Describing Morphologically-rich Languages using Metagrammars: a Look at Verbs in Ikota

Denys Duchier[1], Brunelle Magnana Ekoukou[2], Yannick Parmentier[1], Simon Petitjean[1], Emmanuel Schang[2]

(1) LIFO, Université d'Orléans - 6, rue Léonard de Vinci 45067 Orléans Cedex 2 – France
(2) LLL, Université d'Orléans - 10, rue de Tours 45067 Orléans Cedex 2 – France

prenom.nom@univ-orleans.fr

Abstract

In this paper, we show how the concept of metagrammar originally introduced by Candito (1996) to design large Tree-Adjoining Grammars describing the syntax of French and Italian, can be used to describe the morphology of Ikota, a Bantu language spoken in Gabon. Here, we make use of the expressivity of the XMG (eXtensible MetaGrammar) formalism to describe the morphological variations of verbs in Ikota. This XMG specification captures generalizations over these morphological variations. In order to produce the inflected forms, one can compile the XMG specification, and save the resulting electronic lexicon in an XML file, thus favorising its reuse in dedicated applications.

1. Introduction

Bantu languages form a large family of languages in Africa. In this family, Chichewa and Swahili are the most well-studied, and are used as benchmarks for assessing the expressivity and relevance of morphological theories (Mchombo, 1998; Stump, 1992; Stump, 1998; Stump, 2001) and their implementation (Roark and Sproat, 2007). Ikota (B25) is a lesser-known language of Gabon and the Democratic Republic of Congo. Language of the Bakota people, with an estimated 25000 speakers in Gabon (Idiata, 2007), Ikota is threatened with extinction mainly because of its abandon for French (the official language of Gabon). It manifests many grammatical features shared by the Bantu languages (Piron, 1990; Magnana Ekoukou, 2010):

- Ikota is a *tonal language* with two registers (High and Low):

  (1) a. ìkàká "family"
      b. ìkákà "palm"

  (2) a. nkúlá "year"
      b. nkúlà "pygmee"

- Ikota has ten *noun classes*,[1] see Table 1.

- Ikota has a *widespread agreement in the NP*:

  (3) **b**-àyítò **bá**-nɛ́nì **b**-á Ø-mbókà **bà**-tɛ́ **b**-à-ɠá
      2-women 2-fat 2-of 9-village 2-DEM 2-Prst-eat

      "These fat women of the village are eating"

- Yet, unlike Swahili for instance, Ikota does not have a slot for object agreement.

In this paper, we will consider verbal morphology.

[1]The number of the class in the table corresponds to Meinhof's numbering.

Table 1: Ikota's noun classes

Noun class	prefix	allomorphs
CL 1	mò-, Ø-	mw-, ǹ-
CL 2	bà-	b-
CL 3	mò-, Ø-	mw-, ǹ-
CL 4	mè-	
CL 5	ì-, ɠ-	dy-
CL 6	mà-	m-
CL 7	è-	
CL 8	bè-	
CL 9	Ø-	
CL 14	ò-, bò-	bw-

Production of a lexicon of inflected forms. Our purpose is twofold: first to provide a formal description of the morphology of verbs in Ikota; second, to automatically derive from this description a lexicon of inflected forms. To do so, we propose to adopt the concept of a metagrammar, which was introduced by (Candito, 1996), and used to describe the syntax of Indo-European languages, such as French, English or Italian. Lexicalized wide-coverage tree-grammars for natural languages are very large and extremely resource intensive to develop and maintain. For this reason, they are often automatically produced by software from a highly modular formal description called a metagrammar. The metagrammar is much easier to develop and to maintain.
We propose to adopt a similar strategy to capture morphological generalizations over verbs in Ikota. The outline of the paper is the following. In Section 2., we give a detailed presentation of the morphology of verbs in Ikota. Then, in Section 3., we introduce eXtensible MetaGrammar (XMG), a formal language, used to describe and combine reusable descriptive fragments. In Section 4., we show how to use the XMG framework to describe the morphology of verbs in Ikota. Concretely, we present a metagrammar of verbs in Ikota, which we have also coded in the XMG language, and which can be automatically processed to produce a lexicon

of fully inflected verb forms in Ikota. Finally, in Section 5., we conclude and present future work.

