LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 2, No. 2 (December 2008), pp. 296-324
http://nflrc.hawaii.edu/ldc/

Electronic Reference Grammars for Typology: Challenges and Solutions

Sebastian Nordhoff
University of Amsterdam

Electronic publication offers new possibilities for the creation and exploration of grammatical descriptions. This paper lists values influencing the structure of electronic grammatical descriptions. It then investigates challenges and solutions for a grammar authoring software trying to adhere to these values in the domains of data quality, creation of the description, and exploration of the description. The paper closes by discussing possibilities for the standardization of grammatical descriptions on a macroscopic level, complementing the standardization efforts on a more fine-grained level like GOLD or CRG.

1. INTRODUCTION.[1] Grammatical descriptions of well-known or under-studied languages are one of the primary data sources for linguistic typology. The number and quality of grammatical descriptions (GDs) have peaked at times, but the (meta)theoretical discipline of grammaticography had by and large been left unexplored until the turn of the millennium (Mosel 2006a).[2] Recent years have seen an increase in relevant publications, such as Payne and Weber 2006, Barwick and Thieberger 2006, Gippert et al. 2006, and Ameka et al. 2006, that try to assist aspiring authors of grammatical descriptions to structure their work. Along with new theoretical approaches to grammaticography, the explosion of information technology provides new means to generate and present linguistic content. This paper will explore the intersection of grammaticography and electronic publishing, and discuss values, problems, and possible solutions.

2. SCOPE OF THIS PAPER. This paper deals with the intersection of grammaticography and information technology. In doing this, it limits itself to the coverage of reference grammars of the kind used in language typology. Traditionally, such grammars are published on paper and contain detailed descriptions of all areas of grammar from phonology to discourse. Their primary target audience is the academic community of linguists (Noonan 2006:353), and the concepts (for example, voice onset time, clitics, syntactic ergativity…) and terminology (ablaut, trochee, circumfix...) often make these grammatical descriptions hard to access for lay persons. There are other types of grammars, such as pedagogical grammars or learner's grammars, that target other audiences (Payne 2006:382). These kinds of grammars will not be dealt with in this paper, but some of the aspects presented below might be applicable to them as well. It is true that some published grammatical descriptions serve a

[1] I would like to thank Eva van Lier, Rachel Selbach, and Robert Cirillo for comments on earlier versions of this paper. I would also like to thank three anonyous reviewers whose very constructive comments have helped to improve this paper significantly. The usual disclaimers apply.

[2] Earlier treatises include Zaefferer 1998a, and Christian Lehmann has also written extensively on the topic (1989, 1993, 1998, 2002, 2004a, 2004b).

E-ISSN 1934-5275

double function as a reference work for typologists and speakers alike (see the discussion of Valentine 2001 in Rice 2006:412). The integration of this wider target audience could possibly build on the ideas presented here, but will not be pursued in this paper.

Dictionaries are a related form of publication. Of course, they are concerned mainly with lexical content, while GDs are more concerned with grammatical content. However, sometimes, lexical content can be used for a particular grammatical function, e.g., past time reference coded by an adverb (Mosel 2006:46, 61; Weber 2006a:422). There are thus intersections between GDs and dictionaries; an integrated treatment of electronic GDs and dictionaries is a possibility that should be looked into in more detail in future research.

3. LINEAR AND NON-LINEAR ELECTRONIC PUBLICATIONS. The internal structure of electronic publications can be linear or nonlinear. Linear publications follow the traditional structuring schema of a book and are divided into numbered chapters, which usually have a well defined page layout with numbered pages. Every page has the same size. A scan of a paper grammar would fall within this category, but so would the pre-print pdf file of a grammatical description. These linear electronic GDs are electronic in the sense that they can be made available online, but they do not use hypertext.[3] Of the over 1000 electronic grammars collected by Harald Hammarström, all are linear according to him (p.c. 2008). Whereas for descriptions on paper there is a certain linear structure that must be followed when building up the content, this is not the case for electronic grammars (Comrie 1998:14), which can be structured in a web-like fashion. A web-like structure has no linear order, but extensive linking between pages.[4] Each page covers a certain topic, but some pages may be longer than others. Each page links to other pages which are relevant to it or otherwise related. New means of electronic publishing present opportunities for this kind of GD, but also some challenges. Some readers might find advantages in such a nonlinear approach, such as easy navigation through links and the possibility of accessing background information (e.g., measurements of voice onset time) only if required, and skipping it otherwise. On the other hand, there might be some drawbacks, like the impossibility of presenting data in the didactically most useful way.

In this paper, I will explore the possibilities and challenges of nonlinear GDs. Since different readers have different expectations and preferences, a general categorization of advantages and drawbacks cannot be made. The favorite feature of one user might be undesirable for another one. For this reason, I will use the approach adopted by Bird and Simons (2003), who first define values to which users might subscribe, e.g., accuracy or actuality. Based on these values, relative preference can be given ("If you value X, then A is better than B," (cf. Bagish 1983). To use an example from Bird and Simons 2003:

---

[3] Some might use links for quick cross-referencing, but this is still done within the linear framework of numbered pages; i.e., it sends the reader to another point on the line.

[4] Comrie does not describe in detail how links would be constructed. There are two possibilities: either make use of markup in the document, which is more elegant, or do the linking by hand, by selecting a start point on a page and manually inserting a target.

SAFETY. We value ongoing access to language resources over the very long term. Thus the best practice is one that stores copies of resources in multiple locations so as to ensure against catastrophic damage to a single repository.

This recommendation holds only for people who share the value SAFETY.

In the next section, I will enumerate some values that authors or users of grammatical descriptions might hold. Electronic nonlinear grammars as a publication format will have to be judged with regard to how they respond to these values. Given that electronic GDs are at the intersection of linguistics and computer science, I will focus on values relevant for this intersection. Values that pertain to both paper grammars and electronic grammars alike (CLARITY, CONSISTENCY, ACCURACY) will not be discussed (for a list, see Rice 2006b). Neither will I discuss values that pertain generally to all forms of web publication. Examples are LONGEVITY, dealing with the decay of physical storage or SAFETY, or dealing with protection of the resources from catastrophes. For the more technical aspects of electronic documents, to which electronic GDs must respond as well, the reader is referred to the discussion of the seven dimensions of portability in Bird and Simons 2003.

4. VALUES. In this section, I will describe some values which creators or readers of GDs might hold. The next section will present some responses to these values. In her article "A typology of good grammars," Rice 2006b lists criteria that are used to assess the quality of a GD. These criteria are reflected in the values presented below. I have also included some values emanating from questions posed at several presentations of the GALOES grammar authoring platform (Nordhoff 2007a,b,c).

Some of these values can be conflicting; e.g., one user might prefer the most recent, albeit sketchy analysis of a phenomenon, while another might prefer a more conservative approach. It is not the purpose of this list to resolve these conflicts and define what GDs should be like. Rather, the purpose is to present a list of expectations and values that different people might have with regard to GDs. These expectations are given as numbered maxims (cf. Lehmann 2004b on the categorial imperative of language description).[5] Every maxim applies only if the value stated at the beginning of the section containing it is shared. Each maxim is normally in the form *If we hold the value given in the title of the section, a GD doing X is better than a GD not doing X*. For ease of reading, this is generally shortened to *X should be done*, but the long version stated above should be borne in mind. The short form should not lead to the interpretation that the maxims are inviolable. Constraints of time and money will influence the maxims that a description project can respect, and probably very often the description will aspire to respect the maxims judged most important rather than to respect all of them. In the words of the Resource Creation Group (2003), "If the choice is between having nothing and having less-than-best-practice, the choice must be to have something."

The expectations or values can be grouped in three domains: data quality, creation by the author, and exploration by the reader. These domains will now be discussed in turn.

---

[5] As a general disclaimer, I do not want to endorse any of these maxims in this paper. My purpose is only to present them.

4.1. DATA QUALITY.

4.1.1. ACCOUNTABILITY. We value application of the scientific method.

(1) Every step of the linguistic analysis should be traceable to a preceding step, until the original utterance of a speaker is reached.

