sections focus on three kinds of grammatical variation:

1. the grammatical variation in spontaneous oral texts and the edited versions of these texts (§4);
2. the preference for certain grammatical constructions in particular text varieties (§5);
3. the pervasive use of certain constructions in texts on special themes (§6).

My experiences suggest that the four phases of the grammar writing process — text recording in the field, corpus compilation and annotation, data analysis and description — are so closely interrelated that they should be integrated into a holistic methodology.

2. CORPUS LINGUISTIC METHODS IN GRAMMATICAL ANALYSIS AND GRAMMAR WRITING. The use of text collections as the basis of grammatical analysis makes the writing of grammars of previously undescribed languages a kind of corpus linguistic enterprise, although it is impossible to meet the demands of quantitative corpus linguistics and investigate grammatical variation on the basis of a corpus of millions of words as it is nowadays done for the compilation of grammars of European languages, e.g. Biber et al. (1999). But what seems worth doing is to gather a corpus that comprises texts of various kinds, analyze and describe grammatical categories and constructions, identify linguistic variation across text varieties and interpret the preferences for certain linguistic features in relation to the contexts where they occur. As for the terms text and text variety, I follow Biber and Conrad (2009). While the term text refers to 'natural language used for communication, whether it is realized in speech or writing', text varieties are defined by their situational characteristics, which include the channel, relations among participants, production circumstances, communicative purposes and the topic (Biber & Conrad 2009:5, 40).

Linguistically significant variation is especially noticeable in comparable corpora where two text varieties only differ with respect to one or two variables as, for instance, the transcription of a spontaneously narrated legend and the edited version of this transcription (see §3), or a narrative about the butchering of a chicken and a procedural description of how people butcher chickens (see §5.3).

A corpus gathered in the course of fieldwork is certainly not representative for the language as such, but only for a few selected text varieties. As will be further elaborated on in §3, fieldwork corpora differ from conventional corpora in that the selection of texts is not primarily guided by linguistic or demographic criteria. Rather, especially in the beginning of the research project, the sampling is determined by the external conditions of the fieldwork site, and consequently, classifies as "haphazard, convenience, or accidental sampling" (Kalton 1983:90, quoted in Meyer 2002:43).

For writing a grammar the most useful kind of text collection has the form of a digitalized annotated corpus that links audio or video recordings to transcriptions and translations, provides for each text metadata, is accessible via the internet (Austin 2006),

and allows the user to search with a query language like Regular Expressions (see below §2.2).

2.1. ANNOTATION. The most sophisticated tool for compiling a corpus of a previously undescribed language is ELAN which besides or in combination with Toolbox is widely used in language documentation projects.

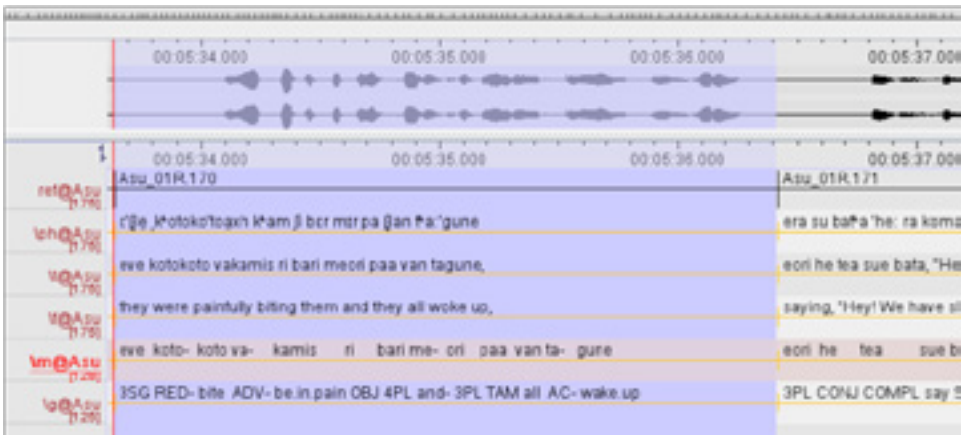

FIGURE 1. Annotation of a Teop audio file in ELAN

ELAN allows a text with various kinds of annotation on separate tiers to be presented, as illustrated in Fig. 1, which shows a narrow phonetic and an orthographic transcription, a free translation, morphological segmentation, and interlinear morpheme glossing. An annotation like the transcription of clauses can be time-aligned to the sound file, whereas other annotations as, for instance, the translation of the clauses are aligned to the corresponding transcription. The annotations can be exported as Toolbox, Praat or text files. For an excellent overview of annotation systems used in documentary linguistics see Schultze-Berndt (2006).

In grammars, individual examples and text samples, which are usually presented in an appendix, are provided in a three-tiered format, a transcription with morphological segmentation, interlinear morpheme glossing, and a free translation in order to show how the meaning of a construction that is rendered in the free translation relates to the constituent parts of the construction. For the grammatical analysis of texts, however, it may be useful to have additional tiers that provide information on the constituent structure of phrases and clauses and thus facilitate the exploration of syntactic structures and the interface of discourse and syntax. Such a system called GRAID (Grammatical Relations and Animacy in Discourse) has been recently developed by Haig & Schnell (2011) and tested for five genetically and typologically divergent languages (Haig, Schnell & Wegener 2011).

2.2. CORPUS-BASED GRAMMATICAL ANALYSIS AND DESCRIPTION. What makes ELAN most suitable for the grammatical analysis of a text corpus is that it facilitates complex searches with the query language Regular Expressions on multiple tiers.

With Regular Expressions you can search not only for all occurrences of a particular word, but also for discontinuous sequences of particular words or bound morphemes, for two or more alternative expressions at the same time, for a particular expression with the exclusion of other expressions, and even for reduplicated word forms.

A typical example for a complex construction in Teop that can easily be searched for with Regular Expressions is the negation of predicates which is expressed by the discontinuous morpheme *saka*/*sa* ... *haa*-. The first part has the variants *saka* and *sa*, the second part may stand by itself or have a suffix.

The search finds 338 tokens in the current Teop corpus of 258,866 words and presents them in a concordance. As Fig. 2 shows, the construction *saka*/*sa* ... *haa*- accommodates nouns, verbs and adjectives. Further searches show that with 229 tokens the first element *sa* is much more frequent than *saka*, and that contrary to our expectations both forms are equally distributed in spontaneous spoken and edited written text varieties.