2. Verbs in Ikota

Verbs are constituted by a lexical root (VR) and several affixes distributed on each side of the VR. For the sake of clarity, we will focus here on the basic verbal forms, leaving aside Mood and Voice markers.
Let us now describe infinitival form and the three verbal classes of Ikota.
Verbs in Ikota are distributed in three classes depending on the form of Aspect and Active markers. Infinitive in Ikota is a hybrid word class. It is composed of a noun class prefix (class 14) and a verbal element (VR+Prog+Active).

(4) a. bòʤákà "to eat"
b. bòwétʃè "to give"
c. bòbɔ́nɔ́kɔ̀ "to choose"

Examples (4) illustrate the three verb classes.
Indeed, it seems that the suffix (Prog+Active) has a subjacent form VkV. In the Makokou variant of Ikota, /k/ is realized by [ʧ] when the vowel is [ɛ]. In Standard Ikota, the form is ékè.
At a subjacent level, the structure of the infinitival suffix boils down to AKA, with three distinct surface realizations ákà, étʃè, ɔ́kɔ̀.
Examples below illustrate the conjugation of bòʤákà "to eat", a typical example of the *aka* verb class:

(5) m-à-ʤ-á ǹlɛ́sì
1sg-Prst-eat-Act rice
"I'm eating rice" (Present)

(6) a. m-à-ʤ-á-ná yàná
1sg-Past-eat-Act-Prox yesterday
"I ate yesterday" (Past (yesterday))

b. m-à-ʤ-á-sá kúlá mwáyèkànàmwɛ́
1sg-Past-eat-Act-Prox year last
"I ate last year" (Distant Past)

c. m-é-ʤ-á ǹlɛ́sì
1sg-Past-eat-Act rice
"I ate rice" (Recent Past)

(7) a. m-é-ʤ-àk-à ǹlɛ́sì
1sg-Fut-eat-Asp-Act rice
"I'll eat rice" (Medium Future)

b. m-é-ʤ-àk-à-ná yàná
1sg-Fut-eat-Asp-Act-Prox tomorrow
"I'll eat tomorrow" (Future (tomorrow))

c. m-é-ʤ-àk-à-sá kúlá
1sg-Fut-eat-Asp-Act-Prox year
mwáyàkàmwɛ́
next
"I'll eat next year" (Distant Future)

d. m-ábí-ʤ-àk-à òsátè
1sg-Fut-eat-Asp-Act soon
"I'll eat soon" (Imminent Future)

As can be deduced from the examples above, Ikota's verbal affixes ordering can be defined as position classes. From the left to the right:

- the class of Subject agreement prefixes occupies the leftmost, word-initial position.
- Tense prefixes (or what can roughly identified as related to Tense) appears at the left of VR.
- the (aspectual) progressive marker is on the immediate right of VR.
- Active suffix occupies the slot to the left of Proximal. It has two values: Active and Passive (to wit: -Active). Applicative and Causative are kept for further studies.
- the Proximal/Distal suffixes occupy the rightmost position.

Table 3 gives an outline of the VR and its affixes and table 2 exemplifies this schema with bòʤákà "to eat".

Table 3: Verb formation

Subj-	Tense-	VR	-(Aspect)	-Active	-(Proximal)

3. eXtensible MetaGrammar

eXtensible MetaGrammar (XMG) refers both to a formal language (*a kind of* programming language) and a piece of software, called a compiler, that processes descriptions written in the XMG language (Crabbé and Duchier, 2004). XMG is normally used to describe lexicalized tree grammars. In other words, an XMG specification is a declarative description of the tree-structures composing a grammar. This description relies on four main concepts: (1) **abstraction**: the ability to associate a content with a name, (2) **contribution**: the ability to accumulate information in any level of linguistic description, (3) **conjunction**: the ability to combine pieces of information, (4) **disjunction**: the ability to non-deterministically select pieces of information.
Formally, one can define an XMG specification as follows:

$$\begin{aligned} \textit{Rule} &:= \textit{Name} \rightarrow \textit{Content} \\ \textit{Content} &:= \textit{Contribution} \mid \textit{Name} \mid \\ &\quad \textit{Content} \vee \textit{Content} \mid \textit{Content} \wedge \textit{Content} \end{aligned}$$