As an example, analysis of topic and focus might rely on morpheme analysis, which relies on phoneme analysis, which relies on phonetic transcription, which relies on the utterance of the speaker. The documentation behind the description should be provided (Bird and Simons 2003:13, 17).

(2) Every phenomenon described should be sourced using an actual utterance.

For example, only stating that *questions are formed by inversion of subject and verb* is not sufficient. An illustrating example should be provided, preferably from naturalistic speech.

(3) More sources for a phenomenon are better than fewer sources. (Rice 2006:395; Noonan 2006:355)

Ample provision of examples allows for testing of competing analyses. If only one example is provided, or examples of a very similar nature, the validity of the presented analysis is difficult to check. If, on the other hand, several, diverse examples are provided, the relative power of competing analyses or generalizations is easier to assess (Weber 2006b:446).

(4) The context of the utterance should be retrievable. (Weber 2006b:450)

Linguists who assume that language is used to reach communicative goals will be interested in knowing about the context of the utterance in terms of who spoke to whom, where, when, and about what.

4.1.2. ACTUALITY. We value scientific progress.

(5) A GD should incorporate provisions to incorporate scientific progress.

A GD is never finished. This statement is even more accurate for a grammar of an underdescribed language. It is just a mere sketch of some more or less incomplete observation (cf. Payne 2006:369ff., Weber 2006a:418, Rice 2006b:396, Cristofaro 2006:139).

In other words, a grammatical description must be maintained ("grown" in the terms of Weber 2006a). New insights must be added, old mistaken analyses commented upon and corrected, more material added, etc. (see also Comrie 1998, Zaefferer 2006:115).. This also means that very short observations can be "jotted down" (Weber 2006a:435) as soon

as they are made. In this manner, we can start with a very sketchy overview and add more information to it as time progresses (cf. Mosel 2006a:52).

(6) The GD should present state-of-the-art analyses.

The GD should keep up with the times and incorporate contemporary analyses if they are widely accepted. It should not keep on using analyses that have been disproven since their formulation. It should not pretend that no information about a phenomenon is available when such information does exist.

4.1.3. HISTORY. We value the recognition of the historic evolution of ideas.

(7) The GD should present both historical and contemporary analyses (Noonan 2006:360).

Historical or competing analyses of a phenomenon can be enlightening to the reader and are relevant for the appreciation of the grammatical system.

4.2. CREATION. In the domain of creation, we can distinguish aspects that are relevant for all authors, and aspects which are relevant only for GDs co-authored by a number of collaborators. For the purpose of this discussion, we assume that there is a grammar-authoring platform (GAP), a kind of software that assists the authors of GDs in creating a grammar.

4.2.1. LAYOUT ASSISTANCE AND TEMPLATES. We value speed of creation and comparability.

(8) Layout should be automatic as far as possible.

Authors should not have to control line spacing, italics, numbering, and tab alignment by hand. Pre-defined styles which can be applied to certain passages are to be preferred. These styles then automatically assign the layout to the relevant elements. This saves time for the authors, which they can devote to tasks of a more linguistic nature.

(9) A GAP which provides templates is better (Weber 2006a:430, 434).

As Comrie (1998:8) notes, field workers are often left to their own devices and must start from scratch when deciding on the macro- and the microstructure of their grammar. This yields a variety of structuring schemas, varying greatly in quality, which is very unfortunate for the typologist. The existence of templates for the microstructure could provide a solution.

In this context, a template would be a page with some sections and dummy content, which the describer expands and replaces with actual content. Templates for description of recurrent phenomena (e.g., personal pronouns) save time for authors, like serial letter

templates save time for secretaries. Such a template could include preformatted headers for allomorphs, position, function, etymology, sociolinguistic remarks, etc.

These templates would have the additional advantage that they could carry semantic markup which could be used for further data manipulation by third parties (cf. Bird and Simons 2003:14, 18; Simons et al. 2004; Farrar and Langendoen 2003). The following fragment from a pronoun description in Sri Lanka Malay gives an illustrative example of a GD fragment with semantic markup.[6] Other structuring schemas are possible, and this illustrative schema should not preclude more detailed investigation into the requirements for an actual schema serving in an implementation.[7]

```
<page type=morpheme title=go>
 <lemma>
 goo
 </lemma>
 <gloss>
 1s
 </gloss>
 <description>
 <objectlanguage value="SLM"> goo </objectlanguage> is a personal pronoun
referring to the  <technicalterm type=general> speaker </technicalterm> .
 The genitive is <objectlanguage value="SLM"> goppe </objectlanguage>,
 the dative <objectlanguage value="SLM"> godang </objectlanguage>.
 ...
 <example>
 ...
 </example>
 ...
 <sociolinguistics>
 <objectlanguage value="SLM"> go </objectlanguage> is the
 normal form used in the South, while in the North it is
 considered to be a sign of impoliteness. Its use is
 stigmatized, and it is a frequent  topic of
 metalinguistic comment. ...
 </sociolinguistics>
 <etymology>
 ...
 </etymology
</page>
```

The schema should support the linguist in structuring data, while staying clear of areas of which the author probably has more knowledge than the designer of the schema. Components of a page are indicated (examples, etymologies, tags) and can be worked out to have more internal structure in the future (e.g., based on Drude 2003). Models of

[6] This schema can be seen as an expansion of Good 2004, figure 14. Good uses the term "annotation" for what is called "page" here.

[7] Of the four major domains covered by the GOLD ontology (expressions, grammar, data constructs, and metaconcepts (Farrar and Langendoen 2003), "data constructs" seems to be the one which would cover this.

a schema of interlinearized examples can be found in Peterson 2002 or Bow et al. 2003. A mechanism for making complex queries to an XML-document through the Resource Description Framework (RDF) is discussed in Simons et al. 2004:19–20.

4.2.2. CREATIVITY. We value the individual mind's expressive abilities.

(10) A GAP that does not interfere with the creativity of the author is better (Weber 2006a:430).

This is in many respects the opposite of the preceding maxim. Authors might prefer to be left to their own devices instead of having to fill in mandatory fields in a template.

4.2.3. COLLABORATION. We now turn to the aspects that are relevant only for collaborative work. This is especially interesting when several authors collaborate but are not based in the same location (Weber 2006a:419, 422).
We value collaboration and the recognition of the respective contributions of the collaborators.

(11) A GAP that does not require the writers to be present at the same place is better (Weber 2006a:422).

It is common that in a research team, one member is located in one country and other members are located in different countries or continents. A GAP should allow for collaboration over the internet for this situation. The GAP should not be tied to a physical location. Concurrent use should be possible.

(12) A GAP should show which collaborator has contributed what.

Identification of authorship should be possible, as well as finding out to what extent collaborators have amended passages.

(13) A GAP which can be used both online and offline is better (Balthasar Bickel, p.c.).

Given that in many field sites access to the internet is difficult, the GAP should be available offline as well as online.

4.2.4. BACKUP. We value safety of the data.

(14) A GAP should provide the author with regular automated backups (Weber 2006a:418, 434).

Not all linguists are equally computer-literate, and a platform that automates backup routines will help prevent data loss.

4.3. EXPLORATION. Exploration of the grammars translates to user-friendliness, an important aspect of good GDs (Rice 2006:395). Users consulting a GD have a variety of needs, which can be captured in part by the following maxims.

4.3.1. EASE OF FINDING. We value ease and speed of retrieving the information needed.

(15) A GD which has a table of contents, an index, and full text search is preferable (Weber 2006a:432; Noonan 2006:355; Cristofaro 2006:147).

The tools listed in (15) help the users to navigate the grammar. A full text search allows users to locate passages that they remember to be present in the grammar even when they have forgotten the exact page number. For instance, readers might recall the existence of that very enlightening example about a man killing a lion, but not remember its exact location. By using the full text search for "lion", they can retrieve all occurrences of "lion," which should include the example they were interested in.

It is also useful to have an index of topics that are *not* found in the GD. The existence of this negative index allows the users to ascertain quickly whether a further investigation of this language is useful for them (Noonan 2006:356; Haspelmath 1993).

(16) A GD that does not require internet access is preferable (Bird and Simons 2003:12).

