The Bantu Verb Phrase: implications for data structure and terminology

Don Salting

North Dakota State University

donald.salting@ndsu.nodak.edu

This paper will offer a preliminary XML Schema data structure for the Bantu Verb Phrase (hereafter BVP). The primary finding is that XML affords a domain-specific data structuring based on a given level of understanding of the domain. This structure can, in turn, serve as an analytical tool towards the development of a more comprehensive structure. The findings will be based on field recordings in OluSaamia[1], but will apply to most Bantu languages. The BVP is, in short, a morphological forest of variables and affords a rich and active arena for analysis. A case is made for a high degree of detail in the coverage. A principal consideration will be the conflicting issues of detail versus manageability for the archivist. In concert with this concern I will address the interface of Bantu-specific categories and terminology with the GOLD ontology and the ontology of the FIELD program. This research indicates that, while issues of universal terminology are fundamental, there are dimensions that are best described in family-specific terms – in this case, noun class.

1. INTRODUCTION

This paper is not an example of depth; it is a time-piece of preliminary knowledge of primary data and its interface with rudimentary knowledge of XML markup and the broader issues of Best Practice. While the detail of knowledge about the language is only at a gross level, a case is made for this methodology in that, with limited time and money resources, a solid start can be made on a project, and the product thereof serve as an extensible platform for further study and detail. To anyone who's interested, it may also serve as an artifact of the learning curve for XML. The ultimate goal is to create a searchable database with access to recordings of any entry.

1.1 Motivation for Subject Area

For the last two years, I have been developing knowledge and skills towards earning research grants to conduct fieldwork on a dialect of the Luyia family – OluXaayo. Reasons for choosing this language are (i) it is undocumented, (ii) I know people from this region and thus, have a point of contact with the community, (ii) I have some knowledge of Bantu linguistics. The work is made even more important by the grim predictions of language death in the coming century. As to OluXaayo's status in the endangered list, it doesn't register on any screens such that one could ascertain its status without actual time spent in the community – a conundrum for grant writing.

The motivation for creating a database as the first goal and to make it accessible to other linguists is borne of the same frustrations and issues mentioned in Bird & Simons (2003). As a phonologist I have studied vowel harmony for a number of years. My primary sources for data have largely been other people's treatments (usually articles) of the phenomenon from a language they had primary or secondary access to and knowledge of. Almost without fail, I experienced what I call the "But what about…" phenomenon created by, what were to me, obviously missing elements in the data. The importance of access to primary data is patently self evident.

[1] OluSaamia is a member of the Luyia family (Guthrie E.34).

1.2 The Language

The abstract mentions research on OluSaamia while the paragraphs above refer to OluXaayo. OluXaayo and OluSaamia are two closely related variants in the Luyia family of western Kenya. Their individual status and their exact relationship are not clear. A year ago, *Ethnologue* cited OluSaamia as a distinct language with its own code, and cited OluXaayo as a dialect of it. Why one is a dialect of the other is not clear – they have similar populations. At this writing, *Ethnologue* cites both as dialects of Luyia. OluSaamia has lost its status as a language.

I am focusing grant applications on OluXaayo because, while there are scattered but substantive bits of research on OluSaamia (Poletto 1998, Marlo 2003a,b and Botne 2004), there appears to be nothing on OluXaayo. Further, a common remark in attempts to rigorously classify the Luyia languages concerns the paucity of data on all the variants (Bennett 1973:6, Mould 1981:186, Kanyoro 1983:24,86). A solid body of OluXaayo data from multiple informants would serve as a valuable resource toward defining the Luyia group.

2. METHODOLOGY

In May of 2003 I was awarded a small startup grant from my university for developing the project to conduct fieldwork in Kenya. The primary informant – Dr. Redempta Kegode – grew up in BuXaayo, but her mother is AbaSaamia. Thus Dr. Kegode is bi-dialectal. When I informed Dr. Kegode of the award, she mentioned that her aunt - an OluSaamia speaker - was visiting for several weeks. Her aunt graciously consented to be recorded as well, so I decided to record them both in OluSaamia for the sake of having multiple informants for one data set. It proved to be a good decision. Firstly, the available time and resources did not allow for sufficient data to attempt a comparison of the two languages, and secondly and most importantly, variations between the informants helped solidify definitions of morpho-phonological elements, contributing to ACCOUNTABILITY as defined in Bird & Simons (2003: 569).

In order to capitalize equally from the two informants, the decision was made to focus on recording as much as possible in the short available time. In addition, the grant funds and personal logistics of the PI and Dr. Kegode were not amenable to regular and extended interview time, the traditional format for fieldwork, so other means of data gathering had to be devised.

2.1 Data Collection

The first problem was finding equipment to ensure high quality recordings. The startup grant did not allow for the purchase of digital recording equipment, and the university did not have a laboratory facility for that. However, a staff member of our university's public radio station, Merrill Piepkorn, graciously allowed us access to their recording studio during off-hours. Recordings were made directly into a computer using *CoolEdit*. The PI later received a faculty development grant for a laptop in January 2004 and subsequent recording has been done with that using the new version of *CoolEdit* -- *Adobe Audition*. Initial recordings had a sample rate of 22khz, and subsequent recordings were at 44khz.