An abstraction is expressed as a rewrite rule that associates *Content* with a *Name*. Such content is either the *Contribution* of a fragment of linguistic description (e.g. a tree fragment contributed to the description of syntax) or an existing abstraction, or a conjunction or disjunction of contents.
One abstraction must be specifically identified as the axiom of the metagrammar. The XMG compiler starts from this axiom and uses the rewrite rules to produce a full derivation. When a disjunction is encountered, it is interpreted

Table 2: Verbal forms of bòʤákà "to eat"

Subj.	Tense	VR	Aspect	Active	Prox.	Value
m-	à-	ʤ		-á		present
m-	à-	ʤ		-á	-ná	past, yesterday
m-	à-	ʤ		-á	-sá	distant past
m-	é-	ʤ		-á		recent past
m-	é-	ʤ	-àk	-à		medium future
m-	é-	ʤ	-àk	-à	-ná	future, tomorrow
m-	é-	ʤ	-àk	-à	-sá	distant future
m-	ábí-	ʤ	-àk	-à		imminent future

as offering alternative ways to proceed: the compiler successively explores each alternative. In this fashion, the execution of a metagrammar typically produces many derivations. Along one derivation, contributions are simply accumulated conjunctively. At the end of a derivation, the accumulated contributions are interpreted as a specification and given to a solver to produce solution structures. The collection of all structures produced in this manner forms the resulting grammar. It can be inspected using a graphical tool, or exported in an XML format.

The XMG compiler is freely available under a GPL-compliant license, and comes with reasonable documentation.[2] It has been used to design various large tree-based grammars for French (Crabbé, 2005; Gardent, 2008), English (Alahverdzhieva, 2008) and German (Kallmeyer et al., 2008).

XMG was expressedly designed for writing wide-coverage high-level modular tree-grammmars covering both syntactic expression and semantic content. While XMG was never intended for expressing morphology, our current project demonstrates that it can successfully be repurposed for the task, at least in the case of the agglutinative Ikota language.

4. Metagrammar of Ikota verbal morphology

Our formalization of Ikota verbal morphology borrows the notion of *topological domain* from the tradition of German descriptive syntax (Bech, 1955). A topological domain consists of a linear sequence of fields. Each field may host contributed material, and there may be restrictions on how many items a particular field may/must host. For our purposes, the topological domain of a verb will be as described in Table 3, and each field will hold at most 1 item, where an item is the *lexical phonology*[3] of a morpheme.

Elementary blocks. The metagrammar is expressed in terms of elementary blocks. A block makes simultaneous contributions to 2 distinct dimensions of linguistic description: (1) lexical phonology: contributions to fields of the topological domain, (2) inflection: contributions of morphosyntactic features. For example:

2 ← é
tense = past
proxi = near

contributes é to field number 2 of the topological domain, and features tense = past and proxi = near to the inflection. Feature contributions from different blocks are unified: in this way, the inflection dimension also acts as a coordination layer during execution of the metagrammar. As Table 2 illustrates clearly, ikota morphology is not cleanly compositional: instead, the semantic contributions of morphemes are determined by mutually constrained coordination through the inflection layer.

Morphosyntactic features. We use p and n for *person* and *number*; tense with possible values past, present, and future; proxi for the *proximal marker* (none, imminent, day, near, far); vclass for the verbal class (g1, g2, g3); and two polar features: active for *voice* and prog for the *progressive aspect*: prog=- marks an eventuality yet unrealized.

Lexical phonetic signs. Careful consideration of Ikota data suggests that regularities across verbal classes can be better captured by the introduction of a *lexical* vowel A which is then realized, at the surface level, by a for vclass=g1, ɛ for vclass=g2, and ɔ for vclass=g3, and lexical consonant K which is realized by tʃ for vclass=g2, and k otherwise.

