Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 12: Special Issue on Language Documentation and Archiving*. Editors: David Nathan & Peter K. Austin

Archiving grammatical descriptions

SEBASTIAN NORDHOFF, HARALD HAMMARSTRÖM

Archiving grammatical descriptions

Sebastian Nordhoff and Harald Hammarström

Max Planck Institute for Evolutionary Anthropology, Leipzig

1. Introduction

Language documentation projects produce and collect audio, video, and textual data, which they usually deposit in archives. Documenters' understanding of best practices in archiving the primary content of their domain has made considerable progress over recent years. Methods for archiving derived content, such as dictionaries and especially grammatical descriptions, have received less attention. In this paper, we explore what the goals of archiving grammatical descriptions are, and what tasks an archive has to fulfill. We first discuss a number of parameters which help us to classify archives with regard to the objects they host and the roles they play in their user community. We argue that the text of grammatical descriptions should be archived in a fashion which allows retrieval of individual elements such as sections, paragraphs, and examples, and that for this to work, grammatical descriptions have to be provided with semantic markup. We discuss the Text Encoding Initiative (TEI), originally a philological enterprise, and the TEI tools which are useful for this purpose. Grammatical descriptions contain a number of elements which are not yet found in TEI, and we identify and describe these. We then discuss how annotation of both legacy and future grammatical descriptions can be accomplished, and report on some preliminary work on this.

2. Archives

There are several aspects to archiving; their importance varies according to the discipline which an archive serves. Archives originally dealt with *physical objects* (e.g. vases, axes, books), but in recent years there has been a move towards digital archiving, where archive collections consist of *representations of objects* as digital surrogates of physical artifacts (e.g. 3D models of vases or axes, or scans of books). Additionally, there is archiving of *content*, such as the text of documents, which is concerned with the symbols that make up a document's content rather than the document's physical form (e.g. paper type or size, or formatting and layout).

Sebastian Nordhoff and Harald Hammarström 2013. Archiving grammatical descriptions. In David Nathan and Peter K. Austin (eds.) *Language Documentation and Description*, *Vol 12,* 164-186. London: SOAS.

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

- a printed book (the *physical object*). Here, archives need to take care of environmental conditions such as humidity, temperature, exposure to sunlight etc.
- a set of scans of a printed book, in the form of digital image files (e.g. TIFF files) representing each page. This is a *representation*, or surrogate, of the physical object. Non-visual information, such as the texture or smell of the book, is not captured
- a file or files containing the sequence of characters which make up the *content* of the book. Here, visual information is also lost (typography, color, page layout etc.), although some can be retained as meta-information.[1]

For grammatical descriptions, archiving the content is clearly the most important aspect. While there are some aesthetic masterpieces in the grammatical literature, grammars produced in language documentation normally excel by virtue of their content rather than their typography, layout, or paper choice. What is to be preserved for future generations is a grammar's content, not the physical arrangement of letters on a page. This is even more the case for a 'born-digital' document, where the physical object does not exist until the user prints the document. References to some visual characteristics of the document can also be stored as data, but these are not central.

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

- without internal structure, that is, as a string of characters; or
- as a set of structural elements which constitute the text (headings, sections, footnotes, cross-references etc., cf. Gippert 2006)

The second, more granular, approach allows for more accurate search for particular items and easier modification of the text. It also allows for better integration into the Semantic Web, a framework for making explicit the nature of the links between pieces of information (Berners-Lee et al. 2001, Shadbolt

[1] Combinations such as image plus text are also possible. This is, for instance, the case for scans with optical character recognition (OCR).

et al. 2006, Auer and Hellmann 2012). For instance, a web page about a novel could link to another web page which has information about the novel's author. This is of course already commonly done, but standard web-authoring (HTML) does not allow one to distinguish a link to an author from a link to a place, or a link to anything else. In the Semantic Web, the relationships between these elements would be explicit, thereby allowing users to find and combine information more easily and accurately.

In the remainder of this paper, we advocate such a granular and semantically-driven approach for archiving grammatical descriptions. We first give an illustrative example and later, following a suggestion by Gippert (2006), relate the needs of linguistics to work which has been done in the framework of the Text Encoding Initiative (TEI; Sperberg-McQueen and Burnard 2010), which has been more concerned with philological questions.

