Advances in the accountability of grammatical analysis and description by using regular expressions

Ulrike Mosel - Kiel University

Drawing on my experiences as a grammaticographer for the past thirty years, I want to discuss the question to what extent digital grammaticography contributes to the accountability of grammatical descriptions by comparing my earlier traditional methods of grammatical analysis with those that I have recently started practicing in the Teop language documentation project[1]. Teop is an Austronesian Oceanic language spoken in the Autonomous Region of Bougainville, Papua New Guinea, and genetically related to the other two languages, Tolai and Samoan, for which I wrote grammatical descriptions (Mosel 1984, Mosel & Hovdhaugen 1992).

On the basis of a language documentation corpus of approximately 260 000 words, which is slowly, but constantly growing, I am currently writing a non-electronic Teop reference grammar that hopefully could one day be transferred to an electronic format and be linked to our Toolbox lexical database (SChwartz et al. 2007) and the ELAN text corpus (Mosel et al. 2007). The grammar starts with an introductory phonology chapter and an overview of the structure of phrases, clauses and complex sentences, and then proceeds in the traditional ascending linear fashion from word classes and morphology to simple clauses and complex sentences (Mosel 2006a). Each chapter is selfcontained and can be read by itself once the reader has read the introductory overview. The language documentation (LD) on which the grammar is based is archived in the DoBeS archive[2].

10.1 Language documentation and corpus linguistics

Writing a grammar on the basis of a language documentation is a corpus linguistic enterprise, but as LD corpora are quite different from the large corpora of European languages (see Table 1 for a summary), LD grammaticography has to develop its own corpus linguistic approach aiming at grammatical descriptions that are reasonably complete from a typologist's point of view, but would also be respresentative for the linguistic data contained in the corpus. This latter kind of coverage is called coextensitivity by Good (this vol. §1.5.1).

10.2 The notions of completeness, coextensitivity and sample representativeness

To what extent a LD grammar is a comprehensive grammar depends on the “documentary coverage”, i.e. “the extent to which a documentary corpus actually includes the information

[1] The Teop Language Documentation project was funded by the DoBeS programme of the Volkswagen Foundation from 2000 to 2007. Subsequently the project "A corpus-based grammar of Teop, an Oceanic language of Bougainville, Papua New Guinea" was funded by the Deutsche Forschungsgemeinschaft from 2008-2011.

[2] http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI77915%23

needed to create a complete grammar of the language." (Good this vol. §1.5.1) In contrast to this generally defined notion of completeness, the notion of coextensitivity relates the content of a grammar to the particular linguistic data that are available in the corresponding LD corpus. A grammar that is based on a small corpus may not be "complete" in a typological sense, but its coextensitivity would be adequate, if it covers all the information that an analysis of the corpus can provide.

Linguists working on LD grammars agree that a grammar should be data-driven and accompanied by a corpus in order to facilitate the verification or falsification of the analysis (see Nordhoff 2008 for a summary of what is considered a good grammar by typologists, and Bender & Ghodke this vol.). But what is less clear is how the degree of coextensitivity of a grammatical description can be made evident by the grammaticographer and consequently be skrutinized by the reader.

Good (this vol. §1.5.3) assumes that "it will often simply be not possible" to base a grammatical description "(more or less) on all of the available data. ... Instead, it should be based on a sample of the data that results in a description that is representative of all the collected data." But how we can know to what extent this sample would be representative without analyzing more data remains unclear. Good himself admits "In any event, the question of what kind of sample of documentary materials can be considered representative enough to form the basis of a description that would also cover the remaining materials appears to be an interesing one, and work in this area would be quite useful for developing general methods for assessing adequacy in coextensitivity."