Rules. Figure 1 shows a fragment of our preliminary metagrammar of Ikota verbal morphology. Each rule defines how an abstraction can be rewritten. For example *Tense* can be rewritten as any one block from a disjunction of 5 blocks. To produce the lexicon of inflected forms described by our metagrammar, the XMG compiler computes all possible non-deterministic rewritings of the *Verb* abstraction.

Example derivation. Let's consider how óʤàkànáá (*tomorrow, you will eat*) is derived by our formal system starting from the *Verb* abstraction. First *Verb* is replaced by *Subj* ∧ *Tense* ∧ *VR* ∧ *Aspect* ∧ *Active* ∧ *Proximal*. Then each element of this logical conjunction (order is irrelevant) is, in turn, expanded. For example, *Subj* is then replaced by one block from the corresponding disjunction: the XMG compiler tries all possibilities; eventually it chooses the 2nd block. Figure 2 shows the initial step, a middle step, and the final step of the derivation. The lexical phonology of

[2] See http://spourcesup.cru.fr/xmg

[3] We adopt here the *two-level* perspective of lexical and surface phonology (Koskenniemi, 1983)

Figure 1: Metagrammar of Ikota verbal morphology

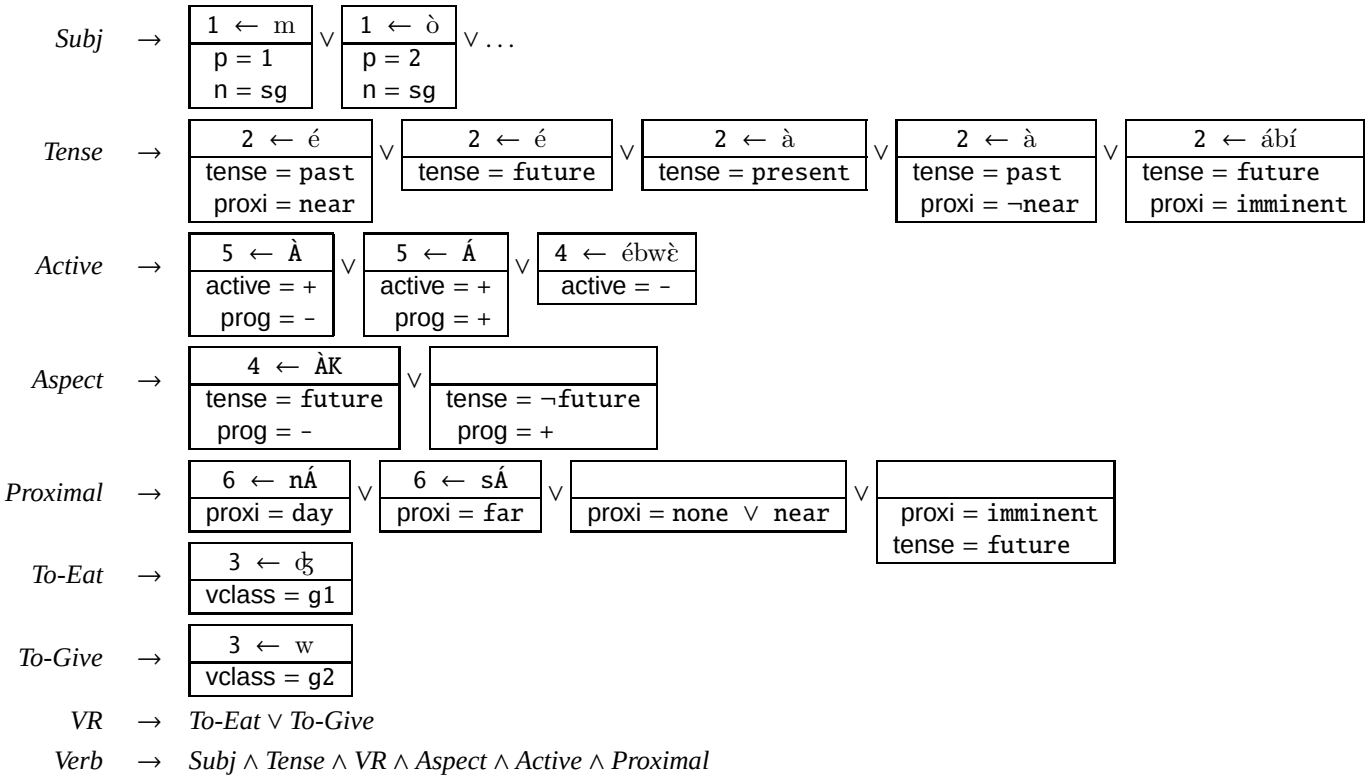
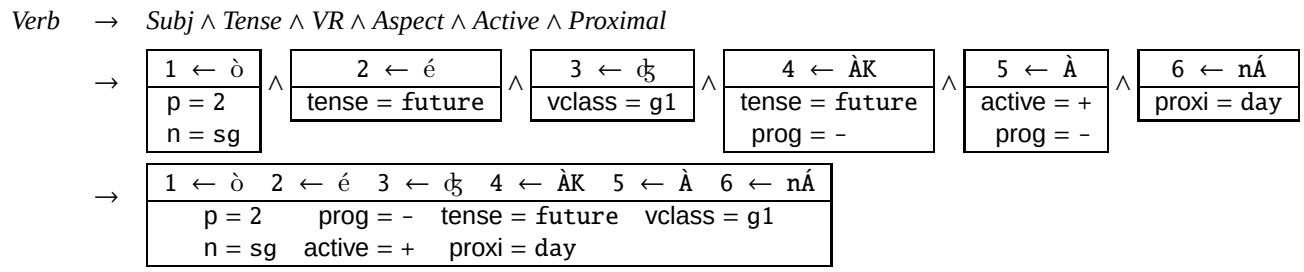