The informants were given lists of prompts in English (both are fluent) which they were to translate into OluSaamia. One source of prompts was the appendix of Kanyoro (1983). These are of two types – individual words and complete sentences. I knew enough about the language and about Bantu in general to know that there would be a very sophisticated and complex Tense/Aspect (hereafter TA) system – both segmental and tonal -- and that the TA system would be just one component of a very complex Verb Phrase morphology. I also decided that being able to study the VP morphology would be of more help for potential fieldwork than compiling a word list. To that end, I hastily compiled sets of Verb Phrase prompts with the goal of getting as many examples of as many forms as possible. A good many distinctions were not elicited, but the resultant data set is a good starting point.

2.2 Data Transcription

To repeat, extensive interview time between the PI and Dr. Kegode was logistically and financially not feasible. Dr. Kegode was working in a post-doc research position as well as raising a young family. The PI was teaching a 3-3 load with committee work and having to maintain a publishing agenda for promotion and tenure considerations. To make the process as convenient and flexible as possible for both, Dr. Kegode was given cdr's of the recordings from which she created orthographic transcriptions. This was a valuable experience of itself in that we developed standards and procedures for future work. When possible, Dr. Kegode would include input forms of contracted elements.

When a set was orthographically transcribed, the PI would create a narrow phonetic transcription using *SoundFileSystems* and *Praat*. Particular attention was paid to tone contours and to segment duration. Naturally, many questions about the language would arise in this process that could only be resolved through interviews with a native speaker. The advantage to the method, however, is that it allowed for very focused and productive interviews, eliminating much of the "How do you say ___?" component. An unexpected outcome of this process was the realization that, given a large corpus of recorded data, initial transcription could be performed by any literate speaker of the target language.

2.3 Initial Database Attempts

With orthographic, phonetic, and gloss text for each token, the next task was to create an informational database that could aid in the analysis and definition of the various morphological components of the VP. Initial attempts were made using FileMaker6 but its inability to interface with Unicode fonts quickly eliminated it as a possible resource. I fell back on Excel 2003 which is usable, but not very flexible for font manipulation and character insertion, though it does allow other fonts. One inconvenience of Excel was the fact that all possibilities, whether optional or obligatory, had to be represented in the initial structure, making for a long and unwieldy line of data. XML has proven superior in its flexibility and extensibility.

3. THE BANTU VERB PHRASE

As mentioned above, the BVP has both properties of a single word and properties of a sentence. Examples of each are given in (1-2) respectively[2].

(1) Joe arisire ngombe amadimwa ‘*Joe fed the cattle corn*’ [aríísir↓é]
Joe fed cattle corn

[a - 0 - r - íís - ir - ↓é]
SP-TA-Root -DSuf-TA2 -FV
3ps -HP - *eat* - CAUS - HP -HP

[2] Unless stipulated, all data in this paper are of OluSaamia from personal field recordings. Also, High tone vowels are marked with the accent, unmarked vowels are Low. The ↓ indicates downdrift.

(2) Hasiyamudekhere ‘*s/he did not cook for her/him (yesterday*)’ [xasíjam↓údeex↓éére]

xasí - j - a - m↓ú - deex - ↓éér - e
NEG - SP - TA - OP - Root - DSUF -FV
NEG -3ps(9)-PHP- 3ps - *cook* - APP - PHP

Given the ambiguity of the BVP vis á vis the terms “word” and “sentence”, it seems logical to categorize it by its standard term: a phrase.

3.1 The Morphemes as variables: an overview

Bantu languages are agglutinating, and each morpheme, at some level, occupies a discrete slice of, or slot in, the larger word. For the purposes of this paper, and as a reflection of my current knowledge of the language, the data-structure will address eight possible morphemes. This is simplified, but again, the extensibility of the format will allow for easy expansion when the information is available. In (1,2) above, the center line lists the morphemes and their possible sequence. Given the items in (1,2) above, a possible morpheme structure sequence would be as in Table.1:

Table 1. Verb Phrase Morpheme Sequence

Sequence	Symbol	Name	Status	Comments
1	NEG	Negative	Optional	minOccurs=”0”, fixed form ‘xasí’
2	SP	Subject Pronoun	Obligatory	one from a set
3	TA1	TenseAspect	Obligatory	one from a set; informs TA2 and FV
4	OP	ObjectPronoun	Optional	Min=0; Max=2, from a set
5	Root	VerbRoot	Obligatory	simple string
6	DSuf	Derivational Suffix	Optional	marks Voice; Min=0; max=3
7	TA2	TenseAspect2	Optional	see §3.8 below
8	FV	FinalVowel	Obligatory	can mark Aspect or Mood, default=’a’

As variables, both the SP and TA1, while obligatory, can contain “0” as their input and output form. The absence of a form indicates a specific category, however – 1ps and HodiernalPast respectively. Another way of arraying the above information might be as in (3) below:

(3) (NEG) + SA + TA1 + (OP) + Root + (DS)+(DS+(DS))) + TA2 + FV
1 2 3 4 5 6 7 8

The arrays above are based on my current understanding of my data, and are simpler than reality. The array for Luyia languages in general in Kanyoro (1983) contains ten slots and is given in (4) below. The elements not addressed above are highlighted in yellow. Slot 9 (ASPect) corresponds to TA2 above, and slot 10 (MODE) corresponds to FV. The choice of the term Aspect for the later slot is not unfounded, and, all in all, seems to be as equally valid and incomplete as “TA2.”

