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

Denys Duchier, Brunelle Magnana Ekoukou, Yannick Parmentier, Simon Petitjean, Emmanuel Schang

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 languages, and are used as benchmarks for assessing the expressivity and relevance of morphological theories (Mchombo, 1998; Stump, 1992; Stump, 1998; Stump, 2001).
Ikota (B25) is a lesser-known language of Gabon and the Democratic Republic of Congo. It manifests many grammatical features shared by the Bantu languages:

- 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* (the number of the class in the table below corresponds to Meinhof's numbering):

Table 1: Ikota's noun classes

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

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

  (3) **b**-àyítò **bá**-nɛ́nì **b**-á Ø-mbókà **bà**-tɛ́ **b**-àʤá
  2-women 2-fat 2-of 9-village 2-DEM 2-eat
  "These fat women of the village are eating"

In this paper, we will consider verbal morphology. 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. The ordering of Ikota's verbal affixes can be defined as position classes, from left to right:

- tense prefixes (or what can roughly identified as related to Tense) appears at the left of VR,
- the class of Subject agreement prefixes occupies the leftmost, word-initial position,
- the (aspectual) progressive marker is on the immediate left of VR,
- the proximal/distal suffixes occupy the rightmost position.

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

Table 2: Verb formation

Subj-	Tense-	VR	-Aspect	-Active	-Proximal

Here, we are interested in defining a formal description of the morphology of verbs in Ikota, which would make it possible to automatically produce a lexicon of verbs in this language. To do so, we propose to reuse the concept of metagrammar, which was introduced by (Candito, 1996), and used to describe the syntax of Indo-European languages, such as French, English or Italian. To get a better view of what a metagrammar is, let us consider formal descriptions of syntax.
A common way to formally describe the syntax of natural language, is to use a formal grammar. Such a grammar corresponds to a mathematical model, that not only defines which sentences belong to a language, but also

Table 3: Verbal forms of bòɔ́ákà "to eat"

Slot 1	Slot 2	Slot3	Slot 4	Slot 5	Slot 6	Gloss
m-	à-	ɔ́		-á		I am eating (present)
m-	à-	ɔ́		-á	-ná	I ate (past, yesterday)
m-	à-	ɔ́		-á	-sá	I ate (distant past)
m-	é-	ɔ́		-á		I've eaten (recent past)
m-	é-	ɔ́	-ák	-à		I'll eat (middle future)
m-	é-	ɔ́	-ák	-à	-ná	I'll eat (future, tomorrow)
m-	é-	ɔ́	-ák	-à	-sá	I'll eat (distant future)
m-	ábí-	ɔ́	-ák	-à		I'll eat (imminent future)
		ɔ́	-ák	-à		eat! (imperative)

what are the relations between the constituents of a valid sentence. Among existing grammatical models, the most renown doubtlessly is context-free grammar (CFG). In a CFG, one defines rewriting rules, which allows to generate the sentences of a language by replacing non-terminal symbols (syntactic categories) with non-terminal symbols or terminal ones (words). As an illustration, consider the toy CFG of Fig. 1.

S	$\rightarrow$	NP	VP	NP	$\rightarrow$	Det	N
VP	$\rightarrow$	V	NP	V	$\rightarrow$	*eats*	
Det	$\rightarrow$	*the*		Det	$\rightarrow$	*a*	
N	$\rightarrow$	*cat*		N	$\rightarrow$	*mouse*	

Figure 1: Toy Context-Free Grammar

Such a grammar describes the syntactic structure of (among others) the sentence "the cat eats a mouse" (see Fig. 2).

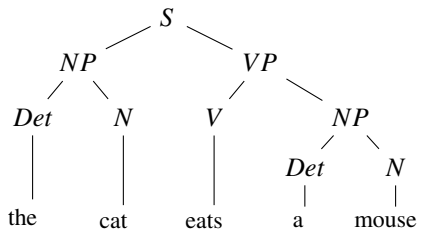

Figure 2: Syntactic structure of the sentence "the cat eats the mouse" using the grammar of Fig. 1

The CFG model is interesting from a computational point of view since it can be efficiently implemented (*e.g.*, the complexity of parsing a sentence with a CFG is polynomial in the size of the sentence). From a linguistic point of view however, it is not satisfactory for it suffers from a lack of expressivity, *i.e.*, there are various linguistic phenomena that cannot be described by CFG, for instance cross-serial dependencies in Dutch (Bresnan, 1982):[1]

(4) "dat An$_1$ Bea$_2$ wil$_1$ kussen$_2$"
"that An Bea wants kiss"

"that An wants to kiss Bea"

[1]This example is taken from (Bouma and van Noord, 1994).

