# A Lexical database of Malagasy adjectives

*Joro Ranaivoarison*
*Department of Malagasy, University of Antananarivo*
*jororanaivo@llm-u-ank.mg*

## Abstract

This paper deals with an electronic resource under construction. The objective is to construct a lexical database of Malagasy adjectives. Malagasy is an agglutinative language of Austronesian origin, spoken in the African island of Madagascar. The method used to construct the resource is adapted from the approach of Gross (1989) to electronic dictionaries. The content of the resource is based on linguistic analysis and encoded so as to be used by language-processing software. However, care has been taken that all linguistic information is easily readable and updatable. The resource allows for morphological analysis and generation of adjectives, removing obstacles to the construction of computer applications to process the Malagasy language. The originality of this paper also comes from our proposal of a distinction between adjectives in the usual sense and adjectival forms of other parts of speech.

Keywords: lexical database, Malagasy, adjective, inflection, derivation

## 1 Introduction

This research takes place in the context of the construction of a general resource for Malagasy language, and focuses on adjectives. This resource is under construction and must be revised and enriched.

The database is designed to be used by language-processing programs. This goal requires a high level of precision and formalization. We borrowed methodological principles from similar successful projects and we used Unitex/GramLab (Paumier, 2003), an open-source, freely available platform of language processing that also offers functionality for constructing language resources.

As compared to traditional grammar and previous work by linguists on adjectives, we distinguish adjectives in the usual sense from adjectival forms of other parts of speech (in this case, nouns).

The next section provides general information about adjectives in Malagasy. Section 3 sums up the approach of Gross (1989) to electronic dictionaries. In Section 4, we describe how (morpheme-internal) allomorphy is represented in the database. Section 5 is about the dictionary of adjectival lemmas in the usual sense, and Section 6 is about the resources that account for adjectival forms of nouns. We report the experiments performed to test the resources in Section 7. The last section contains concluding remarks.

## 2 Adjectives in Malagasy

Language-independent definitions of adjectives, e.g. “a term used in grammatical classification of words to refer to the main set of items which specify the attributes of nouns” (Crystal, 2008: 11), are suitable for Malagasy, but not precise enough, for example, to distinguish adjectives from verbs with the *m-* prefix. For that matter, we follow Rajaona (1972: 404) and Ralalaoherivony (1995) who draw the line by using the following criterion: verbs have forms in the circumstantial voice, while adjectives do not. Refer to Rajaona (1972: 521), Keenan & Polinski (1998: 609), Dalrymple et al. (2005) about circumstantial voice. So, *mandèha* “walk”, *mandrày* “take” are verbs because the circumstantial forms *andehánana, andráisana* are in use; and elements as *makòtroka, màtotra* are adjectives because the circumstantial forms **akotróhana, *atórana* are not attested. (Graphical accents in Malagasy are optional and indicate stress.)

This section reports on the inflection of adjectives and on the distinction between derived and inflected adjectives that stem from nouns.

### 2.1 Inflection of adjectives

Like verbs, some adjectives receive inflectional markers of grammatical tense in Malagasy (cf. Rajaonarimanana, 1995: 64).

Most adjectives combine with *ho* to express future tense, hence *màro* "is/was numerous" /*ho màro* "will be numerous"*, mànta* "is/was raw"/*ho mànta* "will be raw"*,* etc. This marker of the future tense is not an inflectional morpheme because it can be separated from the adjective by other words.

However, other adjectives show morphological alternation associated with tense, e.g. *m-a-fàna* (PRS-ADJZ-heat) "is hot", *n-a-fàna* (PAST-ADJZ-heat) "was hot" and *h-a-fàna* (FUT-ADJZ-heat) "will be hot". Among the about 2,115 adjectives studied for this paper, about 365 take the *m:n:h* alternation.

Traditional grammarians as Malzac (1950), Rajemisa (1969), and linguists like Rajaona (1972), Rajaonarimanana (1995) consider that *m-, ma-, mi-, man-, -ina, -ana* are adjective-forming affixes. Among these, we choose to segment *ma-, mi-, man-* further into the tense markers *m-, n-, h-* and the adjective-forming affixes *a-, i-, an-*, whenever the morphological alternation *mi-, ma-, man-/ni-, na-, nan-/hi-, ha-, han-* correlates with the difference of tense.