With regard to archives, we can make two further distinctions: *orientation* and *perfectivity*. The first axis, orientation, refers to whether the focus of the archive is inward or outward: is it more important to get materials into the archive (i.e. to store materials), or is it more important to get materials out of the archive to interested users (i.e. to serve users)? An example of an inward-oriented archive would be a seedbank which stores seeds of (endangered) plants for future scientific research. The integrity of such an archive is primary, and access is typically restricted. An extreme example of an outward-oriented archive is a public library: its main goal is to get the content out to the public, even if occasionally a book gets lost or damaged. 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* by analogy with the term used in discussing linguistic aspect, refers to the state of completeness sought by an archive. A 'perfective' archive only accepts finished documents, and does not allow modification of already archived documents. Such an archive is, so to speak, read-only, and applies for example to libraries and archives of printed books. A non-perfective archive will store documents in varying states of completion and allow modification of archived content (while keeping version histories to track the evolution of content). An example of this type would be *Living Reviews*, a website of review articles for selected disciplines (Relativity, Solar Physics, Computational Astrophysics, European Governance, Landscape Research, Democracy) where articles are updated as understanding of each field progresses.[2]

[2] www.livingreviews.org

Various authors have observed that a grammatical description is never finished (cf. Payne 2006: 369ff., Weber 2006a: 418, Rice 2006: 396, Cristofaro 2006: 139). Requirements for grammars to be 'finished' have led to delays in publication because authors know they do not get a chance to correct errors after the moment of publication. These facts suggest that archives could benefit from an 'imperfective' approach for storing grammars which allows for ongoing corrections and modifications.

Grammatical descriptions, especially in digital form, are precious resources, but they present little challenge as far as storage is concerned. While fossils, taxidermic specimens or ceramics require the archivist to strike a balance between facilitating public access and minimizing the risk of damage to the objects, this is not the case for digital resources, which can be copied perfectly and disseminated without limit. Therefore, an archive of grammatical descriptions can be outward-oriented.

For the remainder of this paper, we assume that grammatical descriptions should be held in outward-oriented, non-perfective archives.

3. The nature of grammatical descriptions

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. a reader could skip phonology to start with syntax, or jump back from any place to morphology, and/or use the table of contents or index to navigate etc.). Cysouw (2009) argues for the use of atomic linguistic facts as the basis of linguistic knowledge, where facts are expressed as 'micro-publications', very short statements about a linguistic fact for a particular language. Pulling these three authors' ideas together, we can propose a model for a grammatical description: a *non-linear meta-database of micropublications*.

This model recalls the idea of granular representation mentioned above. A granular approach would reflect the model, as content can be added or modified on a local (micropublication) basis, whereas in the conventional approach any modification is tantamount to a global modification of the whole description. The granular approach also allows easier retrieval of and access to particular chunks of information. Finally, a granular approach allows unique identification of chunks and reference to them in the Semantic Web.

4. Granular text

Let us illustrate what we mean by granular representation with an example expressed in extensible markup language (XML) format:

Figure 1: Granular text

```
<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>
      <translation> ... </translation>
    </lgex>
  </div>
</div>
```

The example is in XML, a document markup format which is used to describe the semantics of the text content. XML uses tags (which are delimited by the characters ‘<’ and ‘>’) to indicate where a particular element starts and ends. Furthermore, tags can have attributes which in turn can have values – for example, the tag `<div id="ch3" type="chapter" n="3">` is a ‘div’ tag which has attributes id, type and n, which have the respective values ‘ch3’, ‘chapter’, and ‘3’.

The visual properties of the semantic units – fonts, colors, layout etc. – are not specified in the document but when rendering the document in print or on screen, these can be defined in a style sheet which provides a visual representation for semantic units. For instance, `<head>Nominal morphology</head>` could be specified to be bold, size 14 point, and centred on a separate line.

Note these points:

1. important elements such as headings (`<head>`) and examples (`<lgex>`) are explicitly tagged using XML elements

2. there are unique references to other paragraphs and examples (`id=...`)[3]

3. semantic markup includes references to term definitions. For example, `<gloss type="Leipzig">ACC</gloss>` refers to the Leipzig Glossing Rules[4], and another reference is made in the example to the GOLD ontology.[5] These references allow readers to look up the terms in a central place and can help establish a shared vocabulary across grammars (compare the discussion by Trilsbeek and König (this volume) of the use of ISOCat as a shared vocabulary for metadata descriptions).