In many areas of the world, access to internet is suboptimal. For people dwelling in those areas, accessing the content of a GD without needing a broadband connection is an advantage. This maxim is slightly different from (13) in that for (16), read-only access to the content is sufficient, like on a CD-ROM. (13) requires that the author be able to *edit* at the fieldsite as well.

4.3.2 INDIVIDUAL READING HABITS. We value the individual linguist's decisions as to what research questions could be interesting (Rice 2006:402).

(17) A GD should permit the reader to follow his or her own path to explore it.

Some readers might prefer to approach a GD from a form-to-function-approach (semasiologic: *what does* -ing *mean?*), while others might prefer a function-to-form approach (onomasiologic: *how is progressive aspect expressed?*). Both types of readers should be served by a GD. The formal approach can be associated with the hearer's perspective (Lehmann 2004a) and is more common in language group studies, while the functionalist approach can be associated with the speaker's perspective and is the one used in typological circles (Schultze-Berndt 1998, Cristofaro 2006, Lehmann 1980, 1998, Mosel 2006a, Zaefferer 2006, Jespersen 1924, Payne 2006, Comrie 1998, von der Gabelentz 1891).

A typological GD should acknowledge the interest of and the need for both approaches and thus provide for semasiologically oriented linguists as well as for onomasiologically oriented linguists. This means different ways of organizing, presenting, and structuring things for the two approaches.

Views on the preferred mode of presentation differ. Payne (2006:380ff.) advocates a semasiological approach for some domains, and an onomasiological approach for others, a solution already implemented in Haspelmath 1993. However, it is not clear why the respective other areas of the grammar would not deserve onomasiological or semasiological treatment as well. A GD providing full accounts of all semasiologic aspects and providing full accounts of all onomasiologic aspects is surely better than one that does not. Treating both aspects extensively poses the problem of redundancy, as Comrie (1998) and Cristofaro (2006:147) point out. Proper implementation of both viewpoints can remedy this. When all semasiological topics are covered in the corresponding part, and the same is done for the onomasiological aspects, the only remaining redundancies are examples occurring in both discussions, at the intersection of a formal and a functional domain. However, the different parts will deal with these examples in very different ways. In the formal part, the order of constituents of a construction could be discussed, as well as morphological effects on other items in the construction, etymology, and sociolinguistic considerations. In the functional part, most of this will not be repeated. Instead, this part will explain when a particular construction is used and given preference over the other possible constructions. There is some overlap with the sociolinguistic information provided in the formal part (information on register, for instance), but otherwise the content is complementary. It is therefore not a repetition or redundancy, but a distributed discussion of formal and functional aspects.

(18) A short path between two related phenomena is better.

For the reader of a GD it is convenient to have descriptions of related things grouped together (Cristofaro 2006:147–148; Good 2004). One problem that arises is that the author does not know what the reader is interested in. If the author writes a GD with the reader in mind (Rice 2006b:400), it may not be clear what domains the reader is interested in, and whether the reader prefers a formal or a functional approach (cf. Payne 2006: 382). In the unfortunate case that the structuring ideas of author and reader do not match, the reader has to browse through the whole book to find the relevant information, and is likely to miss it altogether (cf. Cristofaro 2006:142).

This problem has been illustrated by a number of scholars. Payne (2006:377–378), for instance, points out that English past tense is expressed in a morphological way (affixation of *-ed*), while the future tenses are expressed syntactically (by *will* and *gonna*). Having separate parts for the description of morphology and syntax will tear apart these functionally related phenomena and make them non-adjacent, thus losing the notion of a unified tense system. Similar comments can be made about deictics (Payne 2006:378–379):

> [The grammar] may treat some deictic operators in the section on noun morphology, others in the section on verb morphology, and still others in a distinct section on demonstratives. In none of these sections is there likely to be any observations or hypotheses concerning how the various deictic operators function as a system that expresses deictic notions in discourse.

It has been recommended that grammars (should) make extensive use of cross-refer-

encing (Mosel 2006a:43; Noonan 2006:355; Weber 2006a:423). The use of cross-references is of course much easier with electronic publications, since all cross-referenced pages are only one click away (Comrie 1998:14).

4.3.3. FAMILIARITY. We value ease of access.

(19) A GD that is similar to other GDs known to the reader is better.

Many typologists prefer familiar structuring schemas over idiosyncratic ones. Typologists normally do not have the time to study thoroughly each and every one of the several hundred languages in their sample (cf. Comrie 1998:8). They are interested in quickly retrieving (harvesting) the relevant information (Cristofaro 2006:146, 162).

With the number of available GDs increasing, it becomes more and more difficult for typologists to familiarize themselves with a whole new descriptive system for every language in their sample. This means that the GD should be presented in an accessible manner and rely on structuring schemas that can be supposed to be familiar to the users (cf. Comrie 1998:10). It is often argued that every language should be described on its own terms because the categories each language uses are unique to that language (Croft 1990; Keenan and Comrie 1977; Stassen 1985). This is of course correct, and a Procrustean bed that every single language would be fit into would not help the cause of typology. Forcing a particular language into a schema which is not tailored for it would distort the data, and distorted data are far worse an input for typological research than data difficult to access. This does not mean, on the other hand, that the most idiosyncratic description is the best one. As Zaefferer (1998c:1) points out, the really special features of a language can only be appreciated against the background of the features it shares with other languages, its common features. If all the features are special, then they all have the same relevance (i.e., "unique") and the language lacks a profile (also cf. Evans and Dench 2006:5). Writers of grammars should therefore strive to present cross-linguistically common features of the language they describe as common, and cross-linguistically rare features as special.

4.3.4. GUIDING. We value an informed presentation of the data.

(20) The GD should present the data in a didactically preferred way (Rice 2006:401).

The author should take care to present the data in a way that is easy for the reader to understand. This maxim conflicts with maxim (17), where the author is asked to leave the path to the reader.

4.3.5. EASE OF EXHAUSTIVE PERCEPTION. We value the quest for comprehensive knowledge of a language. This is especially true for linguists specializing in one language group who are interested in knowing everything about a language and comparing it to neighboring languages (Cristofaro 2006:162).

(21) The readers should be able to know that they have read every page of the grammar.

In this way the readers will know that they have not missed an important part of the description.

4.3.6 RELATIVE IMPORTANCE. We value the allocation of scarce resources of time to primary areas of interest.

(22) The relative importance of a phenomenon for (a) the language and (b) language typology should be retrievable (Zaefferer 1998c:2; Noonan 2006:355).

Readers should be able to know whether they are dealing with a central or a peripheral aspect of the language, and whether that aspect is cross-linguistically interesting. As an example, vowel lengthening and consonant gemination in Sri Lanka Malay are vital for understanding the nature of prenasalized consonants, palatal affricates, nominalizations, verbal compounds, and foot structure in that language. It is thus a key phenomenon for understanding Sri Lanka Malay phonology. Labialization of /m/ before /o:/ in the same language, on the other hand, is an isolated phenomenon and has no bearing whatsoever on other aspects of the language. This piece of information can safely be ignored without impeding comprehension of other parts of the grammar. Readers might prefer to be informed about the relative importance of these two phenomena (cf. RELEVANCE in Bird and Simons 2003:14).

4.3.7. QUALITY ASSESSMENT. We value indication of the reliability of analyses.

(23) The quality of a linguistic description should be indicated.

Grammar authors might show different levels of confidence with respect to their descriptions. The analysis of the case system might be well-founded and shared by the research community, whereas the analysis of the stress-pattern might be preliminary, sketchy, and debatable. Still, the authors might want to share their impressions of the stress system, sketchy as they are (Munro and Nathan 2005; Holton 2003; Noonan 2006:358). In that case, a reader would appreciate the information that the status of the description of the stress system is not on a par with the description of the case system. In other words, the level of confidence which authors have towards their analyses should be indicated. In prose descriptions of grammatical phenomena, this is frequently done by distantiating expressions such as "It *appears* that X has an iambic system" as compared to "There *are* four cases, nominative, genitive, dative, and accusative in X."

4.3.8. PERSISTENCE. We value citability (Bird and Simons 2003:14).

(24) In order to facilitate longterm reference, a grammatical description should not change over time.