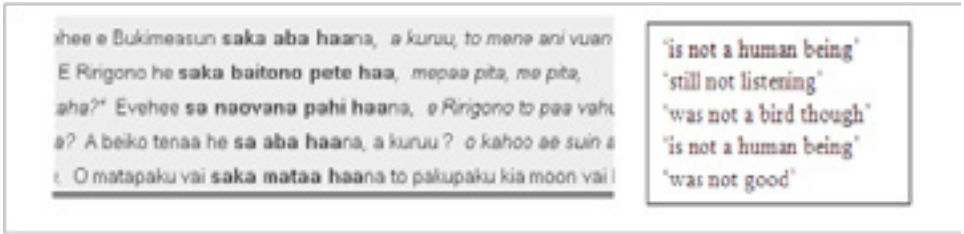

FIGURE 2. Concordance for the negation of predicates

Other typical examples of searches for discontinuous elements include the search for patterns of word-formation by prefixes and suffixes as in English *un*...*able* or the collocation of particular verb forms with temporal adverbs, which is useful for the investigation of aspect and Aktionsart (Van Valin 2005:32-42).

The so-called multilayer search enables you to search on more than one tier. For example you can search for one sense of a polysemous or homonymous lexical item by searching simultaneously on the transcription and the translation or glossing tier. By using the wild card * for the translation tier, the concordance shows the searched Teop items in context on the left-hand side and the corresponding translation on the right-hand side (Fig. 3).

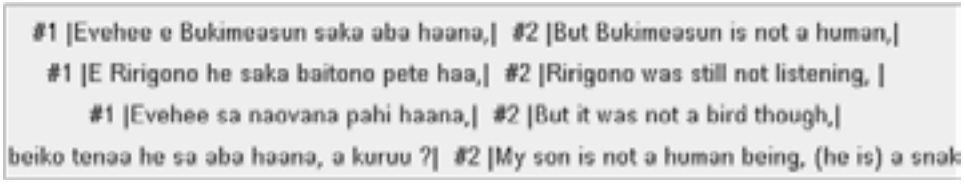

FIGURE 3. Multiple tier searches

Working with a digital corpus that is linked to sound files and facilitates complex searches has a number of advantages over traditional text analysis and thus ultimately leads to more reliable grammars. When browsing texts in search of examples for a particular construction, your attention may be biased; you only note down what seems interesting at the moment, because you cannot make notes for all examples. In contrast, the corpus search gives you all tokens in a concordance and enables you to make explicit statements about the frequency of constructions and their distribution in various text varieties. Since the concordance is linked to the corpus, you can jump from each token to the text file with one mouse click and immediately check the wider context of the token, listen to the sound file and check its annotation.

For the writing of a grammar the corpus-based approach means that linguistic phenomena are described with reference to their context in natural language use and that their frequency can be stated. So instead of giving the vague information that a linguistic phenomenon is rare or more frequent than another one, the grammarian can give exact figures of the number of occurrences in the corpus or a subcorpus, including the search method and the date of access in case the corpus in growing. In Teop, for example, nouns like *aba* 'person, human being' may function not only as the head of a NP, but also as the head of a verb complex (VC) as in

(1) *E* *Magaru* *kou* *na* *aba* *vakis* *nana*
ART Earthquake PART TAM person still IPFV:3SG
te- *a* *taem* *vai*.
PREP ART time DEM
'Earthquake was still a human being at that time.' (Val_02R.31)

But only concrete figures of the functions of prototypical nouns will show how rare the use of nouns as VC heads is (see Table 1).The figures as such are not explanatory, they only show patterns of language use that need to be interpreted. In the case of the distribution of prototypical nouns in the position of VC and NP heads, a probable explanation is that these nouns denote entities that in most contexts are conceptualized as time-stable (cf. Givón 2001:51), whereas the use of a word as the head of a VC implies a change over time.

		VC head	**NP head**
aba	'person'	8	370
beiko	'child'	1	504
moon	'woman'	9	767
naono	'tree'	-	232

TABLE 1. The distribution of prototypical nouns in the Teop Language Corpus (31.12.2011)

2.3. THE ROLE OF METADATA IN GRAMMARS. Metadata are data about data. In the context of language documentation metadata can be classified into

- collection-level metadata giving information on the circumstances of data collection such as the scope and duration of the project and the equipment used for recording;
- item-level metadata giving information on the name of the language, the recording date, the collector and the speaker, the kind of media, the content and the size of the recording;
- biographical information about each participant of the recording session (for details see Conathan 2011:246-248).

In the grammar the information on the circumstances of the data collection and biographical data of the speakers and the people who did the recordings will be given in the introductory chapter, whereas an appendix may contain a list of the names of the primary data and their specific item-level metadata. In addition, each example taken from the corpus should get a label which indicates its source and some information on the text variety in the form of abbreviations. Nordhoff (2009), for example, uses labels that indicate the town and the date of the recording and the text variety, e.g. 'nar' for narrative, 'cvs' for conversation, 'sng' for song etc.

To date many corpus-based grammars of previously undescribed languages do not give any detailed information on the content and structure of the corpus, let alone references or metadata for the examples. The readers of these grammars are not informed whether a particular example has been elicited or comes from a legend, a procedural text, ritual or a certain genre, who the speaker was, and when and under which circumstances the recording was done. Since all languages show variation and grammars never capture the full range of variation, the grammar user should be informed about the text varieties that served as the basis of the grammatical analysis and description.

2.4. ACCESSIBILITY OF THE CORPUS. From a scientific point of view, it should be a matter of course to grant access to the corpus that served as the source for the grammatical analysis because otherwise the grammatical description could not be scrutinized by other researchers (Himmelmann 1998:165). Since grammarians may be misguided by their hypotheses and overlook examples that would not fit their hypotheses, the results of their grammatical analysis remain preliminary as long as they are not replicable by other researchers (for a discussion of replicability in corpus linguistics see McEnery & Hardie 2012:14-16). Consequently, the grammar should either contain a DVD with the corpus or inform the readers how they can access the corpus via the internet (cf. Thieberger 2006, Nordhoff 2009).

