Archiving Grammatical Descriptions

SEBASTIAN NORDHOFF & HARALD HAMMARSTRÖM
Max Planck Institute for Evolutionary Anthropology

Abstract

We argue that grammatical descriptions should be archived not in a monolithic but in a granular way, making use of the framework of the Text Encoding Initiative. We show how tags of the extant framework and additions crafted for descriptive linguistics can be applied to linguistic text. Granular annotation with Uniform Resource Identifiers for all elements will allow grammatical descriptions (and their components) to be part of the Semantic Web. We add to existing formalizations of low-level elements by providing a schema for the higher order organization of grammatical descriptions and discuss how legacy descriptions can be parsed and annotated in a scalable manner.

KEYWORDS: archiving, grammatical description, schema, linked data, Text Encoding Initiative

1. ARCHIVES

There are several aspects to archiving, whose importance varies according to the discipline which an archive serves. Archiving originally dealt with *physical objects* (e.g. vases, axes, books), but in recent years, there has been a movement towards digital archiving, where one archives *representations of objects* (e.g. 3D models of vases or axes, or scans of books) as digital surrogate of physical artifacts. Finally, there is archiving of *content*, such as the text of documents, which is not concerned with physical form (e.g. paper type or size), but rather with the characters which make up a document's content.

When considering the archiving of grammatical descriptions, we can illustrate these three aspects as follows. A grammatical description could be archived as:

1. a printed book (the artifact), taking care of environmental conditions such as humidity, temperature, exposure to sunlight etc.
2. a set of scans (e.g. TIFF files) of a book representing every page in high resolution. This is a representation, or surrogate, of the artifact. Non-visual information, such as the texture or smell of a book is not captured.
3. a file containing the string of characters which make up the content of the book. Here, visual information is also lost (page layout, colour, typography etc.), although some can be recorded as meta-information.

For grammatical description, archiving content is clearly the most important aspect. While there are some masterpieces of editing in the grammatical literature, what is produced in language documentation normally excels by content rather than layout, typography or paper choice. What is to be preserved for future generations is a grammar's content, not the physical arrangement of glyphs on a

support. This is even more the case for documents "born-digital" where the physical support does not exist until the user prints the document. References to the visual characteristics of the document (page breaks, line breaks) can also be stored, but these are not central.

Concerning the archiving of texts, two further distinctions can be made. The text could be archived:

1. without internal structure, that is, a long string of characters; or
2. as a set of elements which constitute the text (headings, sections, footnotes, cross-references, etc)

The second, more granular, approach allows for more accurate search for particular items, easier updates, and better integration into the Semantic Web. In the remainder of this paper, we advocate such a granular approach to archiving grammatical descriptions using the framework of the Text Encoding Initiative (TEI, Sperberg-McQueen & Burnard 2010).

With regard to archives, we can distinguish two further axes: orientation and perfectivity. The first axis, orientation, refers to the focus of the archive being inward or outward: Is it important to get materials into the archive (store), or is it important to get materials out of the archive to the interested users (serve)? A good archive will try to serve both functions as well as possible, but scarcity of resources often means that a choice has to be made. The second axis, which we call perfectivity in analogy to the the term used for linguistic aspect, refers to the state of completeness an archive requires. A 'perfective' archive only accepts finished documents, and does not allow modification of already archived documents. The archive is read-only, so to speak. A non-perfective archive will store documents of varying states of completion, and allow modification of archived content (version history can be kept to assure that the evolution of content can be tracked). We assume that grammatical descriptions should be held in outward-oriented, non-perfective archives.

Good (2004) conceives of a grammatical description as a meta-database of nested 'annotations', a collection of short independent chunks with explicit relations to each other. Nordhoff (2008) argues that grammatical descriptions are best seen as non-linear texts, i.e. hypertexts, where each individual reader follows their own path (e.g. skip phonology, start with syntax, jump back to morphology as required, use table of contents or index to continue etc.).

Cysouw (2009) argues for the use of atomic linguistic facts as the basis of linguistic knowledge, using 'micro-publications', very short statements about a linguistic fact for a particular language. Pulling these three authors' ideas together, we can say that a grammatical description is a non-linear meta-database of micropublications. This also recalls the idea of granular representation mentioned above.