This maxim conflicts with the maxim of actuality. How can a GD be persistent so as to be citable and incorporate recent changes as well? A solution for this could be a versioning system, to be discussed below. A grammar that is based on a proper corpus should also be able to cite that corpus, which is provided with persistence by being located in an archive of some kind.[8]

4.3.9. MULTILINGUALIZATION. We value the interest of every human in a given language, especially interest from the speakers of the language in question.

(25) A GD should be available in several languages, among others the language of wider communication of the region where the language is spoken (Weber 2006a:433).

It would be nice if a grammar of Quechua were available not only in English, but also in Spanish, which would allow greater access. Many members of the speech community will be able to access the content in Spanish, but not in English. This maxim is likely to be impossible to follow in many cases, but it is still a goal which can guide the creation of a GAP.

4.3.10 MANIPULATION. We value portability and reusability of the data.

(26) The data presented in a GD should be easy to extract and manipulate.

Linguists might want to extract data from the GD and manipulate them by scripts (Weber 2006a:432). This is easier if the content is stored in an open format in a well-defined encoding (Bird and Simons 2003). Furthermore, semantic markup would make this task easier as well.

4.3.11. TANGIBILITY. We value the appreciation of a grammatical description as a comprehensive aesthetic achievement.

(27) A GD that can be held in the hand is better.

Bibliophiles might prefer the holistic experience of accessing the GD with all senses. This maxim cannot be met by electronic GDs, and will not be discussed any further.

5. POSSIBILITIES. Having outlined the values which might guide our choices in conceiving an electronic reference grammar for typology, I will now sketch some solutions that respond to the requirements formulated above. Most of the discussion will be general, but

[8] See http://www.language-archives.org/archives.php4, http://www.hrelp.org/archive/ and http://corpus1.mpi.nl/ds/imdi_browser/ for archives hosting language materials.

some possibilities will be exemplified by the SIL Fieldworks Language Explorer (FLEx)[9] or the grammar authoring platform GALOES (Nordhoff 2007a, b, c). Reasons of space forbid a full discussion of either system here. They are used only to show how responses to some of the maxims above could be implemented. Individual choices might be debatable, and better components for the various parts might be available, but this should not preclude using them as sample implementations for illustratory purposes. Occasionally, I will also refer to other software projects that use interesting techniques that could be worth investigating for their use in linguistics.

5.1. DATA QUALITY.

(1) Every step of the linguistic analysis should be traceable to a preceding step, until the original utterance of a speaker is reached.

This maxim is good practice in paper grammars, but has to stop short of the original recording, since sound cannot be rendered on paper. This problem does not arise with electronic presentation, where audio or video files can be linked from the example or the describing text (Austin 2006; Noonan 2006:361).

(2) Every phenomenon described should be sourced using an actual utterance.

There are several possible ways to include media files in an electronic text document. They can either be embedded in the appearance of the page and be played from within the browser through a plug-in, or the media files can be passed on to an external player. While embedding makes for a more coherent appearance of the grammar, it is more difficult to implement, and compatibility problems among browsers, media plug-in, and operating system are common. Providing content for stand-alone applications is less integrated but gives the reader more control of what to do with the file. An audio file could, for example, not only be played back, but also passed to Praat10 for acoustic analysis.

Media files can be downloaded or streamed. In the former case, the user must get the whole file before playing the media, whereas streaming allows instant playback of the content. Given that media files can become very large, streaming would save time compared to the downloading solution. Still, download facilities should be provided for later use of the media when not connected to the internet. An implementation of streaming for media linked to text is ANNODEX (Schroeter and Thieberger 2006:115, http://www.annodex.net/)

In principle, the files can either be served from the same location where the document resides, or they can be referenced in an archive where text corpora (or other types of data such as media, maps, historical records, etc.) of the language are stored. Connection to an archive is advisable, given the following maxim,[11]

[9] http://www.sil.org/computing/fieldworks/flex. Note that FLEx does not aim to provide a non-linear GD, but some of its components could serve as models for software which does have that aim.

[10] http://www.fon.hum.uva.nl/praat/

[11] For audio files, this has been done by Thieberger (2006) for example, where the soundfiles used in

(3) More sources for a phenomenon are better than fewer sources.

In this case one must make sure that the reader does indeed have access to the archive referenced; otherwise the linked media file could not be used for the purposes of accountability. An archive could also contain additional texts from which examples could be retrieved. In the context of a GD, two types of examples can be distinguished. The first type is a "raw" example, an extraction from a transcribed and glossed text. The second type is a didactic example, which is used to illustrate one particular point (cf. Lewis et al. 2006:10–11; Good 2004). This didactic example can contain additional information, such as a short illustration of the context, formatting like boldface to highlight relevant aspects, brackets to show syntactic grouping, indices, and the like. It is also possible for the didactic example to omit some detail from the raw example that would distract the reader from the main point it aims to illustrate. Cases in point would be false starts, very heavy NPs that could be shortened, grammatical detail in the glosses not necessary to illustrate the phenomenon at hand, etc. While didactic examples need to be chosen and prepared by hand, readers might want to access additional, raw examples from a corpus. To this end, it is helpful if the corpus has a well-defined structure that allows the extraction and presentation of relevant examples as well as the serving of the media files linked to the transcription.[12] Such a corpus has the additional advantage of being useful for the author of the GD when selecting raw examples for turning them into didactical examples, a requirement for building a GD (Good 2004).

(4) The context of the utterance should be retrievable.

The context of an utterance should be described in metadata according to an accepted standard such as IMDI[13] or OLAC.[14]

In the domain of actuality, electronic publishing has distinct advantages over publishing on paper. New findings can be incorporated rather easily and do not necessitate printing errata or a revised edition. Furthermore, the machinery to publish something on the web is inexpensive (a personal computer is enough) and does not require a lot of training. Good web design, on the other hand, is not an easy task, and making a website that appeals to the reader involves knowledge and skills that should not be underestimated. While some people have skills in both web design and linguistics, this is not the case for everybody. Normally, linguists should take care of the content, while specialists take care of the form of the page. This separation of content providers and designers is common in many ar-

the grammar are stored in a persistent archive.

[12] A recent version of ELAN, for instance, can be used for these ends (http://www.lat-mpi.eu/tools/elan/).

[13] http://www.mpi.nl/IMDI/

[14] http://www.language-archives.org/

eas of web publishing, and specialized tools—so-called Content Management Systems (CMSs)—have been developed to meet these needs.

Wikis are a special type of CMS used on collaborative web projects. They have some extra features designed to assist the process of collaborative development. Note that some wikis are editable by anyone, while others require authentication to edit or even read. Wikis are thus not synonymous with a free-for-all approach, although some of them might tend in that direction.

After this short introduction to CMSs, let us now turn to the maxims formulated for the domain of actuality.

(5) A GD should provide means for incorporating scientific progress.

(6) The GD should present state-of-the-art analyses.

Given the possibility of changing content in a CMS even for non-technical users, advances in the knowledge of a language can find their way into a GD quickly (note that a CMS will normally allow write-access only to selected users so that disruptive edits by lay people inserting bogus data are not possible). For instance, in the description of Sri Lanka Malay, it is conventional wisdom that all adpositions are postnominal. However, recently the possibility of using *sangke* 'until' in prenominal position has been found by two researchers independently. This can be added quickly to the description of adpositions on the relevant pages.

Always having the most cutting-edge analysis might make the grammar fickle, while some users may prefer a more conservative approach, an issue addressed in

(7) The GD should present both historical and contemporary analyses.