3. BUILDING UP A GRAMMAR WRITER'S CORPUS. Linguists who intend to write a grammar of a previously undescribed language will use a variety of field methods to collect data (Bowern 2008, Chelliah & De Reuse 2011, Mosel 2006, Mosel 2012), and sooner or later start collecting texts. What kind of texts are recorded and in what format they are published depends in the first place on the speech community's interests and values

(Woodbury 2011:180-182). The request for recording "a variety of informal communicative events ... to give an authentic impression of the language" (Seifart 2008:69) cannot always be responded to. Haig, Nau, Schnell & Wegener (2011:4) observe that in most projects of the Dokumentation Bedrohter Sprachen (DoBeS) program, "it still tends to be more traditional monologues than everyday conversational interactions that find themselves as fully-annotated records in the archive".

The speech community's right to set the agenda can mean that linguists who intend to collect texts as their data basis for a grammatical description might have to change their plans and adapt their project to the decisions of the speech community. If, for example, the speech community is only interested in documenting their traditional oral literature, then obviously the grammar that is based on these texts only represents this text variety, and from the point of view of grammaticography there is nothing wrong with this. It goes, however, without saying that the more diverse text varieties the corpus contains, the greater are the chances that the grammar can comprehensively represent the language (Foley 2003:95).

The speech community may also, as it happened in the Teop language documentation project, insist on editing the texts before publication, which clearly contradicts the aim of language documentation to record "the linguistic practices and traditions of a speech community" (Himmelmann 1998:166). As will be shown below, the editing of texts will provide the grammaticographer with a new quality of data because it shows what native speakers are actually doing when transforming speech to writing and thus not only contributes to the analysis of the particular language in question, but also to research on the differences between spoken and written language. The drawback of editorial work, however, is that it is time consuming and requires a well organized workflow as, for example, the work of the Dauenhauers on the documentation of Alaskean Tlingit oral culture shows (Dauenhauer & Dauenhauer 1996). A brief description of my own experiences is given in Mosel (2006:82f).

In order to minimize the danger that editorial work in our Teop project was influenced by the native speakers' knowledge of English text varieties, they were advised to keep the original speakers' way of expression, their phraseology and discourse structure, and thus avoid the dangers of westernizing traditional oral literature. Each edited text was independently checked by at least two other native speakers. Both the edited texts and the original recordings are archived in the DoBeS Archive, but the original recordings with their transcriptions and translations are only accessible under the condition that the users register themselves. A comparison of the spoken and edited text varieties shows that in spite of my advice the editors made quite a number of changes (see §4). The edited version has been printed without translations and is now used in schools (Magum et al. 2007) and is also available in the DoBeS archive.

After they had done transcriptions and editorial work during several fieldwork seasons, some Teop research assistants started writing example sentences for the grammar and the dictionary, stories, and descriptions of animals, plants, artifacts, and everyday activities. These are definitely not traditional, but innovative text varieties. But this does not mean that they are less authentic than, for example, spoken legends or conversations, as long as the linguist does not teach the native speakers what in his or her view a good story is. Furthermore, when speech communities want their language to become a written language and the means of instruction in primary schools, it certainly belongs to the responsibilities

of linguists to help them create it by keeping the uniqueness of their language, but also avoiding a rigid purism that would put off younger speakers. Linguistically these new text varieties are interesting because they allow us to observe the process of putting a previously unwritten language into writing.

In all publications on the Teop language, the references for examples indicate whether the example is taken from the recordings of spontaneous speech (abbr. R), from an edited version (abbr. E) or from a written text that has not been derived from a transcription (abbr. W). In the grammar I try to present wherever possible examples from spoken and edited or written text varieties for each grammatical phenomenon.

In the reminder of this chapter, I show how useful even a relatively small but diversified corpus can be and what special kinds of grammatical constructions are offered by different text varieties.

4. VARIATION IN THE GRAMMAR OF ORAL LEGENDS AND THEIR EDITED VERSIONS. When we analyzed the two subcorpora of spoken and edited Teop legends, which comprise 31,909 and 31,294 words, respectively, we could identify four types of syntactic changes in the edited versions: elaboration, linkage of paratactic clauses, compression of paratactic clauses, and decompression of complex constructions (Mosel 2008).

All constructions found in the edited versions are also found in the oral versions, but the two registers differ in the frequency of certain constructions:

- In the edited versions, the replacement of paratactic constructions by compressed constructions is more frequent than the reverse kind of replacement.
- Elaboration often results in complex structures (e.g. adjectival attributes, serial verb constructions, relative clauses, clausal adjuncts).
- The edited versions make more use of explicit clause linkage.

Table 2 gives a summary of the observed changes in edited narratives, which on the whole result in more complex structures.

Strategy	**Syntactic change**
Elaboration	addition of linguistic units (words, phrases, clauses)
Linkage of paratactic clauses	1. linkage by cross-clausal dependency without embedding (chained Tail-Head-Linkage, adjoined adverbial clauses) 2. integration by embedding (relative and adverbial clause constructions) 3. interlacing by raising in complement constructions
Compression of paratactic clauses	1. serial verb constructions 2. nominalizations 3. ditransitive constructions
Decompression	resolution of complex constructions into paratactic constructions

TABLE 2. Syntactic changes in edited narratives

The following citations from an oral legend (2) and its edited counterpart (3) illustrate the replacement of two coordinate clauses by a so-called Tail-Head construction and a few other changes.

(2) *Me-* *paa* *vahuhu* *bona* *taonim* *a* *si* *iana*.
and- TAM give.birth.to ART five ART DIM fish
'And gave birth to five little fish.'

Me- *a* *taonim* *a* *si* *iana* *bona* *vue*
and ART five ART DIM fish DEM particular
'And these five little fish'

na *vaatii* *roho* *e* *te-* *a* *boon ...*
TAM put first 3SG PREP- ART mangroves ...
'she put in the mangroves.' (Ata_01R.01)

(3) *Me-* *paa* *vahuhu* *bona* *taonim* *a* *si* *iana*.
and TAM give.birth.to ART five ART DIM fish
'And gave birth to five little fish.'

Vaahuhu *vakavara* *vai* *ri* *bari*
give.birth finish then 3PL.OBJM 4PL
'Having given birth to them,'

me *paa* *varavihi* *ri* *bari* *koma-* *n-*
and TAM hide 3PL.OBJM 4PL inside- 3SG.POSS-
a *boon*
ART mangroves
'hid them inside the mangroves.' (Ata_01CE1.01)

In the spoken version (2) the second clause is joined to the first one by the conjunction *me* 'and', the repetition of 'five little fish' and the anaphoric demonstrative *bona* , whereas in the edited version the verb of the first clause (the tail) *vahuhu* 'give birth' is repeated at the beginning (the head) of the second clause. While the clause linkage in the spoken version is similar to an English type of clause linkage, the one of the edited version is not. Furthermore, the editor exchanged the verb *vaatii* 'put' for the semantically more specific verb *varavihi* 'hide', inserted an object marker and the pronoun *bari* 'them' and replaced the multipurpose preposition *te* PREP by the more elaborate locative construction *koma na boon* 'inside the mangroves'.