A granular approach agrees well with this model, as items can be added or modified on a local basis, whereas in the monolithic approach, every local modification would also be a global modification. The granular approach also allows easier serving of particular chunks, because of easier retrieval. Finally, a granular approach will allow to uniquely identify chunks and refer to them in

Semantic Web frameworks such as RDF[1] or OWL.[2]

2 GRANULAR TEXT

Let us illustrate what we mean by a granular text with an example:

```
<div id="ch3" type="chapter" n="3">
  <head>Morphology</head>
  <div id="ch3s1" type="section" n="1">
    <head>Nominal morphology</head>

    <p>
      In contrast to <ref target="#verbalmorphology">verbal
      morphology</ref> Nominal morphology is very important in
      <languagename iso6393="qqq">Ugubugu</languagename>. This
      can be seen in example <ptr target="#ch3s1ex1" />,
      especially the <technicalterm ontology="GOLD" value="case
      marker">case  marker</technicalterm> <phraseglosspair>
      <phrase iso6393="qqq">ka </phrase> <gloss type="Leipzig">
      ACC</gloss></phraseglosspair> is important here.
    </p>
    <lgex id="ch3s1ex1" number="1">
      <sourceline> ... </sourceline>
      <interlinear> ... </interlinear>
      <translationline> ... </translation>
    </lgex>
  </div>
</div>
```

Note these points:

1. important elements such as headings, examples, and cross-references are explicitly tagged
2. there are unique references to paragraphs and examples
3. some terms are enclosed in semantic markup
4. semantic markup includes references to term definitions, in this case the GOLD ontology[3] and the Leipzig Glossing Rules.[4]

Such markup combined with unique references allows for integration into the semantic web. One could then retrieve all sections of grammatical descriptions which refer to the concept ‘case marker’ as defined in GOLD, even if the actual terms employed differ.

3 LINKED DATA

We will briefly illustrate the usefulness of such a granular resource within the

---

[1] http://www.w3.org/RDF/

[2] http://www.w3.org/TR/owl-features/

[3] http://linguistics-ontology.org/

[4] http://www.eva.mpg.de/lingua/resources/glossing-rules.php

Linked Data paradigm (Heath & Bizer 2011).[5] The five steps for Linked Data are (Berners-Lee 2006):

* make materials available on the Web (in whatever format) under an open license

** provide materials as structured data (e.g. spreadsheet, Excel instead of image scan of a table)

*** use non-proprietary formats (e.g., CSV rather than Excel)

**** use URIs to identify things so that others can link to them

***** link your data to other data to provide context

The granular approach allows chunks to be referred to as parts of linked semantic statements, so we can say 'this section covers a topic which is also found in GOLD' or 'this section is a close, but not perfect, match to what is found in …'.

Three things are required in order to arrive at such formulations: the arguments of the linked relation must be identifiable, they must have Uniform Resource Identifiers (URIs); the relation must be defined; and the formalism to link the relation and the arguments must be established.

URIs can, for the purpose of this paper, be equated with web addresses of a certain format. Relations are defined in a number of vocabularies, of which RDFS[6], OWL, Dublin Core[7], SKOS[8], GOLD and lexvo[9] are the most important ones for descriptive linguistics. The formalism to link these together is RDF[10] here represented in a variant called N3.[11] An example will show the general approach:

```
(1)<grammararchive:123/chapter/4/section/5> <rdfs:seeAlso> <gold:Infix> .
```

RDF predications are written in an SVO notation. In this example, the 'subject' is Chapter 4, Section 5 of the book with ID 123; the 'object' is *gold:Infix*; and the two are linked by the predicate *rdfs:seeAlso*. Crucially, all of the three items are dereferenceable, which means that a definition of what they mean can be looked up on the Internet.[12]

---

5 Linked Data for Linguistics is discussed more extensively in Chiarcos et al. 2012 and Good 2012.

6 http://www.w3.org/TR/rdf-schema/

7 http://dublincore.org/

8 http://www.w3.org/2004/02/skos/

9 http://www.lexvo.org/

10 http://www.w3.org/RDF/. Here, RDF is represented in a variant called N3 (http://www.w3.org/DesignIssues/Notation3.html)

11 http://www.w3.org/DesignIssues/Notation3.html

12 The actual Internet addresses have been abbreviated here; one would look up:
http://www.grammararchive.org/grammar/123/chapter/4/section/5
http://www.w3.org/2000/01/rdf-schema#seeAlso

4 TEXTUAL ELEMENTS IN LINGUISTICS

In order to create a schema for grammatical descriptions, one needs to take stock of the elements which are found in this type of work. Some are shared with other types of texts, such as paragraphs, headings, and cross-references. For those, the TEI vocabulary can be used. In linguistic texts, there are also a number of elements not listed in TEI. These will be discussed below.