	Conventional corpus	**LD corpus**
language	well-researched standardized European language	unresearched, non-standardized non-European
texts	available in print and on-line	recorded, transcribed, translated
corpus builder	team of professional native speakers	single non-native speaker in cooperation with non-professional native speakers
size	millions of words	below 1 million
purpose	linguistic research	language preservation linguistic and anthropological research

Table 10.1: Conventional vs. language documentation corpora

Since representativeness can hardly be achieved in the small opportunistic corpora of LDs and the assessment of representativeness is still a matter of debate (Clancy 2010:86-87, McEnery & Hardy 2012:10), I would restrict the notion of coextensitivity to the relationship between the description and the selected text collection on which the description is based and strictly distingish it from the notion of representativeness which in corpus linguistics refers to the relationship between corpora and language varieties. Thus the description of a grammatical phenomenon is adequate in coextensitivity, if it accounts for all its occurences in the selected text collection, irrespective of the size and the kind of the collected texts. But this text collection may not be a representative sample for a particular register or genre. Conversely, a text collection may be considered representative for a register or genre, if it covers all or most of its grammatical phenomena, even if its grammatical description misses some grammatical phenomena and consequently lacks a high degree of adequacy in coextensitivity. The writer of a LD grammar should aim at adequacy in coextensitivity, which is solely his responsibility, whereas the representativeness of the text collection is beyond his control because it depends on the kind, size and number of texts the speech community supplies(Mosel 2006b).

10.3 Corpus building in a LD project

As mentioned above, LD corpora are opportunistic corpora, i.e. corpora that “represent nothing more or less than the data that it was possible to gather for a specific task.” (McEnery & Hardie 2012:11) In other words, the building of LD corpora does not follow previously specified corpus design criteria and hence would not qualify as corpora but merely as "electronic text libraries" for some corpus linguists. (Atkins, Clear & Ostler 1992:1) But this does not mean that the texts of a LD could not be classified with respect to genres, themes and situation characteristics and accordingly organized into subcorpora. At least for frequently occuring grammatical phenomena the division of the corpus into such subcorpora may reveal regular patterns of contrasting constructions that are significant for distinct registers. While, for instance, in Teop narratives a sequence of actions such as the making of a fishing net or the butchering of a chicken is expressed by simple paratactic or coordinated clauses, the very same kind of sequences of actions is expressed by complex sentences with adverbial clauses in comparable procedural texts (Mosel forthoming).

For the grammatical analysis and description of the Teop LD corpus, which now (May 2012) comprises approximately 260 000 words, we classified the texts according to their mode of production and genre into 11 subcorpora:

1. recordings and transcriptions of oral legends
2. edited versions of the transcriptions of the oral legends
3. written legends
4. recordings and transcriptions of personal narratives
5. edited versions of the transcriptions of personal narratives
6. written personal narratives
7. recordings and transcriptions of encyclopedic descriptions of things and activities
8. edited versions of the transcriptions of encyclopedic descriptions of things and activities

9. written encyclopedic descriptions of things and activities
10. interviews on cultural practices that have not been edited yet
11. example sentences that were provided by two native speakers for the Teop Lexical Database

A finer subclassification did not seem suitable because the more diversified the subclassification of a rather small text collection is, the more difficult it becomes to recognize regular patterns of language use.

The corpus is compiled in Elan, which facilitates simultaneous searches with the query language of Regular Expressions on several tiers such as the transcription and the free translation tier. Although parts-of-speech tagging and morphological glossing would make the grammatical analysis easier, we decided to gloss only a few texts because we wanted to record, transcribe and translate as many texts as possible. Consequently we only created three tiers:

1. the reference tier which gives each annotation a label that identifies the text and the number of the annotation, e.g. Aro_05R.003 for third annotation of the fifth recorded spoken text of Arovina Magum;
2. the transcription tier;
3. the free translation tier.

Since Teop is nearly an isolating language and the corpus is accompanied by a lexical database in Toolbox and a sketch grammar (Mosel with Thiesen 2007), it is possible to understand the grammatical constructions and do the glossing in the future even if no native speakers are available. In the grammar the labels on the reference tier are used to indicate the source of all cited examples and thus allow to quickly retrieve them in their original context although the grammar is not linked to the corpus.

10.4 Accountability of grammatical descriptions

Quoting Nordhoff (2008), Rice (2006:395), Noonan (2006:355), Weber (2006:450), Bender & Ghodke this vol. §8.4.1) postulate three maxims of best practices for the accountability of grammatical descriptions:

1. “If we value the application of the scientific method, more sources for a phenomenon are better than fewer sources.”
2. “If we value the application of the scientific method, every step of the linguistic analysis should be tracable to a preceding step, until the original utterance of the speaker is reached.”
3. “If we value the application of the scientific method, the context of the utterance should be retrievable.”

