

DABA: A MODEL AND TOOLS FOR MANDING CORPORA

Kirill Maslinsky

Sociology of Education and Science Laboratory
National Research University Higher School of Economics
Saint-Petersburg, Russia

TALAf 2014
Marseille, 1 juin 2014



OUTLINE

A PORTRAIT OF A MANDING CORPUS

DABA SOFTWARE PACKAGE: TASKS AND TOOLS

Overview

Morphological analysis

Corpus annotation



BAMABARA REFERENCE CORPUS: A MODEL

SOCIOLINGUISTIC SITUATION

- ▶ Linguistic research mostly outside of native-speaking community
- ▶ Low level of written language standardization



BAMABARA REFERENCE CORPUS: A MODEL

CORPUS USE-CASES

- ▶ Field linguists as corpus designers and primary corpus users
 - ▶ linguistic research
 - ▶ language teaching by non-native speakers (European and US universities)
- ▶ corpus use by native speakers
 - ▶ ...is anticipated



BAMBARA REFERENCE CORPUS: A MODEL

REQUIREMENTS

- ▶ Modelled on a large ‘National’ corpora: BNC, Russian National Corpus
- ▶ Freely accessible online
- ▶ Adapted for the needs of linguists and language learners:
 - ▶ Linguistic annotation adhering to the international standards
 - ▶ Glosses in a European language
 - ▶ Search with or without tones
 - ▶ Normalized orthography for searching and statistics
 - ▶ Source orthography preserved for the purposes of linguistic study



INITIAL LINGUISTIC RESOURCES

- ▶ Charles Bailleul dictionary: 10,000+ words
 - ▶ in electronic form in the SIL Toolbox format
- ▶ Other dictionaries and grammatical descriptions of Bambara
- ▶ No linguistically annotated data
- ▶ 100,000 word electronic text collection provided by Gerard Dumestre



OUTLINE

A PORTRAIT OF A MANDING CORPUS

DABA SOFTWARE PACKAGE: TASKS AND TOOLS

Overview

Morphological analysis

Corpus annotation



CORPUS TASKS

1. Automated morphological annotation of all texts. (PARSER)
2. Manual disambiguation of selected texts. (DISAMB)
3. Adding metadata to all texts. (META)
4. Creating an online search interface with flexible possibilities for concordance building. (CONVERTER)



PARSER GUI

GPARSER

The screenshot displays the GParser GUI interface. On the left is a text editor window titled "Efile" containing Bambara text. On the right is a sidebar with three sections: "Available Dictionaries", "Available Grammar", and "Available Orthographic Converters".

Text Editor Content:

```
1
Súrukuba ní Je`
Nsírín
N y`à tà kà à dá Súrukuba. ní Je`dè lá.
Súrukuba ní Je`òlú. be`táa sitomayorá` lá.
"Boñ" Je`be`Súrukuba. den. tà kà bòlí n`ò yé kà bòlí n`ò yé kà bòlí n`ò yé.
Súruku be`sí `toñkòñ kà tila, sú.be`kó kà kó beé`ke;` à te`Je`yé à te`Je`n`à
dén.yé. Á be`bòlí dámíne.` Á be`táa sé`dùgú. do`lá à b`à fò`kó:
Áw má Je`yé ? Jě jě !
Áw má Je`yé ? Jě jě !
Dén.b`à kò`lá dén kálanman.
Bàmun fítinin. b`à kán. ná`mérú mérú.
Òlú. b`à fò`kó : << eé ! Kó`jehín. temená`sísan, kó`ù be`bòlí n`ò yé kà`táa
>>.
Á b`à fò` :
Áw má Je`yé ? Jě jě !
Áw má Je`yé ? Jě jě !
Dén.b`à kò`lá dén kálanmán.
Bàmun fítinin. b`à kán. ná`mérú mérú.
```

Available Dictionaries:

- bam jamuw 0.3 [Remove]
- bam bamadaba 0.5.4 [Remove]
- [Add dictionary]

Available Grammar:

- bamana.gram.txt
- [(Re)Load grammar]

Available Orthographic Converters:

- apostrophe
- bamlatinold
- vydrine
- bailleul
- nko



MANUAL DISAMBIGUATION GUI

GDISAMB

File Settings Debug

< > Save results

Walasa ka geleya ninnu ðɔn a ɲɛma, ani ka fɛɛɛw tige an ka jamana haminakobaw la, ladamuniko siratige la, foroba jɛkafɔ kɛra, min kɛnɛ kan, wele bilara jamana kɔɔ magan kɛbagaw ani jerejekulu bɛɛ ma.

wálasa (conj) ka geleya ninnu ðɔn a ɲɛ

pour.conj

ká (conj) POSS	geɛ.ya (n) duretɛ	ðɔ́n (n) gui	á (intj) ah!
ká (pm) QUAL.AFF	geɛ (vq/adj) dur	ya (mrph) ABSTR	ðɔ́n (n) connaissance
ká (pm) OPT	geɛ.ya (v) durɔr	à (pers) 3SG	ɲɛ
ká (onomat) complètement.sec	geɛ (vq/adj) dur	ya (mrph) DEQU	ðɔ́n (v) connáitre
kà (pm) INF			ðɔ́n (v) danser
kà (v)			ðɔ́n (n) danse



