2

Language Documentation & Conservation Special Publication No. 3 (August 2012):
Potentials of Language Documentation: Methods, Analyses, and Utilization,
ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann,
Dagmar Jung, Anna Margetts, and Paul Trilsbeek, pp. 7–16
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4511

Prospects for e-grammars and endangered languages corpora

Sebastian Drude

Max Planck Institute for Psycholinguistics, Nijmegen

This contribution explores the potentials of combining corpora of language use data with language description in *e-grammars* (or *digital grammars*). We present three directions of ongoing research and discuss the advantages of combining these and similar approaches, arguing that the technological possibilities have barely begun to be explored.

1. INTRODUCTION: GRAMMARS AND LANGUAGE DOCUMENTATION. Grammars, in the sense of comprehensive descriptions of the structural properties of a language, have been at the core of linguistics since its very beginning. All studies that aim to understand the patterns and limits of the variability of the human speech faculty need thorough accounts of more languages than just the few which are more well known (Zaefferer 1998). Grammars are the principal component of the so-called *Boasian triad* (i.e.: grammar, dictionary, texts; cf. Grinevald 2001) that is the customary result of linguistic fieldwork. As such, they are a well-established genre of scientific texts, a genre which recently has gained attention on its own (Ameka et al. 2006, Lehmann 2004a,b, Payne & Weber 2007).

Grammars consist mostly of prose text organized in a hierarchy of sequential chapters and sections. Certain special elements, however, distinguish grammars from other scientific books, in particular *exemplars* (in the sense of Good 2004) such as words, phrases, sentences of the language studied. These exemplars usually come with a translation and some additional analysis; in recent grammars this is often in the form of basic glossings (an interlinearized rendering of the morphs or words of the object language, as standardized by the Leipzig Glossing Rules, Comrie et al. 2008). Other typical elements found almost exclusively in grammars and other linguistic texts are, for instance, paradigms (and similar tables), and, depending on the linguistic theory which underlies the description, formal rules, or structure graphs such as trees indicating the constituent structure of sentences.

Although sometimes idealized as "theory-neutral", all descriptions of languages necessarily rely on general linguistic theories. These provide, in particular, the technical terms which are applied in the description. Often, the writer of a grammar cannot take it for granted that the underlying theoretical concepts are known to the readers. Therefore it is characteristic of many grammars to contain interspersed paragraphs explaining the underlying theory fragments and terms before they are applied to the language described.

A collection of (analysed and translated) texts, another component of the *Boasian triad* mentioned above, is now complemented or even superseded by the outcome of language documentations (in the modern sense as established by Himmelmann 1998). In this sense, documentations consist crucially of multi-purpose digital corpora of data of naturalistic language use, that is, annotated primary multi-media data. One of the uses of documentations is, of course, the study and analysis of the structure of the language, the results of which are traditionally presented in descriptive grammars (see above).

Language documentation, in this sense, crucially depends on new digital technologies (the resulting corpora with their recordings, annotations, and metadata are digital); for this reason among others linguistics has been a pacesetter in the developing field of *digital humanities*. The question now is how grammars can benefit from these new technologies and from the digital corpora that are being compiled. The following sections discuss three recent approaches to this question.[1]

2. HYPERTEXT GRAMMARS. One obvious way to enhance language descriptions with digital technology is to make use of hyperlinks, i.e. the interlinking of different locations within or across documents and other resources, giving rise to what can be called *hypertext grammars*. Already in many paper-based grammars, cross-references abound due to the truly systemic integrated character of the structure of any language. Hyperlinks are an excellent means of making these relations explicit and easy to access.

Perhaps the most needed links, however, are those linking points in the grammar text where a certain phenomenon is described with examples of this phenomenon in the corpus. As has recently been stated many times (e.g., Bird & Simons 2003: 563, Himmelmann 2006: 6), being able to illustrate statements in a grammar (or, say, a typological study) with recordings of language use would make linguistics much more accountable, providing a much more solid basis for the empirical claims and generalizations.

