The Descriptive Grammar as a (Meta)Database

Jeff Good

University of Pittsburgh and the Max Planck Institute for Evolutionary Anthropology

0. Introduction

This paper presents a general model for the structure of the traditional descriptive grammar based on a survey of four printed grammars, each of which was chosen as representative of a different "genre": a "best-practice" grammar, Haspelmath's (1993) Lezgian grammar; a grammar representing the traditions of a specific area/family, Maganga and Schadeberg's (1992) grammar of Kinyamwezi, a Bantu language; a grammar from the Routledge Descriptive Grammars series, Huttar and Huttar's (1994) grammar of Ndyuka; and a "legacy" grammar, Williamson's (1965) grammar of Ijaw, which remains an important resource for the language despite making use of a dated syntactic formalism.

This study is intended to be exploratory more than definitive. Its primary goal is to stimulate debate on the way information found in descriptive grammars is structured, with the ultimate goal of developing a workable model for the digital analog of printed descriptive grammars. Section 1 will present important features found in all of the grammars surveyed. Section 2 will discuss interesting features specific to individual grammars. Section 3 will present a general model for the traditional descriptive grammar, understanding it to be a series of annotations over a lexicon and set of texts. Section 4 will give a possible XML representation of that model. Section 5 will offer a brief conclusion and discuss possible future directions for research on modeling grammars.

1. General features of descriptive grammars

1.0 Four basic features

Four main features were found that were common to all of the grammars in the survey which appear to form the "core" of the traditional descriptive grammar:

- A basic structure consisting of a series of (possibly nested) **sections**
- The use of **descriptive prose** in the sections
- The use of **exemplars** of grammatical phenomena in the sections
- Extensive use of **ontologies** throughout

The two pages given below in figures 1 and 2 from Huttar and Huttar's (1994:85–6) Ndyuka grammar clearly show these four features. This grammar is from the Routledge Descriptive Grammar series whose structure is based on the Comrie and Smith (1977) *Questionnaire*. Because of this, it contains more articulated sectioning than most grammars, with extensive nesting of sections. (This is discussed in more detail in section 2.2.) Each section contains a title, and, in the two example pages, each section title begins with *Indirect commands*, a term employed generally in linguistic description and which can be understood as being implicitly drawn from a general linguistic ontology—the sort of ontology being formalized by the E-MELD project in the GOLD ontology (Farrar and Langendoen 2003). The header at the top of the page, *Syntax*, is also an implicit reference to some general ontology. In this particular case, the header indicates, among other things, that *Indirect commands* are a type of syntactic phenomenon.

The pages in figures 1 and 2 further illustrate that the basic structure of a section is that it can (i)

contain other sections, (ii) contain descriptive prose relating to the phenomenon being described, and (iii) contain examples illustrating that phenomenon. In this particular case, the examples are in the form of interlinear text. As will be discussed in section 1.4, I refer to the examples used in descriptive grammars as exemplars, to emphasize their status as examples specifically chosen to illustrate some phenomenon.

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Figure 1: Ndyuka Indirect Commands (page 85)

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Figure 2: Ndyuka Indirect Commands (page 86)

In the next four sections, I will discuss each of the four basic features of descriptive grammars listed above in more detail. In section 1.5, I will briefly cover another common feature of the grammars, less widely used, but important in some areas, *structured description*.

1.1 Three types of ontologies

In the last section, the fact that the Ndyuka grammar made implicit reference to a *general* ontology for linguistic description was mentioned. The use of ontologies in descriptive grammars, however, goes far beyond this. At least two other types of linguistic ontologies are regularly employed which, here, will be called *subcommunity* and *local* ontologies. Brief descriptions of these three types of ontologies are given immediately below.

- **General ontologies:** Sets of terms, and the relationships among them, assumed to be understood by the whole linguistics community
- **Subcommunity ontologies:** Sets of terms, and the relationships among them, assumed to be understood by a specific subcommunity of linguists
- **Local ontologies:** Sets of terms, and the relationships among them, only taken to be meaningful in the context of the description of a particular language

In addition to referring to a general ontology, with the use of a term like *indirect command*, the excerpt from the Ndyuka grammar in figures 1 and 2 also illustrates the use of a local ontology, which can be seen in the titles for the three different classes of indirect commands—those that use *meke* 'make', those that use a *taki* as a complementizer, and those that use *fu* 'for'. Since these classes are characterized with reference to particular words from Ndyuka, they are clear instances of the use of a local ontology.

The use of local ontologies is particularly noticeable in cases like that seen in the excerpt from Williamson's (1965:28–29) Ijaw grammar given in the figures 3 and 4 below where the label for some grammatical class is completely arbitrary (see, specifically, the section labeled "1.7.2 Tone classes"). In this particular case, different tonal classes of words found in Ijaw are given the labels Class I, Class II, Class III, Class IV, and Class V. It might, of course, be the case that an analysis could be applied to these classes which would allow the words in each class to be given a label drawn from a general or subcommunity ontology. However, in this particular description, as it stands, the classes are given labels drawn from a local ontology.

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Figure 3: Ijaw Tone Classes (page 28)

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Figure 4: Ijaw Tone Classes (page 29)

Something that is important to note about the use of local ontologies—which can be seen in both the Ndyuka and the Ijaw excerpts—is that they tend to be used to subdivide phenomena which are classified using a term drawn from a more general ontology. For example, Williamson's Classes I–V for Ijaw are designated as being *tonal classes*—a concept which would be needed for languages other than Ijaw. It's also important to note that while the labels for the tone classes might be part of a local ontology, a prose description of those labels may draw on terms from a more general ontology. In the case of these Ijaw tone classes, for example, terms like *low*, *rising*, and *isolation*, which have general currency, are employed in characterizing the various classes. These aspects of the use of local ontologies have an important implication: In grammatical description, ontologies are not used in a self-contained way. Rather, terms drawn from different ontologies can be intermingled in the description of some phenomenon.

We have yet to see an example of the use of a subcommunity ontology. An example of this can be seen in figure 5, an excerpt from the Kinyamwezi grammar, which indicates the form of noun class prefixes in the language.

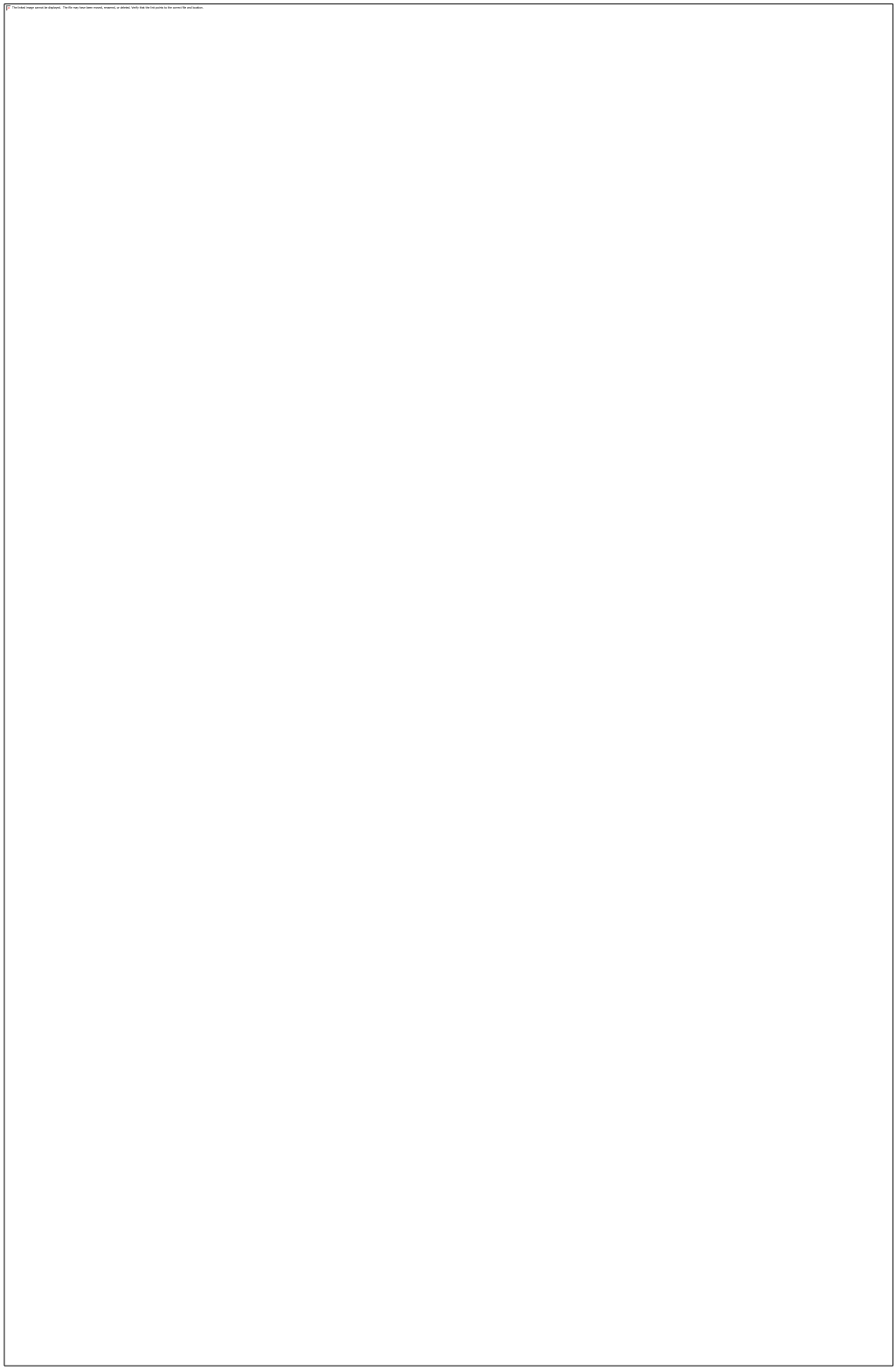
The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Figure 5: Kinyamwezi Noun Classes (page 57)