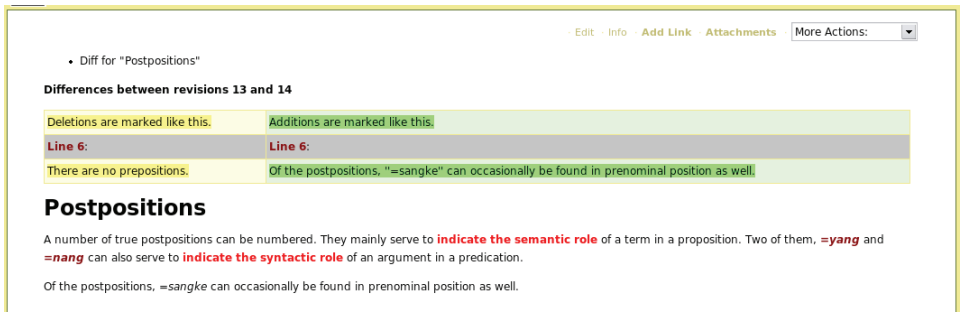

FIGURE 1: Differences between version 13 and version 14 of the page "Postpositions" shown and highlighted in the grammar authoring system GALOES. (http://www.galoes.org/slm/Postpositions?action=diff&rev2=14&rev1=13)

To this end, some CMSs provide a *version history* feature (see figure 1), where older versions of a page can be seen and compared with the current version. In this way, earlier states of analysis are available. It is also possible to remove changes made to a page and revert to a previous page if necessary.

This version history is also important for citation purposes. Content in a CMS can change quickly. When a document is cited, it is therefore important to include the time at which the document was retrieved. The combination of document name and time then allows other people to access a historic version of the site which shows the content the way it was presented when the page was cited.

5.2. CREATION. We now turn to the creation side of a GD, the way authors can feed content into an electronic description. As already stated above

(8) Layout should be automatic as far as possible.

The separation between content and presentation is one of the main features of CMSs.[15] At the very beginning of a publishing project, the content provider makes a list of semantically relevant elements (headings, text body, links, quotations, examples, images, menus, disclaimers, etc.). Ideally, the content provider and the designer then decide how these elements should be rendered using a style sheet. Headings could be bold and centered, links brown and underlined, quotations indented and in italics with the first letter in a slightly bigger font, images always on the right with a frame around them, etc.

Once this style sheet is designed, the content provider merely assigns semantic labels to parts of the text: this is a heading, this is a link, this is a quotation. The typographically

[15] This separation should actually be basic to all data creation. FLEx for instance is not a CMS in the strict sense, but uses the separation of content stored in a database and presentation instruction to generate morphological sketches.

acceptable presentation thereof is taken care of by the stylesheet[16].

For GDs, such a list of semantically relevant items does not exist yet (but see Good 2004). Obvious candidates include the aligned and numbered linguistic example, the paradigm table, the *word* 'gloss' pair, but there are surely many others, which should be a worthwhile topic for future research.

Having discussed the main features of a CMS, we now turn to a CMS used as a GAP[17] and the additional requirements this entails. While low-level layout is taken care of by the CMS, higher level organization of the text is still up to the user. In the discussion of the adposition *sangke* 'until', does the author first speak about the form (prenominal or postnominal) or about the function (indicating terminal boundary in time), or does he or she start with the etymology or sociolinguistic deliberations? There might be reasons to choose the same order for the discussions of all postpositions. In this case, the CMS should provide templates for the discussion of recurring phenomena in line with

(9) A GAP that provides templates is better.

Such templates could be TemplateGrammaticalMorpheme,[18] as sketched above, TemplatePhoneme, TemplateClauseType etc. While for reasons of consistency, use of templates can be advocated, there might be cases where deviation from the standard order might serve the goal better, according to the maxim

(10) A GAP that does not interfere with the creativity of the author is better.

This means that templates should be offered, but not enforced. In the hypothetical case that the GAP is hosted by an institution like a university or a publishing house, this institution might want to enforce templates so that all the GDs which they host would share a standard appearance.

(11) A GAP that does not require the writers to be present at the same place is better.

When a CMS runs on a server, collaboration over the internet is possible. Different geographical locations and time zones are then not a problem for working together on a GD.

[16] It is also possible to propose several style sheets among which the users can choose. For an example of this see http://www.csszengarden.com/

[17] Weber (2006a:430ff) uses the more general term "Authoring Environment," which would also include stand-alone applications like MS Word.

[18] The FLEx program can generate a morphological sketch based on a lexicon with morphological annotations, morphological templates, categories, and rules. The resulting descriptions follow a uniform pattern regardless of the language described—in other words, a template. An example of such a description can be seen at http://www.sil.org/computing/fieldworks/flex/ExampleSketch.htm. The generated description is not fit for publication as is, but has to be further edited by hand. (http://www.sil.org/computing/fieldworks/flex/sketch.html). It is unclear to what extent the templatic structure will be preserved after manual editing.

FLEx has such a network capability, which Butler and van Volkinburg (2007) cite as its most important advantage. Along with network capability, GALOES also features a version history, as discussed above, which allows one to return to a version prior to unfortunate amendments by a researcher if closer analysis shows that the changes were mistaken. Furthermore, it is possible to have talk pages associated with a page where the best way to tackle the description can be discussed.

(12) A GAP should show which collaborator has contributed what.

Ideally, the respective contributions of researchers should be visible. If Jane Doe has contributed the bulk of the description of a morpheme, whereas her co-author John Doe only corrected punctuation, it would be nice if this were retrievable from the GD.

One example of a CMS which provides this functionality is *wikipresto* (figure 2), but unfortunately the software is not open source.[19]

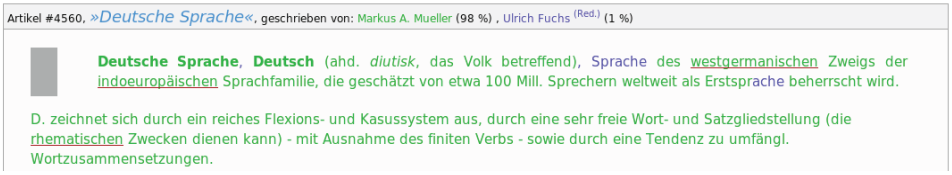

FIGURE 2: Authors and their contributions to an article hosted on a wikipresto system. The contributions by the author Markus A. Mueller are in green (most of the content), while the contributions by the author Ulrich Fuchs are in blue (last occurrence of *Sprache* on the first line and *Erstsprache* on the second line). It is immediately visible that Mueller contributed the bulk of the article, while Fuchs only did some copy-editing. (http://www.wikiweise.de/Frontcontroller? command=displayArticle&article=Deutsche%20Sprache&version=0&section=0&encuser=true)

An important consideration for the creation of an electronic GD is that internet connectivity is not very good in many remote field sites (Amazonia, Himalayas, the Pacific, etc.) so that a GAP that is available only online would not be useful for researchers doing fieldwork at those locations. It is important that the GAP software be able to run locally on a desktop computer.

(13) A GAP which can be used both online and offline is better.

The easiest approach would be to use the server version at the home university, then copy all the description files to a laptop, take it to the field, edit the pages in the field, and then

[19] http://www.wikiweise.de:8080/wikipresto-software-wiki/

play back the new versions upon returning. However, if in the meantime someone has changed the files on the university computer, there is a version conflict and some information is at risk of being lost. Routines for checking out data, checking in data, and merging therefore have to be provided to prevent this data loss. Such a check-in routine would only re-check-in files that have indeed been edited in the field, and it would ask which version to take if a concurrent modification to a file has been made at home in the meantime.

(14) A GAP should provide the author with regular automated backups.

Another advantage of an institution hosting the GAP is that expert help for data security and safety can be provided to prevent accidental data loss (protected server rooms, backup). If this is not the case, a periodical backup script should run on the local computer with the GAP.

5.3. EXPLORATION. We have discussed the way in which an author can make use of an electronic GAP, and we now switch viewpoints and take the perspective of a reader who is exploring the GD to find information.[20]

(15) A GD which has a table of contents, an index, and full text search is preferable.

Most paper GDs have a table of contents and index, but due to the nature of the medium, full text search is not possible in paper works. Most CMSs provide the possibility of including a search field, which can then be used to search for the occurrences of a string in the GD. Some CMSs also offer the possibility of advanced queries with regular expressions, so the query *baa.*a* would find *baaaa, baaba, baaca, baada*, among others. Indices and tables of contents can also be included in some CMSs. As for negative indices, it would be possible to have a page "Unattested Phenomena," which is referred to in the index for all phenomena not found in the language.

(16) A GD which does not require internet access is preferable.

There should be a provision to make the GD accessible offline. This can be done via a dump of the database and the burning of a DVD or another suitable medium. Or one could linearize the description and make a print-out on paper. The dump of a database to use on local computers is provided on some CMSs, but the linearization necessary for a print-out is a more difficult issue (see below).