Some adjectives in Malagasy take imperative suffixes (cf. Rajaona, 1972; Catz & Catz, 2017) such as *-a* in *m-a-heréz-a* (PRS-ADJZ-power-IMP) "be powerful" vs. *m-à-hery* (PRS-ADJZ-power) "powerful", *m-a-zoto-a* (PRS-ADJZ-diligence-IMP) "be industrious" vs. *m-a-zòto* (PRS-ADJZ-diligence) "industrious". For now, the imperative mood is not taken into account in the resource.

### 2.2 Derived or inflected adjectives

Many Malagasy adjectives are derived from nouns. As is usual in derivational morphology, the meaning or the syntax of the derived adjective is not entirely predictable. In Table 1, for example , the base nouns *kòtroka*, *hìdy*, *jèmby* have the following derivations: *m-a-kòtroka, m-a-hìdy, jembé-na.*

*Table 1: Derived forms with unpredictable meaning or syntax*

| **Base noun** | **Derived adjective** | **English gloss** |
|---|---|---|
| *kòtroka* "thunder" | *m-a-kòtroka* | warm |
| *hìdy* "lock" | *m-a-hìdy* | selfish |
| *jèmby* "confusion" | *jembé-na* | very dark |

Other adjective-forming affixes have been described in Malzac (1950), Rajemisa (1969) and Rajaona (1972)[cf. Subsection 5.5 below].

In Malagasy, however, the grammatical relation between nouns and adjectives can also fall under inflectional morphology. For about 5% of the adjectives we studied in this research, the meaning and syntax of the denominal adjective are entirely predictable based on the noun stem and the morphological process involved, as in Table 2.

*Table 2: Inflected forms with predictable meaning and syntax*

| **Base noun** | **Inflected adjective** | **English gloss** |
|---|---|---|
| *hànitra* "fragrance" | *m-ànitra* | fragrant, aromatic |
| *tànjaka* "strength" | *m-a-tànjaka* | strong |
| *fàika* "dregs" | *faiká-na* | dreggy |

Such pairs are so regular that the process can be regarded as inflectional, even if the base form and the derived form belong to distinct parts of speech. In other words, these adjectives can be considered as inflected adjectival forms of the corresponding nouns, just like participles are inflected adjectival forms of verbs in English: on the one hand, they may behave as adjectives (cf. the term 'participial adjective' used for example by Kennedy & McNally (1999)), but on the other hand they belong to verb conjugation.

The resource includes derived adjectives such as *m-a-kòtroka* (PRS-ADJZ-thunder) "warm", adjectival inflected forms of nouns like *m-ànitra* (ADJZ-fragrance) "fragrant", and base adjectives like *àvo* (ADJ) "high".

The distinction between derived and inflected adjectives is an inescapable reality of Malagasy, but the main affixes serve both as derivational and inflectional:

- *i-*, in *m-i-kodiadìa* “big and fat (of a child)”, from *kodìa* “wheel”, is derivational, but in *m-i-kitoantòana* “rough, uneven, craggy”, from *kitoantòana* “uneven ground, rough place”, *i-* is inflectional;
- as for *a-* in *m-a-tètika* “frequent”, from *tètika* “ornamental scarification; cutting up small pieces”, it is derivational, but in *m-a-fàna* “hot”, from *fàna* “heat”*, a-* is inflectional;
- as for *m-* in *m-èndrika* “fit, proper, worthy”, from *èndrika* “face, likeness, image”, *m-* is derivational, but in *m-ànitra* “fragrant”, from *hànitra* “fragrance”, *m-* is inflectional.