Bantu languages are famous for their rich noun class system, which has a number of grammatical reflexes, including a system of nominal prefixes. The noun classes of Bantu languages are consistent enough across the family that they are reconstructible for Proto-Bantu, and there is a generally agreed upon terminology set for referring to them using the numbers one through twenty one (only noun class numbers one through eighteen are seen in figure 5). The form of the particular noun classes will differ from language to language. However, by identifying some noun class prefix in a given Bantu language with a number between one and twenty one, a descriptive claim is made that that prefix is etymologically related to a prefix given the same number in another Bantu language. So, unlike the numbered tone classes of Ijaw, the numeric designations for the noun classes seen for Kinyamwezi in figure 5 are not drawn from a local ontology since the same numbering system is used for other languages. Rather, they are drawn from a subcommunity ontology, in particular the Bantu subcommunity.

As with local ontologies, the terms in subcommunity ontologies may be partially defined with terms from a general ontology. In the Bantu case just described, while the use of the numbers one through twenty one to designate noun classes may be peculiar to the Bantu subcommunity, the term *noun class* has a broader use as does the term *prefix*. We see, then, that local ontologies and subcommunity ontologies are not defined in a conceptual "vacuum". Rather, they tend to build on concepts which are part of general ontologies.

1.2 Nested sections

Most book-length documents are divided into sections of one type or another. So, it is not particularly surprising that grammars are similarly divided. Importantly, the sections found in grammars tend to show a relatively high degree of standardization. For example, all of the grammars in the survey contain a section titled "Phonology".

A general pattern seems to be that the sectioning of grammars is sensitive to, but not dictated by, some sort of general linguistic ontology. This ontology is not explicitly employed, of course. Nevertheless, some notion of appropriate categorization of different phenomena clearly informs the way the sections are organized. Consider, for example, the subdivisions of the chapter entitled "Verbal inflection" of the Lezgian grammar given below in table 1.

Verbal inflection
Introduction
The three stems of strong verbs
Verbal inflectional categories
Forms derived from the Masdar stem
Forms derived from the Imperfective stem
Forms derived from the Aorist stem
Secondary verbal categories
Prefixal negation and the Periphrasis forms
Illustrative verbal paradigms
Irregular verbs
The copulas
Verbs lacking a Masdar and Aorist stem
Secondary verbal categories
Verbs with root in ä(g)-
Functions of basic tense-aspect categories

- *Imperfective*
- *Future*
- *Aorist*
- *Perfect*
- *Continuative Imperfective and Continuative Perfect*
- *Past*

Functions of non-indicative finite verb forms

- *Imperative*
- *Prohibitive*
- *Hortative*
- *Optative*
- *Conditional*
- *Interrogative*

Functions of non-finite verb forms

- *Masdar*
- *Infinitive*
- *Participle*
- *Infinitive (Imperfective converb)*
- *Aorist converb*
- *Specialized converbs*

Archaic verb forms

- *Archaic Preterite*
- *Archaic Future*
- *Archaic Past Future*
- *Archaic Imperfective participle*

Table 1: Subdivisions of chapter on verbal inflection in Lezgian grammar

Some aspects of the sectioning given in table 1 of the chapter on verbal inflection in the Lezgian grammar are clearly driven by a sense of an appropriate grouping of concepts relating to verbal inflection. The sections entitled "The three stems of strong verbs", "Verbal inflectional categories", and "Irregular verbs" are, presumably, clustered together at the beginning chapter because they all relate to morphological aspects of verb inflection. The sections entitled "Functions of basic tense-aspect categories", "Functions of non-indicative finite verb forms", and "Functions of non-finite verb forms" are similarly clustered, presumably because they all relate to semantic aspects of verbal inflection. And, focusing on these ontologically-driven "clusterings" within the chapter leaves out, of course, the most obvious way this chapter is organized with respect to an ontology—each of these sections relates in one way or another to the concept of "verbal inflection", which, unlike the implicit logic in the grouping of the subsections, is explicitly indicated in the title of the chapter.

It's important to point out that, even if an ontology drives some aspects of the organization of the sections in a grammar, the exigencies of producing a coherent linearly-organized document like a book will sometimes force deviations and compromises from a purely "ontological" organization. In table 1, we see, for example, the inclusion of a section called "Illustrative partial paradigms". While paradigms presumably have a place in an ontology of concepts relating to verbs, "illustrative" paradigms would not seem to belong to such an ontology. Rather, they are included for the convenience of the reader trying to understand the morphological patterns in the Lezgian verb. Similarly, the appropriateness of a section on "Periphrastic tense-aspect categories" is questionable in

a chapter on verbal *inflection*. The inclusion of this section reflects a tension between organizing the chapter along the narrow concept of inflection and bringing together grammatical forms with similar semantic functions. There is a type of ontological "clash" in the language where similar function is not expressed by similar form. In the production of a physical grammar, the clash must be resolved somehow and, in this instance, we see that functional-based grouping was favored over formally-based grouping.

Ontologically-sensitive sectioning was a feature of all the grammars surveyed. This is not surprising, of course, given that ontologies are a reflection of how linguists categorize grammatical phenomena and a grammar is intended to be a description of a wide range of phenomena of a given language.

A final characteristic of the sectioning of grammars which should be mentioned is that each section is typically given a unique numeric label which is used for referring to that section. This feature of descriptive grammars is an apparent reaction to the fact that, while a physical descriptive grammar is a linearly arranged book, the structure of the information in a grammar is not linear. Rather, it is highly interconnected, and a given piece of information doesn't necessarily belong in only one "place". When the organization of a grammar forces the primary descriptions of related phenomena to appear separately from each other, section references can be used in the prose to connect them descriptively.

1.3 Descriptive prose

Some important features found in the descriptive prose in grammars, in addition to free-form prose itself are given below:

- References to lexical items
- References to other sections
- References to terms drawn from ontologies
- References to exemplar data

Each of these four kinds of references typically has its own standard format. References to particular lexical items, for example, are usually along the lines of *orthographic_form* 'gloss'. References to other sections are done by the unique identifiers discussed above in section 1.2. References to terms drawn from ontologies are generally implicit, though the use of various typographical conventions like italics or an initial capital letter might be employed in some instances to make such references explicit. Finally, references to exemplar data, like references to sections, are typically made using a unique identifier. Section identifiers and exemplar identifiers are typically sufficiently distinct that there is no ambiguity as to which type of reference is being made. For example, in the excerpt from the Lezgian grammar given in figure 6, a typical convention is employed wherein section identifiers consist of a string of numbers connected by dots while exemplar identifiers consist of one number surrounded by parentheses.

Importantly, while it can be the case that descriptive prose is explicitly associated with an exemplar via a reference, sometimes the association is merely implicit. (Exemplars are discussed in more detail in the immediately following section.) This can also be seen in the excerpt from the Lezgian grammar given in figure 6. In section "10.4.1.3", the first chunk of descriptive prose makes explicit reference to the exemplar data in "(396)". The second chunk of prose is only implicitly associated with the exemplar data in "(397)" by virtue of directly preceding that exemplar.

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Figure 6: Lezgian Descriptive Text (page 175)

The linking of particular chunks of prose to particular exemplars can create something like subsections to a given section since a standard way of indicating this linking is to place a set of exemplars immediately after the relevant prose, with multiple distinct sets of exemplars contained within one section. However, this exemplar-based sectioning is not explicit, in contrast to the sort of sections described above in section 1.2.