4.1 Named entities

Named entities refer to concepts which have a definition external to the text. Examples include countries, cities, persons, as well as languages, books and linguistic concepts. Named entities are enclosed in the relevant semantic markup, as in the following example:

```
<p>
  <language iso639-2="cpp"> Diu Indo-Portuguese
  </language> is spoken in the <city geonamesID="1272502">city of
  Diu</city> in the <country ISO-3166-1="IN">Indian</country>
  <province ISO-3166-2="IN-DD">territory of Daman and
  Diu</province>. <person gnd="118611046">Hugo Schuchardt</person>
  and <person gnd="115161023">Sebastião Dalgado</person> provided
  the first description of this dialect; the latest work is
  Cardoso's <book ISBN="978-90-78328-87-2">A grammar of
  Indo-Portuguese</book>.
</p>
```

The extraction of named entities from a given text (Named Entity Recognition) is an important subfield of computational text linguistics (e.g. Borthwick, 1999).

Linguistic concepts can be be treated as named entities if they are defined outside the text, for example in the ISO register,[13] lexvo, or GOLD.

```
<p>
  <languagename iso639-3="tgl">Tagalog</languagename> has
  <technicalterm GOLD="Infix">infixes</technicalterm>.
</p>
```

4.2 Object language

In linguistic texts, words written in languages which are not the metalanguage of the text are rather frequent. These are commonly typeset in italics. A semantic representation would be as follows.

```
<p>
 Italian <objectlanguage iso639-3="ita">cinque</objectlanguage>
```

---

http://linguistics-ontology.org/gold/2010/Infix

[13] http://www.sil.org/iso639-3/

```
 corresponds to Spanish <objectlanguage iso639-3="spa">cinco
 </objectlanguage>.
</p>
```

The attribute iso639-3 refers to the ISO 639-3 code[14] of the language.

4.3 Phrase-gloss pairs

Grammatical descriptions very often provide object language terms immediately followed by translations:

```
<p>
  Spanish <phraseglosspair><phrase iso639-3="spa">dolor</phrase>
  <gloss iso639-3="eng">pain</gloss><phraseglosspair> preserves
  Latin intervocalic l, while Portuguese <phraseglosspair><phrase
   iso639-3="por">dor</phrase> <gloss iso639-3="eng">pain</gloss>
  <phraseglosspair> does not.
</p>
```

4.4 Exemplars

The most salient text element found in grammars is the example, of which the traditional three-line interlinear glossed text is the best known. The three-line model does, however, not provide a complete model; in Marian Klamer's (2010) *A Grammar of Teiwa,* for example, only 40% of the examples conform to a rigid specification of a three line text with the same number of tokens in the first and second lines. The other 60% deviate structurally from this model in one way or another (subexamples, missing lines, extra lines, missing quotation marks around translations etc.).

For these reasons, we have to specify an *example container*, which provides the numbering and a paragraph where the actual linguistic content can be found. This content can be of varying nature. We have found the following to be recurring types, but there might be more:

- three-liners with interlinear morpheme translation
- two liners with lexeme and gloss
- two liners with tones and lexemes
- two-liners for minimal pairs
- one liners with lexeme<etymology
- ungrammatical one-liners
- ungrammatical one-liners with IMT but no translation
- three liners with lexeme, phonetic transcription, gloss, no translation
- four-liners with orthographic text, phonetic text, IMT, gloss
- four-liners with orthographic text, morphemes, IMT, gloss
- four-liners with intonation, text, IMT, gloss
- ...

We will not provide a full specification for all the subtypes of examples listed here. Schemas for some of those types can be found in Bow et al. (2003).

