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

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

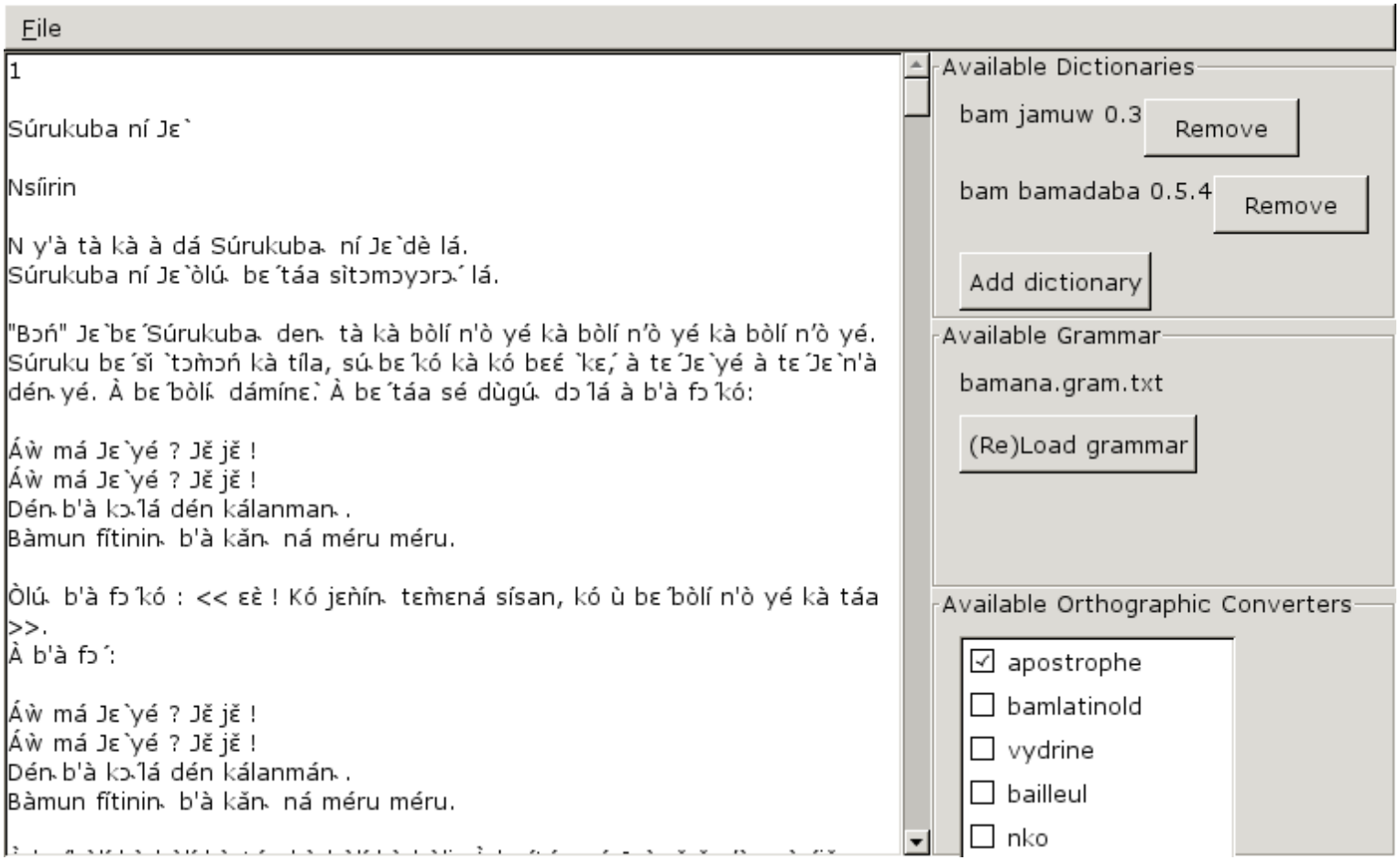
Outline

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)