Although many linguists agree on these maxims and the retrievability of examples in grammatical descriptions has a centuries long tradition in Classical Greek and Latin linguistics, most typological grammar writers do not bother to explicitely state the sources of their examples and thus bring discredit upon linguistic typology as a science of language. Having been educated in the old fashioned philological tradition, I tried whereever possible to quote examples from published original materials in my grammatical descriptions of Tolai and Samoan (Mosel 1984, Mosel & Hovdhaugen 1992) so that in principle most examples are retrievable. But as these publications are only available in a few public libraries, most readers don't have the chance to skrutinize my data and analyses. An exception is Gillian Sankoff (1993) who on the basis of my text collection (Mosel 1977) discovered that I overlooked the similarity between the Tolai focus particle *iat* and its Tok Pisin equivalent *yet* in my comparative study of Tolai and Tok Pisin (Mosel 1980). Due to the rise of documentary linguistics, however, it will soon become a standard that grammars are accompanied by digital corpora, provide easy access to the sources of the examples and thus fulfil the first and the third maxim quoted above.

The second maxim that each step of the linguistic analysis should be tracable is impossible to follow in traditional grammaticography. When analysing Tolai and Samoan texts, I wrote thousands of quotes on cards and stored them in shoeboxes. So I had boxes for alphabetically sorted functional words, for grammatical constructions, and interesting phenomena such as noun/verb distinction or idiomatic phrases for the expression of time, but when I recently realised that they won't be of any use for me or other linguists in the future, these boxes went into the recycling bin.

Electronic grammars which facilitate the retrieval of the examples from a corpus by one or a few mouse clicks reach a higher degree of accountability for practical reasons, but with respect to the second maxim they are in principle not much different from traditional grammars, if it is only the easy retrievability of examples that makes the difference. If one takes the request for scientific accountability and coextensitivity seriously, one could go a step further and inform the readers of how the particular example was found and how other, if not all examples of the grammatical phenomenon in question can be retrieved from the corpus. This is at least to some extent practicable if one uses Regular Expressions for one's seraches and documents their particular formulas.

As illustrated by the examples given below and in the appendix, regular expressions facilitate:

1. searching for discontinuous sequences of words;
2. searching for two or more alternative expressions at the same time;
3. searching for some expression with the exclusion of other expressions;
4. searching for reduplications.

Using Regular Expressions for searches in an ELAN corpus also allows us to simultaneously search on several tiers and explicitely state

1. how many examples the corpus or a selected subcorpus provides for a particular construction;
2. if certain strings of words in the example represent frequent collocations or colligations;
3. how a grammatical formative or construction in question relates to alternative expressions in terms of frequency and/or register.

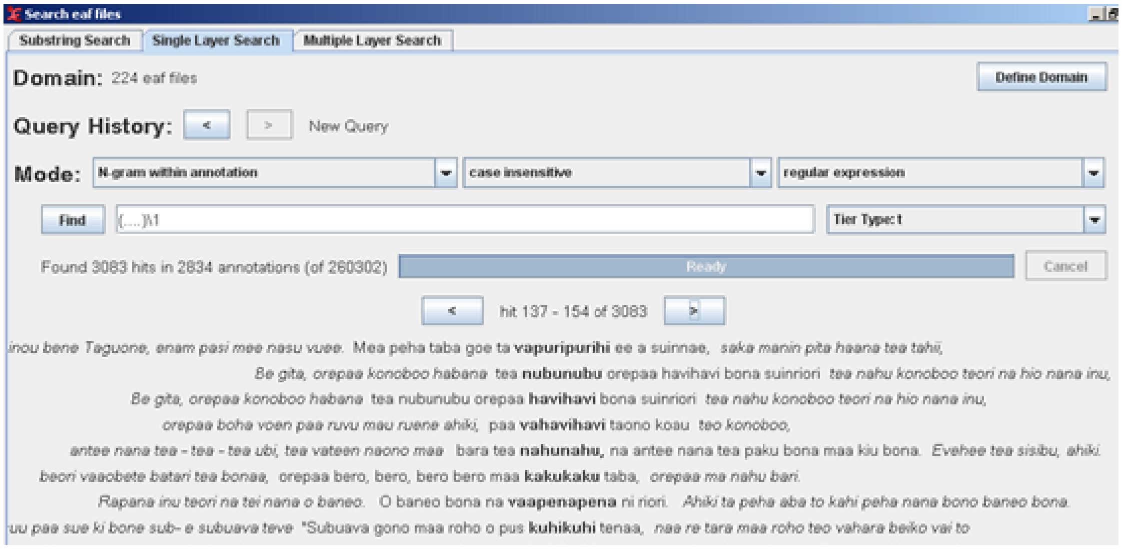
In sum, the use of regular expressions for searches and the documentation of these seraches in the grammatical description advances the degree of accountability of a grammar and allows more explicit statements about its coextensitivity.

