The Malagasy language in the digital age

Challenges and perspectives

Joro Ny Aina RANAIVOARISON

University of Antananarivo
Centre Interdisciplinaire de Recherche Appliquée au Malgache, Madagascar
jororanaivo@yahoo.fr

Abstract

The creation of useful tools such as spell checkers, or machine translation systems, which would introduce less-resourced languages into the era of new technology and encourage users to use them more, is usually the work of specialists of Natural Language Processing (NLP). For Malagasy, an agglutinative language, the collaboration between specialists of NLP and linguists is required. This paper surveys tools and resources that have been constructed for Malagasy, and, among others, a project (Ranaivoarison *et al.*, 2013) based on the DELA framework (Laporte, 1993) to construct NLP dictionaries of Malagasy by using conventional dictionaries and converting them into a structured, but readable and manually updatable resource usable by Unitex (Paumier, 2003) for morphological analysis. We report on the ongoing construction of an NLP dictionary of simple words (nouns, adjectives, adverbs, grammatical words) with the same DELA methodology and we describe the next steps required.

Keywords: Less-Resourced Language, Malagasy, NLP dictionary, tools, transducer, morphology, stem class, affix class, nouns, adjectives, adverbs, grammatical words.

1. Introduction

The Malagasy language is the national language of Madagascar (about 400 km East of Africa) whose official language is French. It is spoken by 23 millions of people. In the 19th century, missionaries from England and France came to Madagascar and studied this language. Conventional dictionaries such as Freeman (1835), Weber (1853), Malzac (1888), and Malagasy grammar books in English, French or Malagasy, such as Griffiths (1854), Cousins (1894), Andrianony (1960), have been then published. Since Rajaona (1972), Malagasy linguists have produced more scientific, richly documented studies of the Malagasy language. In the 90s, with Rabenilaina (1989), followed by Ralalaoherivony (2004), the Malagasy language entered an era of formalized study of language, with the introduction of the concept of NLP dictionary.

For Malagasy language, some tools exist but they are not usable by the general public; digital medias and devices are used by journalists but information written in Malagasy language is not explored efficiently. In this paper, our goal and motivation are presented in section 2, existing tools for Malagasy language are presented in section 3, and existing resources in section 4. Section 5 reports on the general structure of the Malagasy NLP dictionary, and on the ongoing construction of NLP dictionaries of Malagasy nouns and adjectives. Finally, global perspectives about Malagasy language in the digital age are discussed.

2. Motivations and goals

Resources and tools are required for the processing of Malagasy. Krauwer (2003) cites some resources and tools as BLARK, and Enguehard and Mangeot (2014) cite others: adapted keyboards, spell checker, speech synthesis, machine translation, etc. Among all of them, we chose to construct a monolingual dictionary because it is likely to contribute to practically all other objectives mentioned by these authors.

In fact, words of agglutinative languages usually have several morphemes and their roots have several variants. Morphological analysis delimits roots and affixes. The availability of a dictionary facilitates the implementation of a morphological analyser, a spell checker, and indirectly the annotation of corpora, i.e. several BLARK items. Constructing a dictionary for an agglutinative language is a major scientific challenge, and the first milestone in order to build tools and more advanced natural language applications.

3. Previous work on tools

Some researchers and programmers working on Malagasy language have already constructed or developed tools. A program of concordance exists for example with Pr. Jean-Yves Morin at the University of Canada but features of this program cannot be clarified (information about the product is not available). Researchers and developers at the Institut Supérieur Polytechnique de Madagascar (ISPM) led by Pr. Julien Raboanary realized for example a program of machine translation and spell checker. A demonstration of machine translation has been organized by the ISPM but no literature about the system has been found. The spell checker is developed in JAVA and it proposes corrections for errors found in the text (Raboanary *et al.*, 2008). It has no grammar checking module. Dalrymple *et al.* (2006), in the framework of the Parallel Grammar Project (PARGRAM), have built a morphological analyzer for Malagasy language at Xerox. A program of recognition of named entities has been constructed with Poibeau *et al.* (2003). These tools are not widely used and the dictionaries are not available for research.