Figure 2: A successful derivation



the resulting lexicon entry is obtained by concatenating, in the linear order of the topological domain, the material contributed to the various fields; here: ò+é+ʤ+ÀK+À+nÁ.

Figure 3 shows an example of a failed derivation, i.e. one which does not lead to the production of a lexicon entry. The failure is due to clashing values for feature tense (future and ¬future) and also for feature prog (+ and -).

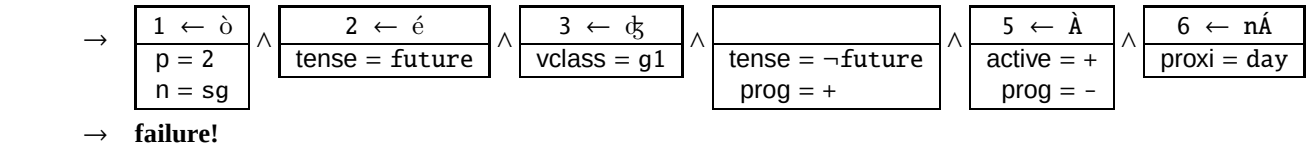
Surface phonology. At present, our metagrammar models only the lexical level of phonology. The surface level can subsequently be derived by postprocessing. For our example, since vclass=g1, the lexical A becomes a on the surface, and K becomes k. Thus we obtain: ò+é+ʤ+à+à+ná, and finally (through vowel deletion) óʤàkànà.

XMG's constraint-based approach makes it ideally suited to a seamless integration of e.g. *two-level phonology* since the latter is precisely a constraint between lexical and surface phonology (Koskenniemi, 1983). This extension of XMG is a planned milestone of an ongoing thesis.

Caveats. Our formalization of Ikota morphology is very preliminary. As we progress, questions arise for which we do not yet have sufficient data. For example, as can be readily deduced from Figure 1, our current metagrammar (deliberately) omits the "passive future" pending further evidential data from native speakers.

Also, it is too early for us to suggest even a tentative account of Ikota's tonal system and its implications on e.g. the prosodic contours of verb forms. As a consequence, in the interest of accurate descriptive morphology, we have been forced to adopt some tricks, in the formal description, as a practical recourse rather than as a theoretical proposal: such is the case of the tonal alternation in the active voice.

Figure 3: A failed derivation: clashes on tense and on prog



5. Conclusion and future work

In this article, we proposed a formal, albeit preliminary, declarative description of verbal morphology in Ikota, an arguably minority African language. In so doing, we illustrated how the metagrammatical approach can usefully contribute to African language technology.

