Language and Technology Conference
Poznań, November 28, 2015

The Malagasy language in the digital age
Challenges and perspectives

Joro Ny Aina RANAIVOARISON

University of Antananarivo – jororanaivo@yahoo.fr

The Malagasy language

Austronesian origin
23 million speakers in Madagascar
Agglutinative
Written in the Latin alphabet
Word order: verb object subject
Nearly no commercial language processing
Nearly no language resources

Previous work on tools and resources on Malagasy

- Concordancer (Morin, J.Y.)
- Program of recognition of named entities (Poibeau *et al.*, 2003)
- Morphological analyzer based on two-level morphology (Dalrymple *et al.*, 2006)
- Machine translation and spell checker (Raboanary *et al.*, 2008)
- Unitex compatible electronic dictionary of Malagasy simple verbs (Ranaivoarison *et al.*, 2013)

Outline

- Dictionary of verbs
- Objectives
- Unitex
- Methods
- Dictionary of nouns
- Other dictionaries
- BLARK about Malagasy
- Discussion and conclusion

Dictionary of verbs

- Formalized resource of Malagasy simple verbs
- It can be tested and is available with Unitex on http://igm.univ-mlv.fr/~unitex
- The dictionary has 3200 lemmas
- 1810 are distributed with Unitex including frequent verbs
- Potential applications: program of information retrieval system, spell checker for example

Objectives

- Construct all the electronic dictionaries of Malagasy (simple words and multi-word units)
- So that developers or specialist of NLP can construct efficient tools on Malagasy

Unitex

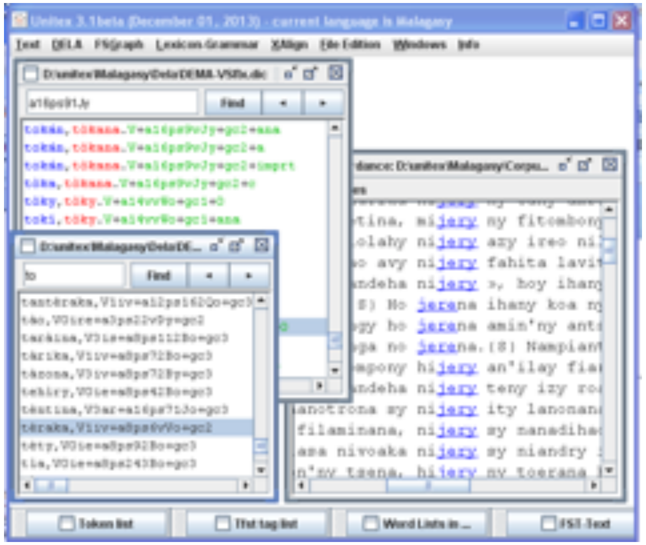
An open-source corpus processor based on language resources

- Dictionary manager
- Morphological analyzer
- Graphical grammar editor
- Concordancer
- Automatic annotator

23 languages

Runs on Linux, Windows, OS X

Paumier (2003)


Unitex 3.1beta (December 01, 2013) - current language is Malagasy
Text
DELA
FSGraph
Lexicon-Grammar
XAlign
File Edition
Windows
Info
D:\unitex\Malagasy\Dela\DEMA-VSfx.dic
a16ps91Jy
Find
D:\unitex\Malagasy\Dela\DE...
to
Find
...dance: D:\unitex\Malagasy\Corpu...
Token list
Tfst tag list
Word Lists in ...
FST-Text

tokàna,tòkana.V+ai4ps9vJy+gc2+ana
tokàna,tòkana.V+ai4ps9vJy+gc2+a
tokàna,tòkana.V+ai4ps9vJy+gc2+imprt
tòka,tòkana.V+ai4ps9vJy+gc2+o
tòky,tòky.V+ai4vv9o+gc1+0
toki,tòky.V+ai4vv9o+gc1+ana

tantèraka,V1iv+ai2ps162Qo+gc3
tào,V0ire+a3ps22v9y+gc2
taràina,V3ie+a8ps112Bo+gc3
tàrika,V1iv+a8ps72Bo+gc3
tàzona,V3iv+a8ps72By+gc3
tebiry,V0ie+a8ps42Bo+gc3
tèatina,V3ar+ai6ps71Jo+gc3
tèraka,V1iv+a8ps6vVo+gc2
tèty,V0ie+a8ps92Bo+gc3
tia,V0ie+a8ps243Bo+gc3