4. Previous work on resources

Diwersy (2009) collected a corpus of modern Malagasy newspaper texts, which is freely available under LGPL-LR license.

As for NLP dictionaries, those constructed for existing tools are not available for research. The only available one

is Ranaivoarison *et al.*'s (2013) NLP dictionary of Malagasy simple verbs. It is a structured, but readable and updatable lexical resource based on peer-reviewed morphosyntactic information. It was inspired by Berlocher *et al.*'s (2006) efficient, large-coverage morphological analyzer for Korean, which is an agglutinative language like Malagasy. Berlocher *et al.*'s analyzer is based on readable and updatable resources: an NLP dictionary of stems, finite-state transducers of suffixes and finite-state transducers of generation of allomorphs. This method was adapted to Malagasy, and Unitex (Paumier, 2003) performs morphological analysis of Malagasy verbs with the resource. The construction of the dictionary involved defining stem and affix classes, constructing transducers of variation of allomorphs and combination of morphemes, and populating the dictionary (Ranaivoarison, 2014). We will outline these 3 processes before discussing the advantages of this method for an agglutinative language as Malagasy compared to two-level morphology.

4.1. Collection of empirical data

Due to the large number of lexical entries, the construction of the dictionary required a huge amount of linguistic information, which was first organized into a table identifying the numerous morphemes of verbs for voices, aspects and modes (Fig. 1).

Entrées	Articles	SI(a)	a	c			o	PAzaccito	SI (c,i,i,o)	Numéro inflect	Numéro DELA	Groupes	Modèles
adàla	1. Séduire, duper, tromper, faire perdre la tête, se jouer importuner, embarrasser. 2. Tenir trop, être passionné pour.	zéro	i/an	i/an	-	-	ina	aha/voa/tafa	o	0iv	z16ps42J	gc3	mod32
àdana	1. Aller, faire ou parler lentement, ralentir. 2. Être dans l'aisance.	a	i/an	i/an	a	-	ina	aha/voa/tafa	o	3iv	a16ps142	gc3	mod69
àdy	Combattre, battre.	a	i	i	-	-	-	aha/voa/tafa	o	0av(1)	a8ps4vBo	gc1	mod2

Fig. 1: Table of verbs
(Source: Ranaivoarison, 2014)

The 7 columns from the 3rd column give morphemes attached to each root and the last column shows how verbs vary in contact of morphemes.
This data classifies verbs according to 2 criteria: firstly, morphemes attached to each root and their combinations with each other; and, secondly, the way roots change in contact of morphemes. Thereafter, combinations of morphemes define classes named "affix classes" and types of variations of form of roots define "stem classes".

4.2. Stem classes and affix classes

The table was used to identify stem classes and affix classes. On the one hand, stem classes give the variations of roots when they are adjacent to morphemes; on the other hand, affix classes provide different morphemes attached to the roots. These 2 classifications cross-classify and made up a complex of linguistic data. The first classification enables Unitex to generate variants of roots and the second allows for recognizing morphemes combined with them. This functionality uses the transducers encoded for each class. With these resources, Unitex performs the morphological analysis of inflected forms of verbs in a text.

4.3. Transducers of generation of allomorphs and combination of morphemes

A finite-state transducer of generation of allomorphs is associated to each stem class and specifies the formal variants of roots found in conjugated forms; in parallel, a finite-state transducer of combination of morphemes is associated to each affix class and specifies which morphemes combine with the roots. We encoded these 2 types of transducers graphically with the graph editor of Unitex.