Other links can enhance a grammatical description. For instance, for occurrences of individual (forms of) words or morphemes discussed in the text, one would like to be directed immediately to a corresponding entry in a lexical database (or electronic dictionary). Also, details of the underlying theoretical framework could, at last, be presented apart from the description that applies these concepts, but static links or intelligent search mechanisms between both resources could provide the needed contextualization. Figure 1 (cf. Drude forthcoming) demonstrates these key features and links of a hypertext grammar.

As explained in what follows, the boxes in this figure mark elements that presuppose certain assumptions for a concrete implementation, applying certain solutions for some of the technical challenges of conceiving and implementing hypertext grammars, especially solutions which are being developed at The Language Archive (at the Max-Planck-Institute for Psycholinguistics in Nijmegen). In particular, the external resources, corresponding to the other components of the Boasian triad (and to theoretical work explaining the underlying theories and applied terms) could be instantiated by existing LEXUS, ANNEX, and ISOcat tools belonging to Language Archiving Technology (LAT, developed at the MPIPL).

---

[1] These and several other recent developments and projects were presented at the symposium on electronic grammaticography, organized by Sebastian Nordhoff, as part of the 2nd Conference on Language Documentation and Conservation 2011 in Hawai‘i.

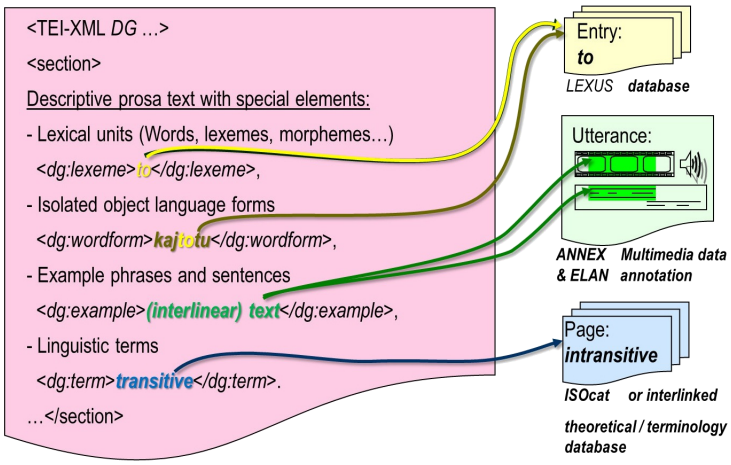

FIGURE 1: Elements and relations in a page in a hypertext grammar

For instance, an online lexical database can be developed or at least rendered with LEXUS.[2] The text corpus (containing examples) can be a collection of ELAN[3] annotation files together with the underlying primary data. If these are provided by a LAT repository,[4] the corresponding examples can be visualized with the ANNEX online service.[5] The technical terms in a description can be linked to their definition in the ISOcat registry[6] (for more comprehensive explanations of theoretical concepts, a theory-specific wiki or something similar could be employed, which still can make use of ISOcat as a point of reference).

Crucially, all elements of the external resources need some kind of persistent identifier that guarantees that the links remain stable over time even if physical locations and infrastructure change. LAT provides Handle Digital Object Identifiers.[7]

As indicated by the boxes on the left-hand side, the text body itself would probably best be encoded in some XML-based format, where the XML-tags allow specification of the hyperlinks from examples or other particular elements in the text. The norms proposed by the Text Encoding Initiative (TEI, cf. TEI Consortium 2009) could provide the basis for such a format, but they probably would have to be extended in order to serve the special needs of a digital grammar.

There are other aspects of hypertext grammars that have been explored in particular by Nordhoff (2007, 2008), for instance the conceptualization of a grammar as a living document such as a wiki. Also, others are working on certain aspects of related technology which could be integrated into a comprehensive system; an example is the EOPAS system

[2] Cf. http://tla.mpi.nl/tools/tla-tools/lexus.

[3] Cf. http://tla.mpi.nl/tools/tla-tools/elan.

[4] Cf. http://tla.mpi.nl/tools/tla-tools.

[5] Cf. http://tla.mpi.nl/tools/tla-tools/annex.

[6] Cf. http://tla.mpi.nl/tools/tla-tools/isocat.

[7] Cf. http://www.doi.org/about_the_doi.html.

being developed by Thieberger and others (Schroeter & Thieberger 2006), which provides easy direct access to spoken language examples.