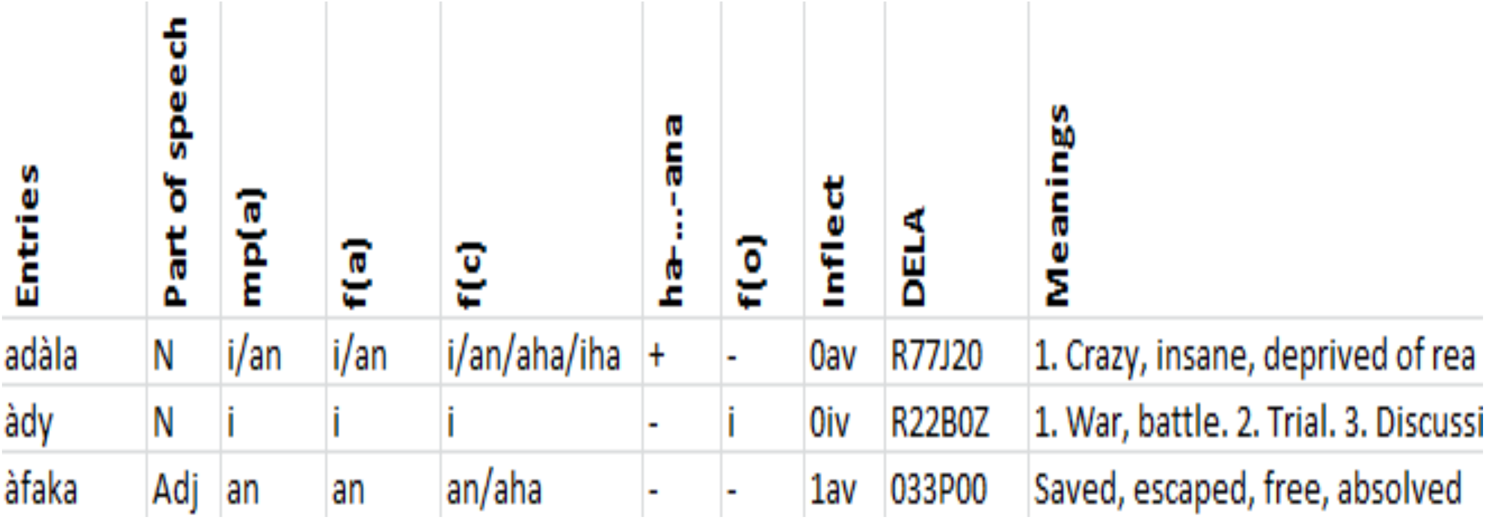
...tina, mijery ny fitombon...
...olahy nijery azy ireo ni...
...o avy nijery fahita lavi...
...ndeha nijery », hoy ihan...
...8) Ho jerena ihany koa n...
...gy ho jerena amin'ny ant...
...pa no jerena.(8) Nampian...
...mpony hijery an'ilay fia...
...ndeha nijery teny izy ro...
...anotrona ny nijery ity lanonan...
filaminana, nijery ny nanadiha...
...ana nivoaka nijery ny niandry ...
...n'ny tsena, hijery ny toerana ...

Methods

- Two-level morphology (Koskenniemi, 1983, 1988; Dalrymple, 2005)
    - Each rule can be applied a priori to each entry
    - Updating rules may damage the good running of the system
- DELA format dictionary (Berlocher, *et al.*, 2006)
    - The dictionary indicates by a number what rule is applied to a given word
    - Easily updatable lexical resource