[20] Given the workload that the production of a GD imposes on the author, the "goodies" for exploration by the reader may not ask for very much additional work from the author. Rather, they should either be done automatically (such as a table of contents, which requires only minimal extra work by the author), or emerge from something that is also beneficial to the author such as semantic markup tied to templates). Work that to a large extent is beneficial only to the reader (such as links or tags) should be made as easy to accomplish as possible. This is a task for the designer of the GAP user interface.

(17) A GD should permit the reader to follow his or her own path to explore it.

By using links generously, the author can facilitate individual exploration by the reader and provide easy access to definitions, background information, or in-depth theoretical discussions that the reader can choose to follow or ignore.

(18) A short path between two related phenomena is better.

A short path is very easy to implement through the use of hyperlinks.

(19) A GD which is similar to other GDs known to the reader is better.

Readers of a GD might appreciate similar structuring schemas employed in the GDs they read. Compatibility among those schemas will facilitate the typologist's work. The morphological sketches generated by FLEx are uniform, and readers will quickly grow accustomed to them after they have read several of them. One drawback of the FLEx sketches is that the system provides no means for variation of the sentences used for presentation. This makes for a very repetitive reading experience[21] and might be a reason why the designers recommend manual editing of the sketch for publication. A different solution is chosen by the GALOES system, which provides backward compatibility to the Lingua Descriptive Studies (LDS) questionnaire (Comrie and Smith 1977) through tags. Similarly, typologists used to LDS grammars should have easier access to GDs written in GALOES. It is in principle possible to read a GALOES grammar in the same order as an LDS grammar.

(20) The GD should present the data in a didactically preferred way.

While the GD should not force the reader to read it following a certain path, some readers might find it useful to have such a thread to direct their exploration. Such a thread or path could either be static or dynamic. A static path would mean that for every page, there is a link to one page that logically precedes it and a link to one that logically follows it. By clicking on the [NEXT] link at the bottom of a page, the reader can get to the next page in the grammar, following a linear path. This static approach is problematic if new pages are added, because the thread must be cut, as it were, and the new page inserted and reconnected to the open ends. The problem is that there might not be a perfect place to fit in the new page, as well as the possibility for errors when reconnecting pages, leaving some pages outside the path, or, even worse, creating a loop from which the reader cannot exit.

A dynamic approach would not preconceive a particular linear order, but would in-

[21] An example of this can be found in the sample sketch provided by SIL (http://www.sil.org/computing/fieldworks/flex/ExampleSketch.htm):

4.53 Stative: The stative category has 1 template: Stative verb. (See instances in the lexicon in appendix B.49.)

4.54 Transitive verb: Verb with two participants The transitive verb category has 1 template: Transitive verb. (See instances in the lexicon in appendix B.53.)

4.55 Bitransitive verb: Verb with three participants The bitransitive verb category has 1 template: Bitransitive verb. (See instances in the lexicon in appendix B.8.)

dicate the domain and relevance of a page. For instance, the page discussing the vowel inventory would belong to the domain Phonology and have top priority. The page discussing vowel length would also belong to Phonology, but have a lower priority. By ordering the pages of a domain according to priority, a linear order can be achieved. Some problems with this approach are obvious, namely that rather unconnected areas, like intonation contours and consonant inventories, are discussed on adjacent pages on the path if they are of the same priority. A solution to this problem would be to make the domains very fine-grained, i.e., instead of DomainPhonology, one would have DomainVocalicPhonology, DomainIntonationContour, DomainVerbalPrefixes, etc. The dynamic approach thus asks for more tagging on the part of the creator, but avoids the insertion problem posed by the static path approach.

The solution employed in the GALOES system is to tag each page for the LDS questions it answers. The page on vowel lengthening in Sri Lanka Malay, for instance, provides answers to the LDS question "*3.3.1.1. Are there distinctive degrees of length in vowels?*" By following the questions of the LDS questionnaire and retrieving all pages tagged for the current question, a reader can be presented with a linearized path, albeit incurring some redundancy, since pages may provide answers to more than one question. An implementation of this tagging system to provide a path can be seen at http://www.galoes.org/trees.

A problem related to the presentation of the grammatical information in a didactically motivated linear path is the problem of exhaustivity. How can the readers know that they have visited every page and read all the information contained in the GD?

(21) The readers should be able to know that they have read every page of the grammar.

For books, this is easy to answer: when the readers have reached the back cover, they have read everything that is contained in the book. For hypertextual grammars, this is more complicated. If there is a preconceived static path, the solution is easy: when the reader has reached the final page, he or she is done. If there is no such path, a solution would be to make use of a different display for visited and non-visited links in the browser. A list of all pages in the GD can be displayed in the browser, and pages that have been visited would be in normal font-family, while pages not visited would be in bold. An example of this can be seen in the TitleIndex of the GALOES system (figure 3). While this solution works in theory, it is quite cumbersome to the reader. As it is now, this solution could only be used toward the end of the reading to make sure that no page has slipped through the fingers of the reader. Another problem is that pages read on a different computer will not show up as visited links, but as new links. The problem of exhaustivity thus needs more thought.

FIGURE 3: The TitleIndex gives a list of all pages in the GD and links to unvisited pages in bold (*-an, =dheri, =dring, =ke*), while links to visited pages are in normal font (*-king, =jo, =ka*) (http://www.galoes.org/slm/TitleIndex).

(22) The relative importance of a phenomenon for (a) the language and (b) language typology should be retrievable.

Readers might appreciate some information about the central or peripheral status of a phenomenon. This can be achieved by tagging the page for relevance. Very relevant pages would have a tag Relevance5, while peripheral pages would have Relevance2 or Relevance1. This could trigger changes in formatting, or pages below a certain relevance level could be excluded from searching.

In a similar fashion, impressionistic statements about the cross-linguistic rarity of a feature could be given, thereby directing typologists to an unusual feature of the GD, whose pages would have a tag like Commonness1. The domains for which tags can be given should probably be restricted, so that only Relevance, Commonness, Certainty, and some others would be admissible as tags.

(23) The quality of a linguistic description should be indicated.

To address this issue, every page could receive one of several quality tags (Note, Sketch, Draft, PeerReview, ...). This allows the author to include some cursory notes of phenomena he or she has come across but has not yet investigated or understood completely. This is useful for the authors to store and organize their knowledge, but it can also be useful for the readers because they know which information can be trusted and which information has to be treated with caution. Still, unsure information is valuable, too, since it provides a starting point from which the users can begin to build their own interpretations. Authors who are all too unsure about their findings could configure the CMS in such a way that

preliminary analyses are not accessible to the general public. This could be done by setting a tag or a flag on a page, or by the use of Access Control Lists.

(24) In order to facilitate longterm reference a grammatical description should not change over time.

The process described above is ever-evolving and in constant flux. This has the advantage of presenting cutting-edge research, but the disadvantage that frequently changing content is not very well suited for some scientific tasks, like citations. For these ends, which require permanence, snapshots can be made. These represent the state of the electronic GD at one point in time. This snapshot remains the same, even if the "main" GD changes thereafter, just like a photographic snapshot remains the same, even if the object photographed changes. It is possible to declare some snapshots as milestones. These milestones are internally consistent, factually correct, and show what is secured knowledge about the language at that time, while excluding more speculative content. These milestones can then be submitted to the normal process of scientific peer review, and references to them are unique. In order to refer to content outside of a milestone, one can use the version history and refer to a specific version, e.g., version 11 (2007-07-12 14:53:05) of the page on postpositions in Sri Lanka Malay (http://www.galoes.org/slm/Postpositions? action=recall&rev=11).

(25) A GD should be available in several languages, among others the language of wider communication of the region where the language is spoken.

Some CMSs provide possibilities for multilingualization. Depending on the language settings of a browser accessing the content, a version in a different language can be served. In this way, a user with his browser set to prefer English language pages will be presented the content in English, while a user who has set his browser to prefer Spanish will be served Spanish pages, if available. The drawback of this is of course that additional resources are needed to provide the translations of the content into other languages. Most projects will not have these resources, but a GAP with the option of multilingualization would be an advantage.