(4) From Kanyoro (1983)

(NEG^{i})+(TN^{i})+(SA)+(NEG^{ii})+(TN^{ii})+(OP)+ROOT+(DS)+(DS^{ii})+(DS^{iii})+(ASP)+MODE
(REL) (INF)
1 2 3 4 5 6 7 ←------- 8 ------→ 9 10

Based on available information, this study will discuss the simpler array as in (3). When data and knowledge are sufficient to expand the markup closer to the array in (4), there will be a foundation on which to build. In the next sections I address each of the morphemes as variables and offer suggestions for their representation in an XML markup. Before doing so, however, I offer a potential mockup of a markup for the Morphology component of a BVP database:

(5) BVP Morphology Markup

```
<xs:element name=”morphology”>
    <xs:complexType mixed=”true”>
        <xs:sequence>
            <xs:element name=”NEG” type=”xs:string” fixed=”xasí” minOccurs=”0”/>
        +   <xs:element name=”SubjectPronoun” type=”pronounType”>
        +   <xs:complexType name=”TenseAspect”>
        +   <xs:element name=”SubjectPronoun” type=”pronounType” minOccurs=”0”
                maxOccurs=”2”>
            <xs:element name=”VerbRoot” type=”xs:string”/>
        +   <xs: complexType =”DerivSuff” type=”Voicetype” minOccurs=”0” maxOccurs=”3”>
            <xs:element name=”Voice” substitutionGroup=”DerivSuff”>
        +   <xs:complexType name=”TenseAspect2” minOccurs=”0”>
        +   <xs:complexType name=”FinalVowel”>
            <xs:element name=”notes” type=”xs:string” minOccurs=”0”/>
        </xs:sequence>
    </xs:mixedType>
</xs:element>
```

In the next sections, I attempt an XML markup for the BVP slots above. The primary focus of this paper will be the Subject and Object Pronoun categories and they are addressed in §4 below. In §5 I address the other slots. Some slots are quite straightforward and require little explication. Interestingly, these are the least varied – NEG – and the most varied – VerbRoot. The Tense Aspect system is very complex and will only be introduced along with some implications for markup.

4. PRONOUNS: Subject and Object

4.1 Terminology

The area treated in the greatest depth in this paper will be Subject and Object Agreement. The first issue to address is one of terminology. For both Subjects and Objects, the marker can serve as a pronoun for a referent in another clause. However, for object NP’s, if the referent is in the same clause, an Object Pronoun is disallowed (Mutonyi 2000:39)[3]. In terms of data structure, this renders the Object Pronoun optional (minOccurs=”0”). On the other hand, regardless of the physical presence or absence of a referent NP, the Subject Pronoun is obligatory. This means that, while the Object Marker can be considered a pronoun and left at that, the Subject Marker is sometimes a pronoun, and sometimes an agreement marker (not unlike the ‘s’ in ‘he runs’). Both

[3] Mutonyi (2000) describes Bukusu, a near but not immediate relative in the Luyia family. This page also states that more than one Object marker is disallowed, but my Saamia data have instances of two Object Markers.

Mutonyi (2000) and Kanyoro (1983) use the shorthand symbol *P* for Pronoun in labeling the slots (4 above), so, for the sake of consistency, I will adopt their term and refer to the slots as SP and OP for Subject Pronoun and Object Pronoun respectively.

The next issue to address is the fact that Bantu languages have extensive Noun Class systems. This dimension would fall under the GOLD category of Gender. For the most part, a given noun will belong to two classes: one for its singular (7a) and one for its plural (7b), and will usually carry the prefix associated with its class:

(7) (a) omu-ndu '*person*' (b) aβa-ndu '*people*' (c) esi-ndu '*thing*' (d) eβi-ndu '*things*'
1 – entity 2 – entity 7-entity 8-entity

While some noun classes contain a preponderance of one type of noun – Class 1/2 is reserved for humans and some animals – most noun classes are not definable in semantic terms[4]. Within Bantu studies, however, a standard has evolved which uses numbers, and these numbers are generally migratable. That is, if one is discussing Class 5 nouns in a given Bantu language, the structure and general makeup will transfer to studies of other Bantu languages[5]. I offer an example of this migratability in §4.2 below, but before doing so, I offer an explication of the possible pronouns for OluSaamia in Table 2 below.