Such markup combined with unique references allows for integration into the Semantic Web. A user could retrieve all sections of a given grammatical description which refer to the concept ‘case marker’ as defined in the GOLD ontology, even if a different term is used in the particular grammatical description. Note that the link to GOLD is mainly useful for *retrieving* the section in question for further inspection by a human reader. While Zaefferer (2006) suggests doing automated reasoning across grammars using ontologies like GOLD, the problem of cross-linguistic categories, and how they can be established and mapped together, is, however, a very difficult one (Haspelmath 2007, 2010), and we would advise against using GOLD for *automated reasoning* (see Nordhoff 2012 for further discussion). Even in the absence of hard and fast cross-linguistic categories, GOLD can improve discoverability of relevant sections of a grammatical description and suggest further reading in other grammatical descriptions. A 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:

[3] *Unique* is used in its computer science meaning: every reference refers to one and only one referent. In other words, a unique reference is unambiguous and clearly identifies one other element in the text.

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

[5] linguistics-ontology.org

I. the *arguments* of the linked relation must be identifiable; they must have Uniform Resource Identifiers (URIs). This Uniform Resource Identifier uniquely identifies a web resource. URIs can, for the purposes of this paper, be equated with web addresses in a certain format

II. the *relation* must be defined. Ideally, one uses relations already defined in widely-used vocabularies such as RDFS,[6] Dublin Core,[7] SKOS,[8] GOLD, and lexvo[9]

III. the *formalism* to link the relation and the arguments must be established. The formalism for linking these together is the Resource Description Framework (RDF),[10] here represented in a variant called N3.[11]

An example will show the general approach:

```
<grammararchive:123/chapter/4/section/5> <rdfs:seeAlso> <gold:Infix>
```

RDF predications are written in a Subject-Verb-Object 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 what we call *dereferenceable*, which means that a definition of what they mean can be looked up on the internet.[12]

[6] www.w3.org/TR/rdf-schema

[7] dublincore.org provides terms for metadata about documents.

[8] www.w3.org/2004/02/skos provides terms for describing concepts and their relation.

[9] www.lexvo.org provides terms for describing language names and script names.

[10] www.w3.org/RDF

[11] www.w3.org/DesignIssues/Notation3.html

[12] The actual Internet addresses have been abbreviated here; one would look up:
www.grammararchive.org/grammar/123/chapter/4/section/5
www.w3.org/2000/01/rdf-schema#seeAlso
linguistics-ontology.org/gold/2010/Infix

The website www.grammararchive.org was created by us and has about 450 grammars from the 19th century whose copyright has expired.

Another example is:

```
<grammararchive:123/chapter/4/section/5#example6> <dublincore:references>
<grammararchive:987/chapter/6/section/5> .
```

In this example, we use the Dublin Core ontology[13] to express metadata about documents. Dublin Core provides the predicate ‘references’, which we use here to assert that the first work references the second. This use of a common ontology is considered to be a best practice for assuring interoperability between resources. For instance, the World Atlas of Linguistic Structures (WALS[14]) and Glottolog[15] also use Dublin Core to represent their metadata.

5. Textual elements in linguistics

In order to create a schema for representing grammatical descriptions, one needs to take stock of the elements which are found in this type of work. Since the 1980s, the *Text Encoding Initiative* (TEI) has worked on schemas for representing the content of texts, mainly in the humanities. TEI is based on the idea that texts consist of recurring elements, which can be labeled. To exemplify, we can mark up a poem using the tags 〈l〉 for *line* and <lg> for *line group*.[16] Note that line groups can be of different types, and that they can be nested.

```
<lg type="stanza">
 <lg type="sestet">
  <l>In the first year of Freedom's second dawn</l>
  <l>Died George the Third; although no tyrant, one</l>
  <l>Who shielded tyrants, till each sense withdrawn</l>
  <l>Left him nor mental nor external sun:</l>
  <l>A better farmer ne'er brushed dew from lawn,</l>
  <l>A worse king never left a realm undone!</l>
 </lg>
 <lg type="couplet">
  <l>He died — but left his subjects still behind,</l>
  <l>One half as mad — and t'other no less blind.</l>
 </lg>
</lg>
```

[13] dublincore.org

[14] www.wals.info

[15] www.glottolog.org

[16] Example from www.tei-c.org/release/doc/tei-p5-doc/en/html/VE.html

The Text Encoding Initiative provides markup (tag) vocabularies for various domains of the humanities, and also a mechanism for creating specialized schemas for new domains. When creating new schemas, one should strive to use existing vocabularies to the extent possible.

