Computational morphology: XMG and verbal morphology of Ikota

Simon Petitjean

Heinrich-Heine-Universität Düsseldorf

May 29, 2019

Introduction

- XMG ([Crabbé et al., 2013], [Petitjean et al., 2016]): description tool for various types of linguistic resources
- Interfaces between different levels of description (syntax, semantics, morphology)
- Today: constraint based description of morphology
- Application: verbal morphology in Ikota (based on [Duchier et al., 2012])
- Implementation of a metagrammar for a small lexicon

Plan

Plan

XMG

- eXtensible MetaGrammar: programming language and compiler for this language
- Input: abstract description → Metagrammar
- Output: linguistic resource (grammar, lexicon, . . .)
- Website: `http://xmg.phil.hhu.de/`
- Documentation, online compiler, visualization of outputs

- Metagrammar: abstract and compact
- Based on constraints and logic
- Idea: instead of describing complex rules, describe fragments of rules and combine them
- Abstraction (classes): capture redundancies, allow generalization
- Classes: organized in a hierarchy
- Compilation: accumulation of all (valid) descriptions

The control language

XMG descriptions:

- Associate a content to an identifier (abstraction)
- Describe structures inside dimensions, with dedicated languages
- Use other abstractions (classes)
- Combine contents in a disjunctive or a conjunctive way

$$
\begin{aligned}
\mathit{Class} &:= \mathit{Name} \rightarrow \mathit{Content} \\
\mathit{Content} &:= \langle \mathit{Dimension} \rangle \{ \mathit{Description} \} \mid \mathit{Name} \mid \\
&\quad \mathit{Content} \vee \mathit{Content} \mid \mathit{Content} \wedge \mathit{Content}
\end{aligned}
$$

Dimensions

- Dimensions: describe different types of structures with different languages
- Examples: syn (tree description language), sem (predicate logic), frame (frame semantics), morph (morphology)
- We will use two dimensions: `morph` and `iface` (feature structures)
- Example:

```
class Subject_Clitic
{
  {
    <morph>{f_subj <- "m"}
    ;
    <iface>{[p=1, n=sg]}
  }
}
```

The morph dimension

- Three types of constraints, combined with conjunction (`;`) and disjunction (`|`)
- `field F`: defines a field (morpheme, lexeme, character)
- `F1 >>F2`: adds a linear precedence constraint between fields `F1` and `F2`
- `F <- STR`: affects the string `STR` into the field `F`

The iface dimension

- Encoding of feature structures, for instance for morpho-syntactic information
- `[a1 =v1, ... an =vn]`: defines a feature structures with attribute-value pairs `a1:v1...an:vn`
- Accumulation in the interface: unification with current interface
- Failing unification: incompatible features, the current description is dropped

Compiling a metagrammar

- Different ways to use the compiler (local installation, virtual disk, online)
- Online compiler: (`http://xmg.phil.hhu.de/index.php/upload/workbench`)
- Upload a file and edit it online
- Result of the compilation: log on the right and generated XML on the left
- Download the XML file or directly view it

Plan

Ikota

- Bantu languages: large family of languages in Africa
- Chichewa and Swahili: most well-studied
- Ikota (B25): lesser-known language of Gabon and the Republic of Congo
- Language of the Bakota people, with an estimated 25000 speakers in Gabon
- threatened with extinction, mainly because of its abandon for French

Ikota

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

(1) a. ìkàká ”family”
b. ìkáká ”palm”

(2) a. nkúlá ”year”
b. nkúlà ”pygmee”

- Ikota has ten *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-, ʤ-	dy-
CL 6	mà-	m-
CL 7	è-	
CL 8	bè-	
CL 9	Ø-	
CL 14	ò-, bò-	bw-

Ikota

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

(3) **b**-àyítò **bá**-n´ɛnì **b**-á Ø-mbókà **bà**-tɛ́ **b**-à-ʤá
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.

- Our focus: verbal morphology

Verbal morphology

- Verbs: lexical root (VR) and affixes distributed on each side of the VR
- Verbs are distributed in three classes depending on the form of Aspect and Active markers.
- Infinitive: hybrid word class, composed of a noun class prefix (class 14) and a verbal element (VR+Prog+Active).

(4)
a. bòʤákà “to eat”
b. bòwétʃɛ̀ “to give”
c. bòbɔ́nɔ́kɔ̀ “to choose”

Examples

(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)

Examples

(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 mwáyàkàmwɛ́
1sg-Fut-eat-Asp-Act-Prox year 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)

Ikota and topological fields

- Verbal morphology can be described using topological fields (following [Bech, 1955])

Table: Verb formation

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

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

Ikota: Metagrammatical description

- Fields are defined and ordered globally
- Classes can affect strings into fields
- Morphosyntactic features: interface dimension

Subj → 1 ← m : p = 1, n = sg ∨ 1 ← ò : p = 2, n = sg ∨ ...

Tense → 2 ← é : tense = past, proxi = near ∨ 2 ← é : tense = future ∨ 2 ← à : tense = present ∨ ...

Active → 5 ← À : active = +, prog = - ∨ 5 ← Á : active = +, prog = + ∨ 4 ← ébwɛ̀ : active = -

Aspect → 4 ← ÀK : tense = future, prog = - ∨ ∅ : tense = ¬future, prog = +

Proximal → 6 ← nÁ : proxi = day ∨ 6 ← sÁ : proxi = far ∨ ∅ : proxi = none ∨ near ∨ ...

...

Verb → Subj ∧ Tense ∧ VR ∧ Aspect ∧ Active ∧ Proximal

Plan

Output and applications

- What is generated? A lexicon of inflected forms, with associated semantics
- Visualization in the WebGUI
  (`http://xmg.phil.hhu.de/index.php/upload/upload_viewer`)
- In a better world: usage as a lexicon for syntactic parsing

Exercise

- We will try to recreate a small part of the lexicon for Ikota
- More precisely, the 8 inflected forms of slide 19
- To get started: download the partial metagrammar (`https://user.phil.hhu.de/petitjean/teaching/hhu-dusseldorf/computational-morphology/`)
- Fill in the blanks (marked `[1]`, `[2]`,...)

Conclusion

- XMG: describing linguistic resources with logic and constraints
- `morph` dimension: describing morphology using constraints on fields
- Input: Metagrammar
- Output: lexicon of inflected forms

Bech, G. (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).

Crabbé, B., Duchier, D., Gardent, C., Le Roux, J., and Parmentier, Y. (2013).
XMG : eXtensible MetaGrammar.
Computational Linguistics, 39(3):1–66.

Duchier, D., Magnana Ekoukou, B., Parmentier, Y., Petitjean, S., and Schang, E. (2012).
Describing Morphologically-rich Languages using Metagrammars: a Look at Verbs in Ikota.
In *Workshop on "Language technology for normalisation of less-resourced languages", 8th SALTMIL Workshop on Minority Languages and the 4th workshop on African Language Technology*, page _, Istanbul, Turkey.


Petitjean, S., Duchier, D., and Parmentier, Y. (2016).
XMG 2: Describing Description Languages.
In *Logical Aspects of Computational Linguistics. Celebrating 20 Years of LACL (1996–2016) 9th International Conference, LACL 2016, Nancy, France, December 5-7, 2016, Proceedings 9*, pages 255–272. Springer Berlin Heidelberg.