4.2 Pronouns as Variables

As variables, the Subject and Object Pronouns comprise a set of fixed elements, and one from that set will be chosen for the appropriate slot in the BVP. This is complicated by the fact that, while most Subject and Object forms are the same for a given noun class, several have different forms depending on the slot. These are highlighted in Table 2. For markup, the issue is, do we create a separate set of all appropriate forms for each of the Pronoun slots, or do we create one base of the regular forms and extend it with slot-specific forms?

There is an additional wrinkle in OluSaamia Subject pronouns. The 3ps/Class 1 subject prefix is /a-/. A number of Tense markers (the next slot) consist of or start with /a/ as well, which leads to input strings of /aa/ or more, creating potential ambiguity. To compensate for this, OluSaamia will substitute the Class 9 prefix, giving /y-a/, thus eliminating ambiguity. The substitution of Class 1 with Class 9 is not uncommon in Bantu languages[6], an example of the, if not universality, 'familiality' of the noun classes.

[4] Exceptions would be the prepositional classes ('in', 'on', 'at,to') and the abstract class (see Table 2).

[5] This is not to say that a given noun will always be in the same class from language to language. Nonetheless, general properties obtain across languages.

[6] Larry Hyman, personal communication.

Table 2. OluSaamia Subject and Object Pronouns

CLASS	Subj.	Obj.		
1ps	ndi-	-ndi-		
1pp	xu-	-xu-		
2ps	o-	-xu-		
2pp	mu-	-mu-		
1/3ps[7]	a-	-mu-		
2/3pp	βa-	-βa-		
3	ku-	-ku-		
4	ki-	-ki-		
5	li-	-li-		
6	ka-	-ka-		
7	si-	-si-		
8	βi-	-βi-		
9	yi-	-i-		
10	chi-	-tsi-		
11	lu-	-lu-	11/10	things having length
12	xa-	-xa-		dim. or derog.
13	tu-	-ru-		
14	βu-	-βu-	14/6	abstract
15	xu-	-xu-	15/6	inf. / gerund
16	ha-	-ha-		loc. 'at'
17	xu-	-xu-		loc. 'on'
18	mu-	-mu-		loc. 'in, into'
19				
20	ku-	-ku-	20/4	aug., sometimes derog.

4.3 Data Categories for Pronouns

The theta-role[8] of the Pronoun can be significant in analyzing Bantu. Notice in (8) below that the vowel of the penultimate syllable is longer in (a) than in (b). This marks a significant difference in the meaning of the two utterances. Item (a) means '*She cooked for him*' while item (b) means '*She cooked him*'. The object marker /-mú-/ in (a) has a Benefactive theta-role, while the same in (b) is a Patient.

(8) a. [j-a-mú-deex-↓ééré] b. [j-a-mú-deex-↓éré]

Segmentally and tonally, the string /-ér-/ could be either a TenseAspect marker or a Derivational Suffix ('Applied') indicating a Benefactive or Recipient theta-role. The two potentially ambiguous suffixes can not be agglutinated, however, as this would render a different

[7] In that Class 1/2 is reserved almost exclusively for humans, the singular (Class 1) covers the same semantic domain as *s/he* in English, and the plural analogy obtains for Class 2 and *they*. Thus, the two are collapsed.

[8] For this paper, I will use theta-role categories for the pronouns, but with the understanding that they are interchangeable with Grammatical case categories in GOLD.

Derivational Suffix indicating intensive and repetitive action. The very necessary acoustic distinction for the semantic difference resides in the vowel length. Thus a case is made for including theta-role in the morphological description, as well as for detail in the phonetic transcription.

That said, the data elements for each of the pronoun slots should be: (i) The Theta role of the referent, (ii) the Class of the referent, and (iii) the form of the Class marker appropriate to the slot. In the next section I offer a possible markup for each of the slots. While each slot will be able to draw on a uniform set of possible Class markers, we will see that the basic design of the markup for the two slots will be fundamentally different.

4.4 Common Elements for Subject and Object

As we saw in Table 2., the forms for most class markers are the same for subject and object slots. I will consider these to be unmarked, and to reside in a set that is accessible for both slots. They reside in a group and are given in (9) below.

(9) <! – Unmarked, Regular Pronouns -- >

```
<xs:group name=”RegProAgr”>
    <xs:choice>
        <xs:element name=”1ps” type=”xs:string” fixed=”ndi”/>
        <xs:element name=”1pp” type=”xs:string” fixed=”xu”/>
        <xs:element name=”2pp” type=”xs:string” fixed=”mu”/>
        <xs:element name=”3ppClass2” type=”xs:string” fixed=”βa”/>
        <xs:element name=”Class3” type=”xs:string” fixed=”ku”/>
        <xs:element name=”Class4” type=”xs:string” fixed=”ki”/>
        <xs:element name=”Class5” type=”xs:string” fixed=”li”/>
        <xs:element name=”Class6” type=”xs:string” fixed=”ka”/>
        <xs:element name=”Class7” type=”xs:string” fixed=”si”/>
        <xs:element name=”Class8” type=”xs:string” fixed=”βi”/>
        <xs:element name=”Class11(10)” type=”xs:string” fixed=”lu”/>
        <xs:element name=”Class12dim.” type=”xs:string” fixed=”xa”/>
        <xs:element name=”Class14(6)abstract” type=”xs:string” fixed=”βu”/>
        <xs:element name=”Class15(6)inf.” type=”xs:string” fixed=”xu”/>
        <xs:element name=”Class16‘at’” type=”xs:string” fixed=”ha”/>
        <xs:element name=”Class17‘on’” type=”xs:string” fixed=”xu”/>
        <xs:element name=”Class18‘into’” type=”xs:string” fixed=”mu”/>
        <xs:element name=”Class20(4)aug.” type=”xs:string” fixed=”ku”/>
    </xs:choice>
</xs:group>
```