4.4. Dictionary populating

Roots as lexical items are inserted in the dictionary with their stem classes and affix classes, which are lexical information. These 2 pieces of information are represented by 2 codes in the dictionary entries. These 2 codes also identify the transducers of 4.3. The dictionary is freely available.
Unitex imposes conventions of codification of entries, e.g. the 2 codes must be separated by the sign "+" without any space. However, the dictionary can be edited manually for extension and corrections. Then, Unitex produces a dictionary of stem forms. This dictionary is then compressed into a binary file with which Unitex can automatically recognize inflected forms of verbs in a text.

4.5. Two-level morphology

An alternative to the model of handcrafted transducers and DELA dictionaries (Gross, 1989; Berlocher *et al.*, 2006), might be the two-level morphology model (Koskenniemi, 1983), which has been used to deal with agglutinative languages such as Finnish (Koskenniemi *et al.*, 1988), Turkish (Oflazer, 1994) and Malagasy (Dalrymple *et al.,* 2006).
However, the resources of a two-level morphology system are less readable and less easy to update because most rules are very abstract and a priori applicable to any word. Updating one rule may affect a priori any lexical item, endangering the performance of the language system. Since adding new entries in the system may involve changes in the rules, the processing of pre-existing entries can become incorrect.
In contrast, experiments with Korean dictionary DECO (Nam, 1994) showed that dictionaries are easy to maintain and update. With DELA dictionaries, every word is explicitly assigned a specific rule, i.e. a transducer. As a result, updating a transducer in the system may only affect the corresponding words. This makes the system more reliable and the construction of the resources can be cumulative.

5. Malagasy nouns and other parts of speech

Nouns and adjectives in the Malagasy language are often structured as verbs. They have stem classes and affix classes. Rajaona (1972) presents morphemes for both parts of speech. On the model proposed in section 4, we undertook to construct dictionaries of nouns and adjectives allowing for morphological analysis, a basic component of future tools. In this section, we introduce the general structure of an inflectional and morphological NLP dictionary of Malagasy, and we deal with challenges about the dictionary of simple words and multi-word units.

5.1. General structure of inflectional and morphological NLP dictionaries of Malagasy

On the model of Nam's (1994) Korean NLP dictionary, we foresee the structure of a project of a Malagasy morphological NLP dictionary (Fig. 2).

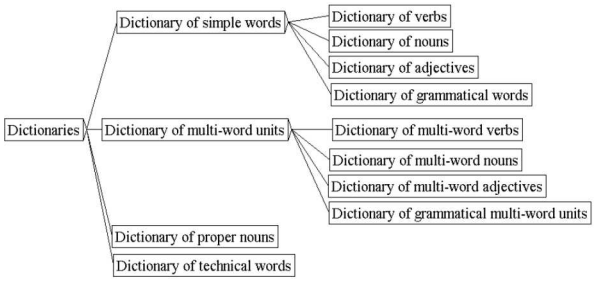

Fig. 2: Structure of a Malagasy NLP dictionary

5.2. Dictionary of simple words

The dictionary of simple words will be composed of dictionaries of verbs, nouns, adjectives, adverbs and grammatical words. A dictionary of verbs is available and distributed with Unitex on http://igm.univ-mlv.fr/~unitex. A large number of inflected forms of verbs are covered and recognized with this resource. We aim to construct other NLP dictionaries: nouns, adjectives, adverbs, grammatical words on the same model of Berlocher *et al.* (2006) and Ranaivoarison *et al.* (2013), so that tools for spell checking, speech synthesis, machine translation and other services may become available more easily.

5.2.1. Dictionaries of nouns and adjectives

Technically, dictionaries of nouns and adjectives have similarities. Both parts of speech combine with specific morphemes and the forms of their roots change in contact with these morphemes. In this paragraph, the description of the dictionary of nouns is presented but the dictionary of adjectives has quite similar features.

The collection of empirical data into a table of nouns is the first step to construct an NLP dictionary of nouns. The table is organized as in Fig. 3.

