Daba software for written corpora of underresourced languages

Kirill Maslinsky ^{1,2} Valentin Vydrin ^{3,4}

¹Higher School of Economics, St. Petersburg, Russia ²Institute of Russian Literature of Russian Academy of Sciences, St. Petersburg, Russia ³Inalco, Paris, France ⁴LLACAN (CNRS, UMR-8135), Paris, France

Building a corpus with daba: A typical workflow



Software implementation

- Cross-platform: Linux, Windows, MacOS
- Python 2.7, 3; wxPython GUI library
- Open source: github.com/maslinych/daba

Corpora built with daba

Corpora of Mande languages built as a part of Corpora Mandeica project: http://cormand.huma-num.fr/mandeica/

- Corpus Bambara de Référence (11M)
- Corpus Maninka de Référence (3.5M)
- Corpus du dan de l'Est (0.5M)
- Corpus Mwan (0.05M)
- ...further corpora in progress

Daba's habitat

- limited amount of texts available, especially in electronic formats. As a rule, such languages are underrepresented on the Internet;
- low level of standardization, sometimes competing orthographic systems;
- often insufficient grammatical and lexicographical description;
- corpora built and used mostly by non-native speakers (linguists, language learners etc.) glosses in a European language required.

Daba features

- A framework to create rule-based morphological parsers — allows bootstrapping corpora without annotated training data.
- Support for hierarchically glossed texts.
- Graphical user interface for manual disambiguation.
- Integration with legacy formats ubiquitous in field linguists' data — SIL Toolbox.

corpus publishing environment. Uses manatee indexes built at the corpus build stage. Requires Linux environment, web server, manatee and bonito from NoSketchEngine to be installed. Can be wrapped in a **docker container**

Daba caveats

- A Linguist + Daba ≠ Corpus. Software developer is still required.
- Each corpus is different and requires writing new code: pattern rules, orthographic converters, Makefile for building a corpus etc..
- Work required to publish a smaller corpus is typically larger due to even more inconsistency and instability of writing and linguistic description characteristic of underresourced languages.

Legend:

- code writing requiredGraphical user interface available
- Command line interface only
- Linguist is Ok with data and daba
- Some programming is inevitable
- Happens automatically



NoSketch	mogo (Q corbama-net-non-tonal		defaults
Home Search Word list	Query mogo 12,183 > GDEX 12,183 (7,022.03 per million) Page 1 of 610 Go Next Last		
Corpus info My jobs	03dennyuman_ni_kononin.dis.html	kono sara . Ala fana y' o mogo oiseau mourir-PFV.INTR . Dieu aussi PFV.TR ce homme	sugu dan ka bila espece creér INF mettre
User guide	bailleul-sagesse_bambara_01.dis.html	mogo kana tige dine na , mogo homme PROH couper monde a' , homme	kana da diɲε na . 123) PROH poser monde a' . CARDINAL)
Save Make subcorpus View options KWIC Sont Left Node References Shuffle Sample Filter Sub-hits 1st hit in doc Frequency Node tags Node forms Doc IDs Collocations Visualize	bailleul-sagesse_bambara_01.dis.html) Ala ka kunkolo dalen , mɔɡɔ) Dieu POSS cráne creér-PTCR.RES , homme	te se ka o ci . 43) IPFV.NEG arriver INF ce frapper . CARDINAL)
	bailleul-sagesse_bambara_01.dis.html tow fa le.reste-PL per	jira u la . 31) se ts mɔgɔ montrer 3PL a`. CARDINAL) arriver CORNEG homme	bolo , a bε Ala bolo . 32 CNTRL , 3SG etre Dieu CNTRL . CARDINAL ,
	bailleul-sagesse_bambara_01.dis.html	masa ye Ala ye . 30) ni məgə roi EQU Dieu PP . CARDINAL) si homme	kelen y' a fɔ : « n buran un PFV.TR 35G dire : « 15G beau-parent
	baabu_ni_baabu_04tarawele-daga_ni_kolon.dis.html	kolon fana ko : - daga , ni m⊃g⊃ mortier aussi QUOT : - canari , si homme	min fana ye kolonkala jit REL aussi PFV.TR pilon fois
	bailleul-sagesse_bambara_01.dis.html	ye tomonkolon ye m⊃g⊃ EQU ruines-usage´ PP homme	kana tigɛ diɲɛ na , mɔɡɔ PROH couper monde a` , homme
	bailleul-sagesse_bambara_01.dis.html) Ala ye ko min latigs , mጋgɔ RDINAL) Dieu PFV.TR affaire REL faire_passer , homme	ts se K o tips . 37) IPFV.NEG arriver INF ce gâter . CARDINAL)
	bailleul-sagesse_bambara_01.dis.html	bon bama ye . 300) fɛn tɛ mɔɡɔ grand crocodile PP . CARDINAL) chose CORNEG homme	<mark>dɔ bolo , a jigi ka ɲi .</mark> certain bras , 35G espoir QUAL.AFF bon .
	bailleul-sagesse_bambara_01.dis.html man QUAL.N	go . 77) ko min bɛ mɔɡɔ i deśagredble . CARDINAL) affaire REL IPFV.AFF homme	sOrO , o ka kegun n' i ye . obtenir , ce QUAL.AFF malin et ZSG PP .
	bailleul-sagesse_bambara_01.dis.html	t a teli ni nObO ye . 91) ni mOgO UALLAFF rapide et suivre PP . CARDINAL) si homme	garij&g& banna yɔrɔ min chance terminer-PFV.INTR lieu REL
enu position	bailleul-sagesse_bambara_01.dis.html a 356	a . 35) Ala ye ko min da mɔgɔ · . CARDINAL) Dieu PFV.TR affaire REL poser homme	ye , a lamin£ i la , kojugu PP , 35G maintenir 25G aʻ, affaire-mechant
Enhanced bonito interface			

ketch

NoSketchEngine

https://github.com/maslinych/daba

LIFT 2019, Orléan

kmaslinsky@hse.ru vydrine@gmail.com