4.5 The Subject Slot

As we saw in Table 2., several Classes have specific forms for Subject versus Object slots. We will have to include these with the items in (9) above for the marker. Unlike the Object slot, the Subject slot’s theta-role specification will not directly specify any given Class member. Thus, the selections of theta-role and class marker will be separate, sequential choices. A possible markup is offered in (10):

(10) <!—SUBJECT AGREEMENT -- >

```
<xs:complexType name=”SubjectPronoun”>
    <xs:complexContent>
        <xs:sequence>
            <xs:complexType name=”ThetaRole”>
                <xs:choice>
                    <xs:element name=”Agent” type=”xs:string” fixed=”Agent”/>
                    <xs:element name=”Patient” type=”xs:string” fixed=”Patient”/>
                </xs:choice>
            </xs:complexType>
            <xs:complexType name=”SubjectAgree”>
                <xs:choice>
                    <xs:group ref=”RegProAgr”>
                    <xs:element name=”2psSubj” type=”xs:string” fixed=”o”>
                    <xs:element name=”3psClass1Subj”>
                        <xs:complexType>
                            <xs:choice>
                                <xs:element name=”Class1Subj” type=”xs:string” fixed=”a”/>
                                <xs:element name=”Class9Subj” type=”xs:string” fixed=”yi”/>
                            </xs:choice>
                        </xs:complexType>
                    </xs:element>
                    <xs:element name=”Class9Subj” type=”xs:string” fixed=”yi”/>
                    <xs:element name=”Class10Subj” type=”xs:string” fixed=”chi”/>
                    <xs:element name=”Class13Subj” type=”xs:string” fixed=”tu”/>
                    <xs:element name=”Class 14 (6) ‘abstract’” type=”xs:string” fixed=”βu”/>
                </xs:choice>
            </xs:complexType>
        </xs:sequence>
    </xs:complexContent>
</xs:complexType>
```

The user would be given choices for the theta-role and for the Class of the Subject; the form would automatically be filled in. Notice in the “SubjectAgree” section that, if 3ps is chosen, it will involve a further choice depending on the segmental nature of the TenseAspect marker (§4.2 above.). With programming, this choice could be made automatically in a situation where the user chose 3ps for the class marker and one of a specified set of tenses.

Consideration was given to creating a group for the theta-roles as well, but as we shall see in the Object Pronoun section, that has a wrinkle that, at this point in my skills with markup, cannot be manipulated.

4.6 The Object Slot

The specifications for the Object slot are the same as for the Subject slot: theta-role, noun class, and form. As with the Subject slot, the Object slot can also access the group of regular forms. For this paper, we will address three possible theta-roles in the Object slot – Patient, Benefactive and Reflexive. Patient and Benefactive, like the Theta roles for the Subject slot, both access the same set of Class markers and forms. To facilitate this, I have created a new group of Object Pronouns “ObjProAgr” (11 below) which ref’s the Regular set and introduces Object-specific forms. This way, each of the possible theta-roles can be explicated with a minimum of lines.

The complication for the Object slot is that there is a permissible theta-role that has its own, fixed form. The Reflexive OP input form is /-(n)e-/. Thus, accommodations within the choice of Theta roles will have to allow for a special class and form for Reflexives, while opening up the entire range of Classes for other possible theta-roles. A further wrinkle is that the Reflexive theta-role would be considered a member of the Voice category in the FIELD verb ontology. Of the categories given for Voice, the majority are marked by the Derivational Suffixes in Bantu (see §5.4 below). This means that the majority of Voices are marked by a suffix, while one is marked by an Object Pronoun, giving one semantic realm several syntactic manifestations. The final treatment of this issue is beyond the scope of this paper, but the reader is referred to Mchombo (1993) for a detailed discussion of this issue in ChiChewa.