A cross-serial dependency refers to subordinate clauses, where the arguments of the verbs appear in the same order as the verbs.

To deal with such linguistic phenomena, extensions of CFG have been defined in order to increase the expressivity of formal grammars. A particularly interesting family of such "extended" formal grammars corresponds to tree-based grammars, *e.g.*, Lexicalized Tree-Adjoining Grammar (LTAG) (Joshi and Schabes, 1997). In an LTAG, one associates predicative words with elementary trees describing the different syntactic behaviors of these words. For instance, the word *mange* in French (*eats* in English) is associated with the structures of Fig. 3 (among others). The tree on the left describe the canonical behavior of a transitive verb, the second tree an extracted object, and the last tree an extracted subject.

LTAG is a particularly interesting formalism since it exhibits advantageous computational and linguistic properties. First, it is polynomially parsable, and secondly, it allows to express wihtin a single grammar rule, relations between words that are far away from each other in the sentence. As one may imagine, in a large LTAG, there is a huge redundancy since there are a large number of elementary tree, among which many share common sub-parts. To deal with this redundancy, Candito (1996) proposed to define an abstract description of a tree-based grammar, where one would define reusable tree fragments, which would be combined to produce the fully redundant grammar.

Here, we propose to adopt a similar strategy to capture mophological generalizations over verbs in Ikota. The outline of the paper is the following. In Section 2., we give an detailed presentation of the morphology of verbs in Ikota. Then, in Section 3., we introduce the eXtensible MetaGrammar (XMG) formalism, which is a formal language, used to describe reusable tree fragments. In Section 4., we show how to use the XMG language to describe the morphology of verbs in Ikota. Concretely, we present a metagrammar of verbs in Ikota, which is written in the XMG language, and which can be processed by the XMG compiler, to produce a lexicon of verbs in Ikota. Finally, in Section 5., we conclude and present future work.

2. Verbs in Ikota

3. eXtensible MetaGrammar

Let us now describe the eXtensible MetaGrammar (XMG). In fact, by XMG, we refer to both a formal language (*a*

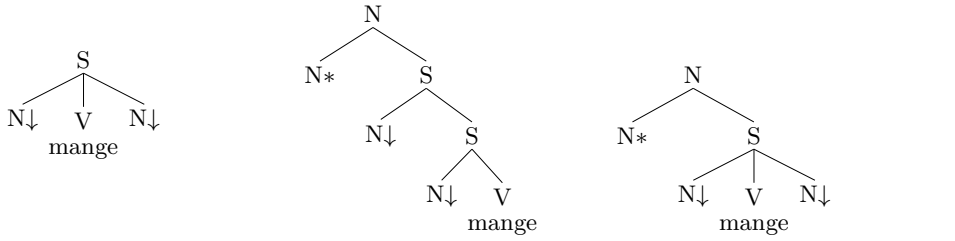

Figure 3: Entries of an LTAG lexicon

kind of programming language) and a software to process a description written in the XMG language (such a software is usually called a compiler).

As mentioned above, XMG is used to describe tree grammars. In other words, an XMG specification is a declarative description of the 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:

$$\begin{aligned} \mathit{Class} &:= \mathit{Name}[p_1, \ldots, p_n] \rightarrow \mathit{Content} \\ \mathit{Content} &:= \langle \mathit{Dim} \rangle \mathrel{+}= \mathit{Desc} \mid \mathit{Name}[\ldots] \mid \\ &\quad \mathit{Content} \vee \mathit{Content} \mid \mathit{Content} \wedge \mathit{Content} \end{aligned}$$

Abstraction is provided by means of parametrized *classes*, which encapsulate different types of information (called here *content*). This information can either be a linguistic description (*e.g.*, a tree description), belonging to a given dimension (*e.g.*, syntax), or an existing abstraction (aliasing), or a conjunction or disjunction of contents.

When describing LTAG, the descriptions encapsulated within metagrammatical classes are tree descriptions defined using a tree description logic. As mentioned above, these descriptions can be accumulated. In the end, the metagrammar defines classes that may reuse other classes. In other words, some classes contain partial information, and some others complete tree descriptions, whose tree models are the described grammar rules. These complete description are called axioms. When one wants to compile an XMG metagrammar, the compiler evaluates the axioms of the metagrammar, and computes the corresponding tree models. These can either be displayed graphically (to inspect the described grammar), or saved into an XML file.

