Collecting data for grammars of previously unresearched languages

Ulrike Mosel, Universität Kiel

Draft for the International LingDy Symposium on Grammar Writing,, Tokio 8th.-10th Dec., 2009

1 Introduction

Most publications on linguistic field methods emphasise that a collection of recorded, transcribed and analysed texts is the most important source for the grammatical description of a previously unresearched language. (Bright 2007:16, Chelliah 2001, Crowley 2007:121, Dixon 2010:321 among many others) But only two field manuals give some information on what constitutes a good corpus for grammaticographers and how the texts that are typically collected during fielwork can be classified (Samarin 1967:55-68, Rivierre 1992:56-63), while the crucial question of what kind of grammatical information can be gained from the analysis of various types of text seems to have been totally neglected in field linguistics. Therefore I would like to start a discussion on this topic by presenting a few results of my attempts to build up a corpus of the Teop language that both satisfies the expectations of the speech community and at the same time provides a useful database for the planned Teop Reference Grammar. Teop is an Oceanic language spoken in Bougainville, Papua New Guinea.

The use of text collections as the basis of grammatical analysis makes writing grammars of previously unresearched 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 (Biber et al. 1999). But what seems worth doing is to gather a corpus that comprises a small number of text types in the broadest sense and then identify and describe any kind of observed linguistic variation. Linguistically significant variation is especially noticable in parallel corpora where two types of text only differ with respect to one variable as, for instance, the transcription of a spontaneously narrated legend and the edited version of this transcription (see §4), or a narrative about the slaughtering of a chicken and a procedural description of how people slaughter chickens (see § 5.3).

In the following , I will first in §2 give an overview of data collection methods and crriticise the way data from text collections are usually presented in grammars. Then, in §3, I will comment on what makes a good text collection in a language documentation project. The remainder of the paper draws on my experiences in the Teop Language Documentation Project and describes grammatical differences in spontaneously spoken legends and the edited written versions of these legends (§ 4), how the different genres of narratives, dictionary definitions and procedural texts differ with respect to the use of certain the grammatical constructions (§5), and what kind thematically defined types of texts provide good examples for certain grammatical phenomena (§6). I assume that similar kinds of data across different types of text can be collected for most language.

2. Kinds of data collection

There are various methods of data collection, which can be roughly classified into four types:

1. language learning and participant observation
2. translational elicitation
3. non-translational elicitation
4. collection of texts (in the widest sense)

As the advantages and disadvantages of these types of method have been discussed in various fieldwork guides and are reviewed in Mosel (forthcoming b), a brief summary with references must suffice here.

2.1 Language learning

In collaborative fieldwork, i.e. fieldwork that involves members of the speech community as active partners in the collection and analysis of linguistic data, the linguists should try to learn to understand and speak the language as best they can, provided that the speech community appreciates the researcher's ambitions. (Abbi 2001:146, Bowern 2008:9-10, Crowley 2007:155, Everett 2001, Hill 2006, Kibrik 1977:52, Mosel 2006b:73-74, Samarin 1967:49-55)

2.2 Elicitation

In its narrow sense elicitation means stimulating native speakers to produce certain kinds of utterances like word lists or example sentences for particular words or constructions, in a wider sense it would also subsume the collection of texts. In the following the term elicitation will be used in its narrow sense and distinguished from the collection of texts. Elicitation techniques can be classified into translational and non-translational ones. Although the problems of translational elicitation are well known (Abbi 2001:88, Bowern 2008:85-87, Chelliah 2001:154-158, Samarin 1967:58-59, 114), it is widely practiced and even fieldwork guides that mention these problems recommend the translation of wordlists in the contact language and present a sample in their appendices. Others totally rely on translational elicitation for grammatical research. (Bouquiaux and Thomas 1992) But translations from the contact language into the target language can be avoided from the very beginning of gathering linguistic data by asking the native speakers to teach us linguists frequently used expressions for persons, things, activities and properties and then show us how meaningful utterances are formed with these words (Mosel 2006b), Samarin 1967:83). As soon as the first simple utterances have been noted down, non-translational techniques like substitution, paraphrasing, or sentence completion can be applied (for a list of such techniques and references see Mosel forthcoming b).