Consequently, during the construction of the resource, we face difficult decisions in classifying denominal adjectives as derived or inflected, especially when the prefix forming the adjective is *m-*. In such cases, we analyse the adjective as an inflected form of a noun only when we identify a pair of syntactic constructions such as *Misy hànitra* $N_0$ “$N_0$ has fragrance” = *Mànitra* $N_0$ “$N_0$ is fragrant”, where *mìsy* “there is” or *mànana* “have” is a support verb (Ranaivoson, 1996; Lakoarisoa et al., 2011; Jaozandry, 2014; Hamitramalala, 2017), and $N_0$ is an accepted subject noun. In the case of *m-ày* “on fire”, *m-àfy* “hard”, respectively from *hày* “burning”*, hàfy* “hardship”, there are no such pairs of syntactic constructions, since the nominal construction is not in use:

*Mày* $N_0$ “$N_0$ is on fire”
**Mìsy/Mànana hay* $N_0$

*Màfy* $N_0$ “$N_0$ is hard”
**Mìsy/Mànana hàfy* $N_0$

Thus, we encoded them as derived adjectives in the resource. Such lexicological decisions border on the arbitrary, but formal criteria are the best way we know to make them reproducible.

Verbs with a resultative prefix, e.g. *mahatalànjona* “amazing”, *mahavàriana* “stunning” are considered as verbal forms of *talanjona* “amazed”, *varìana* “stunned” (cf. Rajaona, 2004:58) and encoded in our database in the framework of conjugation (Ranaivoraison et al., 2013).

## 3 Electronic dictionaries

The method used to construct the resource for adjectives of Malagasy is adapted from the approach of Gross (1989) to electronic dictionaries. This approach recommends several methodological safeguards.

First, for lexical databases to be usable by programs, all data must be explicit. This situation contrasts with that of dictionaries for human readers, where some information may remain implicit, since readers rely on their linguistic proficiency to infer it.

Next, lexicological and lexicographical decisions are based on the observation of a sufficiently large number of lexical entries. Entries are systematically inventoried and decisions are based on this inventory, not on sporadic observations on a limited sample of entries, a practice that would be more likely to necessitate revisions of these decisions.

The resources must be readable, so that they can be updated.

Finally, modes of inflection are defined explicitly, so that the inflected forms of a stem can be generated automatically. A mode of inflection is the set of morphological changes underwent by a stem to generate its inflected forms. Several lexical entries can share the same inflectional mode, as *ring* and *sing* in English. The method requires that inflectional modes are defined independently of one another, in order to avoid constructing a hierarchy of general rules and exceptional rules, since such hierarchies are usually complex to maintain later, as the database undergoes updates. As each inflectional mode is independent, updating one does not require updating others. In consequence, we assign an identifier to each inflectional mode and we mark in each entry of the dictionary the identifier of the type of inflection applicable. Thus, knowing if a rule is applicable to a lexical entry does not require applying it. This contrasts with Two-level morphology (Koskenniemi, 1983), where rule scope is encoded in rules, so that knowing if a lexical entry is affected by a change in a rule requires applying both versions of the rule. In addition, two-level rules are ordered, so that knowing if a lexical entry is affected by a change in rule order requires applying both versions of all rules. These features are practical obstacles to updates, corrections and extensions of two-level databases.

Thus, our approach encodes the generation of inflected forms from their stems, providing a formal link between them, e.g. between *sing* and *singing* in English. However, the approach does not do the same between derivatives and their bases, as *speak* and *speech*, since semantic and syntactic irregularities reduce the potential applications of such generation. Thus, derived words such as *mèndrika* "fit, proper, worthy", *m-i-kodiadìa* "big and fat (of a child)" are encoded as stems.

Gross' approach, devised for inflectional languages, has been extended to agglutinative languages (Berlocher et al., 2006; Ranaivoarison et al., 2013) by distinguishing two levels of morphological changes:

- morpheme-internal allomorphy, e.g., the noun *sómotra* "beard" becomes *somór* immediately before some suffixes,

- affixations, e.g., *-ina*, a morpheme of formation of adjectival forms, can be suffixed to *somór*, giving *somórina* "bearded".

Through the two steps corresponding to these two levels, the inflected form *somòrina* can be generated from the lemma *sómotra*, or conversely, the lemma can be recognised from the inflected form.

The Unitex open-source platform of language processing (Paumier, 2003) is compatible with resources devised according to this approach and encoded in the DELA format (Gross 1989: 8). Unitex performs top-quality inflection, compression and lookup (Neme, Paumier, 2019) of lexical databases encoded in this format. In contrast, the Text encoding initiative's dictionary encoding formats are mainly designed for human-oriented dictionaries (TEI Consortium, 2022) and do not address the processing of lexical databases.