(26) The data presented in a GD should be easy to extract and manipulate.

Storing the data in the GAP in an open format facilitates extraction and manipulation of the data by third-party programs. Furthermore, a "dump" functionality would be a useful thing to have. With this, the reader could get a local copy of the underlying data files that he or she could use for further manipulation. FLEx stores its information in an SQL-database or in XML, which should in principle be readable by other programs. Third-party researchers could then design programs to query this database without the need to run the FLEx program itself.

6. STANDARDIZATION. In the long run, it is possible that a standard for GDs may emerge. Note that this is not a standardization of the *structure of language*, but a standardization of the *structure of language descriptions*. Such a standard could contain the following:

1. a list of semantic elements needed for GDs (linguistic example, word-gloss pair, definition, link, prose description, tag, phoneme table ...) . Good 2004 contains discussions of some of these elements.
2. a list of templates containing the elements of (1) in a certain order (TemplatePhoneme, TemplateMorpheme, TemplateConstruction, TemplateClauseType)
3. a list of domains for which pages can be tagged (Relevance, Commonness, Certainty...)

GDs written in this standard could be exchanged, processed, and rendered/displayed by different software architectures.

The institution hosting the GAP server could decide on whether to ignore, recommend, or enforce the use of such a standard. Both a "rough guideline" (i.e., a recommendation) and a "law-and-order" (i.e., an enforcement) approach have advantages and drawbacks. The advantage of the guideline approach is that it keeps the threshold low; people are more likely to adopt a system if they can start right away and see if the system serves their needs without having to study thick manuals. The risk is obviously a lack of cross-GD consistency, thus putting comparability in jeopardy. The advantage of a more rigid approach is improved processability by automated routines, while the drawback is that researchers might be reluctant to use for their descriptive work a system which restricts their choices.

Besides the standardization of the parts of a language description, another type of standardization is conceivable, namely the standardization of the parts of a language, as espoused by the CRG framework (Zaefferer 1998b, Nickles 2001, Peterson 2002, Zaefferer 2006) and the GOLD project (Farrar and Langendoen 2003, http://www.linguistics-ontology.org/gold.html). For the purpose of discussion, I will call the standard for parts of a description a macroscopic standard, while I call the standard for parts of a language a microscopic standard.

These two types of standards would complement each other. It is conceivable to have a GD that respects only one of those standards, both, or neither. The macroscopic standard is targeted at humans and would help humans to formulate semantic queries to retrieve a first set of pages where an answer to their research question might be found. Examples would be *Get all pages that contain reference to Chinese in the etymology section; Get all pages describing a pronoun that has more than 2 allomorphs; Get all pages with more than 5 examples on them*.

Microscopic standards, on the other hand, are targeted at machines with the ultimate aim of making GDs comparable at logical (machine) level (Zaefferer 2006:130) and to facilitate automatic reasoning (Farrar and Langendoen 2003). In this respect, a microscopic standard would facilitate the creation of electronic databases for grammatical description[22]

[22] FLEx actually creates such databases, which can be made readable for humans by generating a morphological sketch out of them.

(Zaefferer 1998c:5, 2005), from which implicational universals could be established. To what extent the two target audiences of human readers and "machine readers" can be addressed successfully in the same document remains a topic for further research.

The macroscopic standard can also be located in the two-layer model of typological research developed in Cysouw 2007. Cysouw argues that typology should make use of two different layers of data presentation and analysis. The first, lower one is the "social layer." On this layer, the author provides descriptions of a phenomenon intended for cooperation and discussion (7). Atop of this layer comes the "personal layer," where individual typologists recategorize the descriptions of the languages they have investigated according to their criteria (Has case? Y/N; Prepositional/postpositional; etc.). Different typologists can use different criteria to categorize the very same data, as they see fit. The individual typologist can draw on the description on the social layer but can interpret it in a way that suits his or her needs. A macroscopic standard would then help structure the description on the social layer and facilitate the retrieval of relevant information and the interpretation of the description for the typologist.

7. CONCLUSION. This paper has sketched values and expectations that authors and readers might hold with regard to electronic reference grammars. Based on these values, 26 maxims were formulated to guide the development of a grammar authoring software platform. Adherence to some of the maxims is easy and can be done with existing technology. This is the case for maxims treating data quality and creation, and some maxims in the domain of exploration. For some other maxims, a suitable approach for implementing them has yet to be found. This applies most notably to the problems of linearization.

Along with adherence to the maxims, a grammar authoring platform should try to adhere to existing standards and be open to the development and incorporation of emerging standards in order to facilitate analysis and portability of the contained data. As a consequence, portability would also be improved by the development of a macroscopic standard for grammatical descriptions.

References

Ameka, Felix, Alan Dench, and Nick Evans, eds. 2006. *Catching language: The standing challenge of grammar writing*. Berlin, New York: Mouton de Gruyter.

Austin, Peter. 2006. Data and language documentation. In *Essentials of language documentation*, ed. by Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel, 87–112. Berlin: Mouton de Gruyter.

Bagish, Henry. 1983. Confessions of a former cultural relativist. In *Anthropology annual editions* 83/84, ed. by Elvio Angeloni, 22–29. Guilford, CT: Dushkin Publishing Group.

Barwick, Linda, and Nick Thieberger, eds. 2006. *Sustainable data from digital fieldwork*. Sydney: University of Sydney.

Bird, Steven, and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3): 557–582.

Bow, Cathy, Baden Hughes, and Steven Bird. 2003. Towards a general model for interlinear text. Proceedings of the EMELD Language Digitization Project Conference 2003. http://www.linguistlist.org/emeld/workshop/2003/bowbadenBird-paper.pdf

Butler, Lynnika, and Heather van Volkinburg. 2007. Review of Fieldworks Language Explorer (FLEx). *Language Documentation and Conservation* 1(1): 100–106. http://hdl.handle.net/10125/1730

Comrie, Bernard. 1998. Ein Strukturrahmen für deskriptive Grammatiken: Allgemeine Bemerkungen. In *Deskriptive Grammatik und allgemeiner Sprachvergleich*, ed. by Dietmar Zaefferer, 7--16. Tübingen: Niemeyer.

Comrie, Bernard, and Norval Smith. 1977. The *Lingua* Descriptive Studies Questionnaire. *Lingua*, 42:1–72.

Cristofaro, Sonia. 2006. The organization of reference grammars: A typologist user's point of view. In Ameka et al. 2006: 137–170. Berlin, New York: Mouton de Gruyter.

Croft, William. 1990. *Typology and universals*. Cambridge: Cambridge University Press.

Cysouw, Michael. 2007. A social layer for typological databases. In *Language resources and linguistic theory*, ed. by A. Sansò, 59-66. Materiali Linguistici Università di Pavia. Milano: Francoangeli.

Drude, Sebastian. 2003. Advanced glossing: A language documentation format and its implementation with Shoebox. Proceedings of the LREC-Workshop in May 2002, Las Palmas: W1: International Workshop on Resources and Tools in Field Linguistics. http://www.mpi.nl/lrec/2002/papers/lrec-pap-10-ag.pdf.

Evans, Nick, and Alan Dench. 2006. Introduction: Catching language. In Ameka et al. 2006: 1–40. Berlin, New York: Mouton de Gruyter.

Farrar, Scott, and Terry Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT International* 7: 200–203.

Gippert, Jost, Nikolaus Himmelmann, and Ulrike Mosel, eds. 2006. *Essentials of language documentation*. Berlin: Mouton de Gruyter.

Good, Jeff. 2004. The descriptive grammar as a (meta)database. Paper presented at the EMELD Language Digitization Project Conference 2004. http://linguistlist.org/emeld/workshop/2004/jcgood-paper.html

HASPELMATH, MARTIN. 1993. *A grammar of Lezgian*. Berlin, New York: Mouton de Gruyter.

HOLTON, GARY. 2003. Approaches to digitization and annotation: A survey of language documentation materials in the Alaska Native Language Center Archive. Proceedings of the EMELD Language Digitization Project Conference 2003. http://www.linguistlist.org/emeld/workshop/2003/paper-Holton.pdf

JESPERSEN, OTTO. 1924. *The philosophy of grammar.* London: Allen and Unwin.