2.3 The collection of texts and their role in grammatical analysis and description

In contrast to the writers of grammars of well-known written languages, the writers of previously unresearched languages cannot gather their data from existing text collections, but have to create their own corpus. Such a corpus should contain:

- video and/or audio recordings with metadata on the speech situation of the recording;
- annotations of the recordings, i.e. transcriptions, translations, comments on the content and the linguistic form with the corresponding metadata on who did the various annotations;
- elicited data.

Furthermore, this corpus should be accessible. Even recent corpus-based grammars of previously unresearched languages do not give any detailed information on the content and structure of the corpus, let alone references of the examples and access to the corpus (Aikhenvald 2003, Dixon 2004, Enfield 2007, Lichtenberk 2008). 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 whatever genre, who the speaker was, and when and under which circumstances the recording was done, nor can the readers go back to the original source and verify the analysis presented in the grammar. In my view, this linguistic practice of not giving information on the origin of data and giving access to the corpus diminishes the scientific value of even the best grammars (Mosel 2006a:53).

3 Building up a corpus in collaborative fieldwork

In close cooperation with the speech community, language documentation projects aim at building up corpora of authentic data of spoken language

- that are not only interesting for linguistics, but also for other disciplines of the humanities and social sciences;
- that are provided with transcriptions, translations and comments on their content so that they can be understood without prior knowledge of the documented language;
- that can be used for language maintenance and revitalisation by the speech community.

These three principles, which guide the Dokumentation Bedrohter Sprachen (DoBeS) programme funded by the Volkswagen Foundation, may give rise to conflicts and require the linguists to find a balance between what is interesting for them and their special field of expertise, what is relevant for other non-linguistic disciplines and what meets the expectations of the speech community. Many linguists emphasise the importance of documenting spontaneous everyday conversations because it is this kind of speech event that language is mainly used for, that shows the greatest range of variation and that is the domain where language change originates. But what might be most interesting for linguists, may be the least the speakers want to have recorded or even published. Consequently, the kinds of linguistic variety represented in the corpus are inevitably determined by what the kind of speech events the speakers offer to be recorded. Furthermore, the corpus should be presented in a form that is appreciated by the speakers as well as researchers from other disciplines and the general public.

3.1 The user friendly corpus

The linguist's documentation consists of recordings that are linked to transcriptions and translations with comments on content and linguistic phenomena, a grammmatical sketch and various kinds of additional materials to warrant understandibility. (Himmelmann 2006) Even if the transcriptions are done in a practical orthography, they are hard to read for non-linguists simply because spoken language is not meant to be read. Transcriptions are a tool for linguists and not an enjoyable read for researchers of other disciplines. The people engaged in natural discourse repeat themselves, stutter, break up utterances, do not care about background information, mix names and make mistakes so that a recording and transcription of a natural speech event may be incomprehensible for the outsider, unless it is accompanied by footnotes, which, of course, make the text even less reader-friendly.

And the speech community? If they are literate in their own or a dominant language, they may not want their recordings be published in the form of transcriptions (as we researchers would not publish linguistically accurate transcriptions of our lectures). In addition, transcriptions cannot be directly used for the production of written materials that the community may want for maintenance and revitalisation measures. In sum, what is linguistically justified and desirable, might be inconvenient for other researchers and unacceptable for the primary contributors to the language documentation.

3.2 New genres in language documentations

In the Teop Language Documentation Project, the conflict arising from the heterogenous DoBeS aims were solved by training the indigenous research assistants in editing transcriptions. They were advised to keep the original speakers' way of expression, their phraseology and discourse structure, and thus avoid the dangers of westernising Teop (Foley 2003). 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 and sign a code of conduct.