1.4 Exemplar data

I use the term *exemplar* here to refer to language data used in grammars to exemplify the phenomena under discussion. An exemplar is understood to be different from an *example* of some grammatical phenomenon in that it is specifically chosen by the author of a grammar to assist in descriptions of that phenomenon. It can generally be assumed that the particular examples chosen to serve as exemplars more clearly illustrate the phenomenon under discussion than many of the other examples would.

Two major types of exemplars were found in the grammars of the survey: *lexical exemplars* and *textual exemplars*. Lexical exemplars take the form of either words or morphemes accompanied by glosses, typically arranged in a table. Textual exemplars typically take on the form of interlinear text. Exemplars may or may not be given unique labels.

Figure 7 is an excerpt from the Lezgian grammar showing both lexical and textual exemplars. The lexical exemplars, appear with the label "(345)". The core of the lexical exemplars consists of a word, a gloss, and a grammatical label. In addition, each exemplar is accompanied by a lexically related form for comparative purposes (in parentheses). There are four textual exemplars in figure 7, the first two, labeled "(343)" and "(344)" are interlinear phrases, and the second two, labeled "(346)" and "(347)" are interlinear sentences.

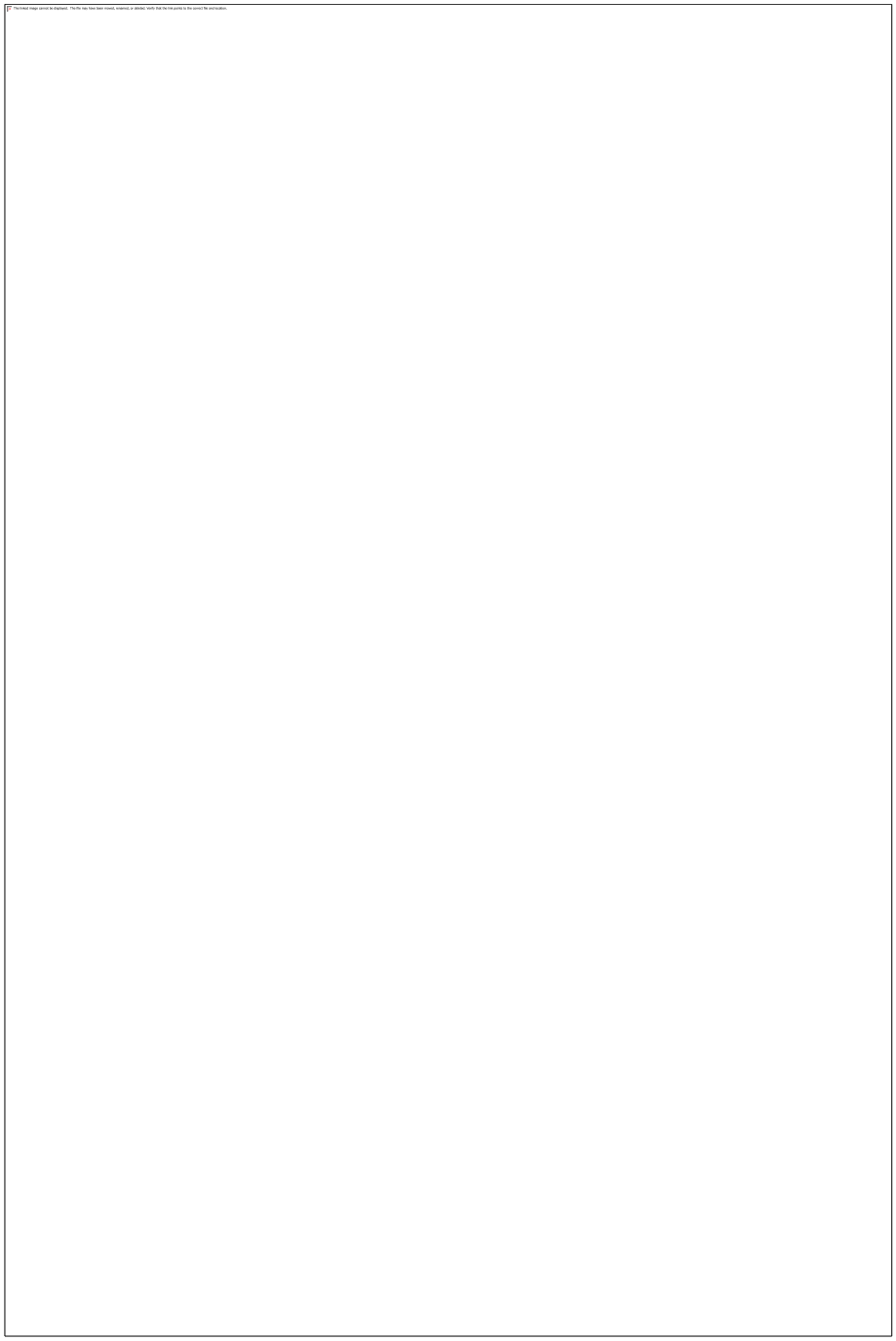

Figure 7: Lezgian Exemplars (page 155; emphasis added)

Some of the exemplars in figure 7 have an important feature: They diverge slightly from a standard presentation format to allow them to better illustrate the phenomena they are exemplifying. This can be seen for the lexical exemplars in that they are accompanied by lexically related forms. One instance of a textual exemplar diverging from standard presentation format can be seen in the part of the exemplar labeled "(346)" that is highlighted in red. The word-by-word glossing has been further annotated for constituency—specifically marking an infinitive clause containing a participial phrase—in order to clarify which part of the sentence is exemplifying the phenomenon under discussion. A comparable device, bolding some words in a textual exemplar, was encountered in the Ndyuka grammar.

Some of the textual exemplars in figure 7 show an additional kind of annotation deviating from strict interlinear format. They also contain external references to the source of the exemplar. These external references are highlighted in blue.

An important aspect of the use of exemplars is that, in some cases, exemplars are grouped with other exemplars. This can be seen, for example, in the set of exemplar data labeled "(345)" in figure 7. The general use of such grouping is to indicate that each member of the set either illustrates the same phenomenon or plays a part in illustrating some phenomenon. However, it is not the case that an ungrouped set of exemplars should be assumed to illustrate different phenomena. The Ndyuka grammar, for example, did not make use of any explicit grouping convention despite numerous instances of exemplars which would have been reasonable to treat as members of a logical set.

1.5 Structured description

A final, less prominent, feature found in grammars of the survey combines some features of descriptive prose and exemplars. This is what I term, *structured description*. This is description, typically in tabular format and offset from the prose, covering a particularly coherent domain of a language's grammar. The most frequently occurring type of structured description is tabular presentations of a language's phoneme inventory. Such an inventory is clearly description, but, unlike descriptive prose, the description has a very particular format grouping segments by generally-accepted phonetic and/or phonological categories.

However, structured description is not restricted to relatively standardized realms like phoneme inventories. It is also used for phenomena which apply to a sufficiently large class of constituents that a generalized schema can be given. The parts of the excerpt from the Kinyamwezi grammar highlighted in red in figure 8 give an example of a structured description summarizing the tone patterns for certain verbal forms using a type of morpheme-to-tone association template.