As far as grammatical descriptions are concerned, we find a variety of recurrent elements. Some are common to other types of texts, such as paragraphs, headings, and cross-references. For those, existing TEI vocabulary can be used. In linguistic texts, we also deal with a number of elements not listed in any TEI schema. These will be discussed below, as will be some special TEI elements which can be adopted for grammatical descriptions.

5.1 Named entities

A named entity is a term which refers to a defined thing such as a country, organization, or person, as well as a language, book or linguistic concept. 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 pnd="118611046">Hugo Schuchardt</person> and
   <person pnd="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>
```

Note that the example above adds additional information to the basic tags. For instance, ISO-3166-1 is a standard for country names. This is used in the tag `<country>` to identify the country India, even though the text in this particular instance is *Indian*. ISO-3166-2 identifies subdivisions of countries, states and territories. *IN-DD* identifies Daman and Diu. ISO 639 is the ISO standard for language names, and ISBN is of course used for books. PND stands for *Personennormdatei* (English: *Person Authority File*), and is used by German libraries to uniquely identify authors. In that standard, Hugo Schuchardt has the identifier 118611046 and Sebastião Dalgado has the identifier 115161023.

The process of identifying named entities in a given text (Named Entity Recognition) is an important subfield of computational text linguistics (e.g.

Borthwick, 1999). Linguistic concepts can be treated as named entities if they are defined outside the text, for example in the ISO register,[17] lexvo, or the GOLD ontology.

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

5.2 Object language

Linguistic texts frequently use words written in languages which are not the language of the main text. These words are commonly typeset in italics. A semantic representation would be as follows.

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

The attribute iso639-3 refers to the ISO 639-3 code of the language. In case ISO 639 codes are not available, Glottolog codes can be used, as these cover over 20,000 ‘languoids’ (Nordhoff et al. 2013).

5.3 Phrase-gloss pairs

Grammatical descriptions often provide object language terms immediately followed by translations:[18]

```
<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>
```

[17] www.sil.org/iso639-3

[18] This element is only possible if the elements are adjacent.

5.4 Linguistic examples

The most salient text element found in grammars is the example sentence, of which the traditional three-line interlinear glossed text is the best known (Bow et al. 2003). The three-line structure, however, does not provide a complete model; in LaPolla's (2003) *A Grammar of Qiang,* for example, only 60% 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 remainder deviate from this model in various ways, e.g. they are subexamples or have missing lines, extra lines, or other variations; some are actually not examples, but lists of people, regions, tables, or other content).

In order to accommodate varying content, within a defined category 'Example', we specify an *example container*, which provides numbering and a paragraph where the actual linguistic content can be found. We have found the following recurring types, but there might be more:

- three-liners with interlinear morpheme translation (IMT)
- two-liners with lexeme and gloss
- two-liners with tones and lexemes
- two-liners for minimal pairs
- one-liners with lexeme<etymology[19]
- ungrammatical one-liners
- ungrammatical two-liners with IMT but no free translation[20]
- three-liners with lexeme, phonetic transcription, gloss, no free 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 types and subtypes of examples listed here. Schemas for some of those types can be found in Bow et al. (2003).

[19] For example, from Davies (2010: 128):
sorop are 'sunset' < *sorop* 'enter' + *are* 'sun', with no following lines

[20] For example, from Epps (2008: 430):
*ʔãh ʔey-tɔʔɔ́h-ɔ́y
1SG call-run-DYNM

A linguistic example can occur within a paragraph (as in the first example 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 containing word/gloss pairs are elements of rows, see Figure 2) or vertically (cells containing words and glosses are elements of implicit columns, see Figure 3). Both solutions have their advantages, and they can be transformed into each other, so the choice is a matter of personal taste.

Figure 2: Representing a linguistic example as rows of word/gloss pairs

Word1	Word2	Word3	…	…	…	…	…	…
Gloss1	Gloss2	…	…	…	…	…	…	…
Translation								

```
<table>
  <tr>
        <td>
          <imtblock>
                <word>Word1</word>
                <gloss>Gloss1</gloss>
          </imtblock>
        </td>
        <td>
          …
        </td>
  </tr>
  <tr>
        <translation> … </translation>
  </tr>
</table>
```

Figure 3: Representing a linguistic example as rows and columns

Word1	Word2	Word3	…	…	…	…	…	…
Gloss1	Gloss2	…	…	…	…	…	…	…
Translation								

```
<table>
  <tr type="sourceline">
        <word>Word1</word>
        <word>Word2</word>
        <word>Word3</word>
        …
  </tr>
  <tr type="imtline">
        <gloss>Gloss1</gloss>
        <gloss>Gloss2</gloss>
        <gloss>…</gloss>
        …
  </tr>
  <tr type="translationline">
        <translation> … </translation>
  </tr>
</table>
```

5.5 References

Linguistic texts can contain references of various types. For references to units within the text, such as to examples or to other chapters, the existing TEI elements <REF> and <PTR> (pointer) can be used.[21] References to items outside the text can be divided into references to the origin of the material (e.g. a corpus, dictionary), or references to academic literature. References to a corpus or a dictionary are formulated as attributes of the element they refer to. References to academic literature can be handled with the existing TEI elements <BIBLSTRUCT> and <LISTBIBL>.