For purposes of the markup, the user will first be given a choice of theta-roles as with the Subject Pronoun. If Reflexive is chosen, the value will automatically be selected. If any other theta-role is chosen, the user will then select the class, which will provide the appropriate form for the slot. This is given in (11) below:

(11) <! – OBJECT AGREEMENT-- >

```
<xs:complexType name=”ObjectPronoun” maxOccurs=”2”>
    <xs:complexContent>
        <xs:group name=”ObjProAgr”>
            <xs:choice>
                <xs:element name=”2psObj” type=”xs:string” fixed=”xu”/>
                <xs:element name=”3psClass1Obj” type=”xs:string” fixed=”mu”/>
                <xs:group ref=”RegProAgr”/>
                <xs:element name=” Class9Obj” type=”xs:string” fixed=”i”/>
                <xs:element name=” Class10Obj” type=”xs:string” fixed=”tsi”/>
                <xs:element name=” Class13Obj” type=”xs:string” fixed=”ru”/>
            </xs:choice>
        </xs:group>
        <xs:complexType name=”ThetaRole”>
            <xs:choice>
                <xs:complexType name=”Reflexive”>
                    <xs:sequence>
                        <xs:element name=”ThetaRole” type=”xs:string” fixed=”Reflexive”/>
                        <xs:element name=”ReflexiveAgr” type=”xs:string” fixed=”ne”/>
                    </xs:element>
                </xs:complexType>
                <xs:complexType name=”PatientObj”>
                    <xs:sequence>
                        <xs:element name=”ThetaRole” type=”xs:string” fixed=”Patient”/>
                        <xs:group ref=”ObjProAgr”/>
                    </xs:sequence>
                </xs:complexType>
                <xs:complexType name=”BenefactiveObj”>
                    <xs:sequence>
                        <xs:element name=”ThetaRole” type=”xs:string” fixed=”Benefactive”/>
                        <xs:group ref=”ObjProAgr”/>
                    </xs:sequence>
                </xs:complexType>
            </xs:choice>
         </xs:complexType>
      </xs:complexContent>
</xs:complexType>
```

For the Object slot, the Class set is determined by the theta-role choice. However, as near as I can tell, the data entry process should be the same for Object and Subject slots.

In the two sections above, I have attempted to configure a structure that will account for all significant detail with minimal lines of script and in a manner that is consistent between the two slots. In addition, I have attempted to minimally represent unmarked forms in the hope of possibly capturing some linguistic reality around retrieval dynamics. In that the possible entries for all the categories in this section are members of a fixed set, the <choice> function offers a ready tool for the database. The above is a rough and first attempt at this, and I invite comments and suggestions, especially around choice between Types, Elements and Groups.

5. Other Slots

5.1 Negative

As a variable, NEG is quite straightforward. It is optional, and when it occurs has an invariant input shape /xasí/. In Schema markup, it could be represented as in (12):

(12) <xs:element name=”NEG” type=”xs:string” fixed=”xasí” minOccurs=”0”/>

5.2 Tense / Aspect and the Final Vowel

This domain is probably the most complex and will be the least treated in this paper for several reasons. While my research affords a rudimentary understanding of the intricacies of the OluSaamia Tense/Aspect system (hereafter TA), a definitive explication and incorporation into XML is well beyond my current skills. I do, however, look forward to the challenge when the time comes.

As a brief example, OluSaamia has six distinct Tenses, three past and three future[9]. However, in terms of possible grammatical time references, the number of possibilities is somewhere between eight (Kanyoro 1983: 105) and eighteen (Mould 1981: 206; Mutonyi 2000: 48). The segmental components of the three past tenses, following Mutonyi (2000: 48) are given in Table 3., below, with their morphological position.

Table 3. OluSaamia Past Tenses

Tense ↓SLOT →	**TA1**	**TA2**	**FV**
Hodiernal Past	0	-ir- /-er-[10]	-e
Pre-hodiernal Past	-a-	-ir- / -er-	-e
Remote Past	-á-	0	0

The description in Table 3 can be misleading in that the surface forms do not always exemplify the discrete slices indicated. The verb /deexa/ ‘cook’, adheres to the pattern, and the

[9] Simple Present tense is marked by a 0 morpheme, but the absence of a morpheme is still an indicator of tense.

[10] These forms will vary as products of Vowel Harmony with the Verb Root vowel. Roots with mid vowels will exhibit the -er- form, while roots with high or low vowels will exhibit the -ir- form.

TA2 forms will segmentally manifest as /deex-er-e/. Other verbs, however, will internalize the morphemes. Also, each Past Tense has its own tone pattern. Examples of the Past tenses for two phrases are given below. In the first set, the Subject Pronoun is /ndi-/ ‘I’, and the verb root is /βwaao/, ‘leave’. All three would be translated as ‘I left’.

Table 4. ‘I left’. Verb root= /βwaao/ ‘leave’

TENSE ↓	**Orthographic**	**Phonetic**	**Gloss**
Hodiernal Past (HP)	mbwereo	m-bw-éer-↓é-o	*I left*
Pre-hodiernal Past (PHP)	ndabwereo	nd-a-βw-eér-é-o	*I left*
Remote Past (RP)	ndabwao	nd-á-βwááo	*I left*

In Table 5. the root verb is /βona/ ‘see’ with an Object Pronoun /-lu-/. Here we also see the substitution of the Class 9 prefix /j/ for the 3ps /a/.