3. TREEBANKS AND THE GRAMMAR MATRIX PROJECT. Hypertext grammars are conceptually close to paper grammars, although the style of writing and aspects of the content will be affected by the digital medium. But there are also *implemented grammars*, the digital equivalent to grammars as developed by computational linguistics, in particular grammar engineering. Rather than texts directed to a human reader, implemented grammars are procedural representations capable of, for instance, parsing and analyzing written word forms and sentences of a language.

As such, implemented grammars are not just tools for automatic annotation (parsing, glossing, structure assignment, etc.) but also aim at a technical representation of what we understand of the language structure – what the implemented grammar is not able to analyze points at possible gaps in our understanding of the language structure on the respective level, or at least gaps in the technical representation of our understanding. At the same time, the technical formal representation of rules/features allows for a much more precise comparison between languages, for instance for typological or historical-reconstructive research, and the same holds for the much more standardized automatically generated annotations generated by such systems.

Most implemented grammars apply only to one each of a small set of better-studied languages. For the study of linguistic diversity and less well-known languages, generic parser engines are needed that can understand different separately developed sets of rules, tailored to different languages. Kirschenbaum et al. (this volume) present research on machine learning for morphological analysis, both supervised and unsupervised. This research shares the perspective of developing systems for the automatic annotation of text in corpora, but is *per se* not rule based, i.e. it does not presuppose a technical representation of the system of the language.

One particularly promising project is the Linguistic Grammars Online (LinGO) Grammar Matrix project being developed by E. Bender and colleagues (Bender et al. 2010). In this project, a generic program applies language-specific rules to sentences and proposes syntactic trees (according to the HPSG theoretical model), which can be included in a *treebank*, a database of such trees (in recent years, one of the most often used resources for major languages).

To be more specific, the LinGO project does not speak of rules but rather of *signs*, an HPSG term that covers not only lexical units but also grammatical classes or word order patterns, each assigned to a semantic interpretation. So far, traditional descriptions have been translated into such *signs* manually by linguists working in cooperation with information scientists. For each sentence in a text, the parsing mechanism then offers a number of possible trees compatible with the signs known to the system, which is able to learn and remember the tree chosen by a linguist.

In the *Grammar Matrix* (GM), a more recent track of research by Bender and her colleagues (2010), the signs are (semi-)automatically derived from a typological profile of the language, which is elicited from the linguist in the form of a questionnaire. This is a major advance since less technical knowledge (or engineering work of a technical specialist) is needed.

Figure 2 (Bender et al. 2010: 29) represents an overview of the GM project with the elicitation of typological information (left), which enters the creation (right) of a customized grammar (i.e. a steadily improving language specific parsing automatism) together with a language independent core grammar and the analyses for previously parsed sentences.

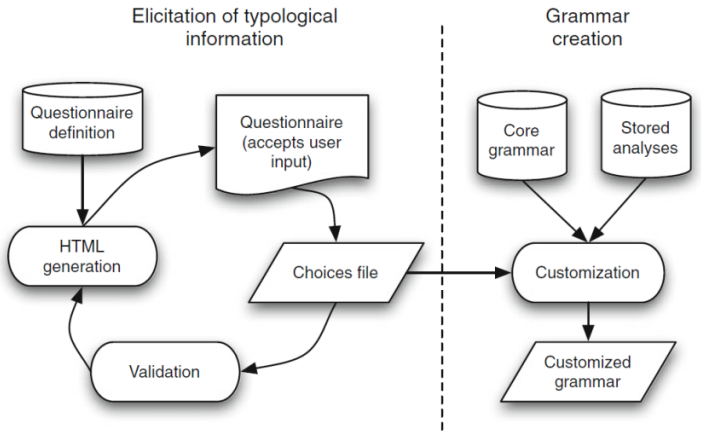

FIGURE 2: Grammar matrix components and workflow

The following screenshot from the LinGO English Resource Grammar (ERG)[8] shows the result of parsing an English sentence with the first two possible tree structures (they are ordered by probability according to what can be deduced from earlier parses of similar sentences). The formalism is, of course, highly theory-dependent, but similar systems could be conceived for other sufficiently formal explicit linguistic frameworks.