Since the Tail-Head construction is typical for Oceanic and Papuan languages, but is not found in English, we were interested in how the editors treated these constructions and counted all Tail-Head constructions in which the Head is modified by *vakavara* 'finish' as in (3). The result was that with 51 tokens the Tail-Head construction is much more frequently used in the edited legends than in the original spoken versions which only show 28 occurrences. Thus we have the impression that the edited version represents a more conservative style of story telling than the original spoken version.

5. GRAMMATICAL VARIATION ACROSS TEXT VARIETIES. The Teop Language Corpus comprises several subcorpora which on the basis of their content and the circumstances of their production can be classified as shown in Table 3. Not unexpectedly, these text varieties do not only differ in their vocabulary, but also in their preferences for certain syntactic constructions, as the remainder of this section will illustrate with examples from legends, dictionary definitions and procedural texts.

Genres	**Themes**	**Production**
legends	fights with giants and witches, bad treatment of children by their stepmothers, controversies between two brothers, origin of natural phenomena and artifacts	spoken and edited; some only written
personal narratives	autobiographies, survival during the Second World War, travel	spoken and edited; two only written
encyclopedic descriptions	plants, animals (mammals, birds, reptiles, fishes, crabs, shells), house and canoe building, fishing, butchering, cooking, cultural practices	descriptions of things only written; procedural texts spoken, edited and written
interviews	young native speakers interviewing elders about customs and the Second World War	spoken and edited
example sentences	not specified	only written

TABLE 3. Text varieties in the Teop Language Corpus

5.1. LEGENDS. Since legends are situated in imaginary worlds where animals can talk or magic allows transformations of things into living beings or living beings into things, they may offer interesting data on the functional flexibility of lexemes (see §2.2) and noun classification. In Teop, for example, animal names belong to the unmarked class of common nouns, but move to the class of personal proper names when referring to one of the protagonists of a legend. This phenomenon, which shows that under certain conditions the classification of nouns is variable, would not be attested and, consequently, not described in the grammar if our corpus did not contain legends.

For the description of argument structure and discourse pragmatics, the beginnings of legends provide easily retrievable data on how new participants are introduced into the discourse. Furthermore, legends may contain direct speech with colloquial expressions of surprise and anger, which are interesting for the description of phraseology and the grammar of interjections, but are difficult to record otherwise (Seifart 2008:73).

5.2 THE GRAMMAR OF DICTIONARY ENTRIES. Since it is impossible to produce a dictionary within a short-term language documentation project, we decided to compile a series of small thematically specialized dictionaries on plants, fishes, house building, cooking, etc. These mini-dictionaries (MD) contain short encyclopedic articles in Teop with an English translation (Mosel 2011, Mahaka et al. 2010). In addition, the dictionaries of the material culture are supplemented by procedural texts which describe selected traditional techniques like thatching the roof of a house, making fishing nets, butchering a pig, etc. Both the definitions and the procedural texts are valuable sources for gathering grammatical data, because they contain some constructions at a much higher rate than narrative texts.

Since the purpose of a dictionary entry is to define the meaning of a word, the entries show a variety of topic constructions that are not encountered in narratives in this density. The definitions of nouns frequently start with a non-verbal clause consisting of a topical subject NP followed by a classifying predicative NP that is modified by an adjectival phrase (AP) or a relative clause:

(4)

SUBJ.NP		PRED.NP		QUALIFICATIVE ATTRIBUTIVE AP	
A	*bokua*	*a*	*iana*	*a*	*beera ...*
ART	bokua	ART	fish	ART	big, ...

'The bokua is a big fish.' (Vaa_09W.068)

(5)

SUBJ.NP		PRED.NP		POSSESSIVE ATTRIBUTIVE AP		
A	*havanao*	*a*	*iana*	*a*	*kapa*	*kikis*.
ART	havanao	ART	fish	ART	skin	strong

'The havanao is a fish with a strong skin.' (lit. '(having) a strong skin') (Sii_11W.039)

(6) SUBJ.NP PRED.NP RELATIVE CLAUSE

O	*poka*	*o*	*hum*	*to*	*vavaobetera-*	*ra-*
shelf	ART	place	REL	put	ART	1PL.INCL.IPFV-
ara		*bona*		*maa*	*taba*.	
1PL.INCL		ART		PL	thing	

'The shelf is a place where we put things.' (MD House, poka)

While the definitions of nouns supply excellent examples for topicalization, non-verbal clauses, adjectival phrases, relative clauses, and the expression of habitual activities (4-6), the definitions of verbs are a good source for nominalizations and complement clauses in predicative function:

(7)

A	*siri*	*atovo*		*ei*	*be-*	*ara*	*gono*
ART	tear	sago.palm.leaf		DEM	when-	1PL.INCL	get
kahi	*o*	*paka*	*bono*	*sikiri*	*na-*	*e*.	
from	ART	leaf	ART	midrib	3SG.POSS	3SG	

'The tearing of the sago palm leaf, this (is) when we remove the midrib from the leaf.' (MD House, siri atovo)

To conclude, although a grammar writer's task is not collecting data for a dictionary, it seems worthwhile asking native speakers to formulate some definitions of animal and plant species, artifacts, and special activities.

5.3. PROCEDURAL TEXTS VS. NARRATIVES. Similar to dictionary entries, procedural texts are not an indigenous, conventionalized genre in Pacific cultures, as people prefer to demonstrate how this or that is done instead of describing it (Mosel 2006:73f). Consequently, the speakers have not yet developed conventionalized ways of describing procedures and seem to be free in their choice of pronouns to refer to generic agents. Some prefer the second person singular, others the first person inclusive plural or the third person plural pronoun. One speaker consistently uses the first person exclusive plural (cf. 13), which the editors of her texts always replace by the first person inclusive pronoun. This variation in the use of pronouns for generic agents is remarkable and needs to be mentioned in the grammar chapter on pronouns.