Entries	Meanings	Entries as nouns	mp(a)	f(a)	f(c)	n	f(o)	Classes	DELA numbers	Inflect numbers
adàla	1. Crazy, insane, deprived of reason, as a little child. 2. Fool.	R	i/an	i/an	i/an/aha/iha	ha...ana	-		R77J20	0av
mpanadala	The deceiver, who unwelcome.	R	an	-	-	-	-		R30000	
mpiadala	One who is fan of.	R	i	-	-	-	-		R20000	
fanadala	A. One who is deceived, unwelcomed. B. The way of embarrassing.	R	-	an	-	-	-		R03000	
fiadala	Too much affection, passion.	R	-	i	-	-	-		R02000	
fiadalana	Passion about, cause of passion, the reason, the location.	R	-	-	i	-	-		R00B00	
fanadalana	The action of fooling, annoying, the means employed to trick, the reason, the location, time.	R	-	-	an	-	-		R00D00	
fahadalana	Madness, stupidity, imbecility.	R	-	-	aha	-	-		R00G00	
fihadalana	The action of going crazy.	R	-	-	iha	-	-		R00B80	
hadalana	Madness, stupidity, imbecility.	R	-	-	-	ha...ana	-		R00020	
ady	1. War, battle. 2. Trial. 3. Discussion, argument. 4. Adjustment, junction, color mixing, agreement of a copy with its original, tuning of instruments or voices.	R	i	i	i	-	i		R22B0Z	0iv

Fig. 3: Table of nouns

The 5 columns from the 4th column give morphemes attached to each root and the last column is for the morphological variation of the roots.

Such linguistic information on nouns consists of their stem and affix classes. The system of stem and affix classes is being organized with precision. The construction of the stem and affix classes of nouns results in the core of an NLP dictionary of simple nouns. They are lexical information and are the basis of the dictionary of nouns.

The codes of stem classes and affix classes are inserted with the entries in the dictionary. The corresponding transducers are being constructed and will allow Unitex to recognize at the same time variants of roots and morphemes of nouns in a text. In fact:

- entries of the dictionary are roots of nouns themselves; and
- lexical information are the code "N" for nouns followed by the codes of stem classes and of affix classes, with delimiters in accordance with Unitex conventions.

The general structure of the dictionary is then: "entry,Nstemclasses+affixclasses". Here are three entries of the dictionary:

àdala,N0av+R77J20
àdy,N0iv+R22B0Z
àfaka,N1av+033P00

For the root *àfaka* "detach, free" which gives nouns as *mpanafaka* "who frees, liberator, savior", *fanafaka* "way of freeing", *fanafahana* "liberation, exemption", *fahafahana* "liberty", the code of stem class "1av" generates the form *afáh* and marks it as compatible with the morpheme *-ana* which is represented by "a" in the code. The letter "v" in the code indicates that the accent in the root *àfaka* shifts forward and gives *afáh.* Figure 4 shows the corresponding transducer N1av.

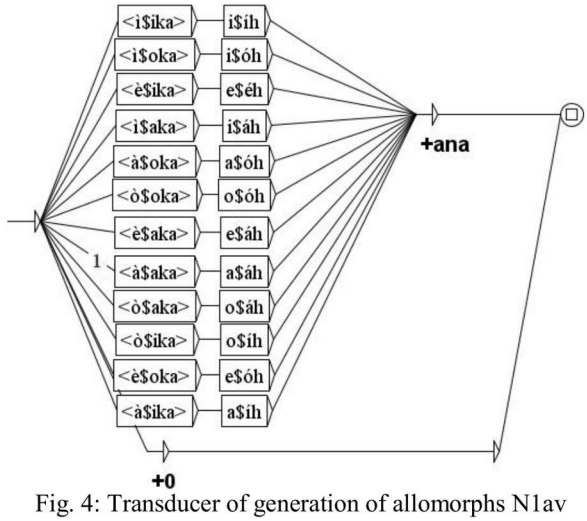

Fig. 4: Transducer of generation of allomorphs N1av

The code of affix class 033P00 indicates that the word *àfaka*:

- is not a noun itself (indicated by 0 at the beginning of the code)
- combines with *an-* (act.-stat.) with *mp-* (indicated by 3 in second position)
- combines with *an-* (act.-stat.) with *f-* (indicated by 3 in third position)
- combines with *an-* (circ.) and *ah-* (circ.) with *f-* (indicated by P in fourth position)
- does not give nouns with *ha-* beginning and the morpheme *-ina* (obj.) indicated respectively by the double 0 at the end of the code.

Associated transducer 033P00 is shown in Fig. 5.

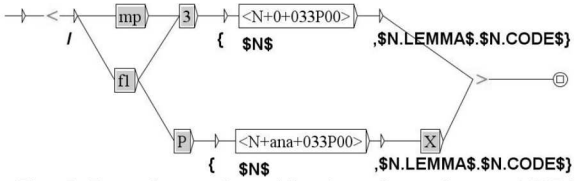

Fig. 5: Transducer of combination of morphemes 033P00

These 2 transducers (Fig. 4 and Fig. 5) operate directly on the dictionary of nouns. The dictionary is inflected to produce the dictionary of noun stems. As regards *àfaka*, the dictionary of noun stems contains the 2 lines below:

àfaka,àfaka.N+033P00+0
afáh,àfaka.N+033P00+ana

After compressing the dictionary and by applying it on a text, Unitex can recognize forms as *mpanafaka, fanafaka, fanafahana, fahafahana,* whereas only the root *àfaka* and its codes have been inserted in the manually-updatable dictionary.
Applied to a portion of Diwersy's (2009) raw corpus, Unitex identified[1]:

- 2 *mpanafaka* (phrases n° 3643, 11 945)
- 0 *fanafaka*
- 1 *fanafahana* (phrase n° 11 944)
- 25 *fahafahana*.

For example, the morphological analysis of the noun *fahafahana* is given in Fig. 6.

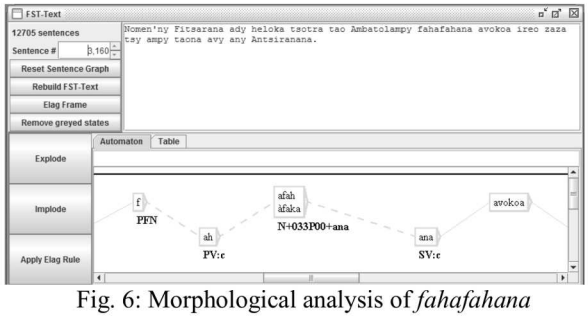

Fig. 6: Morphological analysis of *fahafahana*

5.2.2. Dictionaries of adverbs and grammatical words

Dictionaries of adverbs and grammatical words are slightly different from dictionaries of nouns and adjectives. Malagasy adverbs and grammatical words generally do not combine with specific morphemes as nouns or adjectives do.
Rajaona (1972) analyses examples of adverbs, grammatical words such as personal, interrogative, demonstrative pronouns, interjections, etc. They all have in common that they do not behave as nouns or adjectives: they do not have affix classes and generally their forms do not change.
The collection of these words and their codification into a structured, but readable format are the challenges to construct the dictionaries of adverbs and grammatical words. An excerpt of a dictionary of invariable words under construction is presented in Fig. 7.