A linguistic example can occur within a paragraph (as in the first example

14 http://www.sil.org/iso639-3/

below) or between paragraphs.

```
<p>
 English shows subject-verb agreement as
 <examplecontainer n="1">
  <example type="oneliner">
   <exline type="objectlanguage" iso639-3="eng">The dog
     bark*(s)</exline>
  </example>
 </examplecontainer>
 shows. This is also found in French, Spanish, and
 German (see below).
</p>

<p>
 <examplecontainer n="2">
  <example type="threeliner">
   <exline type="objectlanguage" iso639-3="fra">Tu
   regarde-*(s)</exline>
   <exline type="IMT">2s watch-2s</exline>
   <exline type="TRS"> 'You are watching.'</exline>
  </example>
 </examplecontainer>
</p>

<p>
 <examplecontainer n="3">
  <example type="threeliner">
   <exline type="objectlanguage" iso639-3="spa">(Tú)
    mira*(s)</exline>
    ...
  </example>
 </examplecontainer>
</p>

<p>
 <examplecontainer n="4">
  <example type="threeliner">
   <exline type="objectlanguage" iso639-3="deu">Du
    sieh-*(st)</exline>
   ...
  </example>
 </examplecontainer>
</p>
```

Linguistic examples are technically *n*□*m* tables. These can be serialized either horizontally (cells are elements of rows) or vertically (cells are elements of columns). Both solutions have their advantages, and they can be transformed into each other, so that this is a matter of personal taste. Linguistic example tables, like any other table, take up a lot of space when written in XML; therefore they are not formalized here.

4.5 Paradigms

Another element frequently found in linguistic works is the paradigm. This is a special kind of table. Its formalization is quite complex; see Penton et al. (2004).

4.6 References

Linguistic texts contains many references of different types. For references to units within the text, such as examples or other chapters, existing TEI elements <REF> and <PTR> (pointer) can be used. References to items outside the text can be divided into references to the origin of the material (e.g. corpus, dictionary), and references to the literature. References to the literature can be handled with the existing TEI elements <BIBLSTRUCT> and <LISTBIBL>. References to a corpus or a dictionary are encoded as attributes of the element they refer to.

```
<p>
  The <language iso639-3="sci">Sri Lanka Malay
  </language> word <phraseglosspair><phrase iso639-
  3="sci" src="Nordhoff2009">thaanàm</phrase><gloss
  iso639-3="eng">to plant</gloss></phraseglosspair> is
  also found in the <language linguasphere="110424"> Jakarta
  dialect of Indonesian<language> <bibitem
  src="Adelaar1985" isbn="978-0858834088"/>
  This is discussed in more detail in section <ptr
  target="Jakartan Influence" />
</p>
```

4.7 Tables and Figures

Tables and figures can also be handled according to the general TEI-guidelines, which provide the elements <TABLE> and <FIGURE>. A special kind of table found in linguistic descriptions is the phoneme chart. Due to the lengthy nature of table descriptions in XML, this will not be illustrated here.

5 MARKUP PHILOSOPHIES

There are two types of markup commonly used for describing the structure of a text. The examples above have all used ‘inline’ markup, where the metainformation (in the form of markup) about the text is interspersed with the actual content. The alternative is standoff markup, where the content text is unmodified and the markup is written to another file. The chunk of text to which a particular markup applies is indicated by its character offset; for example [8,23) refers to the chunk of content from the 8th to the 23rd character.

The advantages of standoff markup are that one does not need write access to the content file to be described, and the content and its markup do not have to be in the same file or location. The archive can provide a read-only copy, and the annotation can be stored at another location. Multiple, overlapping, or even conflicting annotations can be provided. The drawbacks of standoff annotations are that they are more difficult to read, and they are vulnerable to changes in the content document. Supposing a text includes the Dutch word *Frankrÿk*, which consists of 8 letters, and later it is decided to use the more standard spelling

Frankrijk, which has one more letter. This changes the offset of all following characters, invalidating any standoff markup after this point. Similarly, globally changing the representation of long /a:/ from <*aa*> to <*ā*> (or vice versa) in a document will affect all standoff markup of it. Our final example is more complicated: in Unicode, you can combine characters with others, to yield composite (but single) glyphs[15] like *ṭ'*. The preceding character is clearly a composite. However, common characters like *é* can be represented in two ways: either as *e* + ◌́, or as *é* and the representation may not be discernible through visual inspection of the human-readable text: the displayed glyphs are identical. But the offsets for all following text will be different in each case.

