Archiving Grammatical Descriptions

O. INTRODUCTION

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

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, color, typography etc.), although some can be recorded as meta-information.[1]

---

1 Combinations like image + text are also possible. This is for instance the case

For grammatical description, archiving content is clearly the most important aspect. While there are some masterpieces of editing in the grammatical literature, the grammars produced in language documentation normally excel 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., cf. Gippert 2006)

The second, more granular, approach allows for more accurate search for particular items, easier updates, and better integration into the Semantic Web. The Semantic Web (Berners-Lee et al. 2001, Shadbolt et al. 2006, Auer & Hellmann 2012) is a framework to make information available on the Internet and the links between the pieces of information explicit. For instance, a page about a novel could link to a page about its author. This is of course already commonly done in web pages, but standard HTML does not allow to distinguish a link to an author from a link to a place or a link to a country. In the Semantic Web, the relations between this elements would be explicit, thereby allowing users to find and combine information more easily.

In the remainder of this paper, we advocate such a granular approach to archiving grammatical descriptions. We will first give an illustratory example and later, following upon 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 & Burnard 2010), which is more concerned with philological questions.

With regard to archives, we can make two further distinctions: 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)? An example of an inward-oriented archive would be a seedbank which stores seeds of (endangered) plants for future scientific research. Integrity of the archive is primordial, and access typically restricted. An example of an outward oriented archive would be a public library. The 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 in analogy to the the term used for linguistic aspect, refers to the

---

for OCRed scans.

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. This is the case for archives of printed books for instance. 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). An example would be *Living Reviews*.[2] Living Reviews is a portal which contains review articles for selected disciplines (Relativity, Solar Physics, Computational Astrophysics, European Governance, Landscape Research, Democracy). These articles are updated as understanding of the field progresses.

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 "finished" grammars have led to a) a delay in publication as the authors knew they would not get a chance to correct errors and b) uncorrected errors persisting in the book as updates are not possible. These facts suggest that archives should use an "imperfective" approach for storing grammars, which allows for modifications.

Grammatical descriptions, especially in digital form are very precious artifacts for linguists, but they present rather small challenges as far as physical storage is concerned. While fossils, taxidermic specimens or ceramics require the archivist to strike a balance between minimizing the risk of physical damage to the artifact and public access, such is not the case for digital grammatical descriptions. There is no reason to keep them in special reading rooms controlled for temperature and humidity or the like. 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.

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 would reflect this model, as items can be added or

2 http://www.livingreviews.org/

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 allows unique identification of chunks and reference to them in the Semantic Web.

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

The example is in XML. This is a storage format which describes the semantics of the document. When rendered on the screen, fonts, colors, layout etc will be chosen according to the semantics. For instance, `<head>` Nominal morphology `</head>` could be bold, slightly larger than surrounding text, on a line of its own and centered. XML uses tags (in black above) to indicate when a particular element starts and ends. Furthermore, tags can have attributes (in green above), which can have certain values (in red). When a text is stripped of all the items just mentioned, the bare textual content remains (in normal color above) :

Note these points:

1. important elements such as headings (`<head>`), examples (`<lgex>`), and cross-references (`<ptr>` [pointer]) are explicitly tagged using XML elements
2. there are unique references to paragraphs and examples (`id=...`)[3]

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

3. some terms are enclosed in semantic markup (`<gloss type="Leipzig">ACC</gloss>`)
4. semantic markup includes references to term definitions, in this case referencing the GOLD ontology[4] and the Leipzig Glossing Rules.[5] These references allow readers to look up the meanings of the terms used in a central place and can help establish a shared vocabulary across grammars.

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.[6]

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:

1. The **arguments** of the linked relation must be identifiable; they must have Uniform Resource Identifiers (URIs). This Uniform Resource Identifier uniquely identifies a variable and allows to look it up on the Internet, providing more information about the value of this variable. URIs can, for the purpose of this paper, be equated with web addresses of a certain format.
2. The **relation** must be defined. Ideally, one uses relations already defined in widely-used vocabularies such as RDFS,[7] Dublin Core,[8] SKOS,[9] GOLD, and lexvo.[10]
3. The **formalism** to link the relation and the arguments must be established. The formalism to link these together is RDF[11] here represented in a variant

---

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

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

[6] Note that the link to GOLD is mainly useful to *retrieve* the section in question for further inspection by a human reader. Zaefferer (2006) suggests to do automated reasoning across grammars using ontologies like GOLD. The problem of cross-linguistic categories, and how they can be established and equated, 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 GD and suggest further reading in other GDs.