```
<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>
```

5.6 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 representation in XML, this will not be illustrated here.

[21] See www.tei-c.org/release/doc/tei-p5-doc/en/html for TEI elements.

6. Proposals

The structure of grammatical descriptions has received detailed treatment in Lehmann (1980, 1989, 1993, 1998, 2004a, 2004b), Lehmann and Maslova (2004), 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 printed book can be dispensed with. There is no reason, for example, to treat relative clauses before purposive clauses, verbal morphology before nominal morphology, phonology before morphology, or consonants before vowels. Of course, for pedagogical reasons, it is often useful to proceed in conventional order: before reading about diphthongs, it might be good to have a basic knowledge of the vowels of the language, and complex clauses should follow simple clauses. However, there are many cases where the dependency is mutual: in order to understand stress assignment, one has to understand syllable structure, but in order to understand syllable structure, the notion of stress is important. In order to understand a certain type of split alignment system (where case marking is determined by clausal tense or aspect), one has to know about tense and aspect, but in order to understand the examples for tense and aspect, one has to be acquainted with the alignment system and so on. In these cases, there is no obvious order for arranging the content.[22]

A second insight is that there are two fundamental perspectives to form-meaning relations (von der Gabelentz 1891; Lehmann and Maslova 2004; Mosel 2006; Nordhoff 2008, 2012): *form-to-function* and *function-to-form*. Dissolving the linear order of a book means that it can be alternatively viewed as a form-to-function (*semasiological*) arrangement or a function-to-form (*onomasiological*) arrangement. This is of course easier to achieve for grammatical descriptions where these perspectives are consistently handled.[23]

[22] For archivists, of course, it is desirable to preserve materials retaining their original structure and linear order, but this is not a consideration for the Semantic Web.

[23] For instance, sections on verbal morphology often contain passages about periphrases (e.g. perfect tense) or serial verbs. These constructions encode content that would otherwise often be dealt with by morphology. However, these constructions are not morphological: a shift in perspective from semasiological (‘What are the verbal affixes for?’) to onomasiological (‘How are tense and aspect expressed in this language?’) has taken place.

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 (cf. Gibbon 2000 for an example from lexicography). Within a document's macrostructure, we can identify, for example:

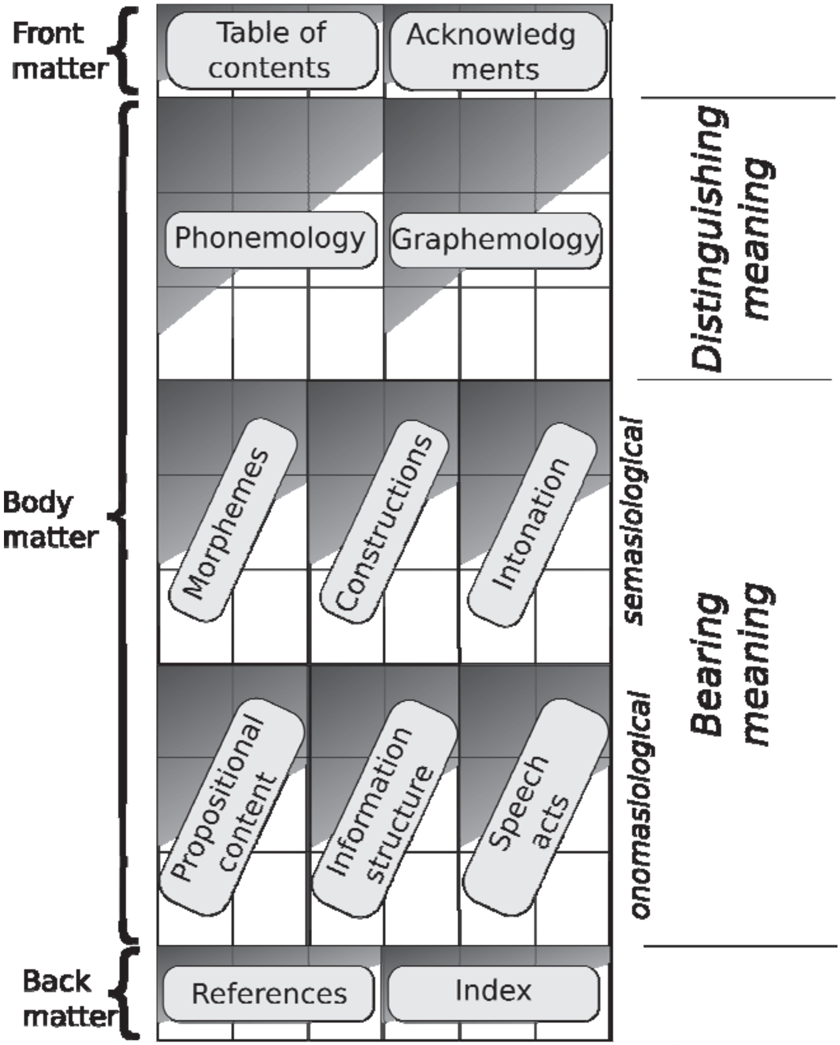
- the front matter, with table of contents, preface, acknowledgments
- the body matter, where the main content resides
- the back matter, with bibliography, index, appendices etc.

Tags such as `<FRONT>`, `<BODY>`, and `<BACK>` are provided by TEI and can simply be adopted for grammatical descriptions. Front and back matter of grammatical descriptions are not very different from other documents, and are not treated here.

As for the body matter, 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 and Maslova (2004), following some basic structuralist principles, propose a division into *expressive* and *significative* subsystems. The expressive subsystem contains segmental phonology and graphology/orthography. The significative subsystem contains the meaning-bearing items, which can be further approached from formal (semasiological) and functional (onomasiological) viewpoints. The semasiological component includes various meaning-bearing items such as 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-specific. The general structure can be modeled using a TEI schema. Note that the structure and order of many existing grammatical descriptions often does not coincide with the structure proposed here. For instance, intonation is commonly treated within phonology, whereas here it is classed as a meaning-bearing entity. As such, it is treated separately 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 box chart in Figure 4.