Another remarkable feature of the procedural texts is that all speakers and writers use the same kind of clause linkage construction when explicitly referring to the regular fixed order of actions. While in Teop narratives the sequence of events is simply expressed by paratactic and coordinate clauses, or the so-called Tail-Head construction (see §3), the procedural texts show constructions with adverbial clauses. Our first example (8) comes from a legend in which a giant scrapes the bark of *kave* vines for making a fishing net. In the Tail-Head construction the narrator repeats the head of the VC *kahu* 'scrape', but modifies it by *vakavara* 'finished' expressing that this action was finished, before he did the next one, i.e. *taatagi* 'prepare'.

(8) *me-* *ori* *paa* *dee* *voosu* *maa,* *me-* *ori* *paa*
and- 3PL TAM carry home DIR and- 3PL TAM
ma *kahu,*
come scrape
'and they carried (the kave vines) home, and they scraped them'

me- *ori* *kahu* *va-* *kavara* *bona* *kano-* *kanono* *te-*
and- 3PL scrape ADV- finished ART RED- rope PREP-
ori,
3PL
'and they finished scraping their ropes,'

a- *maa* *kara* *kave* *te-* *ori,* *me-* *ori* *paa*
ART- PL string kave PREP- 3PL and- 3PL TAM
taatagi *bari,*
prepare 4PL.OBJ
'their kave strings, and they prepared them.' (Sii_06R.56-60)

The second example (9) comes from a written description of how Teop people made nets for catching turtles in former times. Here the fixed sequence of two actions is expressed by a *be-re* 'when-then' construction, which is very frequent in procedural texts.

(9) *Be-* *ve* *obete* *nana* *te-* *o* *kasuana,*
when- 3SG lie 3SG.IPFV PREP- ART ground
'When it is lying on the ground,'

eara *re-* *paa* *kahu* *a* *kapa* *nae*
1PL.INCL then- TAM scrape ART bark 3SG.POSS

bono *kehaa*
ART shell
'then we scrape its bark off with a shell'

to *dao* *ra-* *ara* *bono* *sui.*
REL call 1PL.INCL.IPFV- 1PL.INC ART sui
'that we call sui.'

Be- *ara* *kahu* *vaka-* *va-* *kavara* *e,*
when- 1PL.INC scrape RED- ADV- finished 3SG
'When we have finished scraping it,'

eara *re* *paa* *vaaroava* *e* *bono* *buaku*
1PL.INCL then TAM dry.in.sun 3SG ART two

ge	*o*	*kukan*	*o*	*bon*.
or	ART	three	ART	day

'then we put it into the sun for two or three days.' (Eno_08W.4-6)

Other variants of this construction in procedural texts include:

(10)

be-	AGENT	X	*va-*	*kavara,*	AGENT	*re-*
when	AGENT	X	ADV-	finished	AGENT	then-

paa	Y
TAM	Y

'when AGENT has finished doing X, then AGENT does Y'

(11)

be	*kavara,*	AGENT	*re-*	*paa*	X
when	finished	AGENT	then-	TAM	X

'when it is finished, then AGENT does X'

(12)

be-	AGENT	*tau*	X,	*AGENT*	*re-*	*paa*	Y
when	AGENT	about to	X,	AGENT	then-	TAM	Y

'when AGENT is about to X, then AGENT does Y'

(13)

be-	AGENT	*mei*	*tea*	X,	AGENT	*toro*	Y
when	AGENT	not.yet	COMP	X	AGENT	must	Y

'before AGENT X, AGENT must do Y'
(lit. 'when AGENT has not yet X, AGENT must Y')

In order to get further evidence for the difference in clause linkage in narratives and procedural texts, I bought a rooster from a neighbor and asked him to butcher it while I was taking a series of photographs. Luckily his four years old twins were helping him butcher the rooster, while his wife was watching, so that three months later I could ask her to look at the photographs and narrate the story of how her husband and her children butchered a rooster during my last visit. In addition, I asked another woman to have a look at the photographs and describe how Teop people butcher a rooster.

While in the procedural text nine clauses out of a total 40 clauses are adverbial clauses introduced by *be* 'when' (14), the narrative text, which consists of 53 clauses, has none of these constructions, but uses paratactic clauses instead (15):

(14) Procedural text

Be	*kavara,*
when	finished

'When it is finished,'

be-	*nam*	*pee-*	*pee*	*va-*	*ruta-*	*rutaa*	*va-*
when-	1PL.EXCL	RED-	cut	ADV-	RED-	small	ADV-

kavara	*eve,*
finished	3SG

'when we have finished cutting it into small pieces,'

o-	*re*	*paa*	*vahio*	*bari*	*te-*	*o*	*suraa*.
3PL-	then-	TAM	put	4PL	PREP-	ART	fire.

'they put it onto the fire.' (Hel_13R.33-34)

(15) Narrative text

Eove	*he*	*kaku*	*va-*	*kavara*	*bene*	*toa*
3SG	but	butchered	ADV-	finished	ART	chicken

'But he finished butchering the rooster,'

me-	*ori*	*paa*	*vaa-*	*tei*	*bari*	*te-*	*a*
and-	3PL	TAM	CAUS-	be	4SG/PL	PREP-	ART

sosopene.
saucepan
'and they put it into the saucepan.' (Pau_01R.51-52)

As the preceding examples illustrate, the distinction between Tail-Head constructions and the adverbial clauses is most clearly shown by comparing narrative and procedural texts of the same or closely related contents as, for instance, net making (8,9) or butchering a chicken (14,15). Consequently, it seems practical to include this kind of comparison in a grammar.

6. DIFFERENT THEMES - DIFFERENT GRAMMATICAL PHENOMENA. People talk about different themes in different ways. For the collection of grammatical data this means that some themes will provide more and better data for certain grammatical phenomena than others. Thus inanimate topics are certainly better represented in descriptions of how certain artifacts are manufactured than in autobiographies, whereas ditransitive constructions with agents, recipients and themes are most likely to be found in texts about trading and ceremonial exchanges of food and valuables.