In the future, Bender and colleagues plan to develop a system that automatically derives rules (*signs*) from a sufficiently large set of texts with interlinear glosses (although these only contain morphological information and only incidentally syntactic features which are needed for constructing syntactic trees).

4. INTEROPERABLE GRAMMARS AND LITERATE PROGRAMMING. A drawback of most current parsing mechanisms is that the technical rules are difficult to build and work only with the one generic parsing mechanism they have been developed for. If a new parser is developed, the rules all have to be coded manually again. They are also not easily understood by humans and are not explicitly linked to a traditional description, although individual rules usually correspond to specific parts of a paper grammar or dictionary.

A current project developed by Maxwell and others (Maxwell & David 2008) promises to overcome these shortcomings. They build *Interoperable Grammars* (IGs), which should be comprehensible to both humans and machines. The idea is to apply the concept of *literate programming* (Knuth 1984) to grammar writing, arranging the prose descriptions around technical parts, which translate the described structural properties (rules) into a format that

[8] http://erg.delph-in.net/ (1 November, 2011).

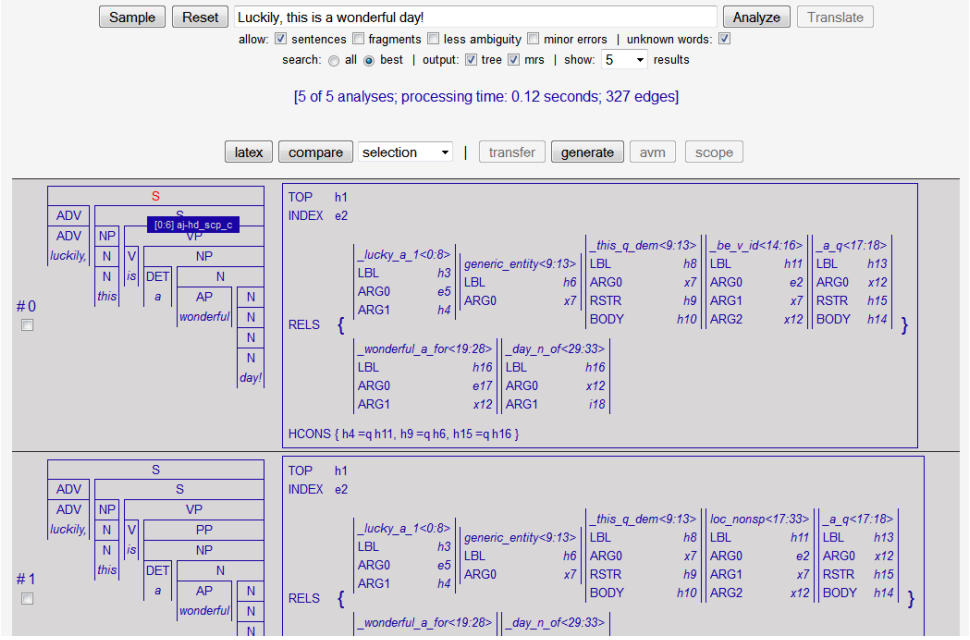

FIGURE 3: Automatically generated tree from the LinGO English Resource Grammar

can be understood by a software system. The text is the explanation, as it were, of the program code.

Crucially, the XML-based format for the rules is independent of any individual parser but allows for automatic transformation into formats usable by the respective parsers. In this manner, the code remains compatible with the evolving and frequently superseded parsers. A new parser will only need a corresponding pre-processor for rendering the rules in its specific format.

The parser applied by the IGs so far deals only with morphology (different from those in GM), and so the rules currently cover mostly morphemes and word-internal phonological variation. The outcome of applying a tailored parser to a text is a morphologically analysed and tagged text (basically, an interlinearized text).

The following diagram (Figure 4, from Maxwell forthcoming) shows the derivations of the different files and components of the IG system, where the publishable (text) grammar and the formal grammar are derived from one and the same XML document. The formal grammar is converted into a parser engine specific format and is combined with lexical information. The resulting specific parser instance can be applied to texts in the target language.