10.5 The use of regular expressions in grammatical analysis and description

Most searches can be performed with quite simple formulas (see the appendix). For example, it is very easy to find all reduplicated wordforms with regular expressions, which is impossible with simple word searches. Once I had observed that in Teop the reduplicated sequence is a prefix consisting of two, three or four letters, I could search for all reduplicated wordforms, as illustrated here for reduplicants of four letters. The regular expression

(1) (....)\1

finds all sets of four letters that are repeated once, altogether 3083 tokens in the Teop Language Corpus, e.g. simple forms like *nubunubu, havihavi* and prefixed forms like *vapuripurihi, vahavihavi* and *vaapenapena.*



Domain: 224 eaf files

Mode: N-gram within annotation | case insensitive | regular expression

(....)\1

Tier Type: t

Found 3083 hits in 2834 annotations (of 260302)

hit 137 - 154 of 3083

inou bene Taguone, enam pasi mee nasu vuee. Mea peha taba goe ta **vapuripurihi** ee a suinnae, *saka manin pita haana tea tahi,*
Be gita, orepaa konoboo habana tea **nubunubu** orepaa havihavi bona suinriori *tea nahu konoboo teori na hio nana inu,*
Be gita, orepaa konoboo habana tea nubunubu orepaa **havihavi** bona suinriori *tea nahu konoboo teori na hio nana inu,*
orepaa boha voen paa ruvu mau ruene ahiki, paa **vahavihavi** taono koau *teo konoboo,*
antee nana tea - tea - tea ubi, tea vateen naono maa bara tea **nahunahu,** na antee nana tea paku bona maa kiu bona. *Evehee tea sisibu, ahiki.*
beori vaaobete batari tea bonaa, orepaa bero, bero, bero bero maa **kakukaku** taba, *orepaa ma nahu bari.*
Rapana inu teori na tei nana o baneo. O baneo bona na **vaapenapena** ni riori. *Ahiki ta peha aba to kahi peha nana bono baneo bona.*
uu paa sue ki bone sub- e subuava teve "Subuava gono maa roho o pus **kuhikuhi** tenaa, *naa re tara maa roho teo vahara beiko vai to*

Figure 10.1: Search for reduplicated wordforms

Using more complex formulas you can restrict the search to forms with or without affixes in general or to forms with particular affixes as briefly illustrated in the appendix. But here I want to present one of my most complicated examples, namely the investigation of the noun-verb distinction in Teop, and discuss its problems.

Oceanic languages are well known for their presumably weak noun/verb distinction (Hengefeld, Rijkoff & Sievierska2004, Hengeveld & Rijkoff. 2005), but the investigation of lexical flexibility in Teop seems the first corpus-linguistic study of this phenomenon. The workflow of this investigation was as follows:

1. Identify the elements constituting the constructional frames[3] of noun phrases (NPs) and verb complexes (VCs) - constructional frame consists of functional morphs and empty, syntactically defined head and modifier positions for content words and stems).
2. Identify the functional elements that directly precede or follow the empty head position for the content word.
3. Construct Regular Expressions for the head position in NPs and VCs.
4. Select a few prototypical frequent action and object words e.g. ‘do, make‘, ‘say‘, ‘person‘, ‘thing‘, and search.