After they had done transcriptions during several fieldwork seasons, some local research assistants started writing example sentences for the grammar and the dictionary, stories, and descriptions of animals, plants, artefacts, and everyday activities. These are definitely not traditional, but innovative genres. 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 speaker 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 genres are interesting because they allow us to observe the process of putting a previously unwritten language into writing.

4 Variation in the grammar of oral legends and their edited versions

When we analysed the two parallel 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.

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 compex structures (e.g. adjectival attributes, serial verb constructions, relative clauses, clausal adjuncts).
- The edited versisons make more use of explicit linkages (e.g. Tail-Head-Constructions).

Table 1: Syntactic changes in edited narratives

Strategy	Syntactic change
Elaboration	addition of linguisitc 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. nominalisations 3. ditransitive constructions
Decompression	resolution of complex constructions into paratactic constructions

(compare Lehmann 1988)

As a consequence, the edited versions show more complex constructions. Hence they are not only useful for the production of educational materials and resources for other disciplines. For the linguist, the parallel corpus of oral narratives and the edited versions:

- gives a fuller picture of the expressive potential of the language;
- shows what native speakers regard as alternative ways of expressing the same content,
- provides a new type of data for research on the differences of spoken and written language as it shows what speakers actually do when they put an oral text into writing.

For details see Mosel 2008.

5. Grammatical variation across genres

The Teop Language Corpus comprises several subcorpora which on the basis of their content and the circumstances of their production can be classified as follows:

Table 2: Genres,. themes and modes of production of texts in the Teop Language Corpus

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 artefacts	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, slaughtering, 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

For the collection of data, including word lists and example sentences, translation from the contact languages English and Tok Pisin was strictly avoided, but translations from Teop to English and Tok Pisin were used to help me to understand and translate the texts. Some translations were also done by Teop people. The only stimuli-based technique we used was asking Teop people to describe objects or procedures while looking at a series of photographs of indigenous plants, animals, artefacts or activities (cf. §5.3).

Not unexpectedly, the spoken, edited or creatively written texts of different genres and themes do not only differ in their vocabulary, but also in their preferences for certain syntactic constructions. To illustrate these differences, the remainder of this section will present some examples from legends, dictionary definitions and procedural texts. The personal narratives have not been analysed yet.

5.1 Legends are better than their reputation in documentary linguistics

In discussions on language documentation one sometimes gets the impression that the focus should lie on recording casual conversations and that the recording of legends belongs to the olden days of Boas, Sapir and Bloomfield. Of course, legends do not present a genre of spontanous speech and also may contain archaic expressions, as many or even most sentences of a legend may be recited from memory by the speaker. But this can also be an advantage at the beginning of a documentation project when the people are shy and still reluctant to speak while being recorded. Legends are situated in imaginary worlds where animals can talk or magic allows transformations of things into living beings or living beings into things, so that the legends may provide interesting data on noun classification and noun-verb distinction as in the example below. Here *magaru* 'earthquake' is treated as a personal name and the noun *aba* 'person' is the head of a verb complex and combined with the realis mood marker *na* and the imperfective aspect marker *nana*.

(1)

NP			VC				
E	***Magaru***	***kou***	***[na***	***aba***	***vakis***	***nana***	**]**
ART	Earthquake	PART	PART	person	still	3SG.IPFV	

'Earthquake was still a human being (at that time).' (Val. 2.31R)

Furthermore, legends may contain direct speech with colloquial expressions of surprise and anger, or even obscene swear words that nobody wants to have recorded in actual interactions.

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 thematically specialised, mini-dictionaries on plants, fishes, house building, cooking, etc. These mini-dictionaries contain short excyclopedic articles in Teop with an English translation. (Mosel et al. 2009) 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, slaughtering a pig, etc. Both the definitions and the procedural texts are a valuable source for gathering grammatical data, because they contain some constructions at a much higher rate than narrative texts, as well as constructions that we have not encountered yet in the narratives.