In the electronic-dictionary approach, lexical databases mainly cover parts-of-speech and inflection. This contrasts with the lexicon-grammar approach, also proposed by Gross (Elia, 1978), which also investigates the applicability of syntactic operations. Due to this difference, sense distinctions are more fine-grained in the latter approach (Laporte, 1991).

## 4 Allomorphy

This section provides information about morpheme-internal allomorphy and how it is encoded so that lexical variants of stems can be produced automatically. We list the phenomena that affect stems and we describe graphs that encode allomorphy.

### 4.1 Phenomena of allomorphy

Immediately before or after affixes, some stems do not vary, as in *èrika* "drizzling rain"/ *m-èrika* "drizzly, misty", *dìo* "cleanliness, purity" /*m-a-dìo* "clean, clear, pure", but most do. They can undergo 5 types of variation:

- prosodic alternations as in *sòmotra/somór* (the accent shifts from the first *ò* to the second *ó*)
- insertion of a letter as in *safòfoka/tsafòfoka* (insertion of *t* before the stem in the inflected form *mantsafòfoka*)
- substitution of a letter *hàtsiaka/gàtsiaka* (substitution of *g* in the inflected form *mangàtsiaka* for *h* in the stem)
- extension as in *nòfo/nofós* (insertion of *s* in the inflected form *nofósana*)
- deletion of a letter as in *vorètra/orètra* (deletion of the letter *v* of the stem in the inflected form *mamorètra*) or in *sòmotra* /

*somór̲* (deletion of the *t* and *a* of *sòmot̲r̲a̲* in the inflected form *somór̲ina*)

All allomorphic stems of adjectives undergo one or several of the phenomena above in their lexical variants.

For a deeper study of the phenomena affecting the stems in Malagasy, refer to Rajaona (2004).

### 4.2 Encoding allomorphy with graphs

To encode the lexical variants of stems, we attach to each entry a code that identifies accurately the applicable variations. These variations are encoded in the form of a graph like that of Fig. 1, which produces automatically the lexical variants of stems like *pànda* "freckles", *kilèma* "deformity", *kìbo* "belly".

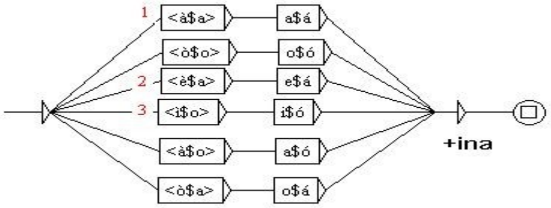


*Figure 1: Graph of allomorphy* '**0v**'

Path 1 encodes *pànda* > *pandá* in *pandáina* "freckly, sunburnt"; path 2, *kilèma* > *kilemá* in *kilemáina* "maimed"; and path 3, *kìbo* > *kibó* in *kibóina* "big-bellied". This technique of lexical marking facilitates updates. The '**0v**' identifier is the name of the graph and is attached to those words to encode these prosodic alternations. So, several lexical entries such as *tràtra* "breast", *tsikòko* "scabs" share the same inflectional mode '**0v**' (identifier) to produce automatically lexical variants as *tratrá, tsikokó.*

Other lemmas have *-ka, -tra, -na* endings, and their final *-a* is missing in most of their allomorphs: the corresponding graphs specify the deletion of this vowel.

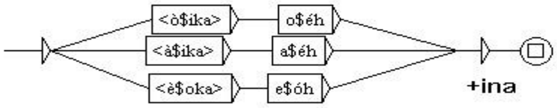


*Figure 2: Graph of allomorphy* '**1v**' *for lemmas in* -ka.

For example, the '**1v**' graph in Fig. 2 deletes the ending *-ka* and inserts a letter *h*, in addition to the prosodic and/or vocalic alternations, encoding variations such as *dònika* > *donéh* in *donéhina* "suffering from mumps", *fàsika* > *faséh* in *faséhina* "sandy", *tsèroka* > *tseróh* in *tseróhina* "dirty, scurfy".