Dictionary of nouns

Constructing a table
The table informs about noun affixes



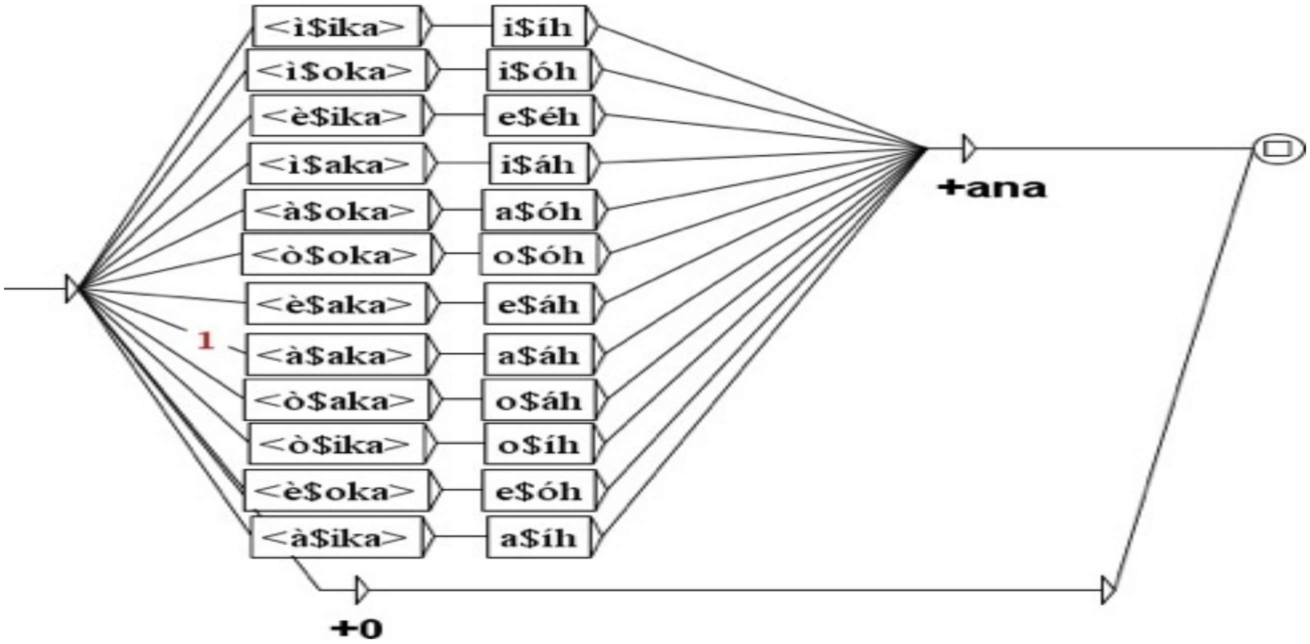
Entries	Part of speech	mp(a)	f(a)	f(c)	ha-...-ana	f(o)	Inflect	DELA	Meanings
adàla	N	i/an	i/an	i/an/aha/iha	+	-	0av	R77J20	1. Crazy, insane, deprived of rea
àdy	N	i	i	i	-	i	0iv	R22B0Z	1. War, battle. 2. Trial. 3. Discussi
àfaka	Adj	an	an	an/aha	-	-	1av	033P00	Saved, escaped, free, absolved

9 forms of nouns recognized with the first entry : *mpiadala, mpanadala, fiadala, fanadala, fiadalana, fanadalana, fahadalana, fihadalana, hadalana.*