6.1. TROPICAL FISHES ARE COLORFUL. The question of whether in Oceanic languages lexemes denoting properties form a word class in its own right, i.e. adjectives, or are better classified as a subclass of verbs is probably as old as Oceanic linguistics itself, but a thorough corpus-based study of property words in any of these languages is still missing. A distributional analysis of dimensional and evaluative adjectives such as *beera* 'big' and *mataa* 'good' in Teop shows that they often occur as the head of VCs, but that they are distinct from intransitive verbs as they never occur as the head of NPs, which all intransitive verbs do, e.g. *a pita* 'the walking', *a mate* 'the dying, death', but not **a beera* 'the being big' or **a mataa* 'the being good'. Secondly, these adjectives differ from intransitive verbs in that they must take the prefix *va-* when modifying a verb, e.g. *vabeera* 'to a great extent', *vamataa* 'well, properly'.

lexeme	**VC head**	**AP head**
beera 'big'	96	295
mataa 'good'	83	133

TABLE 4. Distribution of *beera* 'good' and *mataa* 'good'.

For color words we did not have comparable data until we started compiling a small fish dictionary in which the fish names are defined by descriptions in Teop with English translations. Most descriptions contain color words and clearly show that in Teop color words behave exactly like dimensional and evaluative adjectives. They enter into the comparative construction (cf. 16 and 17) and are transformed into an adverb by the prefix *va-* when modifying a verb, e.g. *tara vamataa* 'look good'.

lexeme		**VC head**	**AP head**
gogooravi	'red'	12	23
kakaavo	'white'	7	25
paru	'black'	5	34

TABLE 5. Distribution of three colour words in the fish dictionary.

(16)

	NP		AP		VC		
evehee	*a*	*toobono*	*a*	*beera,*	*[na*	*beera*	*oha*
but	ART	toobono	ART	big	TAM	big	pass
		NP					
nana]		*bona*		*pasupua*			
3SG.IPFV		ART		pasupua			

'(The toobono looks like the genuine pasupua,)
but the toobono is big, is bigger than the pasupua.' (MD Fishes, toobono)

(17)

NP		AP	predicate		
A	*aranavi*	*[a*	*gogooravi*	*vasihum] ...*	
ART	aranavi	ART	red	a.bit	

'The aranavi is a bit red...'

NP		VC				NP
A	*sinarona*	*[na*	*gogooravi*	*oha*	*nana]*	*bona*
ART	sinarona	TAM	red	pass	3SG.IPFV	ART
aranavi.						
aranavi						

'The sinarona is redder than the aranavi.' (MD Fishes, aranavi)

Similar to *tara vamataa* 'look good', we find derived colour adverbs modifying *tara* 'look':

(18) *Be-* *ori* *hovo* *ruene* *o-* *re* *paa* *tara* *va-*
when- 3PL enter river 3PL- then TAM look ADV-
paru.
black
'(While they are still staying in the ocean, they look white.)
When they enter the rivers, they look black.' (MD Fishes, ovunaa)

As far as we can judge from our limited set of data in Tables 3 and 4, adjectives occur more often as the head of an AP than as the head of a VC, but the difference between these figures is not as marked as those of the distribution of nouns as NP and VC heads. A possible explanation for these findings may be that these adjectives denote less time-stable concepts than nouns.

The preceding examples illustrate that frequency analyses can be helpful in formulating hypotheses about the interaction of lexis and grammar. Munro (2007:72) stresses the importance of dictionary work for grammatical analysis, "Making dictionaries helps in grammatical analysis, and in fact in the absence of dictionary work a grammatical description is very likely to miss important things".

6.2 WHAT TREES ARE GOOD FOR. The Teop language is a verb second language. This means that the verb complex always occurs in the second position of the clause, while the first position is held by the topic of the clause, which can be the subject, a primary object, a secondary object, or an adjunct. If the topic can be recovered from the preceding context, the topic position can be left empty. With ditransitive verbs, Teop shows the following clause patterns:

TOPIC	**VC**	**Argument**	**Argument**
SUBJ (subject)	VC	OBJ1 (primary object)	OBJ2 (secondary object)
OBJ1 (primary object)	VC	SUBJ (subject)	OBJ2 (secondary object)
OBJ2 (secondary object)	VC	SUBJ (subject)	OBJ1 (primary object)

TABLE 6. Clause patterns.

Teop does not have a passive construction. If the agent of an action is not identifiable, the third person plural pronoun functions as a non-topical subject.

The 2007 version of Teop Language Corpus gives the impression that constructions with the subject in the first position represent the dominant word order. For the ditransitive verb *hee*, for example, we find the following frequencies of clause patterns (Mosel 2007, 2010):

clause patterns	**frequency**
SUBJ VC OBJ1 OBJ2	25
OBJ1 VC SUBJ OBJ2	6
OBJ2 VC SUBJ OBJ1	4

TABLE 7. Clause patterns of *hee* 'give' (Sept. 2007).

With *hee* 'give', the primary object (OBJ1) refers to the recipient and the secondary object (OBJ2) to the theme. Other ditransitive verbs like *nahu* 'cook' govern a primary object referring to the patient and an optional secondary object referring to the instrument:

(19) SUBJ:agent VC OBJ1:patient OBJ2:instrument
... a-re *ma* *nahu* *a* *guu* *vai* *bona* *tahii*.
1PL.INCL-then come cook ART pig this ART saltwater
'(You must fetch some saltwater) so that we can cook this pig with saltwater.' (Mat. 1.68R)

When analyzing clauses of this kind, I had the impression again that the dominant, unmarked order was SUBJ VC OBJ1 OBJ2. But when the Teop research assistants collected descriptions of trees and what the parts of trees are used for, I realized that it would only make sense to speak of a dominant word order with respect to a particular text variety. If as in the tree descriptions the topic of discourse is a patient or instrument, the noun phrases denoting these roles function as objects, but occupy the first position of the clause, as the following dictionary entry for *asita* 'putty nut tree' nicely illustrates. The entry starts with the sentence:

(20) OBJ2 VC SUBJ OBJ1
O *asita* *[na* *asi-* *asita* *ri-* *]* *ori* *bono*
ART putty.nut TAM RED- plaster 3PL.IPFV 3PL ART
sinivi.
canoe
'The putty-nut tree, they use it for plastering the canoe.' (i.e. the nuts of the tree) (MD Plants, asita)

In the second clause of the entry (21), the topic position is empty. The topic is still *asita* in the function of a secondary object, but as it is easily recoverable from the context, it does not need to be mentioned.