[7] http://www.w3.org/TR/rdf-schema/

[8] http://dublincore.org/, provides terms for metadata about documents.

[9] http://www.w3.org/2004/02/skos/, provides terms for describing concepts and their relation.

[10] http://www.lexvo.org/, provides terms for describing language names and script names.

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

called N3.[12] 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.[13]
Another example would be

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

In this example, we assert that the first work references the second. We use the Dublin Core ontology, which provides the predicate "references". The Dublin Core ontology[14] is widely used to express metadata about documents. The use of a common ontology is a best practice in order to assure interoperability between resources. For instance, WALS[15] and Glottolog[16] also use Dublin Core to represent their metadata.

3 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. Since the 1980s, the *Text Encoding Initiative* (TEI) works on schemas for representing the content of texts, mainly in the humanities. The idea is that texts consist of recurring elements, which can be labeled as such. As an example, we can take a poem.[17] This can be marked up using the tags `<l>` for *line* and `<lg>` for *line group*. Note that line

[12] http://www.w3.org/DesignIssues/Notation3.html
[13] 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
http://linguistics-ontology.org/gold/2010/Infix
In a first version of this paper, the address www.grammararchive.org was only mentioned as a placeholder for whatever the URL of an archiving institution was. But several reviewers asked about this site, so we went ahead and created such a site with about 450 grammars from the 19th century where copyright has expired. The structure of the site is expected to change as we implement the theoretical considerations developed in this paper, but the web address should remain stable.

[14] http://dublincore.org/
[15] http://www.wals.info
[16] http://www.glottolog.org
[17] Example from http://www.tei-c.org/release/doc/tei-p5-doc/en/html/VE.html

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>
```

The Text Encoding Initiative provides vocabularies for various domains of the humanities, and also the possibility to create a specialized schema for new domains. When doing this, 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 shared with other types of texts, such as paragraphs, headings, and cross-references. For those, existing TEI vocabulary can be used. In linguistic texts, we are also a dealing 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.

3.1 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 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 tags. For instance,

ISO-3166-1 standardizes country names. This is used in the tag `<country>` to identify the country of India, even if the string in this particular instance is *Indian* with a final *n*. ISO-3166-2 identifies subdivisions of countries, states and territories in the case of India. *IN-DD* identifies Daman and Diu. ISO 639 is the ISO standard for language names, and ISBN is of course used for books. PND finally stands for *Personennormdatei* (English *Person Authority File*), and is used by German libraries to uniquely identify authors. Hugo Schuchardt has the identifier 118611046 and Sebastião Dalgado has the identifier 115161023.

The identification of 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 be treated as named entities if they are defined outside the text, for example in the ISO register,[18] lexvo, or GOLD.

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

3.2 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>
 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, which cover over 20000 languoids (Nordhoff et al. 2013).

3.3 Phrase-gloss pairs

Grammatical descriptions very often provide object language terms immediately followed by translations:[19]

```
<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>
```

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

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

3.4 Linguistic Examples

The most salient text element found in grammars is the example, of which the traditional three-line interlinear glossed text is the best known (Bow et al. 2003). The three-line model does, however, 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 other 40% deviate structurally from this model in one way or another (subexamples, missing lines, extra lines, missing quotation marks around translations etc.). Some of the "examples" are actually not examples, but lists of people, regions, tables, or other content.

In order to accommodate varying content, which can be found in the enumerated environment "Example", we 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[20]
- ungrammatical one-liners
- ungrammatical two-liners with IMT but no translation[21]
- 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
- ...

Figure 1 gives a breakdown of the number of lines examples have in the Qiang grammar. 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).

---

[20] E.g. from Davies 2010:128:
sorop are 'sunset' < *sorop* 'enter' + *are* 'sun'
with no following lines

[21] e.g. from Epps 2008:430:
*ʔãh ʔey-tɔʔɔ́h-ɔ́y
1SG call-run-DYNM
with no following translation line

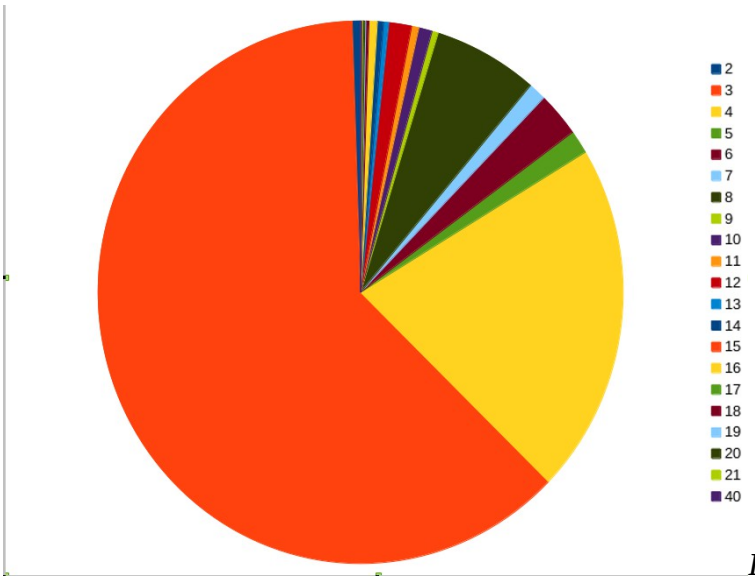


Figure 1: The distribution of examples with n lines in LaPolla's A Grammar of Qiang

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 are elements of rows, see (3)) or vertically (cells are elements of columns, see (4)). Both solutions have their advantages, and they can be transformed into each other, so that this is a matter of personal taste.

(3)

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>
```