Figure 4: Box chart 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, and to annotate more than just a few works, computational techniques would need to be used. We are testing the scalability of computational techniques using 7,500 scanned and character-recognised grammatical descriptions. This corpus consists of about 5,000 grammars written in English and about 2,800 written in other languages (Bulgarian, Dutch, French, German, Indonesian, Italian, Japanese, Portuguese, Russian, Spanish, Swedish).

The first step is to divide each grammar into sections. This can be done using pattern matching, based on the fact that grammars normally use one of two patterns for section titles:

- digits combined with fullstops, e.g. *3.1.2. Some Title*
- a structural term followed by digits, e.g (Chapter|Section|Kapitel|Chapitre|...) [123456789]

When we try to split these grammars into manageable chunks, we arrive at the following results:

Figure 5: Evaluation of recognition process for sections in 7500 grammatical descriptions in English and other languages.

Column1	Column2	Column3	Column4	Column5	Column6	Column7
	in English		in other lg		Total	
Total files	5006		2839		7845	
Good files	1951	39.0%	618	21.8%	2569	32.7%
Bad files (total)	3055	61.0%	2221	78.2%	5276	67.3%
No matches	1812	36.2%	1321	46.5%	3133	39.9%
Too granular	171	3.4%	62	2.2%	233	3.0%
Not granular enough	1072	21.4%	838	29.5%	1910	24.3%
Chunks yielded	122978		20478		143456	
	63.0		33.1		55.8	

About a third of the grammar can be split into chunks using the patterns mentioned above. The chunks are stored as text files for further processing. More than 60% of the grammars do not yield satisfying results. Some have no matches at all for the patterns. Others yield too many sections (several per page), and finally there are some which do yield sections, but the length of the retrieved sections is not typical for a grammatical description. For the purposes of this calculation, we set the lower bound for acceptable average section length to 1,000 characters (slightly less than one page), and the upper bound to 10,000 characters (roughly 7 pages).

Effective recognition of linguistic examples requires more sophisticated pattern matching. The ODIN project[24] has achieved some success in this area.[25]

The second step, which we have not started working on yet, is named entity recognition. The third step is automatic semantic analysis of each section. We started working on Latent Semantic Analysis (Deerwester et al. 1990) of the 120,000 sections we extracted from English grammars with the aim of classifying the documents. However, due to the heterogeneous nature of the documents, the calculation became too complex for the hardware we had at hand. We then switched to Random Indexing (Karneva et al. 2000), but have not yet arrived at a successful classification.

As far as future and yet unwritten content is concerned, the application of the schema we are developing 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 with documents composed using the GALOES grammar authoring platform (Nordhoff 2007a,b,c).[26] GALOES currently stores text files, and can be made to store documents in DocBook format,[27] which is a very good input format for a schematization process. Beermann and Mihaylov (2009), Black and Black (2012), and Maxwell (2012) discuss other projects which output XML, for which prospects for conversion look good. Finally, publishing houses such as Mouton De Gruyter are moving towards an XML-first workflow, which requires that content is available as XML even as it enters the production cycle. Nordhoff is currently designing an XML DTD for grammatical descriptions to which forthcoming books in the *Mouton Grammar Library* series will conform.