(21) VC SUBJ OBJ1
[Na *asita* *ri-* *]* *ori* *[bona* *maa* *panapana]*
[TAM plaster 3PL.IPFV] 3PL ART PL knotholes
'They plaster the knotholes (of the canoe with it).' (MD Trees, asita)

This sentence is then followed by two other sentences of the same structure, while the last sentence shows a construction in which the valency of a ditransitive verb—here *porete* 'treat s.o. with s.th. (some kind of traditional ditransitive verb—is reduced by the particle *ni,* resulting in a transitive construction meaning 'use s.th. as traditional medicine' (Mosel 2010:493).

(22) OBJ VC

Asita	*me*	*[na*	*pore-*	*porete*	*ni*	*ri]-*
plaster	also	TAM	RED-	make.medicine	APPL	3PL.IPFV
SUBJ						
ori.						
3PL						

'Asita is also used for making medicine.' (MD Plants, asita)

8. CONCLUDING REMARKS. The present chapter suggests that the grammaticography of previous undescribed languages can profit from an approach that combines language documentation with corpus linguistic methods. In contrast to traditional grammar writing, the corpus linguistic approach accounts for language internal variation in relation to text varieties. As Conrad (2010:228) puts it, "corpus analyses lead us to describing grammar not just in structural terms, but in probabilistic terms—describing the typical social and discourse circumstances associated with the use of particular grammatical features". The modern technology of corpus linguistics allows us to systematically search for particular lexical items and their collocations as well as for constructional patterns and the lexical items they accommodate, to view all findings in a concordance and to analyze the grammatical structures in their natural context.

This chapter emphasizes the need for a diversified corpus and shows what kind of data is provided by different text varieties. In particular we examined spontaneously spoken and edited versions of legends, procedural texts and dictionary definitions, and discovered that due to their different contents and discourse structure these text varieties provide useful data for various grammatical phenomena:

1. The comparison of oral and edited legends shows what kind of constructions native speakers regard as synonymous, in particular variation in narrative clause linkage.
2. Comparable narrative and procedural texts about the very same topic show how the contrast between specific and habitual sequential actions is expressed.
3. Monolingual dictionary definitions of nouns provide data of how the classification of living beings and things is expressed, which in the case of Teop involves non-verbal predicates, various kinds of adjectival attributes and relative clauses. The definitions of verbs typically contain nominalizations in subject position and complement clauses as predicates.
4. In the descriptions of trees and their parts we find numerous examples for constructions with inanimate topics and the expression of the semantic role of instrument.

The macro-structure of a corpus-based grammar may follow the traditional ascending model starting with a chapter on phonology and concluding with a chapter on complex sentences (Mosel 2006b), but its content would probably differ in the following aspects:

1) The introductory chapter would provide explicit information on
   a) fieldwork methods (cf. §3),
   b) the sociolinguistic profile of the speech community,
   c) the sociolinguistic background of those native speakers who were recorded or otherwise involved in the project (cf. §2.3), and
   d) the genres (§5), the topics (§6) and the size of the texts as well as the technology of recordings and the annotation methods (cf. §2.1)
2) In addition, the appendices of the grammar may supply detailed information on the individual texts and speakers in the form of tables (cf. §2.3).
3) Within the chapters the description of grammatical phenomena would account for variation in linguistic form and function and, wherever it seems reasonable and significant, make statements about preferred structures in terms of frequencies. This may, for example, include
   a) the syntactic distribution of words or word classes (cf. Table 1, 4, 5),
   b) the frequency of clause patterns (cf. Table 7), or
   c) the occurrence of particular constructions in certain text types (cf. §5.3).
4) The examples would get labels that inform the reader on their origin and facilitate their identification in the corpus, which ideally is easily accessible.

In the near future digital linguistics will develop electronic formats of grammars and new tools assisting in grammatical analysis (Evans & Dench 2006:28-30, Nordhoff (ed.) 2012), but the arguments for a corpus based grammaticography as outlined in this chapter will certainly not lose their validity.

References

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, 87-112.

Biber, Douglas & Stig Johansson & Geoffrey Leech & Susan Conrad & Edward Finegan. 1999. *Longman Grammar of spoken and written English*. Harlow, Essex.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge:CUP.

Bouquiaux, Luc and Jacqueline Thomas (eds.). 1992. *Studying and describing unwritten languages*. Dallas: SIL

Bowern, Claire. 2008. *Linguistic fieldwork. A practical guide*. New York: Palgrave Macmillan.

Bright, William. 2007. 'Contextualizing a grammar', in Thomas E. Payne & David J. Weber (eds.). *Perspectives on grammar writing*. Amsterdam: John Bejamins Publishing Company, 11-17.

Chelliah, Shobhana L. 2001. 'The role of text collection and elicitation in linguistic fieldwork', in Paul Newman and Martha Ratliff (eds.). *Linguistic Fieldwork*. Cambridge: Cambrige University Press, 152-164.

Chelliah, Shobhana L. & Willem de Reuse. 2011. *Handbook of descriptive linguistic fieldwork*. Dordrecht etc.: Springer.

Conathan, Lisa. 2011. Archiving and language documentation. In Peter Austin & Julia Sallabank (eds.). *The handbook of endangered languages*. Cambridge: Cambridge University Press, 235-254.

Conrad, Susan. 2010. What can a corpus tell us about grammar? In Anne O'Keeffe & Michael McCarthy (eds.). *The Routledge handbook of corpus linguistics*. Abingdon: Routledge.

Crowley, Terry. 2007. *Field linguistics. A beginner's guide*. Oxford: Oxford University Press.

Dauenhauer, Nora & Richard Dauenhauer. 1999. The paradox of talking on the page: Some aspects of the Tlinglit and Haida experience. In Laura J. Murray & Karen Rice (eds.). *Talking on the page. Editing aboriginal oral texts*. Toronto, Buffalo, London: University of Toronto Press, 3-41.

Dixon, R.M.W. 2010. *Basic linguistic theory*. Vol. 1. Methodology. Oxford: Oxford University Press.

Evans, Nicholas & Alan Dench. 2006. Introduction: Catching language. In Felix Ameka, Alan Dench & Nicholas Evans (eds.). *Catching language. The standing challenge of grammar writing*. Berlin, New York: Mouton de Gruyter, 1-39.

Foley, William A. 2003. Genre, register and language documentation in literate and preliterate communities. In Austin, Peter. *Language Documentation and Description*. vol. 1, pp. 85-98.