(4)

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>
```

3.5 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>
```

3.6 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 descriptions in XML, this will not be illustrated here.

4 PROPOSALS

The structure of grammatical descriptions has received detailed treatment in Lehmann (1980, 1989, 1993, 1998, 2004a, 2004b), Lehmann & 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 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. Of course, for didactic reasons, it is often useful to proceed in a certain manner. Before reading about diphthongs, it might be good to have a basic knowledge of the vowels of the language; complex clauses should follow simple clauses; and the numbers above 10 should be treated after the lower numbers. 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 split alignment system, 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 solution how to arrange the content.[22]

For an archivist, it will of course always be desirable to preserve the linear order of materials received, but for use in the Semantic Web, linearity is not a requirement.

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.[23]

4.1 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 lexicography). Within this macrostructure, we can identify, for example, the frontmatter with table of contents, preface, acknowledgments, the backmatter with

---

[22] Alexis Dimitriadis (p.c. 2009) suggests that the best order from a mathematical point of view would be one which minimizes forward dependencies. Forward dependencies are references to sections which come after the current one. If the network of cross-references is seen as a graph, one can mathematically compute the order(s) for which there are the least forward dependencies. The result, however, might result in the integrity of chapters being torn apart as one subchapter is relocated to the very end in order to get rid of yet another forward dependency contained therein.

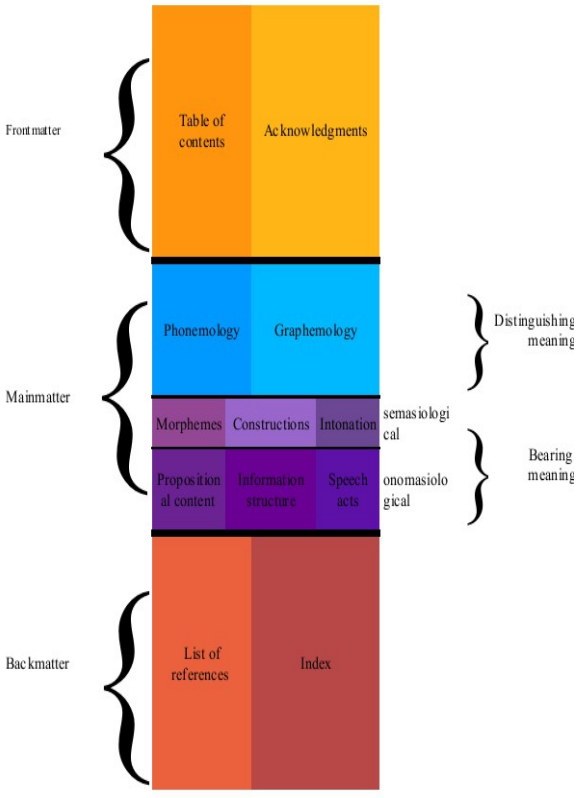
[23] For instance, sections on verbal morphology often contain passages about periphrases (e.g. perfect tense) or serial verbs. These constructions encode content that would otherwise often be dealt with by morphology. However, these construction are not morphological. A shift in perspective from semasiological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

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 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 subsystems. 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 box chart in Figure 2.