Table 5. ‘She saw it’. Verb root=/βon/ ‘see’

TENSE ↓	**Orthographic**	**Phonetic**	**Gloss**
Hodiernal Past (HP)	alubwene	a-lú-bween↓é	*She saw it*
Pre-hodiernal Past (PHP)	yalubwene	j-a-lú-βwééné	*She saw it*
Remote Past (RP)	yalubona	j-á-ĺu-β↓óna	*She saw it*

Citing the many phonological shapes that can obtain for this morpheme Mould (1981: 204) uses the term Modified Base to refer to the TA2 and FV elements in Table 3. It seems advisable given the scant data I have. Also, the TA components /-ir-/ and /-e/ always occur together as a single unit. Kanyoro (1983) labels the TA2 slot Aspect, and this is not unfounded. For example, Habitual aspect is indicated with the suffix /-angá/. However, certain non-tense aspect markers occur in the TA1 slot as well, e.g. *–si-* ‘still’.

In sum, certain TA2 elements (HP, PHP) are determined by the status of the TA1 slot, and need not necessarily require a second specification, while others are independent of TA1 (Habitual). Further, certain TA1 elements are not tense but aspect. This issue speaks to criteria for labeling and seems to parallel Hopi (Farrar et al. 2002). Ideally, a markup would, when applicable, allow for a specification (<choice>) for TA1 apply to TA2 and FV. At this point it appears that treating this morpheme will require extending the Schema, which is a skill I leave for future work.

However, the greatest challenge will be to configure the tonal specifications for each Tense. While each Tense has a tone configuration associated with it, the final location of the tones will depend on myriad factors. As is shown in (13) below, the inclusion of an OP will affect the location of the tones:

(13) a. [a-bẃéen$^{\downarrow}$é] ‘she saw’ b. [a-lú-bween$^{\downarrow}$é] ‘she saw it’

The initial goal of my creating this database was to be able to study tonal patterns within and across tenses to develop a description of them. As I write this, it becomes apparent that an initial database with the Pronouns, Derivational Suffixes and a simple Tense label could serve as a the analytical starting point for a later extended structure which offers more detail of the Tense /Aspect system. That is, this practice accommodates its own development.

5.3 Verb Root

The element with the greatest variety proves the simplest to markup:

(14) <xs:element name="VerbRoot" type="xs:string"/>

It's instance appearance would be: <VerbRoot>deex</VerbRoot>

An additional gloss element could be appropriate as well.

5.4 Voice: Derivational Suffixes

The final slot I briefly address is commonly referred to in Bantu studies as the Derivational Suffix (hereafter DS) category, though it has other names[11]. By and large, it corresponds with the Voice category in the FIELD verb configuration. Examples include Causative (1 above), Reciprocal, Middle[12], Reversive, et al. Of interest for application of this paradigm to other Bantu languages is the suffix referred to in Bantu studies as the Applied. For OluSaamia, it corresponds closely to the Benefactive voice, and a substitutionGroup would allow intuitive study for both Bantuists and non-Bantuists. However, for other Bantu languages, the Applied Suffix can mark other roles in addition to Benefactive. For example, In ChiChewa, it can mark Instrumental case in the noun immediately following the BVP (Alsina & Mchombo 1993). This could be treated with a language specific assertion regarding the suffix. The DerivationalSuffix component of Bantu morphology will offer a rich venue for the development of structures that can, as closely as possible, access universal terminology while capturing the family- and language-specific intricacies they contain. A brief sampling of the 11 – 13 possible Derivational Suffixes for Olusaamia is given in (15) below:

(15) OluSaamia Derivational Suffixes:

	ROOT – DS – FV	
a. Applied	deex – er – a	'cook for smbdy'
b. Causative	r – is – a	'feed' (cause to eat)'
c. Reciprocal	βon – an – a	'see each other'
d. Middle	βon – ex – a	'be visible'
e. Passive	βon – erw – a	'be seen'
f. Reversive	fuumb – ul – a	'close' *vt.* (un-open) (from Marlo 2003)

In addition, these morphemes can be concatenated in certain sequences:

(16) r – is – irw – a 'be fed' (be caused to eat)
eat-CAUS-PASS-FV

There are restrictions on allowable sequences, and they are beyond the scope of this paper and my knowledge[13]. Nonetheless, as a starting point, we can say that the Derivational Suffixes present a fixed set which can occur between 0 and 3 times. A preliminary markup is given in (17):

[11] "Thematic extensions" in Mutonyi (2000: 69) and "thematics" Kanyoro (1983: 112).

[12] Often referred to as 'Stative' in Bantu studies.

[13] See Mutonyi (2000: 85) for discussion of this dimension in Bukusu.

(17) <! – Derivational Suffixes-- >

```
<xs:complexType name=”DerivSuff” minOccurs=”0” maxOccurs=”3”>
    <xs:choice>
        <xs:element name=”Applied” type=”xs:string” fixed=”ir-er”/>
        <xs:element name=”Causative” type=”xs:string” fixed=”is-es”/>
        <xs:element name=”Passive” type=”xs:string” fixed=”irw-erw”/>
        <xs:element name=”Reversive” type=”xs:string” fixed=”ul-ol”/>
        <xs:element name=”Middle” type=”xs:string” fixed=”ix-ex”/>
    </xs:choice>
</xs:complexType>
```

Conditions on allowable sequences will offer an interesting challenge to a more comprehensive markup. The open-ended makeup of the above schema, however, will allow for the creation of a database that will illustrate the possible sequences. Again, the practice accommodates its own development.