The following XML-snippet (Figure 5, from Maxwell & David 2008) exemplifies the generic technical representation of grammatical features, in this case a certain first-person-future-indicative affix with two allomorphs. (The use of UNICODE provides a means to deal with the non-western script without difficulties.)

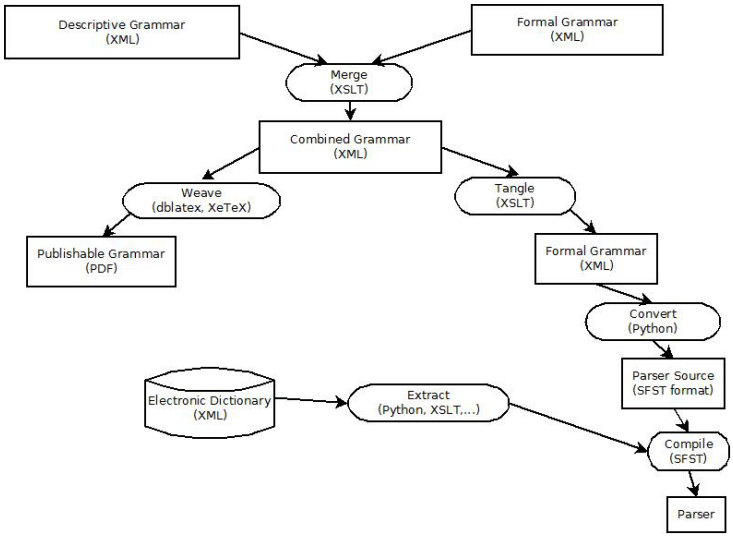

FIGURE 4: Components and workflow in the Interoperable Grammars project

```
<Mo:InflectionalAffix gloss="-1Fut" id="af1Fut">
  <!--The two "allomorphs" are really allographs-->
  <Mo:Allomorph form="(ৎ)ব">
    <!--Spelled 'bo'; usually (not always) after a C-stem -->
  </Mo:Allomorph>
  <Mo:Allomorph form="ব">
    <!--Spelled 'b'; usually (not always) after a vowel stem -->
  </Mo:Allomorph>
  <Mo:inflectionFeatures>
    <Fs:f name="Tense"><Fs:symbol value="Future"/></Fs:f>
    <Fs:f name="Mood"><Fs:symbol value="Indicative"/></Fs:f>
    <Fs:f name="Person"><Fs:symbol value="1"/></Fs:f>
  </Mo:inflectionFeatures>
</Mo:InflectionalAffix>
```

FIGURE 5: XML-snippet showing the representation of a grammatical feature

5. CONCLUSION: PROSPECTS FOR DIGITAL GRAMMARS AND LANGUAGE DOCUMENTATION. There are several other projects on implemented grammars which cannot be discussed any further here, for instance *Grammix* (Müller 2007), *ParGram* (Butt et al. 2002), *Meta-Grammar* (Kinyon et al. 2006), *OpenCCG* (Baldridge et al. 2007), or *KPML* (Bateman et al. 2005). Some of these may contribute features, conceptions, or functionalities which can be picked up by other future more integrated projects and systems. Thus, there is clearly a need for more interaction between field linguists and computational linguists. There is a huge potential in integrating the emerging systems, which are in several aspects complementary to one another, and in others resemble one another.

For instance, most computational approaches to grammar(s) aim at parsing sentences of a corpus; but the IG system is currently restricted to morphology, and the GM framework focuses on syntactic structure. The architecture of both implemented grammars relies on a generic parsing engine configured with customized rules and trained with data from individual languages. This approach is clearly most suitable for describing the plenitude of understudied languages. Advantages include gaining richer textual data and spotting gaps in analysis and description.

However, one point which continues to be problematic for all parsers, so far, is the often elliptical and in many other ways non-standard character of natural spontaneous spoken language. Syntactic parsers can usually only cope with complete grammatical sentences, or at least much better so. It remains to be explored and solved as to how parsers can be adjusted to deal with variability as typical for spoken language. (Some attempts have been made for more well-known languages, e.g. Nivre & Grönqvist 2001 for Swedish.)