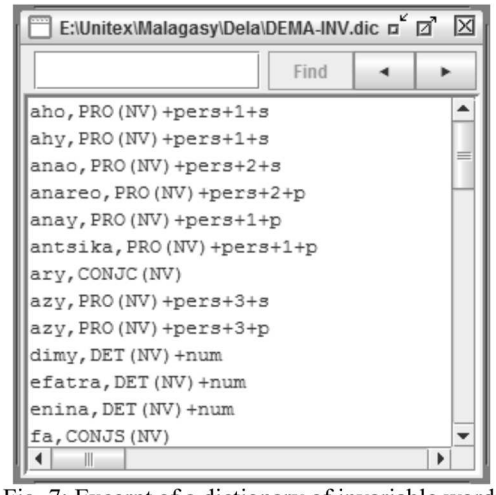

Fig. 7: Excerpt of a dictionary of invariable words

The dictionary of invariable words contains adverbs, prepositions, conjunctions, interjections, articles, personal pronouns, demonstrative pronouns, interrogative pronouns, locative pronouns.

5.3. Construction of a dictionary of multi-word units

The construction of an NLP dictionary of multi-word units is a heavy and complex task, but such a project is not unrealistic. The construction of such a dictionary for Malagasy is different from the similar project implemented for French (Silberztein, 1990) because the entries for Malagasy dictionary are not inflected forms, but roots. However, multi-word units are in principle compositions of inflected forms. Efforts should be made to build this dictionary because it can increase the accuracy of text analysis systems.

6. Discussion

Charon (2011) discusses general information about medias and devices in the digital era. For Malagasy language, journalists as native speakers of Malagasy generally present information in this language. Several types of language data in Malagasy are available (written documents, images, sounds, videos) in substantial quantities in libraries, archives, radio and television networks and even in national or private press centres. They all have in common that they cannot be really explored. A project of Malagasy language data collection can probably encourage research on the processing of this language.
The emerging collaboration between developers and linguists is likely to open Malagasy to the digital sphere and indirectly to enhance the teaching of this language. The construction of NLP dictionaries with root-entries enabling automatic recognition of different inflected forms of simple words (nouns, adjectives, adverbs, grammatical words) would be beneficial for both goals. Spell checking, text processing, information retrieval, information extraction or translation, when available to the general public, would be

[1] This portion of the corpus has about 8764 paragraphs and 12 700 phrases. It is about 2,12 Mb.

useful services. The majority of people who use computers need them. With them, information available on internet could be explored with better precision.

For now, building readable and updatable dictionaries of nouns, adjectives, adverbs and grammatical words is a challenge for research on Malagasy language. Such resources may help developers to construct adaptable tools for this language.

6. Conclusion

The method presented in this paper to construct a dictionary of nouns and adjectives is similar with the method used to build the dictionary of Malagasy verbs. This method offers a large recognition of inflected forms of words in a text after inserting the roots of these parts of speech in the NLP dictionary and encoding lexical information. Malagasy nouns and adjectives are variable words whereas adverbs, grammatical words such as prepositions, conjunctions, demonstrative, personal pronouns are invariable. For the latter, an NLP dictionary different from dictionaries of nouns and adjectives is under construction. These resources (verbs, nouns, adjectives, adverbs, grammatical words) will form the dictionary of simple words of Malagasy language. The main challenges for linguists are to construct systems of stem classes and affix classes for these dictionaries, and the dictionary of multi-word units. Public or private subsidies are needed to help researchers and specialists not only to achieve these objectives but also to develop knowledge about the language and to implement tools to explore it efficiently. In addition to these morphological dictionaries, syntactic dictionaries are also need to be constructed to develop efficient tools. Neither only linguists nor only developers can claim to be able by their own forces and knowledge to build powerful tools able to handle the mass of data relevant to the Malagasy language; only cooperation between both sciences may allow the community to build efficient resources and tools and enter the era of new technology and knowledge sharing.

References

Andrianony (1960). *Gramera na fianarana ny teny Malagasy*. Tananarive: LMS.

Berlocher, I., Huh, H.G., Laporte, É., Nam, J.S. (2006). Morphological annotation of Korean with Directly Maintainable Resources. *Poster session of LREC*. Genoa.

Charon, J.M. (2011). Les medias à l'ère numérique. *Les cahiers du journalisme n° 22/23*.