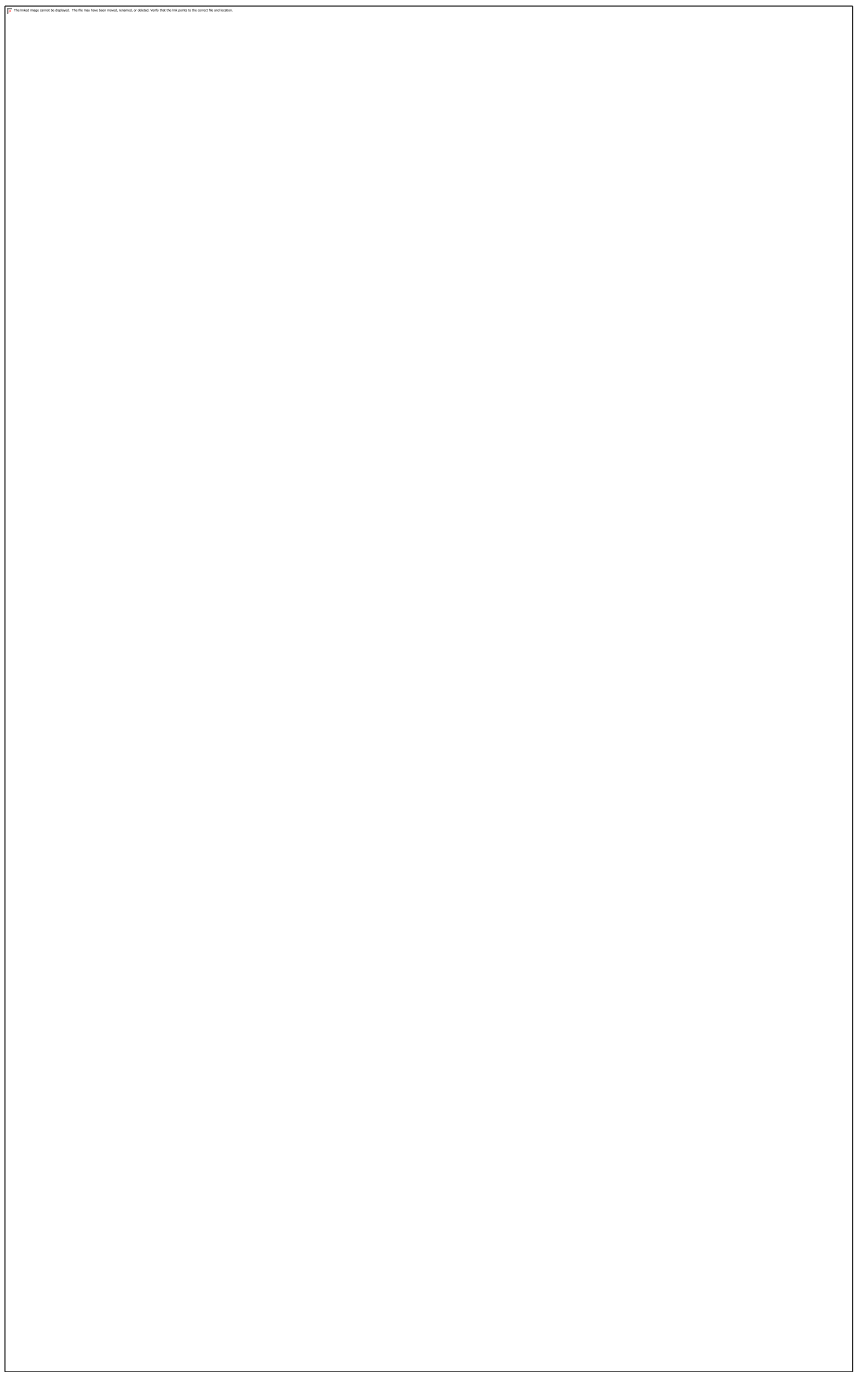

Figure 8: Kinyamwezi Structured Description (page 110; emphasis added)

Some instances of structured description bear resemblance to theoretically-oriented formalizations of grammatical phenomena. The boundaries between structured description and formal "description" are not immediately clear, and, in fact, as will be discussed briefly in section 2.4, one of the grammars in the survey, the Ijaw grammar, made extensive use of formal rules in describing the language's grammar.

2. Particular features of the four grammars

2.1 The Lezgian grammar

Of the four grammars in the survey, I designated the Lezgian grammar as the "best-practice" grammar because it is widely recognized as exceptionally well designed, even including some innovative features which anticipate recent developments in computer-assisted linguistics. Some interesting features of the grammar are given below, and I'll discuss each of them in more detail in turn.

- A subject index which not only refers readers to topics covered by the grammar but also has conventions for explicitly indicating common grammatical phenomena not found in Lezgian
- An index to examples indicating which other examples, anywhere in the grammar, illustrate some phenomenon even if they have not been chosen as exemplars of that phenomenon
- A typographic distinction between language-particular morphological categories and universal and semantic categories

Figure 9 below illustrates how the Lezgian grammar's index indicates that some cross-linguistically common phenomenon is not present in Lezgian. Specifically, a convention of placing a "(-)" after a relevant entry is employed. (Some such entries are highlighted in red in figure 9.) This system of indexing shows a sensitivity to something like a general linguistic ontology, insofar as it presupposes that certain phenomena are important enough to "grammar" that a linguist might expect them to be present in a language even if they were completely unfamiliar with that language. Importantly, a statement on the non-existence of some phenomenon in a language is different from the absence of any discussion of that phenomenon in a grammar. The latter case could be an accidental omission. The former is an actual descriptive statement about the language.

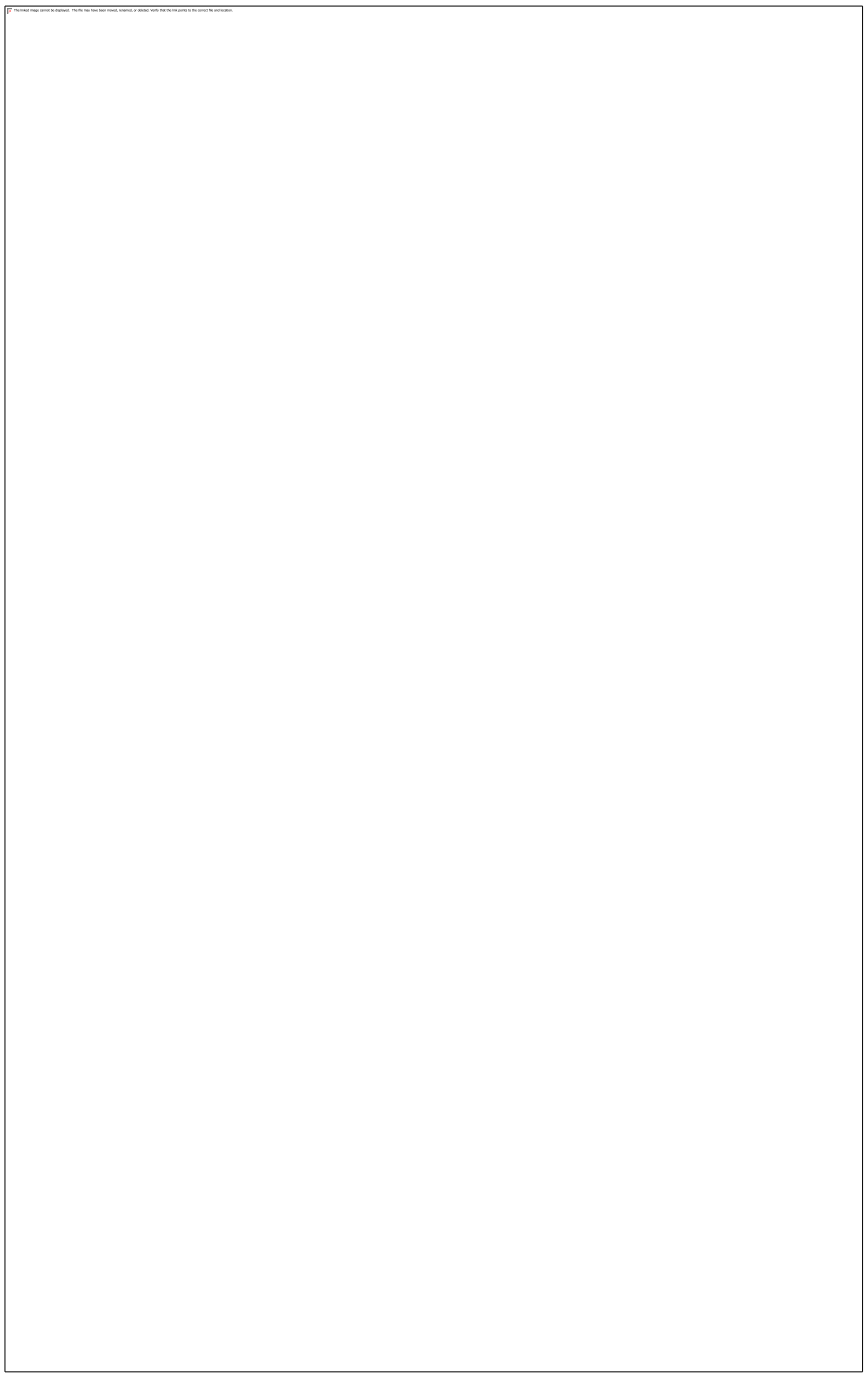

Figure 9: Lezgian Subject Index (page 564)

Figure 10 illustrates a page from the index of example sentences found in the Lezgian grammar. The keys to the index are the numbers for the various examples in the grammar and values associated with each key are other example numbers. Notably, the examples referred to are not limited to the exemplars in the grammar but also include examples of various phenomena found in texts provided with the grammar. This sort of index provides much of the functionality which creators of digital resources hope to make possible by providing online or offline search facilities.

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Figure 10: Lezgian Example Index (page 530)

The final feature particular to the Lezgian grammar I will mention here is the fact that it makes a typographic distinction between what are classified as language-specific categories like "Ergative" case or "Involuntary Agent" construction, which are capitalized and what are considered universal or semantic categories, like "complement clause" or "adverbial modifier".