Hypertext Grammars may seem uninteresting and even boring in comparison to implemented grammars since they do not create any new content or annotation. Still, linguistics needs an easy way to interlink scholarly linguistic work with corpora, lexica, and terminological / theoretical texts. Each of the pioneering works by C. Lehmann, S. Nordhoff, and N. Thieberger and others cover some aspects, but none of them covers all or even a major part of the necessary or desirable aspects which were detailed above in Section 2, and they do not address the integration with implemented grammars exemplified in Sections 3 and 4.

Overall, in a mid or long term perspective it would be ideal to combine all three (and possibly still other) approaches into one integrated framework which allows the linguist to create descriptions that satisfy two important requirements. 1) They match the high standards of traditional grammatical descriptions, including an overall pedagogically informed exposition. 2) They allow the reader to a) access the underlying primary data as well as other resources, and also b) to check the validity of derived technical or formal rules against a corpus of sentences, producing richer annotation. Automated processes would also result in more standardized annotations, which would be more suitable for cross-linguistic comparisons and typological work.

An ideal future framework would be modular; various parsers based on different theoretical frameworks could be chosen for different purposes (like the algorithms presented by Kirschenbaum et al. in this volume, or the IG morphological parser, or the HPSG/based trees created by the GM framework). Core aspects of an integrated authoring and reading environment are the hypertext interlinking with external resources and perhaps the IG *literate programming* conception. Additional possible modules not discussed here include statistical and supervised and unsupervised machine learning methods. The possibilities of combining such different approaches have not yet been explored, even initially.

With such a combination, we would gain: a) more comprehensive, empirically sound and accountable grammatical descriptions, b) more comparable and richly annotated corpora, and as a result c) a deeper understanding of language variation and ultimately even d) a ground-laying conceptual and technical framework for understanding language structure and the meaning of linguistic constructions, a necessary condition for machine-translation, as it were, the holy grail of computational linguistics.

REFERENCES

Ameka, Felix K., Alan Dench & Nicholas Evans (eds.). 2006. *Catching Language: The Standing Challenge of Grammar Writing* Trends in Linguistics. Studies and Monographs 167. Berlin: De Gruyter.

Baldridge, Jason, Sudipta Chatterjee, Alexis Palmer & Ben Wing. 2007. DotCCG and VisCCG: Wiki and programming paradigms for improved grammar engineering with OpenCCG. In Tracy Holloway King & Emily M. Bender (eds.), *Proceedings of the GEAF 2007 Workshop Stanford, CA. CSLI* CSLI Studies in Computational Linguistics ONLINE, 5–25. Stanford: CSLI.

Bateman, John A, Ivana Kruijff-Korbayová & Geert-Jan Kruijff. 2005. Multilingual resource sharing across both related and unrelated languages: An implemented, open-source framework for practical natural language generation. *Research on Language and Computation* 3(2–3). 191–219.

Bender, Emily M., Scott Drellishak, Antske Fokkens, Laurie Poulson & Safiyyah Saleem. 2010. Grammar customization. *Research on Language and Computation* 8(1). 23–72.

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 557–582. http://www.sil.org/~simonsg/preprint/Seven%20dimensions.pdf (5 April, 2012).

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi & Christian Rohrer. 2002. The parallel grammar project. In John Carroll, Nelleke Oostdijk & Richard Sutcliffe (eds.), *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, 1–7. http://www.aclweb.org/anthology/W/W02/W02-1503.pdf (5 April, 2012).

Comrie, Bernard, Martin Haspelmath & Balthasar Bickel. 2008. *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses*. Leipzig: Max Planck Institute for Evolutionary Anthropology & Department of Linguistics of the University of Leipzig. Online at http://www.eva.mpg.de/lingua/pdf/LGR08.02.05.pdf (5 April, 2012).

Drude, Sebastian. forthcoming. Digital grammars: Integrating the Wiki/CMS approach with Language Archiving Technology and TEI. In Nordhoff, Sebastian (ed.), Electronic Grammaticography. University of Hawai'i Press. [Talk given at the grammaticography colloquium at the 2.ICLDC, Hawai'i, Februrary 2011].