DABA: THE CODE

- ▶ Python
- ▶ GPL
- ▶ <http://github.com/maslinych/daba/>

OUTLINE

A PORTRAIT OF A MANDING CORPUS

DABA SOFTWARE PACKAGE: TASKS AND TOOLS

Overview

Morphological analysis

Corpus annotation



TOOLBOX: A PROTOTYPE

- ▶ a morphological parser that can handle almost all types of morphophonemic processes.
- ▶ a word formula component that allows the linguist to describe all the possible affix patterns that occur in words.
- ▶ a user-definable interlinear text generation system which uses the morphological parser and lexicon to generate annotated text.
- ▶ **doesn't have the batch mode parsing with ambiguous results**



INPUT DATA FOR THE PARSER

- ▶ a dictionary (Toolbox lexical DB)

(1) \lx ádamaden
 \va hádamaden
 \ps n
 \ge humain

- ▶ morpheme combination constraints (Word formulas in Toolbox)



ANNOTATION MODEL

INTERLINEAR GLOSSED TEXT

(2) báarabaliw
báara-bali-w
ptcp
travailler-PTCP.PRIV-PL



GLOSS OBJECT

(3) báara:v:travailler []

(4) báarabaliw:ptcp: [báara:v:travailler bali::PTCP.PRIV w::PL]

báarabaliw (ptcp)		
báara (v)	bali	w
travailler	PTCP.PRIV	PL



PATTERN RULE

- ▶ pattern CONTEXT | RESULT

pattern :v/ptcp: [**{|bali|}**::] | :ptcp: [:v: :mrph:PTCP.PRIV]

- ▶ use of regular expressions:

pattern :v: [{<re>.*[aoeuieɛɔ]n</re>|na}::] | :v: [:v: :mrph:PROG]



PATTERN RULE

- ▶ pattern CONTEXT | RESULT
pattern :v/ptcp: [{|bali}::] | :ptcp: [:v: :mrph:PTCP.PRIV]
- ▶ use of regular expressions:
pattern :v: [{<re>.*[aoeuieɔ]n</re>|na}::] | :v: [:v: :mrph:PROG]



PATTERN RULE

- ▶ pattern CONTEXT | RESULT

pattern :v/ptcp: [{|bali}::] | :ptcp: [:v: :mrph:PTCP.PRIV]

- ▶ use of regular expressions:

pattern :v: [{<re>.*[aoeuiɛɔ]n</re>|na}::] | :v: [:v: :mrph:PROG]



PROCESSING INSTRUCTIONS

- ▶ a sequence of pattern rule applications and dictionary lookups
plan
for token:
stage 0 add parallel parse inflection
stage 0 add parallel parse common_derivation
stage 0 add parallel parse participles
stage 0 apply lookup
return if parsed



PARSING EXAMPLE

stage	parses list
start	baarabaliw::
inflection	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL]
derivation	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL]
participles	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL] baarabaliw:ptcp: [baara:v: bali:mrph:PTCP.PRIV w:mrph:PL]
lookup	baarabaliw:ptcp: [baara:v:travailler bali:mrph:PTCP.PRIV w:mrph:PL]



PARSING EXAMPLE

stage	parses list
start	baarabaliw::
inflection	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL]
derivation	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL]
participles	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL] baarabaliw:ptcp: [baara:v: bali:mrph:PTCP.PRIV w:mrph:PL]
lookup	baarabaliw:ptcp: [baara:v:travailler bali:mrph:PTCP.PRIV w:mrph:PL]



PARSING EXAMPLE

stage	parses list
start	baarabaliw::
inflection	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL]
derivation	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL]
participles	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL] baarabaliw:ptcp: [baara:v: bali:mrph:PTCP.PRIV w:mrph:PL]
lookup	baarabaliw:ptcp: [baara:v:travailler bali:mrph:PTCP.PRIV w:mrph:PL]



PARSING EXAMPLE

stage	parses list
start	baarabaliw::
inflection	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL]
derivation	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL]
participles	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL] baarabaliw:ptcp: [baara:v: bali:mrph:PTCP.PRIV w:mrph:PL]
lookup	baarabaliw:ptcp: [baara:v:travailler bali:mrph:PTCP.PRIV w:mrph:PL]



PARSING EXAMPLE

stage	parses list
start	baarabaliw::
inflection	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL]
derivation	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL]
participles	baarabaliw:: baarabaliw:ptcp/adj/n: [baarabali:ptcp/adj/n: w:mrph:PL] baarabaliw:ptcp: [baara:v: bali:mrph:PTCP.PRIV w:mrph:PL]
lookup	baarabaliw:ptcp: [baara:v:travailler bali:mrph:PTCP.PRIV w:mrph:PL]



DECOMPOSITION

TREATMENT OF COMPOSITE FORMS

- ▶ Composition is a productive process in Bambara:

(5) mɔ̀gɔ̀jugu
mɔ̀gɔ̀-juɡu
homme-mauvais

- ▶ Decompose parser rule:

- ▶ Dictionary is represented as a suffix tree (TRIE)
- ▶ Lookup leftmost substring which is a valid word: mɔ̀gɔ̀jugu
- ▶ Check if a remainder is also a valid word: mɔ̀gɔ̀juɡu
- ▶ Check validity of the composite by applying the pattern rule:

(6) pattern :n: [:n: :adj:] | :n: [:n: :adj:]



AUXILIARY TASKS OF THE PARSER

- ▶ tokenizing input text
- ▶ sentence splitting
- ▶ normalizing orthography



OUTLINE

A PORTRAIT OF A MANDING CORPUS

DABA SOFTWARE PACKAGE: TASKS AND TOOLS

Overview

Morphological analysis

Corpus annotation



NOskETCHENGINE

ONLINE CORPUS PUBLISHING TOOL

- ▶ free and open source;
- ▶ a mature project, alive and rather well supported;
- ▶ supports ambiguous values for annotation fields;
- ▶ provides a very flexible query language, CQL
- ▶ NoSketchEngine is an open source variant of SketchEngine



TOKEN ANNOTATION

word	baarasɔɔbaliw
lemma	baarasɔɔ
tag	ptcp PL PTCP.PRIV
form	báara-sòrɔ-bali-w
gloss	travail-obtenir-PTCP.PRIV-PL
parts	baara sɔɔ



TOKEN ANNOTATION

word	baarasɔɔbaliw
lemma	baarasɔɔ
tag	ptcp PL PTCP.PRIV
form	báara-sòɔɔ-bali-w
gloss	travail-obtenir-PTCP.PRIV-PL
parts	baara sɔɔɔ



TOKEN ANNOTATION

word	baarasɔɔbaliw
lemma	baarasɔɔ
tag	ptcp PL PTCP.PRIV
form	báara-sòɔɔ-bali-w
gloss	travail-obtenir-PTCP.PRIV-PL
parts	baara sɔɔɔ



TOKEN ANNOTATION

word	baarasɔɔbaliw
lemma	baarasɔɔ
tag	ptcp PL PTCP.PRIV
form	báara-sòɔɔ-bali-w
gloss	travail-obtenir-PTCP.PRIV-PL
parts	baara sɔɔɔ



TOKEN ANNOTATION

word	baarasɔrɔbaliw
lemma	baarasɔrɔ
tag	ptcp PL PTCP.PRIV
form	báara-sòrɔ-bali-w
gloss	travail-obtenir-PTCP.PRIV-PL
parts	baara sɔrɔ



TOKEN ANNOTATION

word	baarasɔɔbaliw
lemma	baarasɔɔ
tag	ptcp PL PTCP.PRIV
form	báara-sòrɔ-bali-w
gloss	travail-obtenir-PTCP.PRIV-PL
parts	baara sɔɔ



GLOSSED EXAMPLES FROM THE CORPUS

	la	.	Ni	mògò	tè	t'	i	
sùma	lá	.	ní	mògò	té	t'	í	
récolte	dans	.	si	homme	IPFV.NEG	aller	2.SG	
	olu	kan	.	Mògò	minw	minèna		
dá	òlú	kán	.	mògò	mín-w	mìnè-na		
poser	ce_PL2	sur	.	homme	REL-PL	attraper-PFV.INTR		
	Siyaw	bèe	mògòw	ye	balimaw			
.	síya-w	béé	mògò-w	yé	bálima-w			
.	race-PL	tout	homme-PL	EQU	frère-PL			
	fo	ni	dugumògò	bèe	tun	ye		
kónɔ	fo	ní	dùgu-mògò	béé	tùn	yé		
à_l'intérieur	jusqu'à	si	terre-homme	tout	PST	EQU		
	Bakari	somògò	minw	ka	kòrò			
kósɛbɛ	Bákàrí	sómogò	mín-w	ká				
très	NOM	personne_de_la_famille	REL-PL	QUAL.AFF				
buran	ye	Mògòya	sira	la	Bakari			
búran	yé	mògòya	síra	lá				
beaux-parent	PP	humanité	chemin	dans				
	pèwu	Mògò	n' i	ni-mògò				
bán	péwu	mògò	n' í	nìmògò				
terminer	complètement	homme	et	2.SG	beau-parent_cadet			



REPRESENTING VARIANTS

word	ka
lemma	ka k'
tag	OPT
form	ká
gloss	OPT
parts	

word	k'
lemma	ka k'
tag	OPT
form	k'
gloss	OPT
parts	



REPRESENTING VARIANTS

word	ka
lemma	ka k'
tag	OPT
form	ká
gloss	OPT
parts	

word	k'
lemma	ka k'
tag	OPT
form	k'
gloss	OPT
parts	



Weaknesses:

- ▶ simplistic rule-based morphological analysis
- ▶ slow
- ▶ ugly

Strengths:

- ▶ easily adaptable to other languages
- ▶ integration with Toolbox data formats
- ▶ an integration of the interlinear glossed format into the corpus annotation
- ▶ a systematic representation of the lexical variation in the lexical database and in the corpus