In Teop the constructional frames of NPs and VCs are quite complex. But for our purposes it was sufficient to construct the formulas that only include those functional elements that immediately precede either the head of NPs or the head of VCs, i.e. the articles, a plural marker, the numerals for ‘one’ and ’two’, and the diminutive particle for NPs and for the VC the pre-head tense, aspect and mood markers, the conjunction *re* ‘so that, then’, and the relative pronoun *to*:

(2) Regular expression for Teop NPs:

```
(\ba\b|\bo\b|\bbona\b|\bbono\b|\bpeha\b|\bpeho\b|\bbua|\b|\bbuo\b|\bamaa\b|\bmaa\b\|\bsi\b)
\bHEAD\b
```

This formula means:

(3) find any lexical form that is inserted in the HEAD position and preceded by

- the specific basic article *a* or *o* or
- the object article *bona* or *bono,* or
- the numeral *peha, peho* ‘one‘ or *bua, buo* ‘two’, or
- the plural particle *maa* or the complex form *amaa* consisting of the article *a* and the plural marker *maa,* or
- the diminutive particle *si.*

(4) Regular expression for Teop VCs

[3] Compare the notions of collocational frame works, grammar patterns and colligates in Stefanowitsch & Gries 2009:936-937)

(\bare\b|\bkahi\b|\bna\b|\bore\b|\bpaa\b|\bpasi\b|\bre\b|\brepaa\b|\btau\b|\bto\b|\btoro\b)
\bHEAD\b

This formula means:

(5) find any lexical form that is inserted in the HEAD position and preceded by

- one of the six TAM markers *kahi, na, paa, pasi, tau* or *toro* or
- the conjunction *re* ‘then, so that’ or
- the relative pronoun *to*, or
- the conjunction *re* with the 1inc.pl prefix *a-* , i.e. *are,* or
- the conjunction *re* with the with 3sg/pl-prefix *o-*, i.e. *ore,*
- or the conjunction *re* ‘then, so that’ with the suffixed TAM marker *paa,* i.e. *repaa.* [4]

These formulas do not cover the full range of possible contexts of NP and VC heads. The most notable exception is the very first position of the clause. In fast speech the speakers sometimes omit the article of NPs or the tense/aspect particle of a VC. Furthermore, imperatives are always unmarked for tense, aspect and mood and are not preceded by the conjunction *re* or the relative pronoun *to.* Consequently, the search with these formulas cannot reach the highest degree of coextensitivity, but at least the readers are informed about the method and the limitations of the investigation, which contributes to the accountability of the grammatical description.

A further problem is that some functional words are homonyms: the form *na* represents a TAM marker and a portmanteau morph representing the 3pers.sg. possessive marker and the article of the following NP, and *to* the relative pronoun preceding the VC and a rare non-specific article. This means that I either had to exclude these homonyms from my investigation or check all examples containing *to* and *na.* I opted for the latter, which was not too time consuming as the selected object words only rarely occured with *na* or *to.* Had I choosen the first solution, the investigation wohld habe been less coextensive, but would have still sufficed the accountability maxime as long as I documented my choice.

The result of these searches was, as shown in the table below, that action words and object words are flexible with respect to the head positions in NPs and VCs, but that, as expected, action words are much more frequent in the VC head position and object words are much more frequent in the NP head position. Further research which employed the same kind of strategy revealed that action and object words are morpho-syntactically distinct in modifier positions so that verbs and nouns

[4] Unfortunately the current version of ELAN 4.3.2 has problems to digest these long formulas so that I had to split them into two: For the heads of NPs I used (\ba\b|\bo\b|\bbona\b|\bbono\b|\bpeha\b|\bpeho\b) \bHEAD\b and (\bbua|\b|\bbuo\b|\bamaa\b|\bmaa\b\|\bsi\b) \bHEAD\b,
and for the heads of VCs
(\bare\b|\bkahi\b|\bna\b|\bore\b|\bpaa\b|\bpasi\b) \bHEAD\b and
(\bre\b|\brepaa\b|\btau\b|\bto\b|\btoro\b) \bHEAD\b.