Dictionary entries are, for example, interesting, because they show a variety of topic constructions. 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 or a relative clause:

(2) 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.' (MD Fishes, *bokua*)

(3) SUBJ.NP PRED.NP POSSESSIVE ATTRIBUTIVE AP
A booboo a iana a kapa kikia.
ART *booboo* ART fish ART skin strong
'The *booboo* is a fish with a strong skin.' (MD Fishes, *booboo*)

(4) SUBJ.NP PRED.NP RELATIVE CLAUSE
O poka o hum to vavaobete ra- ara
ART shelf ART place REL put 1PL.INCL.IPFV- 1PL.INCL
bona maa taba.
ART PL thing
'The shelf is a thing where we put things.' (MD House, *poka*)

Definitions of this kind supply excellent examples for:

1. non-verbal clauses;

2. topicalisation;

3 qualificative and possessive APs;

4. relative clauses with relativised objects, oblique arguments and adjuncts.

In the definitions of verbs we find nominalisations and complement clauses in predicative function:

(5) *A siri atovo* *ei* *be-* *ara* *gono kahi* *o* *paka*
ART tear sago.palm.leaf DEM when- 1PL.INCL get from ART leaf
bono sikiri *nae.*
ART midrib 3SG.POSS .
'The tearing of the sago palm leaf, this (is) when we remove the midrib from the leaf.'[1] (MD House, *siri atovo*)

5.3 Procedural texts vs narratives

Similar to dictionary entries, procedural texts are not an indigenous, conventionalised genre in Pacific cultures, as people prefer to demonstrate how this or that is done instead of describing it (Mosel 2006b). Consequently, the speakers have not yet developed conventionalised ways of describing procedures and seem to be free in their choice of pronoun to refer to generic agents. Some prefer the second person singular, others the first person exclusive plural or the third person plural pronoun. One speaker consistently uses the first person exclusive plural, which the editors of her texts always replace by the first person inclusive pronoun. But with one exception, all texts, which are spoken or written by various people, show the same kind of clause linkage construction which explicitly refers to a regular fixed order of actions. While in Teop narratives the sequence of events is simply expressed by paratactic and coordinate clauses, and the so-called tail-head construction, the procedural texts show constructions with adverbial clauses. Our first example (5) 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

[1] The VC *gono kahi* 'get from' is ditransitive with the source NP *o paka* functionning as the primary object and the theme NP *bono sikirinae* as the secondary object.

modifies it by *vakavara* 'finished' expressing that this action was finished, before he did the next one, i.e. *taatagi* 'prepare'.

(6)

me-	***ori***	***paa***	***dee***	***voosu***	***maa,***	***me-***	***ori***	***paa***	***ma***	***kahu,***
and-	3PL	TAM	carry	home	DIR	and-	3PL	TAM	come	scrape

me-	***ori***	***kahu***	***va-***	***kavara***	***bona***	***kano-kanono***[2]		***te-***	***ori,***
and-	3PL	scrape	ADV-	finished	ART	RED-	rope	PREP-	3PL

a-	***maa***	***kara***	***kave***	***te-***	***ori,***	***me-***	***ori***	***paa***	***tatagi***	***bari,***
ART-	PL	string	*kave*	PREP-	3PL	and-	3PL	TAM	prepare	4PL.OBJ

and they carried (the *kave* vines) home, and they scraped them[3] and they finished scraping their ropes, their *kave* strings, and they prepared them. (Sii_06R.56-60)

The second example (7) 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.

(7)

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***	***bono***	***kehaa***
1PL.INCL	then-	TAM	scrape	ART	bark	3SG.POSS	ART	shell

then we scrape its bark off with a shell

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

that we call sui.