If the file to be marked up is read-only, it is unlikely to change, and the problems just described may not occur. But if the file is to be updated for any reason, standoff is not a viable option. For this paper, we restrict ourselves to inline markup. The advantages of standoff annotation in allowing for overlapping and conflicting annotations are less likely to be important for linguistic examples, which are hardly ever contentious, or overlapping.

6 PROPOSALS

The structure of grammatical descriptions has received detailed treatment in Lehmann (1980, 1989, 1993, 1998, 2004b), Lehmann & Maslova (2004), Lehmann (2004a), Good (2004, 2012), Drude (2012) and Nordhoff (2008, 2012). A basic insight is that the order of elements in a grammatical description is quite free, so that the linear order forced by a book can be dispensed with. There is no particular reason, for example, to treat relative clauses before purposive clauses, verbal morphology before nominal morphology, phonology before morphology, or consonants before vowels, An archivist, will of course always wish to preserve the linear order of materials received, but for use in the Semantic Web, linear order is a burden.

A second insight is that there are two fundamental perspectives to form-meaning relations (von der Gabelentz, 1891; Lehmann & Maslova, 2004; Mosel, 2006; Nordhoff, 2008, 2012): form-to-function and function-to-form. Dissolving the linear order of a book means that the sections can be regrouped into form-to-function part (semasiological part) and a function-to-form part (onomasiological part). This is of course easier to achieve for descriptions written with this insight in mind than for legacy descriptions, where the authors often changes perspective.

6.1 Macrostructure

We can refer to a document's greater elements, their order and their relations to each other as the 'macrostructure' of the document. Within this macrostructure, we can identify, for example, the frontmatter with table of contents, preface, acknowledgments, the backmatter with bibliography, index, appendices etc., and the mainmatter, where the central content of the document resides. Tags such as **<FRONT>**, **<MAIN>**, and **<BACK>** are provided by TEI and can simply be taken over

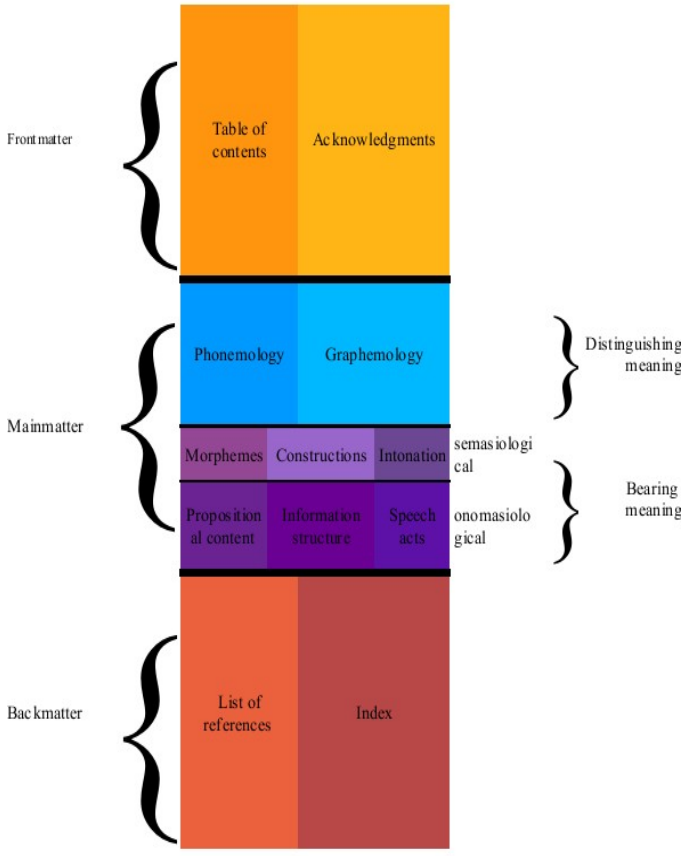
15 A glyph is the visual representation of a single character.

for grammatical descriptions. The markup of frontmatter and the backmatter of grammatical descriptions is not very different from other documents, and is disregarded here.