Cousins, G. (1894). *A concise introduction to the study of the Malagasy language as spoken in Imerina.* Tananarive: LMS.

Dalrymple, M., Liakata, M., Mackie, L. (2006). Tokenization and morphological analysis for Malagasy. In: *Computational Linguistics and Chinese Language Processing* 11 (4), pp. 315-332. Taipei: Institute of Linguistics, Academia Sinica.

Diwersy, S. (2009). *Corpus journalistique du malgache contemporain.* Romance Philology Department University of Cologne.

Enguehard, C., Mangeot, M. (2014). LMF for a selection of African Languages.

Freeman, J. J. (1835). *A dictionary of the Malagasy language: English and Malagasy*. Antananarivo: LMS.

Griffiths, D. (1854). *A grammar of the Malagasy language.* Woodbridge: Edward Pite.

Gross, M. (1989). La construction de dictionnaires électroniques. In : *Annales des télécommunication, tome 44* N°1, 2. Issy-les-Moulineaux/lannion : CNET.

Koskenniemi, K. (1983). *Two-Level Morphology: A general Computational Model for Word-Form Recognition and Production.* Department of General Linguistics, University of Helsinki.

Koskenniemi, K. and Church, K.W. (1988). Complexity, two-level morphology and Finnish. In: *COLLING'88.*

Krauwer, S. (2003). *The basic Language Resource Kit* (BLARK) as the First Milestone for the Language Resources Roadmap. In: *SPECOM.* Moscow, Russia, pp. 8-15.

Laporte, É. (1993). Separating entries in electronic dictionaries of French, in J. Darski and Z. Vetulani, (eds.), *Sprache - Kommunikation - Informatik, Akten des 26. Linguistischen Kolloquiums*, Poznan , Tübingen: Max Niemeyer, pp. 173-179.

Malzac, V., Abinal, A. (1888). *Dictionnaire Malgache – Français.* Tananarive: Imprimerie de la Mission Catholique.

Nam, J. S. (1994). Construction d'un lexique électronique des noms simples en coréen. In : *Lexiques-grammaires comparés et traitements automatiques*. Université du Québec à Montréal : Jacques Labelle, pp. 219-245.

Oflazer, K. (1993). Two-level Description of Turkish Morphology. In: *EACL'06.* Netherlands, Utrecht.

Paumier, S. (2003). *Unitex 3.0. User manual.* Université Paris-Est. English version, Munich : Ludwig-Maximilians-Universität.

Poibeau, T. *et al.* (2003) : The multilingual named entity recognition framework. In : *EACL'03*, vol 2. USA : Association for Computational Linguistics.

Rabenilaina, R.B. (1989). Construction du dictionnaire électronique du malgache parallèlement à celui du français. Communication au Colloque International sur les Industries de la langue, du 21 au 24 Novembre à Montréal, publiée en 1991. In : *Actes du Colloque Tome 1*. Montréal : Office de la Langue Française et Société des Traducteurs du Québec.

Rajaona, S.R. (1972). *Structure du malgache. Études des formes prédicatives.* Fianarantsoa : Ambozontany.

Ralalaoherivony, B. S. (2004). *Quelques problèmes posés par la représentation des unités lexicales dans le dictionnaire morphologique du malgache*. Université d'Antananarivo : CIRAM-DLLM.

Ranaivoarison, J.N.A., Laporte, É., Ralalaoherivony, B. S. (2013). Formalisation of Malagasy conjugation. In: *Language and Technology Conference*. Poznan, Poland. pp.457-462.

Ranaivoarsion, J.N.A. (2014). *Modélisation de la morphosyntaxe du malgache. Construction d'un dictionnaire électronique des verbes simples.* PhD, University of Antananarivo.

Silberztein, M. (1990). Le dictionnaire électronique des mots composés. In : *Langue française,* 87, p. 71.

Weber, J. (1853). *Dictionnaire Malgache – Français.* Île Bourbon : Notre Dame de la Ressource.