[2] REDUPLICATION of a noun, expressing distributional plurality, i.e. the rope(s) each of them had prepared.
[3] I.e. scraped the bark off.

Be- ***ara*** ***kahu*** ***vaka-*** ***va-*** ***kavara*** ***e,***
when- 1PL scrape RED- ADV- finished 3SG
When we have finished scraping it,
eara ***re*** ***paa*** ***vaaroava*** ***e*** ***bono*** ***buaku*** ***ge*** ***...***
1PL.INCL then TAM dry.in.sun 3SG ART two or ....
then we put it into the sun for two ot three days.' (Eno_08W.4-6)

Other variants of this construction in procedural texts include:

(8) ***be-*** AGENT X ***va-*** ***kavara,*** AGENT ***re-*** ***paa*** Y
when AGENT X ADV- finished AGENT then- TAM Y
when AGENT has finished doing X, then AGENT does Y

(9) ***be*** ***kavara,*** AGENT ***re-*** ***paa*** X
when finished AGENT then- TAM X
'when it is finished, then you do X'

(10) ***be-*** AGENT ***tau*** X, AGENT ***re-*** ***paa*** Y
when AGENT about to X, AGENT then- TAM Y
'when we are about to X, we do Y

(11) ***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 constructions of narratives and procedural texts, I bought a rooster from a neighbour and asked him to slaughter it while I was taking a series of photographs. Luckily his four year old twins were helping him slaughtering 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 slaughtered a rooster during my last visit. In addition, I asked another woman to have a look at the photographs and describe how Teop people slaughter a rooster.

While in the procedural text nine clauses out of a total of 40 clauses are an adverbial clause introduced by *be* 'when' (12), the narrative text, which consists of 53 clauses, has not any of these constructions, but uses paratactic clauses instead (13):

(12) Procedural text

Be kavara,
when finished

be-	***nam***	***pee-***	***pee***	***va-***	***ruta-***	***rutaa***	***va-***	***kavara***	***eve***
when-	1PL.EXCL	RED-	cut	ADV-	RED-	small	ADV-	finished	3SG

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

'When it is finished, when we have finished cutting it into small pieces, they put it onto the fire. (Hel_13R.33-34)

(13) Narrative text

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

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

'But he finished slaughtering the rooster,
and they put it into the saucepan. (Pau_01R.51-52)

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 artefacts 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 colourful

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 lingusitcs itself, but a thorough corpus based study of property words in any of these languages is still missing. A preliminary investigation of the distribution of the property words across the Teop Language Corpus (Mosel 2007, 2009) showed that the most frequent property words *beera* 'big' and *mataa* 'good' frequently occur as the heads in VCs and APs, but never as the head of a NP, and that in contrast to activity words (verbals), property words (adjectivals) must take a prefix *va-* when they modify the head of a verb complex, e.g. *vabeera* 'to a great extent, too big, loudly, strongly', *vamataa* 'well', *tara vabeera* 'look big', *tara vamataa* 'look good'.

Table 3: Distribution of *beera* 'good' and *mataa* 'good'

lexeme		hits	VC head	NP head	AP head	juxtaposed modifier
beera	'big'	222	81	-	111	30
mataa	'good'	154	89	-	55	10

For colour words we did not have comparable data. The words *kakaavo* 'white', *paru* 'black' and *gogooravi* 'red' were only attested in two clauses each. But since we have started compiling a small fish dictionary in 2008, our database of colour words is growing and clearly shows that they are similar to *beera* 'big' and *mataa* 'good'. They can function as the head of both VCs and APs, are also used in comparative serial verb constructions and require the prefix *va-* in adverbial position, but they never occur as the head of a NP.