As for the mainmatter, things are more interesting. We can distinguish background chapters, which treat the location, history, demography, and sociology of the language, from structural chapters dealing with phonology, morphology, syntax etc. For structural chapters, Lehmann & Maslova (2004), following some basic structuralist insights, propose a division into expressive and significative subsytems. The expressive subsystem contains segmental phonology and graphology/orthography. The significative subsystem contains the meaning-bearing items, which can be further approached from a formal (semasiological) and from a functional (onomasiological) point of view. The semasiological component includes various meaning-bearing items: morphemes, constructions, and intonation contours. The onomasiological domain covers various types of meaning: propositional content, discourse structures, and pragmatics. All remaining subdivisions would be language-particular. The general structure can be modeled in a TEI-schema. Note that the linear order of many extant grammatical descriptions often does not coincide with the structure proposed here. For instance, intonation is commonly treated within phonology, whereas here it is seen as a meaning-bearing entity. As such, it is separated from phonemes, which distinguish meaning rather than bear it. An application of the schema proposed here thus requires reorganization of the content of a grammatical description.

An overview of the proposed schema is given as a boxchart in Figure 1.

FIGURE 1
A boxchart of the structure of grammatical descriptions



7 INCORPORATION

We have described some of the advantages of detailed annotation of grammatical descriptions. However, it takes much time to manually annotate them, so to annotate more than just a few significant works we have to use computational techniques. We are testing this using 7500 scanned and OCRed grammatical descriptions to test scalability. The first step is to divide the works into sections. This can be done using pattern matching, for which we are getting quite good results. The recognition of linguistic examples requires more sophisticated pattern matching; the ODIN project[16] has achieved some success in this area. The second step, which we have not started working on yet, is named entity recognition. The third step is an automatic semantic analysis of each section. We are currently working on Latent Semantic Analysis (Deerwester et al. 1990) of the 120,000 sections we extracted and will report our results in future publications.

As far as future - yet unwritten - content is concerned, the application of the

16 http://www.csufresno.edu/odin/

schema will be easier if grammar writers use authoring software that is compatible with the semantics and syntax of the schema. We have had good results with the conversion of documents written in LaTeX and HTML, and using the GALOES grammar authoring platform (Nordhoff 2007a,b,c)[17]. GALOES currently stores text files, but can be made to store DocBook, which will be a very good input format for a schematization process. This is on our agenda for the future. Furthermore, Mihaylov & Beermann (2009), Black & Black (2012), and Maxwell (2012) and discuss other projects which output XML. As this output is already available in the right format, prospects for conversion look good.

A direct link between a semantically structured grammar authoring environment and an archive would mean that the task of the archivist changes from an iteration of an 'acquire-conform-incorporate' process to a more continuous model, where granular pieces of archive material automatically find their place in the archive.

For the production of linguistic knowledge, this approach would compress the traditional cycle of gather-process-condense-publish-archive (cf. Good 2012), so that the primary data (the 'gathered'), the transcription (the 'processed') and the analyses (the 'condensed') can be archived as they are ready without the intermediate stage of book publication. This is thus a movement towards micropublications in the sense of Cysouw (2009).

REFERENCES

Ameka, F., A. Dench & N. Evans (eds.) 2006. *Catching Language – The standing challenge of grammar writing*.

Berners-Lee, T. 2006. "Tim Berners-Lee Date: 2006-07-27". URL http://www.w3.org/DesignIssues/LinkedData.html.

Borthwick, A. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, New York University.

Black, C. A. & Black, H. A. 2012. Grammars for the people, by the people, made easier using PAWS and XLingPaper. In Nordhoff, S. 2012b, 103-28.

Bow, C., B. Hughes & S. Bird 2003). "Towards a general model for interlinear text". *Proceedings of the EMELD Language Digitization Project Conference*. URL http://www.linguistlist.org/emeld/workshop/2003/bowbadenBird-paper.pdf.

Chiarcos, C.; Nordhoff, S. & Hellmann, S. (Eds.). 2012. Linked Data in Linguistics. Representing Language Data and Metadata Heidelberg:Springer.

Cysouw, M. 2009. "Micropublication: footnotes for the 21st Century". Paper presented at the workshop "Small Tools for Cross-Linguistic Research", June 2009, University of Utrecht.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer & R. Harshman 1990. "Indexing by latent semantic analysis". *Journal ofthe American Society for Informotion Science*, 41:391–407.