Additionally, from this formal description, using the XMG compiler, we are able to automatically produce a lexicon of fully inflected verb forms with morphosyntactic features. This lexicon can be saved in XML format, thus providing an easily reusable normalization resource for this less-resourced language.

From a methodological point of view, the use of XMG for expressing our ideas has made it easy to quickly test them by generating the predicted verb forms and their features and then validating the results against the available data.

A further advantage of adopting the metagrammar approach is that, using the same tool, we will be able to also describe the syntax of the language using e.g. tree-adjoining grammars (the topic of an ongoing PhD thesis).

6. References

Katya Alahverdzhieva. 2008. XTAG using XMG. Master Thesis, Nancy Université.

Gunnar Bech. 1955. *Studien über das deutsche Verbum infinitum*. Det Kongelige Danske videnskabernes selskab. Historisk-Filosofiske Meddelelser, bd. 35, nr.2 (1955) and bd. 36, nr.6 (1957). Munksgaard, Kopenhagen. 2nd unrevised edition published 1983 by Max Niemeyer Verlag, Tübingen (Linguistische Arbeiten 139).

Marie Candito. 1996. A Principle-Based Hierarchical Representation of LTAGs. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, volume 1, pages 194–199, Copenhagen, Denmark.

Benoît Crabbé and Denys Duchier. 2004. Metagrammar redux. In Henning Christiansen, Peter Rossen Skadhauge, and Jørgen Villadsen, editors, *Constraint Solving and Language Processing, First International Workshop (CSLP 2004), Revised Selected and Invited Papers*, volume 3438 of *Lecture Notes in Computer Science*, pages 32–47, Roskilde, Denmark. Springer.

Benoît Crabbé. 2005. *Représentation informatique de grammaires fortement lexicalisées : Application à la grammaire d'arbres adjoints*. Ph.D. thesis, Université Nancy 2.

Claire Gardent. 2008. Integrating a Unification-Based Semantics in a Large Scale Lexicalised Tree Adjoining Grammar for French. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 249–256, Manchester, UK, August. Coling 2008 Organizing Committee.

Daniel Franck Idiata. 2007. *Les langues du Gabon: données en vue de l'élaboration d'un atlas linguistique*. L'Harmattan.

Laura Kallmeyer, Timm Lichte, Wolfgang Maier, Yannick Parmentier, and Johannes Dellert. 2008. Developing a TT-MCTAG for German with an RCG-based Parser. In *The sixth international conference on Language Resources and Evaluation (LREC 08)*, pages 782–789, Marrakech, Morocco.

Kimmo Koskenniemi. 1983. *Two-Level Morphology: a general computational model for word-form recognition and production*. Ph.D. thesis, University of Helsinki.

Brunelle Magnana Ekoukou. 2010. Morphologie nominale de l'ikota (B25): inventaire des classes nominales. Mémoire de Master 2, Université d'Orléans.

Sam A. Mchombo. 1998. Chichewa: A Morphological Sketch. In Andrew Spencer and Arnold Zwicky, editors, *The Handbook of Morphology*, pages 500–520. Blackwell, Oxford, UK & Cambridge, MA.

Pascale Piron. 1990. Éléments de description du kota, langue bantoue du gabon. mémoire de licence spéciale africaine, Université Libre de Bruxelles.

Brian Roark and Richard W. Sproat. 2007. *Computational approaches to morphology and syntax*. Number 4. Oxford University Press, USA.

Gregory T. Stump. 1992. On the theoretical status of position class restrictions on inflectional affixes. In G. Booij and J. van Marle, editors, *Yearbook of Morphology 1991*, pages 211–241. Kluwer.

Gregory T. Stump. 1998. Inflection. In A. Spencer and A. M. Zwicky, editors, *The Handbook of Morphology*, pages 13–43. Blackwell, Oxford & Malden, MA.

Gregory T. Stump. 2001. *Inflectional Morphology: a Theory of Paradigm Structure*, volume 93. Cambridge University Press.