Table 4: Distribution of three colour words in the fish dictionary

		VC head	AP head	juxtaposed modifier
gogooravi	'red'	2	5	1
kakaavo	'white'	8	5	12
paru	'black'	5	9	14

Compare the function of *beera* (14) and the colour words *gogooravi* 'red'(15) and *paru* 'black' (16) in the following examples:

(14) NP AP VC

evehee a toobono a beera, [na beera oha nana]

but ART *toobono* ART big TAM big pass 3SG.IPFV

NP

bona pasupua

ART *pasupua*

'(The toobono looks like the genuine *pasupua*,)

but the *toobono* is big, is bigger than the *pasupua*.' (MD Fishes, *toobono*)

(15) NP AP$_{\text{predicate}}$

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 aranavi.

ART *sinarona* TAM red pass 3SG.IPFV ART *aranavi*

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

(16) ***Be- ori hovo ruene o- re paa tara va- paru.***

when- 3PL enter river 3PL- then TAM look ADV- black

'(While they are still staying in the ocean, they look white.)

When they enter the rivers, they look black.' (MD Fishes, *ovunaa*)

Since this paper does not deal with parts of speech, these three examples must suffice here to demonstrate that dictionary work can provide very useful data for the classification of parts of speech. 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 postion is held by the topic of the clause which can be the subject, an object or an adjunct. Teop does not have a passive construction. 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:

Table 5: Clause patterns

TOPIC	**VC**	**Argument**	**Argument**
S (subject)	VC	O1 (primary object)	O2 (secondary object)
O1 (primary object)	VC	S (subject)	O2 (secondary object)
O2 (secondary object)	VC	S (subject)	O1 (primary object)

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):

Table 6: Clause patterns of *hee* 'give'

clause patterns	**frequency**
S VC O1 O2	25
O1 VC S O2	6
O2 VC S O1	4

With *hee* 'give', the primary object (O1) refers to the recipient and the secondary object (O2) 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:

(17)

S:agent	VC		O1:patient			O2:instrument	
... a-re	*ma*	*nahu*	*a*	*guu*	*vai*	*bona*	*tahii.*
...1PL.INCL-so.that	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 analysing clauses of this kind, I had again the impression that the dominant, unmarked order was S VC O1 O2. But when the Teop research assistants collected descriptions of trees and what the parts of trees are used for, I realised that it would only make sense to speak of a dominant word order in respect to a particular type of text. If as in the tree descriptions the topic of discourse has the semantic role of a patient or instrument, it will function as an object of transitive and ditransitive clauses, but occupy the first position of the clause, as the dictionary entry for *asita* 'putty nut tree' nicely illustrates. The entry starts with the sentence:

(18)

O2		VC					S	O1	
O	*asita*	*[na*	*asi-*	*asita*	*ri-*	*]*	*ori*	*bono*	*sinivi.*
ART	putty.nut	IPFV	RED-	plaster	3PL.IPFV		3PL	ART	canoe

'The putty-nut tree, they use it for plastering the canoe.' (i.e. the nuts of the tree) (MD Trees, *asita*)

In the second clause of the entry (19), 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.

(19)

VC			S	O1		
[*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 medicine)' is reduced by the particle *ni* resulting in a transitive construction meaning 'use s.th. as traditional medicine'.

(20)	O		VC					S
	Asita	***me***	**[*na***	***pore-***	***porete***	***ni***	***ri*]-**	***ori.***
	plaster	also	TAM	RED-	make.medicine	APP	3PL.IPFV	3PL

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

6 Concluding remarks

Complying with the speech communities demands for educational materials does not need to be counter-productive to the aim of collecting data for a reference grammar. On the contrary, it may provide unexpected kinds of data that prove to be useful for a deeper understanding of the expressive power of the language and help to reduce the application of elicitation methods, which are "artificial even under the best circumstances." (Samarin 1967:59). The paper shows how the collection of various types of text can provide interesting grammatical data such as paratactic vs. embedding constructions, the expression specific and habitual events, the coding of inanimate topics, and the use of non-verbal predicates and various types of attributes:

1. Editing the transcriptions of oral narratives is a very practical method of collecting samples of constructions that native speakers regard as synonymous.
2. Parallel narrative and procedural texts about the very same topic like net making or slaughtering a chicken show how the contrast between specific and habitual actions and between specific and generic agents 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, and various kinds of adjectival attributes and relative clauses. The definitions of verbs, on the other hand, may contain nominalisations in subject position and complement clauses as predicates.