FIGURE 2
A box chart of the structure of grammatical descriptions



5 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 OCR-ed 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 some usable results. The basic insight is that Grammars normally use one of two patterns for section titles:

1. Digits with periods, e.g. *3.1.2. Some Title*
2. A keyword followed by some digits
   (Chapter|Section|Kapitel|Chapitre|...) [123456789]

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

	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	

Illustration 2: Table 1: Recognition of sections in 7500 grammatical descriptions in English and other languages

There are about 5000 grammars written in English and about 2800 written in other languages (Bulgarian, German, Italian, Russian, Swedish, Dutch, French, Indonesian, Japanese, Portuguese, Spanish). Of those, about a third can be split into chunks using the patterns mentioned above. The chunks are stored as text files in the file system for further processing. Above 60% of the files do not yield satisfying results. Some have no matches at all for the patterns. Others yield too many sections (several per page), while finally there are some which do yield sections, but the length of the sections retrieved is not typical for a grammatical description. For the purposes of this calculation, we set the lower bound for acceptable average sections lengths to 1000 characters (slightly less than one page), and the upper bound to 10,000 characters (roughly 7 pages).

The 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 an 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 in order to arrive at document classification. Due to the heterogeneous nature of the documents with very many different strings, 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 - yet unwritten - content is concerned, the application of the schema will be easier if grammar writers use authoring software that is compatible

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

25 Recognition and markup of examples is of course not limited to grammatical descriptions, but can also be useful for other types of data, e.g. corpora. This is, however, outside the scope of this paper.

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)[26]. 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) discuss other projects which output XML. As this output is already available in the right format, prospects for conversion look good. Finally, publishing houses such as Mouton De Gruyter are moving towards an XML-first workflow, which will require that the content already be available as XML as it enters the production cycle. This means that conversion to XML can then be dispensed with. Nordhoff is currently working on designing an XML-DTD for grammatical descriptions to which upcoming books in the Mouton Grammar Library will conform.

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. Such an authoring tool would go beyond the annotation of examples (as is done for instance in Toolbox, FLEx or ELAN) and also provide assistance for textual elements. The semantic annotation of the components would furthermore enhance possibilities of discovering relevant data and harvesting them in order to include them in other projects. Persistent URIs furthermore mean that third party researchers can 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 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).

6 CONCLUSION

Next to primary material, archives should also store derived material, like grammatical descriptions. These grammatical descriptions should be stored in an archive which is outward-oriented and imperfective. The grammatical descriptions should be accessible in a granular fashion, allowing for sections, paragraphs and examples to be retrieved individually. Semantic markup, building upon the Text Encoding Initiative's efforts, will allow for better querying, discoverability and harvesting possibilities and will allow language descriptions to become part of the Semantic Web.

---

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

REFERENCES

Ameka, F., A. Dench & N. Evans (eds.) 2006. *Catching Language – The standing challenge of grammar writing*. Berlin, New York: Mouton de Gruyter.

Auer, S. & Hellmann, S. 2012. "The Web of Data: Decentralized, collaborative, interlinked and interoperable". *LREC 2012*, http://www.lrec-conf.org/proceedings/lrec2012/keynotes/LREC%202012.Keynote%20Speech%201.Soeren%20Auer.pdf

Berners-Lee, T.; Hendler, J. & Lassila, O. 2001. "The Semantic Web". *Scientific American* (284). 34-43

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.

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

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.

Drude, Sebastian. 2012. Digital Grammars -- Integrating the Wiki/CMS approach with Language Archiving Technology and TEI. In Nordhoff, Sebastian. 2013. Electronic Grammaticography. Manoa: University of Hawai'i Presspp. pp. 160-178.

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

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.

Kanerva, P., Kristoferson, J. & Holst, A. (2000): Random Indexing of Text

Samples for Latent Semantic Analysis, Proceedings of the 22nd Annual Conference of the Cognitive Science Society, p. 1036. Mahwah, New Jersey: Erlbaum, 2000.
LaPolla, Randy. 2003. A Grammar of Qiang. Berlin: Walter 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.
Nordhoff, Sebastian & Hammarström, Harald & Forkel, Robert & Haspelmath,

Martin (eds.) 2013. Glottolog 2.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://glottolog.org, Accessed on 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.
Penton, D., C. Bow, S. Bird & B. Hughes 2004. “Towards a general model for linguistic paradigms”. *Proceedings of EMELD 2004*.
Rice, Keren. 2006. A typology of good grammars. Studies in Language 30(2): 385-415.
Shadbolt, N.; Hall, W. & Berners-Lee, T. 2006. “The semantic web revisited”. *Intelligent Systems* 21, 96-101
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.
Weber, David. 2006a. Thoughts on growing a grammar. Studies in Language 30(2): 417-444.
Zaefferer, D. (ed.) 1998. *Deskriptive Grammatik und allgemeiner Sprachvergleich*. Tübingen: Niemeyer.