Givón, Talmy. 2001. *Syntax*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Good, Jeff. 2002. *A gentle introduction to metadata*. Open Language Language Archives Community. www.language-archives.org/documents/gentle-intro.html (31.12.2011)

Haig, Geoffrey & Nicole Nau & Stefan Schnell & Claudia Wegener. 2011. Documenting endangered languages before, during, and after the DoBeS programme. In Geoffrey Haig & Nicole Nau & Stefan Schnell & Claudia Wegener /eds.), *Documenting endangered languages. Achievements and perspectives*,1-14. Berlin: De Gruyter Mouton.

Haig, Geoffrey & Stefan Schnell. 2011. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse). Introduction and guidelines for annotators*. Version 6.0. http://vc.uni-bamberg.de/moodle/course/view.php?id=9488 (31.12.2011)

Haig, Geoffrey & Stefan Schnell & Claudia Wegener. 2011. Comparing corpora from endangered languages projects: Explorations in language typology based on original texts. In Geoffrey Haig & Nicole Nau & Stefan Schnell & Claudia Wegener /eds.), *Documenting endangered languages. Achievements and perspectives*,1-14. Berlin: De Gruyter Mouton, 55-86.

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

Kalton, Graham. 1983. *Introduction to survey sampling*. Beverly Hills, CA: Sage.

Lynch, John & Malcolm Ross & Terry Crowley. 2002 *The Oceanic languages*. Richmond: Curzon.

Magum, Enoch Horai & Joyce Maion & Jubilie Kamai, & Ondria Tavagaga, & Ulrike

Mosel & Yvonne Thiesen, (eds) 2007. *Amaa vahutate vaa Teapu. Teop Legends*. Kiel: CAU, Seminar für Allgemeine und Vergleichende Sprachwissenschaft, and www.mpi.nl/DOBES/projects/teop (31.12.2011)

Mahaka, Mark & Enoch Horai Magum & Joyce Maion & Naphtaly Maion & Ruth Siimaa Rigamu & Ruth Saovana Spriggs, & Jeremiah Vaabero & Ulrike Mosel & Marcia Schwartz & Yvonne Thiesen. 2010- *A inu. The Teop-English dictionary of house building*. Kiel: CAU, Seminar für Allgemeine und Vergleichende Sprachwissenschaft, and www.mpi.nl/DOBES/projects/teop (31.12.2011)

McEnery, Tony & Andrew Hardie. 2012. *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press.

Meyer, Charles. 2002. *English corpus linguistics. An introduction*. Cambridge: Cambridge University Press.

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, 67-85.

Mosel, Ulrike. 2006b. Grammaticography: The art and craft of writing grammars. In Felix Ameka, Alan Dench & Nicholas Evans (eds.). *Catching language. The standing challenge of grammar writing*. Berlin, New York: Mouton de Gruyter, 41-68..

Mosel, Ulrike. 2007. A corpus based approach to valency in a language documentation project. Pre-ALT workshop, Paris, 24.9.2007. http://www.uni-muenster.de/imperia/md/content/allgemeine_sprachwissenschaft/forschen/lingtypolangdoc/handout_mosel_corpus_based.pdf (31.12.2001

Mosel, Ulrike. 2008. 'Putting oral narratives into writing – experiences from a language documentation project in Bouganville, Papua New Guinea'. Paper presented at the Simposio Internacional Contacto de lenguas y documentación, August 2008. Buenos Aires, CAIYT, http://www.linguistik.uni-kiel.de/mosel_publikationen.htm#download

Mosel, Ulrike. 2010. Ditransitive constructions and their alternatives in Teop. In Malchukov, Andrej & Martin Haspelmath & Bernhard Comrie (eds.) *Studies in ditransitive constructions: a comparative handbook*. Berlin, New York: De Gruyter Mouton, 486-509.

Mosel, Ulrike. 2011. Lexicography in endangered language communities. In Peter Austin & Julia Sallabank (eds.). *The handbook of endangered languages*. Cambridge: Cambridge University Press.

Mosel, Ulrike. 2012. Morphosyntactic analysis in the field – a guide to the guides. In Nick Tieberger (ed.) *The Oxford handbook of linguistic fieldwork*. Oxford: Oxford University Press.

Mosel, Ulrike & Even Hovdhaugen. 1992. *Samoan reference grammar*. Oslo: Oslo University Press.

Mosel, Ulrike & Enoch Horai Magum, & Jubilie Kamai & Joyce Maion, & Naphtali Maion & Siimaa Ruth Rigamu, & Ruth Saovana Spriggs & Yvonne Thiesen. 2007. *The Teop language Corpus*. www.mpi.nl/DOBES/projects/teop (31.12.2011)

Mosel, Ulrike & Yvonne Thiesen. 2007. *Teop sketch grammar.* www.mpi.nl/DOBES/projects/teop (31.12.2011)

Munro, Pamela. 2007. Form parts of speech to grammar. In: Thomas E. Payne & David Weber (eds.) *Perspectives on grammar writing*. Amsterdam & Philadelphia: Benjamins, pp. 71-111.

Nordhoff, Sebastian. 2009. *A grammar of Upcountry Sri Lanka Malay*.Utrecht: LOT Publications, http://www.lotpublications.nl/publish/articles/003745/bookpart.pdf

Nordhoff, Sebastian (ed.). 2012. *Electronic grammaticography*. Language Documentation & Conservation Special Publication No. 4. Honolulu: University of Hawai'i Press.

Rivierre, Jean Claude. 1992. 'Text Collection', in Luc Bouquiaux and Jacqueline Thomas (eds.). 1992. *Studying and describing unwritten languages*. Dallas: SIL, 56-63.

Samarin, William J. Field linguistics. *A guide to linguistic fieldwork*. New York etc. Holt, Rinehart and Winston.

Seifart, Frank. 2008. On the representativeness of language documentations. In Peter K. Austin (ed.). *Language documentation and description*. Vol. 5. London: School of Oriental and African Studies, 60-76.

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, 213-251.

Thieberger, Nicholas. 2006. *A grammar of South Efate*. Honolulu: University of Hawai'i Press.

Van Valin, Robert D. 2005. *Exploring the syntax-semantic interface*. Cambridge: Cambridge University Press.

Woodbury, Anthony C. 2011. Language documentation. In Peter Austin & Julia Sallabank (eds.). *The handbook of endangered languages*. Cambridge: Cambridge University Press, 159-186.

umosel@gmx.de