[24] www.csufresno.edu/odin

[25] Recognition and markup of examples is also relevant for other types of documents, e.g. corpora. This is, however, outside the scope of this paper.

[26] See www.galoes.org

[27] For DocBook format, see www.docbook.org

8. Conclusion: implications for archives

An architecture is emerging that could support direct linkage between a structured grammar authoring tool and an archive. Such an architecture would transform the archivist's task from an iterative 'acquire-conform-incorporate' process to that of overseeing a more continuous process where granular chunks of material automatically find their place in the archive. The authoring tool would go beyond the annotation of examples (as is currently done for instance with Toolbox, FLEx or ELAN software) to also provide assistance for composing textual elements. Semantic annotation of these components would also enhance the discovery and harvesting of them by other projects. Use of persistent URIs furthermore enables third party researchers to enrich the data with additional annotations.

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 transcriptions (the 'processed') and the analyses (the 'condensed') can be archived as they become ready, without requiring an intermediate stage of book publication. This thus represents a movement towards micropublications in the sense of Cysouw (2009).

In addition to primary linguistic material, archives should also store derived material including grammatical descriptions. These grammatical descriptions should be stored in an archive which is outward-oriented and imperfective. The archive should enable grammatical descriptions to be accessed in a granular fashion, allowing sections, paragraphs and examples to be retrieved individually. Further, the use of semantic markup, building upon the Text Encoding Initiative's efforts, will allow for more effective querying, discoverability and harvesting and allow language descriptions to join the Semantic Web.

References

Ameka, Felix, Alan Dench and Nicholas Evans (eds.) 2006. *Catching language – The standing challenge of grammar writing*. Berlin: Mouton de Gruyter.

Auer, Sören and Sebastian Hellmann. 2012. The web of data: Decentralized, collaborative, interlinked and interoperable. *LREC 2012*. www.lrec-conf.org/proceedings/lrec2012/keynotes/LREC%202012. Keynote %20Speech%201.Soeren%20Auer.pdf

Beermann, Dorothee and Pavel Mihaylov. 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.

Berners-Lee, Tim, James Hendler and Ora Lassila. 2001. The Semantic Web. *Scientific American* (284), 34-43

Borthwick, Andrew. 1999. *A maximum entropy approach to named entity recognition*. Ph.D. thesis, New York University.

Black, Cheryl A. and H. Andrew Black. 2012. Grammars for the people, by the people, made easier using PAWS and XLingPaper. In Sebastian Nordhoff (ed.) *Electronic Grammaticography*, 103-128. Manoa: University of Hawai'i Press.

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

Cristofaro, Sonia. 2006. The organization of reference grammars: A typologist user's point of view. In Felix Ameka, Alan Dench and Nicholas Evans (eds.) *Catching language – The standing challenge of grammar writing*, 137-170. Berlin: Mouton de Gruyter.

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

Davies, William. 2010. *A Grammar of Madurese*. Berlin: Mouton de Gruyter.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41: 391–407.

Drude, Sebastian. 2012. Digital Grammars – Integrating the Wiki/CMS approach with Language Archiving Technology and TEI. In Sebastian Nordhoff (ed.) *Electronic Grammaticography*, 160-178. Manoa: University of Hawai'i Press.

Epps, Patience. 2008. *A Grammar of Hup*. Berlin: Mouton de Gruyter.

Gibbon, Dafydd. 2000. On Lexical objects and their properties. Paper presented at the workshop on Web-Based Language Documentation and Description, December 2000, Philadelphia, USA