4. The descriptions of trees and their parts, and how they are used for manufacturing artefacts provide examples for constructions with inanimate topics and the expression of the semantic role of instrument.

In general, the analysis of the Teop data shows how important it is to distinguish various text types, and that consequently, the origin of the examples must be specified in any kind grammar: So I hope that this paper instigates further discussions on fieldwork methods and the compilation of corpora for previously unresearched languages, the interaction of lexicography and grammaticography, and the presentation of data in grammars and their accessibility in archives.

References

Abbi, Anvita. 2001. *A manual of linguistic fieldwork and structures of Indian languages.* München: Lincom Europa.

Aikhenvald, Alexandra. 2003. *A Grammar of Tariana, from northwest Amazonia.* Cambridge: Cambridge University Press.

Biber et al. 1999. *Longman Grammar of spoken and written English.* Harlow, Essex.

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 and 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.

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

Dixon, R.M.W. 2004. *The Jarawara language of Southern Amazonia.* Oxford: Oxford University Press.

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

Enfield. 2007. *A grammar of Lao.* Berlin: Mouton de Gruyter.

Everett, Daniel L. 2001. 'Monolingual field research', in Paul Newman and Martha Ratliff (eds.) *Linguistic Fieldwork.* Cambridge: Cambridge University Press, 166-188.

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.

Himmelmann, Nikolaus. 2006. '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, 1-30.

Lehmann, Christian. 1988. Towards a typology of clause linkage. In Haiman, John & Thompson, Sandra A. *Clause combining in grammar and discourse.* Amsterdam/ Philadelphia: Benjamins, pp. 181-225.

Magum, Enoch Horai & Maion, Joyce & Kamai, Jubilie & Tavagaga, Ondria with Mosel, Ulrike & Thiesen, Yvonne (eds) 2007. *Amaa vahutate vaa Teapu.* Teop Legends. Kiel: CAU, Seminar für Allgemeine und Vergleichende Sprachwissenschaft.

Mosel, Ulrike. 2006a. 'The art and craft of writing grammars', in Felix Ameka, Alan Dench and Nicholas Evans. *Catching language.* The standing challenge of grammar writing, 41-68.

Mosel, Ulrike. 2006b. 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. 2006c. 'Sketch grammar', in Jost Gippert, Nikolaus Himmelmann and Ulrike Mosel (eds.) *Essentials of language documentation.* Berlin, New York: Mouton de Gruyter, 301-309.

Mosel, Ulrike with Yvonne Thiesen. 2007. *Teop sketch grammar.* www.mpi.nl/DOBES/projects/teop and http://www.linguistik.uni-kiel.de/mosel_publikationen.htm#download

Mosel, Ulrike. 2007. A corpus based approach to valency in a language documentation project. Pre-ALT workshop, Paris, 24.9.2007. http://www.linguistik.uni-kiel.de/mosel_publikationen.htm#download

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. 2009. Lexical flexibility revisited, a corpus based approach. 11th International Conference on Austronesian Linguistics, 22-26 June, Aussois, France. http://www.linguistik.uni-kiel.de/mosel_publikationen.htm#download

Mosel, Ulrike. forthcoming a. *Lexicography in endangered languages.* In Peter Austin (ed.). *The handbook of endangered languages.* Cambridge: CUP.

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

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.

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.

Vries, Lourens de. 2007. 'Some remarks on the use of Bible translations as parallel texts in linguistic research', in *Sprachtypologie und Universalienforschung.* Vol. 60.2, 148-157.