Good, Jeff. 2004. The descriptive grammar as a (meta)database. In *Proceedings of the E-MELD Workshop 2004: Linguistic Databases and Best Practice, July 15–18, 2004, Detroit, Michigan*, .

Grinevald, Colette. 2001. Encounters at the brink: Linguistic fieldwork among speakers of endangered languages. In Osamu Sakiyama (ed.), *Lectures on Endangered Languages 2*, 285–313. Osaka: ELPR Publications.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–195. Full version online at http://www.hrelp.org/events/workshops/eldp2005/reading/himmelmann.pdf (5 April, 2012).

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

Kinyon, Alexandra, Owen Rambow, Tatjana Scheffler, SinWon Yoon & Aravind K Joshi. 2006. The metagrammar goes multilingual: A cross-linguistic look at the V2-phenomenon. In *Proceedings of the 8th International Workshop on Tree Adjoining Grammar and Related Formalismss*, 17–24. Sydney: Association for Computational Linguistics.

Kirschenbaum, Amit, Peter Wittenburg & Gerhard Heyer. this volume. Unsupervised morphological analysis of small corpora: First experiments with Kilivila.

Knuth, Donald E. 1984. Literate programming. *The Computer Journal* 27(2). 97–111.

Lehmann, Christian. 2004a. Documentation of grammar. In Osamu Sakiyama, Fubito Endo, Honore Watanabe & Fumiko Sasama (eds.), *Lectures on endangered languages: 4. From Kyoto Conference 2001* Endangered Languages of the Pacific Rim Publication Series, C-004, 61–74. Osaka: Osaka Gakuin University.

Lehmann, Christian. 2004b. Funktionale Grammatikographie. In Waldfried Premper (ed.), *Dimensionen und Kontinua: Beiträge zu Hansjakob Seilers Universalienforschung* Diversitas Linguarum 4, 147–165. Bochum: Brockmeyer.

Maxwell, Michael. forthcoming. Electronic grammars: Taking advantage of the possibilities. In Nordhoff, Sebastian (ed.), Electronic Grammaticography. University of Hawai'i Press. [Talk given at the grammaticography colloquium at the 2.ICLDC, Hawai'i, Februrary 2011].

Maxwell, Michael & Anne David. 2008. Interoperable grammars. In *Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, Hong Kong. http://hdl.handle.net/1903/11611 (5 April, 2012).

Müller, Stefan. 2007. The Grammix CD-ROM: A software collection for developing typed feature structure grammars. In Tracy Holloway King & Emily M Bender (eds.), *Proceedings of the GEAF 2007 Workshop Stanford, CA. CSLI* CSLI Studies in Computational Linguistics ONLINE, Stanford: CSLI.

Nivre, Joakim & Leif Grönqvist. 2001. Tagging a corpus of spoken Swedish. *International Journal of Corpus Linguistics* 6(1). 47–78.

Nordhoff, Sebastian. 2007. The Grammar Authoring System GALOES. Talk given at the Workshop on "Wikifying research". MPI Leipzig. http://www.eva.mpg.de/lingua/staff/nordhoff/pdf/The%20grammar%20authoring%20system%20GALOES.pdf (5 April, 2012).

Nordhoff, Sebastian. 2008. Electronic reference grammars for typology: Challenges and solutions. *Language Documentation and Conservation* 2(2). 296–324. http://hdl.handle.net/10125/4352 (5 April, 2012).

Payne, Thomas E & Davis J Weber (eds.). 2007. *Perspectives on Grammar Writing*. Amsterdam: Benjamins.

Schroeter, Ronald & Nicholas Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In *Sustainable Data from Digital Fieldwork. Proceedings of the conference held at the University of Sydney, 4–6 December 2006*, Sydney: Sydney University Press. http://hdl.handle.net/2123/1297 (14 December, 2011).

TEI Consortium. 2009. TEI P5: Guidelines for electronic text encoding and interchange, 1st edn. http://www.tei-c.org/Guidelines/P5/.

Zaefferer, Dietmar. 1998. *Deskriptive Grammatik und allgemeiner Sprachvergleich* Linguistische Arbeiten 383. Tübingen: Niemeyer.

Sebastian Drude
Sebastian.Drude@mpi.nl