Gippert, Jost. 2006. Linguistic documentation and the encoding of textual materials. In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (eds.) *Essentials of language documentation*, 337-362. Berlin: Mouton de Gruyter.

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

Good, Jeff. 2012. Deconstructing descriptive grammars. In Sebastian Nordhoff (ed.) *Electronic Grammaticography*, 2-32. Manoa: University of Hawai'i Press.

Haspelmath, Martin. 2007. Pre-established categories don't exist: consequences for language description and typology. *Linguistic Typology,* 11(1), 119-132

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language*, 86(3), 663-687.

Kanerva, Pentti, Jan Kristoferson and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 1036. Mahwah, New Jersey: Laurence Erlbaum.

LaPolla, Randy. 2003. *A Grammar of Qiang*. Berlin: Mouton de Gruyter.

Lehmann, Christian. 1980. Aufbau einer Grammatik zwischen Sprachtypologie und Universalienforschung. In Hansjakob Seiler, Gunter Brettschneider and Christian Lehmann (eds.) *Wege zur Universalienforschung*, 29-37. Tübingen: Narr.

Lehmann, Christian. 1989. Language description and general comparative grammar. In Gottfried Graustein and Gerhard Leitner (eds.) *Reference grammars and modern linguistic theory*, 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 Dietmar Zaefferer. (ed.) *Deskriptive Grammatik und allgemeiner Sprachvergleich*, 39-52. Tübingen: Niemeyer.

Lehmann, Christian. 2004a. Documentation of grammar. In Osamu Sakiyama, Fubito Endo, Honoré Watanabe and Fumiko Sasama (eds.) *Lectures on endangered languages: 4. From Kyoto Conference 2001*, 61-74. Osaka: Osaka Gakuin University.

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

Lehmann, Christian and Elena Maslova. 2004. Grammaticography. In Geert Booij, Christian Lehmann, Joachim Mugdan and Stavros Skopeteas (eds.) *Morphologie. Ein Handbuch zur Flexion und Wortbildung*, vol. 2, 1857-1882. Berlin, New York: de Gruyter.

Maxwell, Michael. 2012. Electronic grammars and reproducible research. In Sebastian Nordhoff (ed.) *Electronic Grammaticography*, 207-234. Manoa: University of Hawai'i Press.

Mosel, Ulrike. 2006. Grammaticography: The art and craft of writing grammars. In Ameka, Felix, Alan Dench and Nicholas Evans (eds.) *Catching language – The standing challenge of grammar writing*, 41-68. Berlin: Mouton de Gruyter

Nordhoff, Sebastian. 2007a. The grammar authoring system GALOES. Paper presented at the workshop *Wikifying research* at the MPI Leipzig.

Nordhoff, Sebastian. 2007b. Grammar writing in the electronic age. Paper presented at the ALT VII conference in Paris.

Nordhoff, Sebastian. 2007c. Growing a grammar with GALOES. Paper presented at the Dobes workshop, MPI Nijmegen.

Nordhoff, Sebastian. 2008. Electronic reference grammars for typology – challenges and solutions. *Language Documentation and Conservation*, 22, 296–324.

Nordhoff, Sebastian. 2012. The grammatical description as a collection of form-meaning pairs. In Sebastian Nordhoff (ed.) *Electronic Grammaticography*, 33-62. Manoa: University of Hawai'i Press.

Nordhoff, Sebastian and Harald Hammarström, Robert Forkel, and Martin Haspelmath (eds.) 2013. Glottolog 2.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at glottolog.org [accessed 2013-08-25]

Payne, Thomas. 2006. A grammar as a communicative act, or what does a grammatical description really describe? *Studies in Language*, 30(2), 367-383.

Rice, Keren. 2006. A typology of good grammars. *Studies in Language*, 30(2), 385-415.

Shadbolt, Nigel, Tim Berners-Lee and Wendy Hall. 2006. The semantic web revisited. *Intelligent Systems* 21, 96-101

Sperberg-McQueen, C. Michael and Lou Burnard. 2010. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford: TEI Consortium.

von der Gabelentz, Georg. 1891. *Die Sprachwissenschaft. Ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig.

Weber, David. 2006a. Thoughts on growing a grammar. *Studies in Language*, 30(2), 417-444.

Zaefferer, Dietmar (ed.) 1998. *Deskriptive Grammatik und allgemeiner Sprachvergleich*. Tübingen: Niemeyer.