Given the discussion in section 1.1, this is a particularly interesting feature since it shows an explicit recognition that grammars tend to employ terminology drawn from different ontologies. In this particular case, when a category is drawn from a general ontology, no capitalization is used. However, when a category is drawn from a local ontology, capitalization is employed.

In addition, by using capitalization of recognizable terms for "language-specific" terms, instead of, say, constructing new terminology entirely, there is an implicit recommended mapping of the language-specific term to a general term. For example, the fact that the label "Ergative" is employed for a particular case form in Lezgian can be taken as a recommendation that that case be mapped to prototypical "ergative" case. The fact that "Ergative" is capitalized is an indication that arguments marked with this case in Lezgian may not have all the features typically associated with ergative case arguments. However, it can reasonably be expected to have the core properties of such arguments.

2.2 The Ndyuka grammar

The primary feature particular to the Ndyuka grammar which I will discuss is its sectioning, which is is based on Comrie and Smith's (1977) *Questionnaire* for language description.

The *Questionnaire* was designed to set forth a range of questions to ask when working on the grammatical description of a language. The use of the *Questionnaire* for a given language is meant to ensure that the description of that language has adequate grammatical coverage as well as to facilitate cross-linguistic comparison of that language with other languages described using the *Questionnaire*, since grammars based on the *Questionnaire*, for the most part, would be expected to have similar sectioning. For illustrative purposes, the first page of the *Questionnaire* is given in figure 11 below.

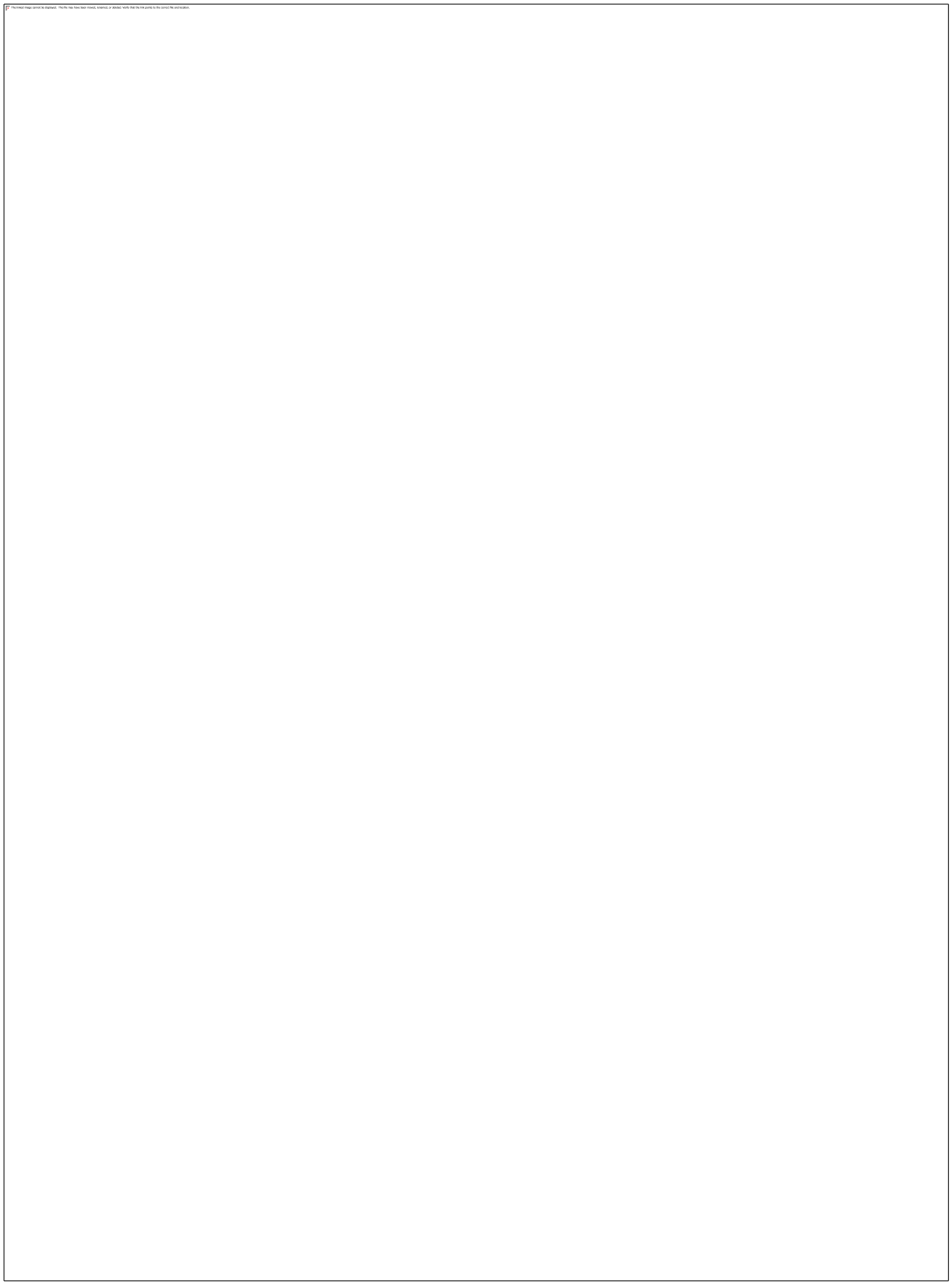

Figure 11: Comrie and Smith (1977) *Questionnaire* (page 1)

Any grammar following the *Questionnaire* should, ideally, use the exact sectioning outlined in the *Questionnaire* itself. So, for example, in the Ndyuka grammar, there is a section with the identifier "1.1.1.2.1" on Yes-no questions.

A feature of the Ndyuka grammar imposed on it by *Questionnaire*-based sectioning is that, like with the Lezgian grammar, there are numerous places where it is explicitly indicated that the language lacks some grammatical phenomena. For example, question "2.1.1.14" of the *Questionnaire* essentially asks if obviation is found in the language. Since obviation is not a grammatical phenomenon in Ndyuka, section "2.1.1.14" of the Ndyuka grammar simply states that nominals are not marked for obviation.

2.3 The Kinyamwezi grammar

The most noteworthy aspect of the Kinyamwezi grammar, with respect to the present survey, is its extensive use of a subcommunity ontology. This grammar, of course, was chosen to exemplify subcommunity grammars—so, this is not surprising.

One need only look at the table of contents of the grammar to see the use of terms which, while being well-known in the Bantuist community, could not be expected to be well-known outside of it. For example, a section of the chapter on consonants is entitled, "Dahl's Rule"—this refers to a particular historical dissimilation process which is important in Bantu historical phonology. The section on nouns has a subsection entitled "The Augment", again a Bantu-specific term. Similarly, the chapter on verbal derivation contains section titles using the word "extension"—a reference to a particular subclass of suffixes found on Bantu verbs.

The use of terms specific to Bantu linguistics, importantly, does not make this grammar unusable to people from outside that community because the descriptive prose, in general, either defines community-specific terms or describes the relevant phenomenon clearly enough that it is not necessary to know the precise definition of the term to interpret the grammatical facts of the language.

Given the large number of Bantu languages and the extensive similarities found among them, the use of a subcommunity ontology plays a similar role to the use of the *Questionnaire* format for the Ndyuka grammar. It facilitates cross-linguistic comparison. However, unlike the Ndyuka case it is not general cross-linguistic comparison which is made easier. Rather, comparison with other Bantu languages is facilitated. We can think of the Bantu term set as an ontology which is "optimized" for one particular language family. In some areas, like verbal suffixes, it makes finer-grained distinctions than a general ontology would, while it entirely ignores phenomena which are not well-represented in Bantu languages.

2.4 The Ijaw grammar

The Ijaw grammar was chosen to be part of the survey for its use of a "legacy" formalism—specifically, notational devices of early transformational grammar. While there are numerous instances of legacy formalisms to be found in linguistic description, I chose this grammar, in particular, because it is still cited fairly frequently as it represents the only grammar (to my knowledge) of a typologically unusual Niger-Congo language.

Some of the effects of the use of a legacy formalism can be seen in the grammar's overall structure. For example, phenomena generally classified as syntactic are spread out over four chapters, "Phrase-structure rules", "Verb phrase transformations", "Noun phrase transformations", and "Sentence transformations". While the idea of having multiple chapters for syntactic phenomena is not

necessarily tied to any particular formal theory of grammar, these particular divisions are clearly derived from early transformation grammar, and it is unlikely that they would be employed in any grammar produced today.