17 See http://www.galoes.org.

Good, J. 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.
Good, J. 2012. Deconstructing descriptive grammars. In Nordhoff, S. 2012b.
Heath, T. & C. Bizer 2011). *Linked Data - Evolving the Web into a Global Data Space*. San Rafael: Morgan & Claypool.
Klamer, M. 2010. *A Grammar of Teiwa*, *Mouton Grammar Library*, vol. 49. Berlin: Mouton de Gruyter.
Lehmann, C. 1980. “Aufbau einer Grammatik zwischen Sprachtypologie und Universalienforschung”. In Seiler, H., G. Brettschneider & C. Lehmann (eds.) “Wege zur Universalienforschung”, Tübingen: Narr. pp. 29–37.
Lehmann, C. 1989. “Language description and general comparative grammar”. In Graustein, G. & G. Leitner (eds.) “Reference grammars and modern linguistic theory”, Tübingen: M. Niemeyer. pp. 133–162.
Lehmann, C. 1993. *On the system of semasiological grammar*, *Allgemein-Vergleichende Grammatik*, vol. 1. Bielefeld: Universität Bielefeld, Universität München.
Lehmann, C. 1998. “Ein Strukturrahmen für deskriptive Grammatiken”. In Zaefferer 1998, pp. 39–52.
Lehmann, C. 2004a. “Documentation of grammar”. In Sakiyama, O., F. Endo, H. Watanabe & F. Sasama (eds.) “Lectures on endangered languages: 4. From Kyoto Conference 2001”, Osaka: Osaka Gakuin University. pp. 61–74.
Lehmann, C. 2004b. “Funktionale Grammatikographie”. In Premper, W. (ed.) “Dimensionen und Kontinua. Beiträge zu Hansjakob Seilers Universalienforschung.”, Bochum: N. Brockmeyer. pp. 147–165.
Lehmann, C. & E. Maslova 2004. “Grammaticography”. In Booij, G., C. Lehmann, J. Mugdan & S. Skopeteas (eds.) “Morphologie. Ein Handbuch zur Flexion und Wortbildung”, , vol. 2Berlin, New York: de Gruyter.
Maxwell, Michael. 2012. Electronic Grammars and Reproducible Research. In Nordhoff, S. 2012b. 207-234.
Mihaylov, P. & Beermann, D. 2009. "TypeCraft: Linguistic data and knowledge sharing.Open Access and linguistic methodology.” Presentation at the workshop Small Tools in cross-linguistic Research. University of Utrecht. The Netherlands. June 2009.
Mosel, U. 2006. “Grammaticography: The art and craft of writing grammars”. In Ameka et al. 2006).
Nordhoff, S. (2007a). “The grammar authoring system GALOES”. Paper presented at the workshop “Wikifying research” at the MPI Leipzig.
Nordhoff, S. (2007b). “Grammar writing in the Electronic Age”. Paper presented at the ALT VII conference in Paris.
Nordhoff, S. (2007c). “Growing a grammar with GALOES”. Paper presented at the Dobes workshop at the MPI Nijmegen.
Nordhoff, S. 2008. “Electronic reference grammars for typology – challenges and solutions”. *Journal for Language Documentation and Conservation*, 22):296–324.
Nordhoff, S. 2012a. “The grammatical description as a collection of form-meaning

pairs". In Nordhoff, S. (2012b). 33-62.
Nordhoff, S. (ed). 2012. Electronic Grammaticography, Manoa: University of Hawai'i Press.
Penton, D., C. Bow, S. Bird & B. Hughes 2004. "Towards a general model for linguistic paradigms". *Proceedings of EMELD 2004*.
Sperberg-McQueen, C. M. & L. Burnard 2010. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford, Providence, Charlottesville, Nancy: TEI Consortium.
von der Gabelentz, G. 1891. "Die Sprachwissenschaft. Ihre Aufgaben, Methoden und bisherigen Ergebnisse". Leipzig.
Zaefferer, D. (ed.) 1998. *Deskriptive Grammatik und allgemeiner Sprachvergleich*. Tübingen: Niemeyer.