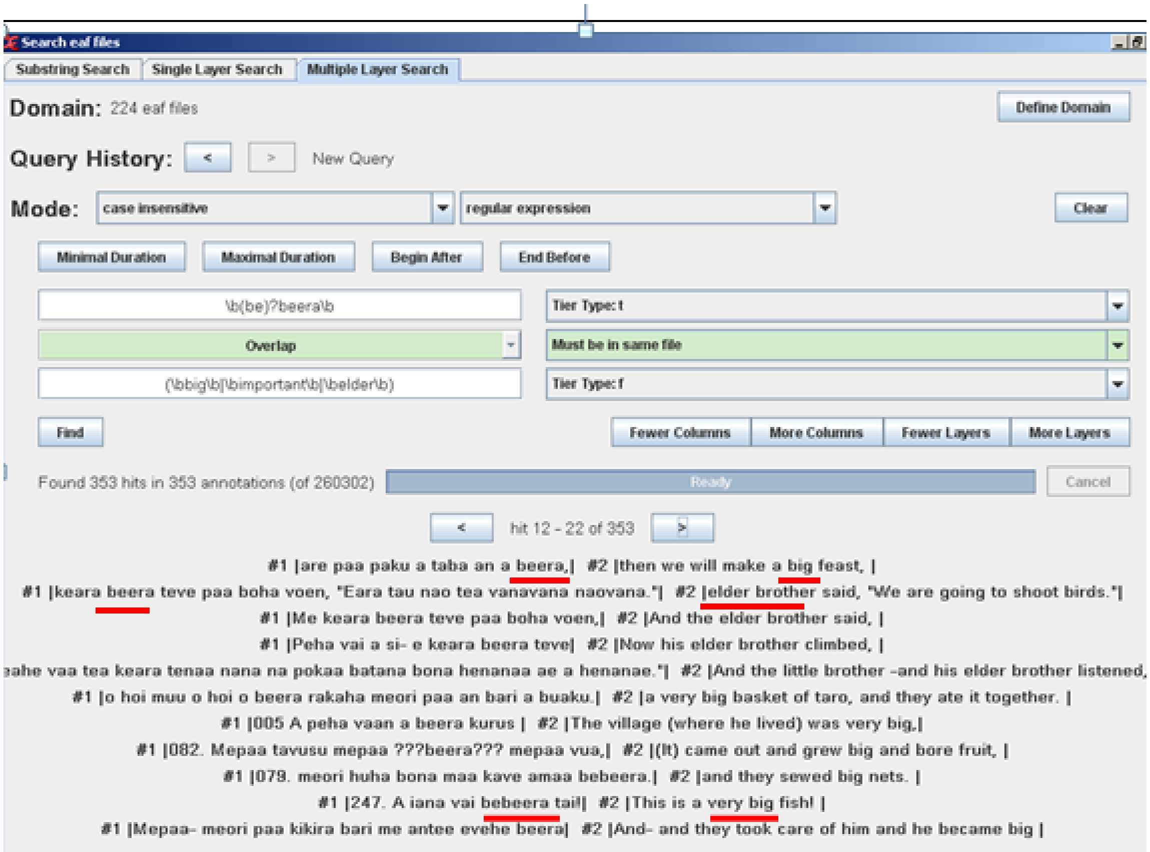
are formally distinct word classes in Teop, but at the same time show flexibility with respect to the NP and VC head positions, which contradicts Hengeveld, Rijkoff and Siewierska's (2004) theory of lexical flexibility.

Since I only investigated a few prototypical action and object words, my grammatical description is strictly speaking not adequate in coextensitivity. This would only be the case if I had analysed and described all action and object words of the corpus, which would have been too time consuming and also unnecessary in my prototype approach. As I document the Regular Expressions of my searches, the readers can re-enact them and test the formulas with other head words.