In addition, the use of this formalism also affects the nature of the description of particular phenomena. Figure 12 gives an excerpt from the chapter on phrase-structure rules wherein aspects of the basic syntax of sentences are discussed. While the discussion is still accessible to a present-day reader, the particular format of the description, using phrasal expansion rules to describe possible sentences, is not in common use today in descriptive grammars.

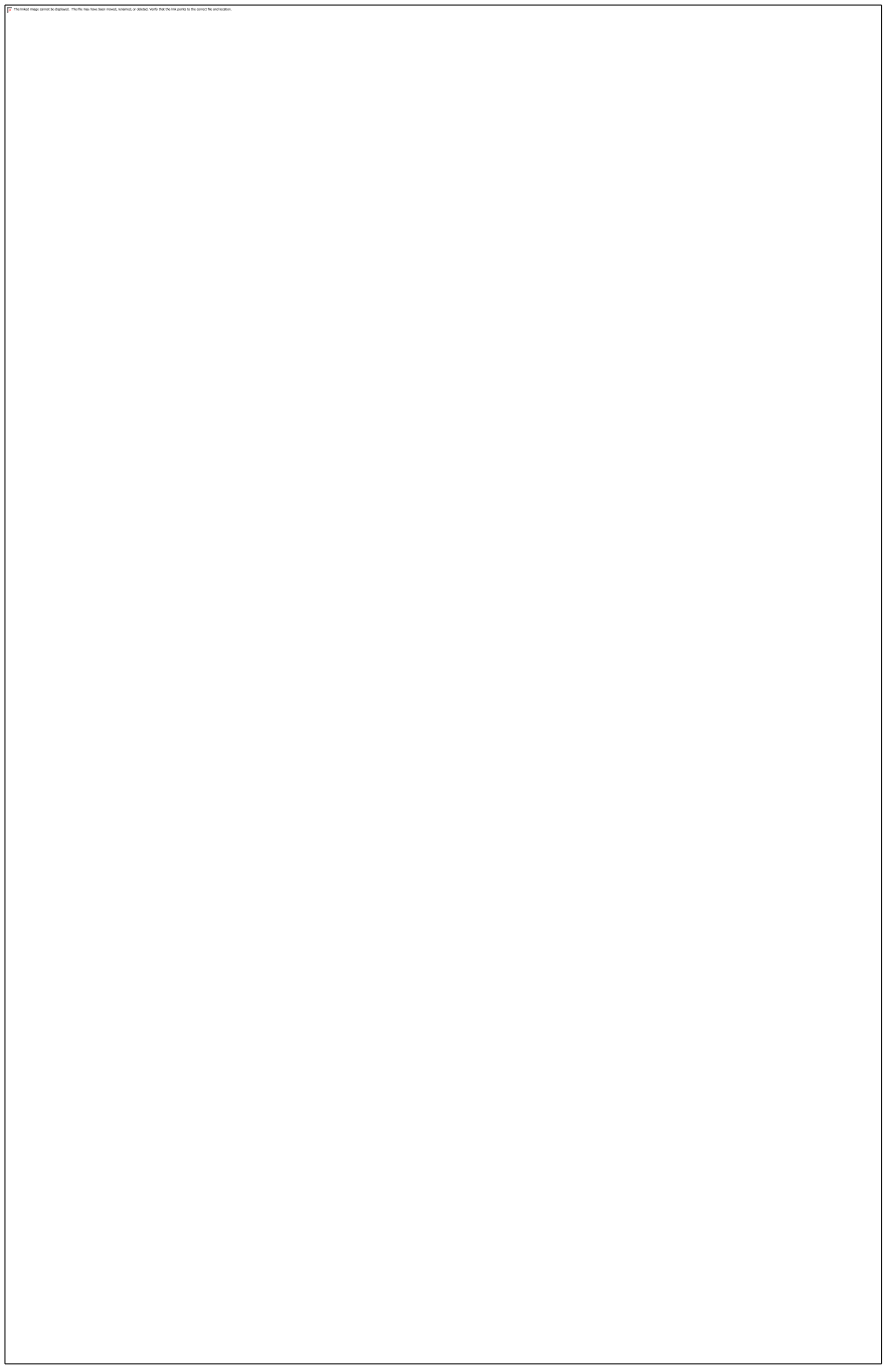

Figure 12: Ijaw Sentence-Level Phrase Structure Rules (page 33)

While formal representations of grammatical phenomena may have a place within language documentation (see for example, Bender, et al. (2004)), they have not been widely used by authors of descriptive grammars. This is presumably because grammatical formalisms have changed too rapidly to be seen as valuable tools for the creation of language documentation which is intended to last for decades—if not centuries.

The existence of legacy formalisms in grammars like the Ijaw grammar should probably not influence best-practice recommendations for the production of new grammars. However, they do need to be considered in the formulation of best-practice recommendations for the conversion of existing print grammars to digital formats.

3. Towards a model of the structure of a descriptive grammar

In this section, I will propose a basic model for the structure of descriptive grammars based on the features found in the grammars in the survey. This model will be incomplete insofar as it will not encompass all of the features found in each grammar. Rather, it will aim to isolate the features common to all of the grammars in order to establish a foundation upon which more particular features can be added. This reflects the fact that the primary aim of this paper is not to present a definitive model for all descriptive grammars but rather to stimulate discussion on standards for encoding the information found in descriptive grammars electronically.

The basic model for a descriptive grammar that I propose here is given in figure 13.

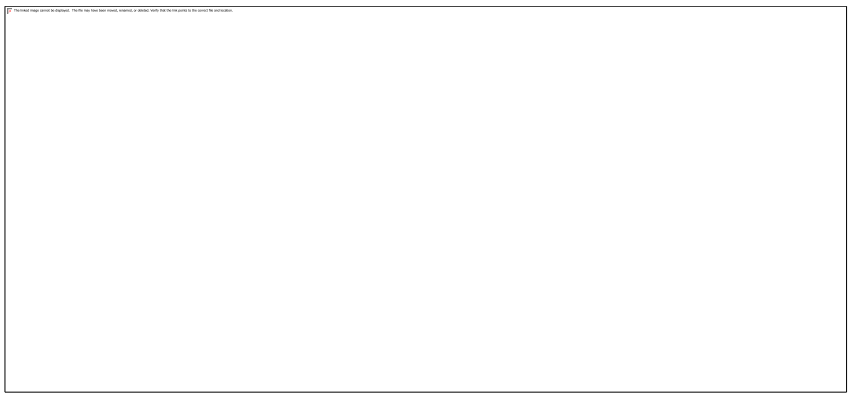

Figure 13: A basic model of the descriptive grammar

The basic model for a descriptive grammar given in figure 13 envisions it as a series of annotations on a lexicon and a set of texts. The presentational analog to an annotation found in the surveyed grammars is the section, described above in section 1.2. While, in some cases, a descriptive grammar is accompanied by a published lexicon and a set of texts (sometimes in the same volume as the grammar itself), I do not mean to imply figure 13 that such documents must physically (or electronically) accompany a grammar. Rather, the existence of a lexicon and body of texts is presupposed by a descriptive grammar, which contains generalizations over the lexicon and collected texts of a language. Furthermore, even if there are no particular resources corresponding to a lexicon or set of texts, a partial lexicon and a set of one-sentence "texts" could always be constructed on the basis of the exemplars in the grammar.

The model for descriptive grammars given in figure 13 treats the annotation, a type of relatively unstructured but highly expressive metadata, as the core of the grammar. I classify annotations as a kind of metadata since their content consists of generalizations over the more primary data found in the lexicon and texts. A proposed model for the annotations found in a descriptive grammar is given in figure 14. This structure is simplified somewhat, and other features of annotations are diagrammed in figure 15.