KEENAN, EDWARD, and BERNARD COMRIE. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8:63-99.

LEHMANN, CHRISTIAN. 1980. Aufbau einer Grammatik zwischen Sprachtypologie und Universalienforschung. In *Wege zur Universalienforschung*, ed. by Hansjakob Seiler, Gunter Brettschneider, and Christian Lehmann, 29-37. Tübingen: Narr.

LEHMANN, CHRISTIAN. 1989. Language description and general comparative grammar. In *Reference grammars and modern linguistic theory*, ed. by Gottfried Graustein and Gerhard Leitner, 133-162. Tübingen: M. Niemeyer.

LEHMANN, CHRISTIAN. 1993. *On the system of semasiological grammar: Allgemein-Vergleichende Grammatik, vol. 1*. Bielefeld: Universität Bielefeld, Universität München.

LEHMANN, CHRISTIAN. 1998. Ein Strukturrahmen für deskriptive Grammatiken. In *Deskriptive Grammatik und allgemeiner Sprachvergleich*, ed. by Dietmar Zaefferer, 39-52. Tübingen: Niemeyer.

LEHMANN, CHRISTIAN. 2002. Structure of a comprehensive presentation of a language. In *Basic materials in minority languages*, ed. by Tasaku Tsunoda, 5–33. Osaka: Osaka Gakuin University.

LEHMANN, CHRISTIAN. 2004a. Documentation of grammar. In *Lectures on endangered languages* 4, Kyoto Conference 2001, ed. by Osamu Sakiyama, Fubito Endo, Honore Watanabe and Fumiko Sasama, 61-74. Osaka: Osaka Gakuin University.

LEHMANN, CHRISTIAN. 2004b. Funktionale Grammatikographie. In *Dimensionen und Kontinua. Beiträge zu Hansjakob Seilers Universalienforschung*, ed. by Waldfried Premper, 147–165. Bochum: N. Brockmeyer.

LEWIS, WILLIAM, SCOTT FARRAR, and TERRY LANGENDOEN. 2006. Linguistics in the internet age: Tools and fair use. Proceedings of the EMELD Language Digitization Project Conference 2006. http://www.linguistlist.org/emeld/workshop/2006/papers/Lewis.pdf

MOSEL, ULRIKE. 2006. Grammaticography: The art and craft of writing grammars. In Ameka et al. 2006: 41–68. Berlin, New York: Mouton de Gruyter.

MUNRO, ROBERT, and DANIEL NATHAN. 2005. Towards portability and interoperability for linguistic annotation and language-specific ontologies. Proceedings of the EMELD Language Digitization Project Conference 2005. http://linguistlist.org/emeld/workshop/2005/papers/Munro-paper.doc

NICKLES, MATTHIAS. 2001. Systematics: Ein XML-basiertes Internet-Datenbanksystem für klassifikationsgestützte Sprachbeschreibungen. München: Centrum für Informations- und Sprachverarbeitung.

NOONAN, MICHAEL. 2006. Grammar writing for a grammar-reading audience. *Studies in Language* 30(2): 351–365.

NORDHOFF, SEBASTIAN. 2007a. Growing a grammar with GALOES. Paper presented at the DoBeS workshop, June 2007, MPI Nijmegen. http://home.medewerker.uva.nl/s.Nord-

hoff/bestanden/Growing%20a%20grammar%20with%20GALOES.pdf

NORDHOFF, SEBASTIAN. 2007b. The grammar authoring system GALOES. Paper presented at the Wikifying Research Workshop, June 2007, MPI Leipzig. http://home.medewerker.uva.nl/s.Nordhoff/bestanden/The%20grammar%20authoring%20system%20GALOES.pdf

NORDHOFF, SEBASTIAN. 2007c. Grammar writing in the electronic age. Paper presented at the Conference of the Association for Linguistic Typology VII, September 2007, Paris. http://home.medewerker.uva.nl/s.Nordhoff/bestanden/ALT2007.pdf

PAYNE, THOMAS. 2006. A grammar as a communicative act, or what does a grammatical description really describe? *Studies in Language* 30(2): 367–383.

PAYNE, THOMAS, and DAVID WEBER, eds. 2006. *Perspectives on grammar writing*. Special issue of *Studies in Language* 30:2. Amsterdam: John Benjamins.

PETERSON, JOHN. 2002. *Cross-linguistic reference grammar (final report)*. München: Centrum für Informations und Sprachverarbeitung.

RESOURCE CREATION GROUP. 2003. Working group report from the EMELD Language Digitization Project Conference 2003. http://www.linguistlist.org/emeld/workshop/2003/resource-creation.doc

RICE, KEREN. 2006a. Let the language tell its story? The role of linguistic theory in writing grammars. In Ameka et al. 2006: 235–268. Berlin, New York: Mouton de Gruyter.

RICE, KEREN. 2006B. A TYPOLOGY OF GOOD GRAMMARS. *STUDIES IN LANGUAGE* 30(2): 385-415.

SCHULTZE-BERNDT, EVA. 1998. Zur Interaktion von semasiologischer und onomasiologischer Grammatik: Der Verbkomplex im Jaminjung. In Deskriptive Grammatik und allgemeiner Sprachvergleich, ed. by Dietmar Zaefferer, 149-176. Tübingen: Niemeyer.

SIMONS, GARY, BRIAN FITZSIMONS, TERRY LANGENDOEN, WILLIAM LEWIS, SCOTT FARRAR, ALEXIS LANHAM, RUBY BASHAM and HECTOR GONZALEZ. 2004. A model for interoperability: XML documents as an RDF database. Proceedings of the EMELD Language Digitization Project Conference 2004. http://linguistlist.org/emeld/workshop/2004/simons-paper.pdf

SCHROETER, RONALD, and NICK THIEBERGER. 2006. EOPAS, the EthnoER online representation of interlinear text. In *Sustainable data from digital fieldwork*, ed. by Linda Barwick and Nick Thieberger, 99-124. Sydney: University of Sydney.

STASSEN, LEON. 1985. *Comparison and universal grammar.* Oxford: Basil Blackwell.

THIEBERGER, NICK. 2006. *A grammar of South Efate: An Oceanic language of Vanuatu*. Oceanic Linguistics Special Publications no. 33. Honolulu: University of Hawai'i Press.

VON DER GABELENTZ, GEORG. 1891/1984. *Die Sprachwissenschaft. Ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Tübingen: Narr.

VALENTINE, RANDY. 2001. *Nishnaabemwin reference grammar.* Toronto: University of Toronto Press.

WEBER, DAVID. 2006a. Thoughts on growing a grammar. *Studies in Language* 30(2): 417-444.

WEBER, DAVID. 2006b. The linguistic example. *Studies in Language* 30(2): 445–460.

ZAEFFERER, DIETMAR, ed. 1998a. *Deskriptive Grammatik und allgemeiner Sprachvergleich*. Tübingen: Niemeyer.

ZAEFFERER, DIETMAR. 1998b. Ein Strukturrahmen für deskriptive Grammatiken: Die Beschreibung sprachlicher Funktionen. In *Deskriptive Grammatik und allgemeiner Sprach-*

vergleich, ed. by Dietmar Zaefferer, 29–38. Tübingen: Niemeyer.
Zaefferer, Dietmar. 1998c. Einleitung: Allgemeine Vergleichbarkeit als Herausforderung für die Sprachbeschreibung. In *Deskriptive Grammatik und allgemeiner Sprachvergleich*, ed. by Dietmar Zaefferer, 1–5. Tübingen: Niemeyer.
Zaefferer, Dietmar. 2005. The place of linguistic concepts within a general ontology of everyday life. Proceedings of the EMELD Language Digitization Project Conference 2005. http://www.linguistlist.org/emeld/workshop/2005/papers/Zaefferer-paper.doc
Zaefferer, Dietmar. 2006. Realizing Humboldt's dream: Cross-linguistic grammaticography. In Ameka et al. 2006: 113–136. Berlin, New York: Mouton de Gruyter.

Sebastian Nordhoff
s.nordhoff@uva.nl