		VC head	NP head			VC head	NP head
asun	'hit, kill'	70	1	*aba*	'person'	6	243
hua	'paddle'	115	4	*beiko*	'child'	1	366
mosi	'cut'	85	9	*iana*	'fish'	-	176
nao	'go'	1002	5	*moon*	'woman'	4	665
paku	'do, make'	996	10	*naono*	'tree'	-	136
pita	'walk'	63	3	*otei*	'man'	-	478
rosin	‘flee’	211		*taba*	'thing'	11	461
sue	'say'	752	10	*vasu*	'stone'	-	48

Table 10.2: The distribution of typical verbs and nouns as heads of VCs and NPs

With lexical items the problem of homonymy can easily be solved by simultaneous seraches on the transcription and the translation tier. In order to distinguish, for instance, the adjective *beera* ‘big, elder, important’ from the noun *beera* ‘chief’ or the adjective *beera* ‘chiefly’, you only need to search for *beera* on the transcription tier and its three translation equivalents on the translation tier.



Domain: 224 eaf files

Mode: case insensitive | regular expression

```
\b(be)?beera\b                              Tier Type: t
Overlap                                     Must be in same file
(\bbig\b|\bimportant\b|\belder\b)           Tier Type: f
```

```
Found 353 hits in 353 annotations (of 260302)

hit 12 - 22 of 353

#1 |are paa paku a taba an a beera,|  #2 |then we will make a big feast, |
#1 |keara beera teve paa boha voen, "Eara tau nao tea vanavana naovana."|  #2 |elder brother said, "We are going to shoot birds."|
#1 |Me keara beera teve paa boha voen,|  #2 |And the elder brother said, |
#1 |Peha vai a si- e keara beera teve|  #2 |Now his elder brother climbed, |
ahe vaa tea keara tenaa nana na pokaa batana bona henanaa ae a henanae."|  #2 |And the little brother -and his elder brother listened,
#1 |o hoi muu o hoi o beera rakaha meori paa an bari a buaku.|  #2 |a very big basket of taro, and they ate it together. |
#1 |005 A peha vaan a beera kurus |  #2 |The village (where he lived) was very big.|
#1 |082. Mepaa tavusu mepaa ???beera??? mepaa vua,|  #2 |(It) came out and grew big and bore fruit, |
#1 |079. meori huha bona maa kave amaa bebeera.|  #2 |and they sewed big nets. |
#1 |247. A iana vai bebeera tail|  #2 |This is a very big fish! |
#1 |Mepaa- meori paa kikira bari me antee evehe beera|  #2 |And- and they took care of him and he became big |
```

Figure 10.2: Search for *beera* or *bebeera* with the meaning ‘big’, ‘important’ or ‘elder’

If in the chapter on word classification I only gave the references for my examples, my description would be much less adequate in coextensitivity, even if all of them were retrievable in the corpus because these few examples merely illustrate lexical flexibility, but do not demonstrate how my analysis and description actually relates to the corpus. Furthermore, the accountability of my description would be less scientific in view of the second maxim quoted above.

In the Teop Reference Grammar (Mosel in prep.) frequently used, relatively simple regular expressions as the one in Figure 10.2 are described in an appendix, whereas the application of complex searches is documented in the endnotes of each chapter. In an electronic grammar each example could perhaps be linked to its particular Regular Expression formula which together with all other formulas would be stored in a single separate file.

10.5 Concluding remarks

Comparing my previous personal reserach methods with those of my current analysis of the Teop language, I am convinced that Whalen (2004) was right when he assumed that “the study of endangered languages will revolutionize linguistics.” It is not only the way how we record speech by audio and video recordings, process the data in annotated digital corpora, make them globally accessible via the internet, and search them by the means of sophisticated query languages like Regular Expressions, but it is also the way how we - literally speaking - look at the data. Browsing through concordances as the ones depicted in Fig. 10.1 and Fig. 10.2 sharpens your eyes for regular patterns of language use and variation. It inspires you to create and test new formulas in your query language to either narrow down or extend your searches and thus explore the complex network of form-meaning correspondances in a hitherto unresearched language. With some practice you eventually "think regular expressions", as Friedl (2006:6) puts it, and undertand how selected lexical units interact with certain constructions and, conversely, under which conditions selected constructions accommodate certain types of lexical units.

Unfortunately digital formats change so rapidly that digital archives cannot give long-term guarantees that they would continously covert electronic grammars to new formats. Therefore I strongly recommend that the developers of electronic grammars provide for a function that facilitates the production of print outs and that these print-outs are stored in traditional libraries.