The XMG compiler is freely available under a GPL-compliant license, and comes with a 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).

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

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 2, 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 ← ɛ́
tense = past
proxi = near

contributes ɛ́ 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 medium during execution of the metagrammar.

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. Figure 4 shows a fragment of our preliminary metagrammar of Ikota verbal morphology.

Surface phonology. At present, our metagrammar models only the lexical level of phonology. The surface level can subsequently be derived by postprocessing. However, XMG's constraint-based approach makes it ideally suited to a seamless integration of *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.

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

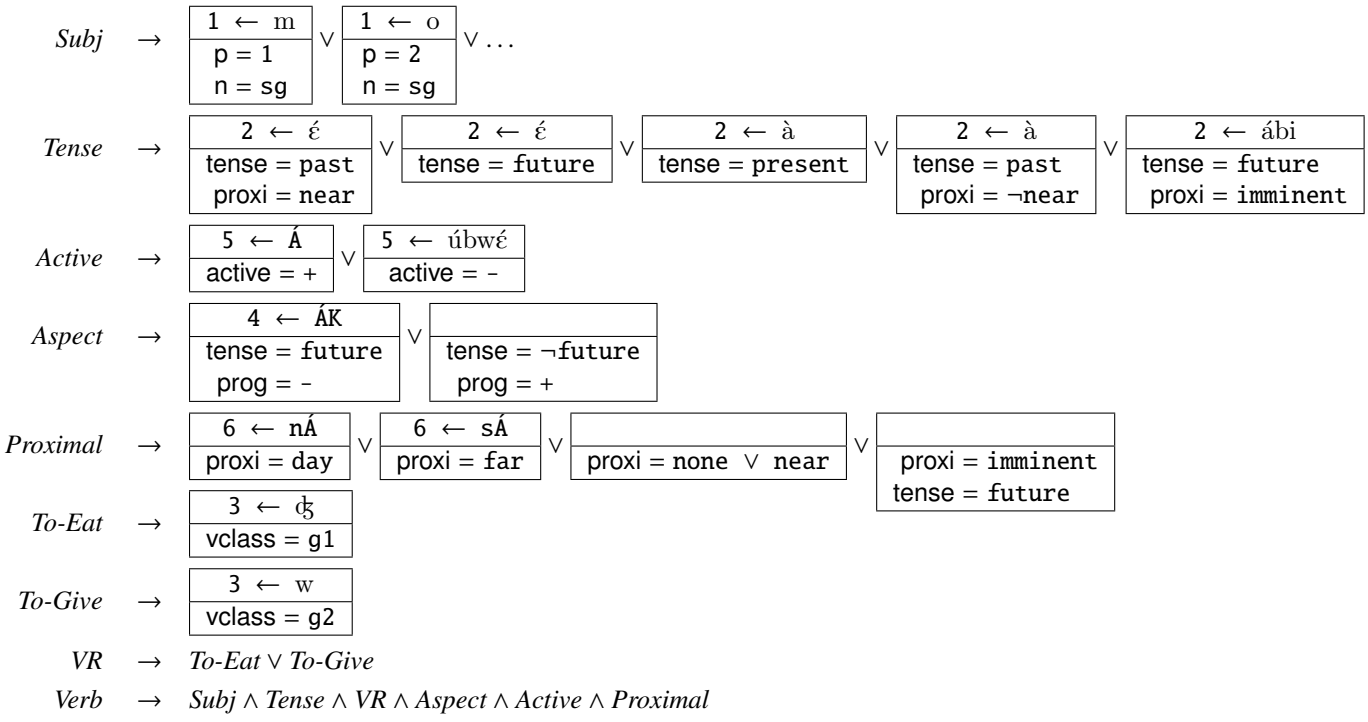

Figure 4: Metagrammar of Ikota verbal morphology

5. Conclusion and future work

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).

Gosse Bouma and Gertjan van Noord. 1994. Constraint-based categorial grammar. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 147–154, Las Cruces, New Mexico, USA, June. Association for Computational Linguistics.

Joan Bresnan. 1982. The passive in lexical theory. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge, MA.

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. 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.

Aravind K. Joshi and Yves Schabes. 1997. Tree adjoining grammars. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*. Springer Verlag, Berlin.

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.

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.

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. Default inheritance hierarchies and the evolution of inflectional classes. In Laurel Brinton, editor, *Proceedings of the XIVth International Conference on Historical Linguistics*, pages 293–307, Amsterdam, Netherlands. John Benjamins.