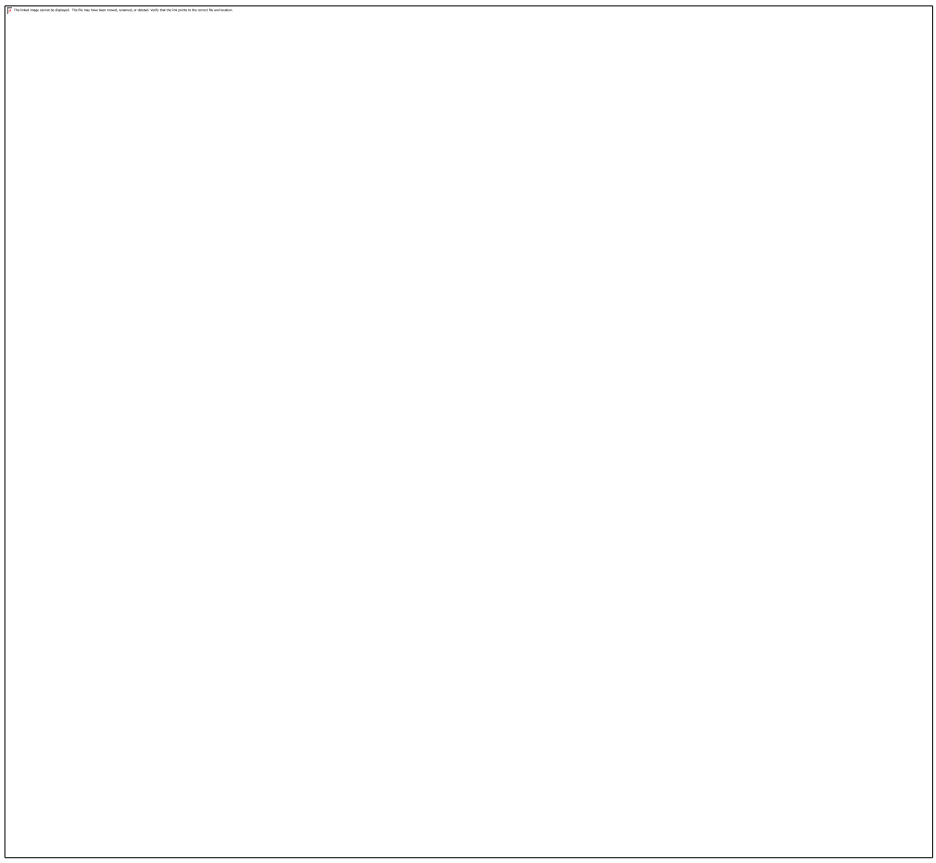

Figure 14: The structure of an annotation

The model for a grammatical annotation given in figure 14 treats the annotation as having descriptive prose at its center with links from the descriptive prose to lexical and textual exemplars as well as to structured description. In addition, exemplars are linked to a (possibly abstract) lexicon or set of texts, and the descriptive prose might contain references to terms drawn from one of the three types of ontologies discussed above. The local and subcommunity ontologies are shown as being linked to the general ontology, indicating that it is common (and presumably best) practice to indicate how local or subcommunity terminology relates to generally understood terminology.

Figure 14 is a simplification of the structure of an annotation in a number of respects. First, it does not indicate whether the components of an annotation are obligatory or optional. In the grammars found in the survey, only the descriptive prose appeared to be obligatory. In addition, figure 14 treats exemplars themselves as part of the structure annotation, while it might be more accurate to consider them as external to the annotation and, instead, only place references to exemplars within the annotation itself. Certainly, from a presentational perspective, exemplars appear to be part of an

annotation. However, it is not clear that the content of exemplars is part of the annotation's logical structure. Figure 14 also implies that links to the ontologies can only be made via descriptive prose while, in fact, links to the ontology could be made from any part of the annotation. (Such links were not presented to minimize visual clutter.) Similarly, figure 14 does not include the fact that structured descriptions can make reference to exemplars, just as descriptive prose can.

Perhaps the most crucial feature of an annotation which is omitted from figure 14 is the fact that an annotation can have two important kinds of relationships to other annotations. First, an annotation can contain other annotations—this is the analog to the nesting of sections in a printed grammar. Second, references to other annotations can be made within annotations. These two kinds of relationships are schematized in figure 15.

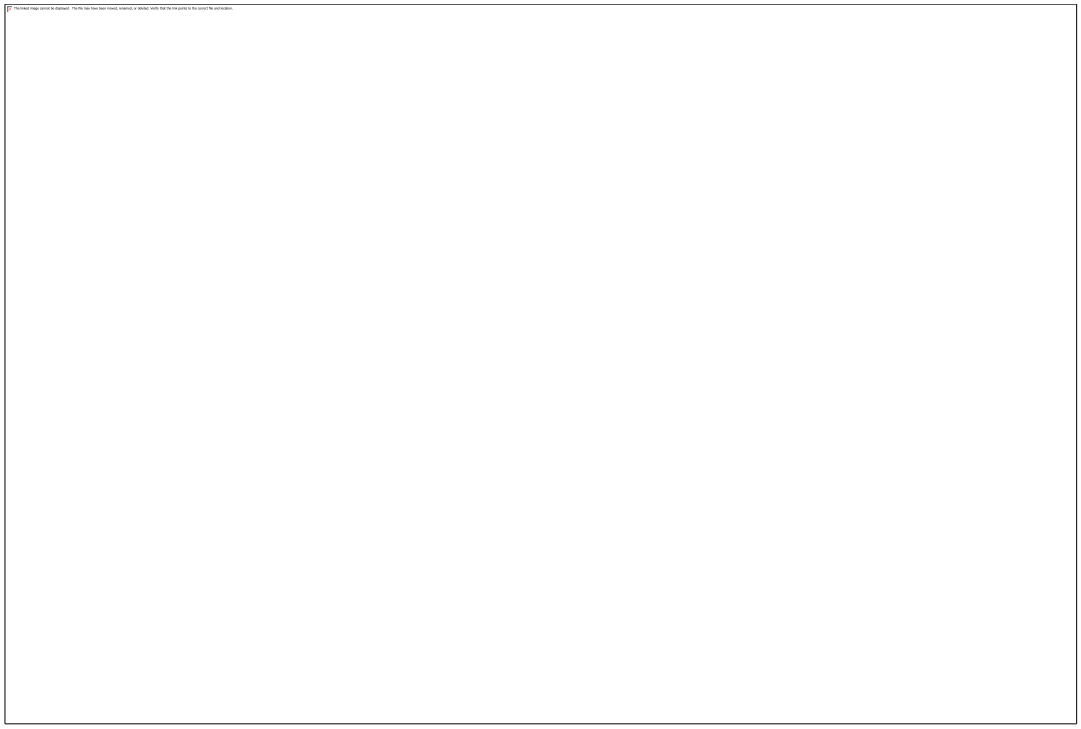

Figure 15: Relationships among annotations

Figures 14 and 15 do not include two features of the sections found in descriptive grammars as part of the annotation: a label (used for referring to the section) and a title. While it is certainly important to represent these pieces of information, they would seem not to be part of the logical structure of an annotation. Rather, they are metadata for the annotation.

Before continuing on to section 4, where I will present a partial XML representation for the information found in descriptive grammars, it would be worthwhile to point out that the model of the descriptive grammar given in figure 13 as a system of annotations over a lexicon and set of texts can be understood as treating a grammar as a sort of metadatabase—that is a database of generalizations over primary data. The only aspect of the model presented here which deviates from the most typical database structure is the fact that annotations can be contained within other annotations. Databases are commonly conceived of as consisting of a series of records with a fixed, non-recursive structure. However, even though it diverges from a prototypical database structure, this aspect of the model makes the structure of a database of grammatical annotations only slightly more complicated than,

say, a database of lexical items. Representing the possible nesting of annotations simply requires the addition of metadata indicating that a given annotation can have another annotation as its "parent".

4. Towards an XML representation of the descriptive grammar

It is not possible here to give a full XML document type definition (DTD) or schema for a descriptive grammar following the model seen in figure 13 since this would require also having a markup schema for lexicons and texts. There has been work on representing both kinds of resources in XML—so, this is, fortunately, not an insurmountable problem. Chapter 12, of the TEI guidelines (Sperberg-McQueen and Bernard 2002), for example, is devoted to markup standards for dictionaries and E-MELD's FIELD lexical input tool can produce richly marked up XML lexicons. For text markup, Bow, Baden, and Bird (2003) provides a system of XML markup which can be applied to interlinear text. These existing standards could easily be adopted in a markup schema for grammars and would be largely sufficient to encode the sorts of lexical and text examples found within them.

The only crucial feature lacking in existing markup systems for lexical entries and interlinear data which would be required to fully represent the lexical and text exemplars in grammars is a system allowing an example to receive special annotation to clearly indicate how it is an exemplar for some particular feature. In the discussion of figure 7 in section 1.4, for example, we saw some cases from the Lezgian grammar where such special annotation was employed. Devising a system of markup for such annotation is outside the scope of the present proposal. It would seem to be a fairly complex task because the nature of the special annotation on exemplars can be quite varied and would require a separate survey in its own right, which should, presumably, include a survey of special exemplar annotation conventions found in theoretical work, where such annotation tends be very widely used (a syntactic tree, for example, could be understood as an example of such annotation).

Another feature of the grammars in the survey which is outside of the scope of this paper is modeling possible types of structured description, discussed in section 1.5. This, too, would seem to require a separate survey in its own right, which should also probably include theoretical work, in addition to descriptive work. First, it would have to be determined how many different types of structured description are used in linguistic analysis and, then, a representation would have to be developed for each type. Below, I give a DTD for the descriptive grammar which allows for the existence of structured description in an annotation. However, I do not give any model for any particular instance of structured description.