Dictionary of nouns

Defining stem classes

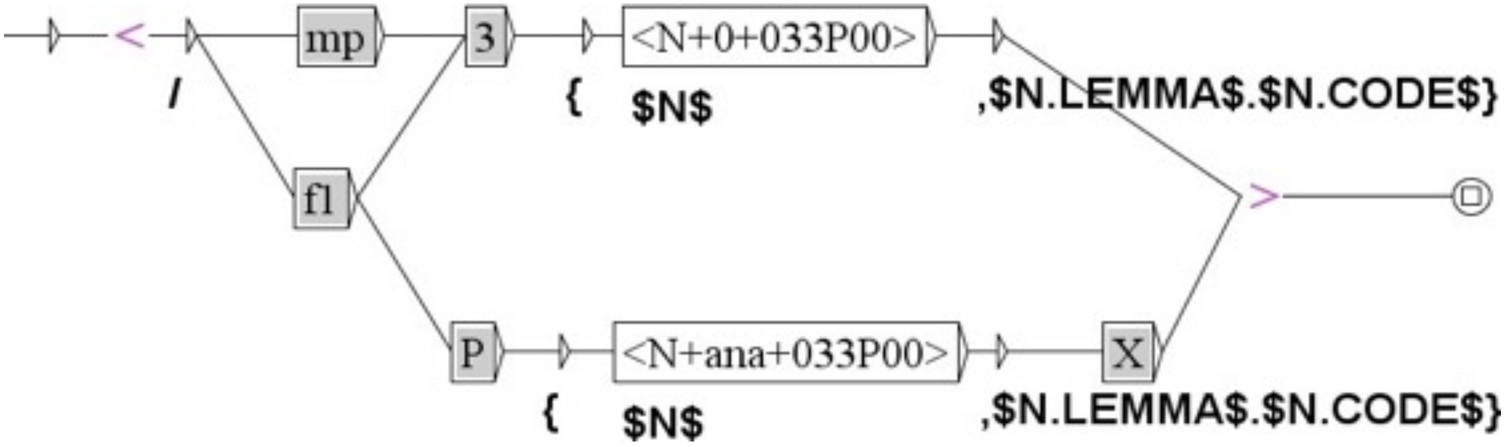
Stem classes describe the morphological variations of a stem when it combines with noun affixes



Dictionary of nouns

Defining affix classes

Affix classes define the combinatorics of affix attached to stems

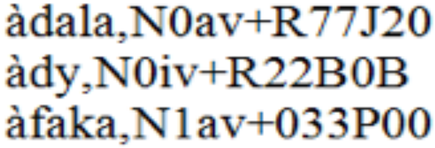


For *àfaka* **adj.** « saved, escaped, free, absolved », such a graph allows to recognize nouns as :

mpanàfaka « liberator, savior, deliverer »
fanàfaka « A. The one that delivers. B. Manner of delivering. »
fanafàhana « inssuance, postage, exemption »
fahafàhana « freedom »

Dictionary of nouns

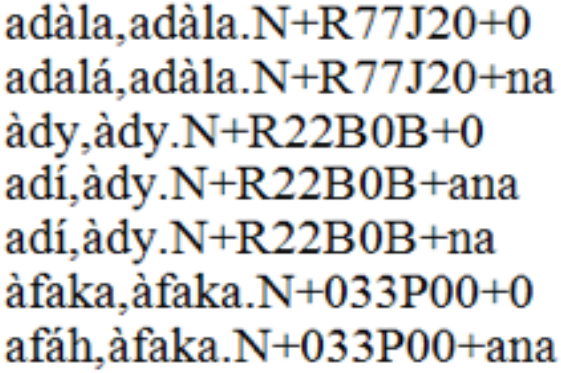
Dictionary of stems



àđala,N0av+R77J20
àđy,N0iv+R22B0B
àfaka,N1av+033P00

The dictionary of stems produces a dictionary of stem variants. Unitex automatically generates the dictionary of stem variants from the dictionary of stems.