6. Conclusion

An exciting component of XML markup is the ability to fashion the structure around the data, rather than shaping the data to the structure, as with over-the-counter databases. An unanticipated result of this project has been the realization that XML markup allows for the creation of a domain specific structure which will serve as a platform for further analysis, thus facilitating the creation of a more comprehensive data structure. That is, the user need only bring current knowledge of the domain – the analysis afforded by the initial markup will foster and feed the next level of detail and comprehensiveness.

As with the specific structures contained in this paper, it is hoped that this paper itself will serve as a starting point for the development of extensible but constrained structures for the study of Bantu morphology. Issues that will bear further consideration include terminology for family-specific domains (noun classes) in a universal ontology. For appropriate detail, it appears that allowances will have to be made to the language family for the sake of specificity.

TO DO: Understand Namespace component. Learn how to validate script. ~~Request dedicated data repository site from university~~. (done)

References **Bantu Verb Phrase**

Alsina, A. & S. A. Mchombo. (1993). Object asymmetries and the applicative construction. In S.A. Mchombo (Ed.), *Theoretical Aspects of Bantu Grammar* (pp. 17-45). Stanford: CSLI.

Bennett, Patrick. 1973. A Phonological History of Northeast Victoria. Paper read at the 4th Annual Conference on African Linguistics. New York. Ms.

Bird, S. and G. Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79: 557-582.

Blois, K.F. de. 1975. *Bukusu Generative Phonology and Aspects of Bantu Structure*. Musee Royal de L’Afrique: Tervuren, Belgique.

Botne, R. 2004. Specificity in Lusaamia infinitives. *Studies in Language* 28:1 (2004), 137–164.

Dalgish, Gerard M. 1976. *The Morphophonemics of the Olutsootso Dialect of (Olu)Luyia; issues and implications*. Ph.D. dissertation, University of Illinois at Urbana-Champaign.

Davy, J.I.M. & Derek Nurse. 1982. Synchronic Versions of Dahl’s Law: the multiple application of a phonological dissimilation rule. *Journal of African Languages and Linguistics*. 4:157-195.

Ethnologue. http://www.ethnologue.com/ Sponsored by SIL International.

Farrar, S., William, D.L., and D. T. Langendoen. 2002. A common ontology for linguistic concepts. In *Proceedings of the Knowledge Technologies Conference*, Seattle, Washington, March 10-13, 2002.

Goldsmith, John. 1992. Tone and accent in Llogoori. In Brentari, Diane et al. (eds.) *The Joy of Grammar: a festschrift in honor of James D. McCawley*. Benjamins: Amsterdam.

Guthrie, Malcom. 1970. *Comparative Bantu*. Gregg International Pub. Ltd.: Hants, England.

Hinnebusch, Thomas H., Derek Nurse & Martin Mould. 1981. *Studies in the Classification of Eastern Bantu Languages*. Helmut Buske Verlag: Hamburg.

Hyman, Larry & Charles W. Kisseberth (eds.). 1998. *Theoretical Aspects of Bantu Tone*. Center for the Study of Language and Information: Stanford CA.

Kanyoro, Rachel Angogo. 1983. *Unity in Diversity: a linguistic survey of the Abaluyia of Western Kenya.* Beiträge zur Afrikanistik: Vienna.

Marlo, Michael. 2003a. Lusaamia verb database. Ms. Indiana University.

Marlo, Michael. 2003b. On the status of word-initial NCs in Lusaamia: conflicting evidence from native speaker intuitions and phonological patterns. Paper given at MCWOP 9, Champaign-Urbana.

Mchombo, S.A. 1993. On the binding of the reflexive and reciprocal in Chichewa. In S.A. Mchombo (Ed.), *Theoretical Aspects of Bantu Grammar* (pp. 181-207). Stanford: CSLI.

Mould, Martin. 1981. Greater Luyia. In Hinnebusch, Thomas H., Derek Nurse & Martin Mould. 1981: 181-235.

Nasiombe, Mutonyi. 2000. *Aspects of Bukusu Morphology and Phonology*. PhD. dissertation. Ohio State University. *Unavailable through library*.

Nurse, Derek & G. Philippson. (1980). Bantu languages of East Africa: a lexicostatistical survey. In Polomé, E.C. & C.P. Hills (eds.) 1980: 26-67.

Poletto, Robert. 1998. Tonal Association in Olusamia. In Hyman, L. & C. Kisseberth (Eds.). 1998: 331-364.

Salting, Don. 1994. ChiChewa conflict resolution. Ms. from *American Conference on African Linguistics*, Rutgers University, March 26.

Salting, Don. 1992. ChiThumbuka vowel harmony. Ms. from *Workshop on Underspecification*, Ohio State University.

Sample, Ward. 1974. The applied extension with dative and benefactive implications in Llogooli. In *Mila: A Biannual Newsletter of Cultural Research*. 4(2): 12-22. Nairobi, Kenya.