Putting the issues of special annotations for exemplars and structured description aside, a possible system for XML markup for descriptive grammars is given below in figure 16, which contains an XML fragment of a markup system consistent with the model of descriptive grammars discussed in section 3.

```
<grammar>

   <ontology id="GOLD" level="general">
   An internal general ontology, or a reference to an external general ontology would be placed
in    this element.
   </ontology>

   <ontology id="MySubcommOnt" level="subcommunity">
   An internal subcommunity ontology, or a reference to an external subcommunity ontology
would    be placed in this element.
   </ontology>

   <ontology id="MyLang" level="local">
   An internal local ontology, or a reference to an external local ontology would be placed in
```

this element.
</ontology>

<ontology id="MiscTerms" level="other">
It might also be worthwhile to allow for other types of ontologies than the three found in the surveyed grammars.
</ontology>

<annotation title="Sample annotation" id="annotation_1">

<ontRef ontologyName="GOLD" ref="some_GOLD_id">
An annotation can be associated with a reference to an ontology. This is a reference to a term from a general ontology.
</ontRef>

<descProse>
Descriptive prose for an annotation would be placed here. In addition, there could be inline references to a lexical itemvia an element like the following
<lexRef ref="some_lexicon_id"/>. There can also be an exemplar set using the markup immediately below. The descriptive prose could also draw a term from an ontology by using an ontology reference asfollows
<ontRef ontologyName="GOLD" ref="some_other_GOLD_id"/>.
</descProse>

<exSet id="exemplar_set_1">

<ontRef ontologyName="GOLD" ref="yet_another_GOLD_id">
Explicit references to ontologies can also be placed within example sets.
</ontRef>

<ontRef ontologyName="MySubcommOnt" ref="some_subcommOnt_id">
Multiple ontology references are allowed. This is a reference to a subcommunity ontology.
</ontRef>

<textEx id="some_text_id">

Content retrieval could be completely automatic or could also be specified within the element.

<ontRef ontologyName="MyLang" ref="some_localOnt_id">
Ontology references can also be directly related to exemplars. This is a reference to a local ontology.
</ontRef>

</textEx>

<textEx id="some_other_text_id"></textEx>

</exSet>

<descProse>
The exemplars above are textual exemplars. Lexical exemplars are also possible, as seen below.
</descProse>

```
    <exSet id="exemplar_set_2">

      <lexEx ref="some_lexicon_id"></lexEx>

      <lexEx ref="some_other_lexicon_id"></lexEx>

    </exSet>

    <annotation title="Sample sub-annotation" id="annotation_2">

      <descProse>
      This is a nested annotation. Here's a reference to the higher-level annotation
      <crossRef ref="annotation_1"/>, and
      here's a reference to the textual exemplar set above <crossRef ref="exemplar_set_1"/>.
      </descProse>

    </annotation>

  </annotation>


  <lexicon>
  An internal lexicon, or reference to an external lexicon, would be placed in this element.
  </lexicon>

  <texts>
  An internal set of texts, or reference to an external set of texts, would be placed in this element.
  </texts>

</grammar>
```

Figure 16: XML fragment for a grammar

A DTD consistent with the XML fragment in figure 16 can be downloaded by clicking here.
(Internet Explorer for Windows users will need to specifically instruct their browser to view the DTD as source.)

The XML fragment in figure 16 formalizes the model discussion in section 3 in the following ways:

- A grammar is understood to consist of four basic kinds of components: ontologies (**<ontology>**), annotations (**<annotation>**), a lexicon (**<lexicon>**), and texts (**<texts>**).
- The annotation is the central element of the grammar, while the other components "support" the annotations.
- Each grammar is assumed to be associated with one lexicon and one set of texts, which can either be internal to the grammar document or be external to it. Hence, both the **<lexicon>** and **<texts>** elements can be specified with a **ref** attribute (not used in figure 16).
- Each grammar can be associated with multiple ontologies. Four kinds of ontologies are specified: general, subcommunity, local, and other. These are specified using the required **level** attribute. Ontologies of the "other" class were not encountered in the survey, but this option, should, perhaps, be included in case there are other classes of ontologies than the three discussed here.
- Annotations can consist of descriptive prose (**<descProse>**), instances of structured description (**<strucDesc>**), exemplar sets (**<exSet>**), references to terms drawn from an

ontology (**<ontRef>**), and nested annotations. Each of these elements can occur multiple times in any order.

- Two types of metadata are permitted for annotations, both encoded as element attributes, an **id** and a **title**.
- Descriptive prose can consist of text as well as references to lexical items (**<lexRef>**), cross references to parts of other annotations (**<crossRef>**), and references to terms drawn from ontologies.
- In the present DTD, structured description is specified as consisting basically of text data and references to terms from ontologies. This is clearly insufficient. However, as discussed above, giving a more detailed specification of possibilities for structured description requires further research. Structured description is associated with **id** attributes for reference purposes. No instance of structured description is given in figure 16 (i.e., there is no instance of a **<strucDesc>** element in the figure).
- An exemplar set can consist of lexical exemplars (**<lexEx>**), text exemplars (**<textEx>**), and references to terms drawn from ontologies. Exemplar sets are associated with **id** attributes for reference purposes.
- References to terms drawn from ontologies, lexical items, cross references, lexical exemplars, and text exemplars all have a **ref** attribute, whose value is an identifier indicating what data they refer to. In addition, references to terms drawn from ontologies must be specified for the **ontologyName** attribute, indicating what ontology they are being drawn from.
- In the DTD employed here, only cross reference elements are specified as having to be empty. All other elements can have character data as part of their content. This may or may not be desirable depending on what implementation is chosen for transforming the base form of a grammar resource into other documents.

As mentioned above, this XML representation is intended to provide only a structure common to more or less all descriptive grammars, rather than being a complete system of representation for any one grammar. Fully representing any existing grammar would, in all likelihood, require defining a number of additional elements.

5. Conclusion and topics for further research

This paper has presented a model, derived from a survey of four printed grammars, of the information found in descriptive grammars wherein they are understood as a series of annotations over a lexicon and texts (figure 13). In addition, it has given a model for the structure of annotations themselves, taking them to consist primarily of descriptive prose, structured description, exemplars, and sub-annotations. The model given for annotations also allows them to contain references to parts of other annotations, to elements in a lexicon or set of texts, and to terms drawn from ontologies. In addition, a possible XML representation for this model was given in section 4.

Section 4 pointed out several ways in which the present model—and the XML representation, in particular—is incomplete. First and foremost, it presupposes that lexicons and texts on which grammatical annotations are based have already been properly modeled. In addition, research is required to determine models for, and representations of, structured description and special annotations on exemplars.

In addition to dealing with these issues, there are several other possible directions for future research on descriptive grammars. Bender et al. (2004) are researching the possibility of bridging the gap between traditional description and formal description so that a machine-readable grammar can be built along side of human readable one. The formalization of grammatical description also provides an excellent test for the utility of ontologies, especially considering that descriptive grammars typically make use of multiple ontologies in an interconnected fashion. Finally, an important area not covered at all here are methods for transforming the basic representation of descriptive grammars provided here into other formats, in particular into human-readable documents.

References

Bender, Emily M., Dan Flickinger, Jeff Good and Ivan A. Sag. 2004. Montage: Leveraging advances in grammar engineering, linguistic ontologies, and markup for the documentation of underdescribed languages. *Proceedings of the Workshop on First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, LREC 2004, Lisbon, Portugal.*

Bow, Cathy, Baden Hughes, and Steven Bird. 2003. Towards a general model of interlinear text. *Proceedings of E-MELD Workshop 2003: Digitizing and Annotating Texts and Field Recordings.* LSA Institute: Lansing MI, USA. July 11–13, 2003. Available at: http://www.emeld.org/workshop/2003/bowbadenbird-paper.html

Comrie, Bernard and Norval Smith. 1977. Lingua descriptive studies: Questionnaire. *Lingua* 42:1–72.

Farrar, Scott and Terry Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT International* 7, 97–100.

Haspelmath, Martin. 1993. A grammar of Lezgian. Berlin: Mouton

Huttar, George L. and Mary L. Huttar. 1994. Ndyuka. London: Routledge.

Maganga, Clement and Thilo C. Schadeberg. 1992. Kinyamwezi: Grammar, texts, vocabulary. Köln: Rüdiger Köppe.

Sperberg-McQueen, C. M., and Lou Burnard (Eds.). 2002. *TEI P4: Guidelines for Electronic Text Encoding and Interchange: XML-compatible edition.* Available at: http://www.tei-c.org/P4X/.

Williamson, Kay. 1965. A grammar of the Kolokuma dialect of Ịjọ. Cambridge: Cambridge University.