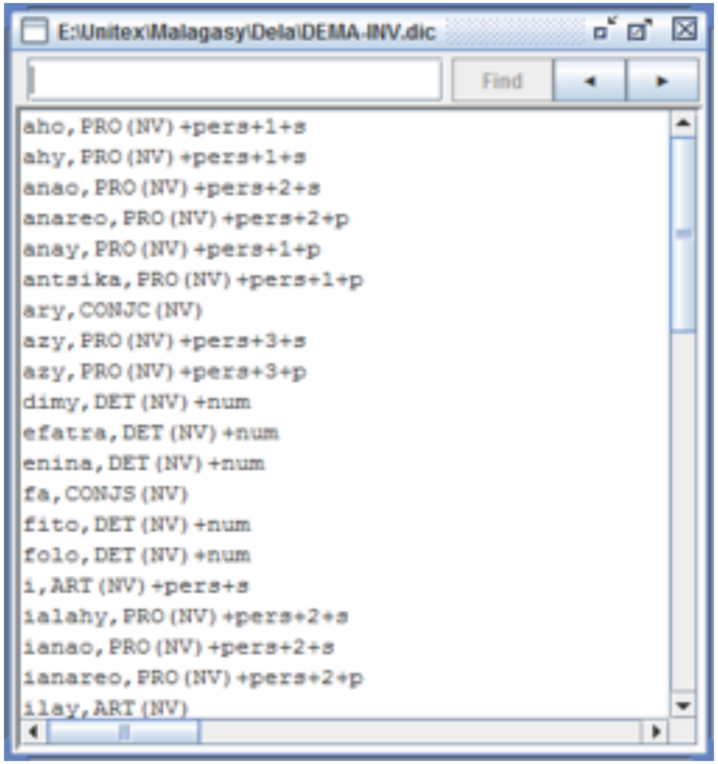
Dictionary of stem variants



adàla,adàla.N+R77J20+0
adalá,adàla.N+R77J20+na
àdy,àdy.N+R22B0B+0
adí,àdy.N+R22B0B+ana
adí,àdy.N+R22B0B+na
àfaka,àfaka.N+033P00+0
afáh,àfaka.N+033P00+ana

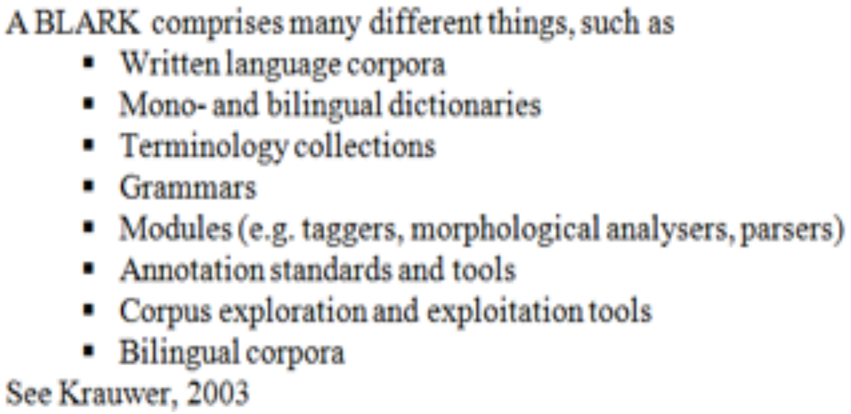
Other dictionaries

- Dictionary of adjectives
- Dictionary of grammatical words
- Dictionary of multi-word units



```
aho,PRO(NV)+pers+1+s
ahy,PRO(NV)+pers+1+s
anao,PRO(NV)+pers+2+s
anareo,PRO(NV)+pers+2+p
anay,PRO(NV)+pers+1+p
antsika,PRO(NV)+pers+1+p
ary,CONJC(NV)
azy,PRO(NV)+pers+3+s
azy,PRO(NV)+pers+3+p
dimy,DET(NV)+num
efatra,DET(NV)+num
enina,DET(NV)+num
fa,CONJS(NV)
fito,DET(NV)+num
folo,DET(NV)+num
i,ART(NV)+pers+s
ialahy,PRO(NV)+pers+2+s
ianao,PRO(NV)+pers+2+s
ianareo,PRO(NV)+pers+2+p
ilay,ART(NV)
```

BLARK about Malagasy



A BLARK comprises many different things, such as

- Written language corpora
- Mono- and bilingual dictionaries
- Terminology collections
- Grammars
- Modules (e.g. taggers, morphological analysers, parsers)
- Annotation standards and tools
- Corpus exploration and exploitation tools
- Bilingual corpora

See Krauwer, 2003

- Among all the requirements and recommendations of BLARK, only a corpus and a dictionary of verbs are available
- Constructing formalized dictionaries is a linguistic challenge
- With dictionaries, creating BLARK items will be easier

Discussion and Conclusion

- Our dictionaries are reliable

With a little corpus that has not been used to construct the dictionary, with 100 verbs

    - 92% are recognized
    - 8% are not recognized

- Resources are easily updatable
    - 1 needs to be injected in the dictionary
    - 2 need to be encoded with another existing transducer of morphological variation
    - 5 need to be set to recognize ellision and hyphen

Updating these kinds of problems are easy with our dictionary

It does not affect the proper functioning of the remains of words in the dictionary.

- Resources or data can help developpers to create engine or tools

Thanks