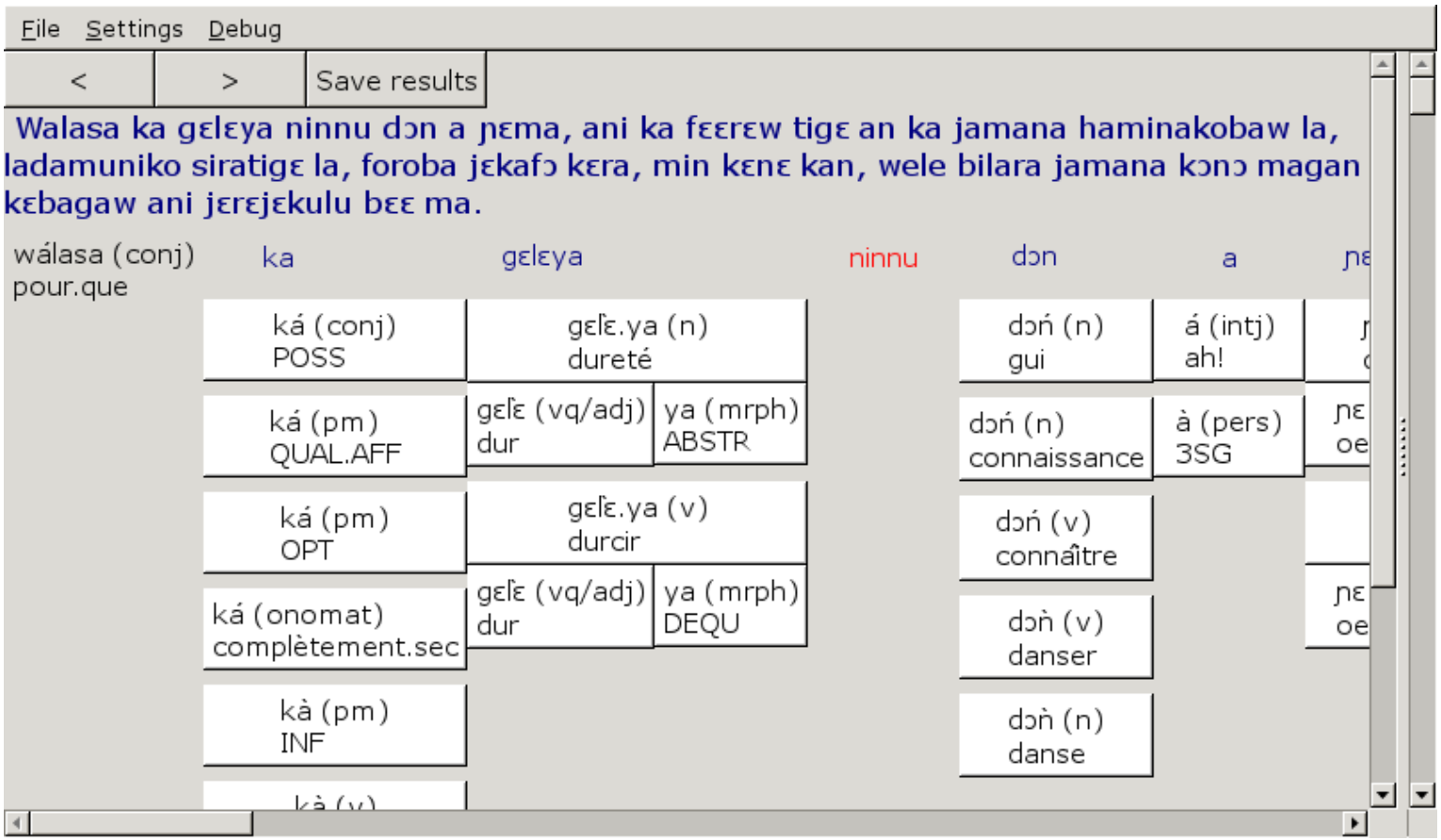
Parser GUI

GParser



Manual disambiguation GUI

GDISAMB



Daba: the code

- Python
- GPL
- http://github.com/maslinych/daba/

Outline

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>.*[aoeuiɛɔ]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]

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) mɔgɔjugu
  mɔgɔ-jugu
  homme-mauvais

- Decompose parser rule:
  - Dictionary is represented as a suffix tree (TRIE)
  - Lookup leftmost substring which is a valid word: mɔgɔjugu
  - Check if a remainder is also a valid word: mɔgɔjugu
  - 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

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

Token annotation

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

Token annotation

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

Token annotation

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

Token annotation

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

Token annotation

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

Glossed examples from the corpus

		la	.	Ni	**mògò**	tè	t'	i
	sùma	*lá*	.	*ní*	**mɔ̀gɔ**	*tɛ́*	*t'*	*í*
	récolte	*dans*	.	*si*	**homme**	*IPFV.NEG*	*aller*	*2.SG*
		olu	kan	.	**Mògò**	minw	minèna	
dá	*òlú*		*kàn*	.	**mɔ̀gɔ**	*mín-w*	*mɲ̀ɛ-na*	
poser	*ce_PL2*		*sur*	.	**homme**	*REL-PL*	*attraper-PFV.INTR*	
		Siyaw	bèe		**mògòw**	ye	balimaw	
		síya-w	*bɛ́ɛ*		**mɔ̀gɔ-w**	*yé*	*bálima-w*	
		race-PL	*tout*		**homme-PL**	*EQU*	*frère-PL*	
	,	fo	ni		**dugumògò**	bèe	tun	ye
kɔ́nɔ	,	*fó*	*ní*		**dùgu-mɔ̀gɔ**	*bɛ́ɛ*	*tùn*	*yé*
à_l'intérieur	,	*jusqu'à*	*si*		**terre-homme**	*tout*	*PST*	*EQU*
			.	Bakari	**somògò**	minw	ka	kòrò
	kósɛbɛ		.	*Bákàrí*	**sɔ́mɔgɔ**	*mín-w*	*ká*	
	très		.	*NOM*	**personne_de_la_famille**	*REL-PL*	*QUAL.AFF*	
	buran		ye	.	**Mògòya**	sira	la	, Bakari
	búran		*yé*	.	**mɔ̀gɔya**	*síra*	*lá*	,
	beaux-parent		*PP*	.	**humanité**	*chemin*	*dans*	,
		pèwu		.	**Mògò**	n'	i	ni-mògò
bán		*péwu*		.	**mɔ̀gɔ**	*n'*	*í*	*nɲ̀mɔgɔ*
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