If the lemma contains inflectional prefixes, the graphs also strip them off.

Thus, graphs of allomorphy for stems in Malagasy are divided in 4 types: for those with *ka, tra,* or *na* endings, and for those with none of these endings. In addition, they specify some of the five phenomena identified in Subsection 4.1. Thirty graphs of allomorphy are used for about 2,115 adjectives.

We will now describe the other resources for lexical entries of adjectives, and then those for adjectival forms of nouns.

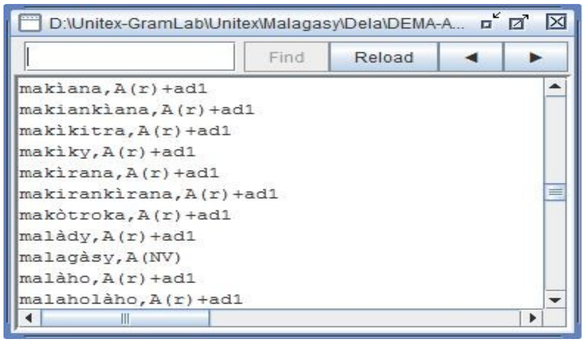
## 5 Encoding lexical entries of adjectives

Base adjectives such as *àntitra* "old", *àvo* "high", and derived adjectives such as *makìkitra* "determined", *makòtroka* "warm" are all encoded in the lexical database as adjectival entries without distinction.

We report on the different components of this resource: lexical entries and graphs of affixation, and we mention the main inflectional and derivational morphemes of adjectival entries.

### 5.1 Lexical entries

Our resource contains about 2,000 adjectival lemmas (Fig. 3). Entries and morphological codes are separated with a comma. The letter "A" is the morphological code of adjectives. The code in parentheses indicates the graph of allomorphy applicable to the entry, and the name "ad1" indicates the graph of affixation for the words that take the *m-:n-:h-* morphological alternation, as in *malàdy* "is quick to hear": *nalády* "was quick to hear": *halády* "will be quick to hear".



```
makìana,A(r)+ad1
makiankìana,A(r)+ad1
makìkitra,A(r)+ad1
makìky,A(r)+ad1
makìrana,A(r)+ad1
makirankìrana,A(r)+ad1
makòtroka,A(r)+ad1
malàdy,A(r)+ad1
malagàsy,A(NV)
malàho,A(r)+ad1
malaholàho,A(r)+ad1
```

*Figure 3: Resource of adjectival entries.*

## 5.2 Graphs of affixation

Graphs of affixation specify which combinations of grammatical and lexical morphemes make up inflected forms of adjectives. In the present state of the lexical database, the graphs of affixation take into account the tense prefixes *m-:n-:h-*, but not yet affixes for imperative.

Take for example the lemmas *marènina* "deaf", *makàka* "spacious". The *m-* in the beginning is the prefix of present tense, so the stems are *arènina, akàka,* which are generated by the graph of allomorphy identified in their entries. The "ad1" graph of affixation (Fig. 4) specifies that this stem takes the three tense prefixes.

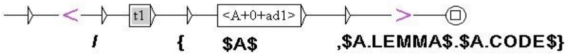


*Figure 4: Graph of affixation "ad1".*

In this graph, the '**t1**' box is a call to a subgraph that represents the *m-:n-:h-* morphological alternation. The **<A+0+ad1>** box stands for the relevant variant of the stem, i.e. *arènina, akàka,* among others.

Among the about 2,000 adjectival entries, about 505 begin with *m*, about 300 of which take the *m:n:h* alternation. Adjectives without the *m-:n-:h-* alternation, such as *mòana* "speechless" (**nòana* and **hòana* are not in use), form a syntactic future with *ho*, as in *ho mòana* "will be speechless".

## 5.3 Inflectional and grammatical codes

The codes of parts of speech used in this part of the database are:

- A for adjective
- T for tense marker

Inflectional codes are:

- r, p, f respectively for present, past, future tenses
- n for indicative mood.

## 5.4 Main inflectional morphemes of adjectival entries

Table 3 contains the main inflectional morphemes used in this part of the database.

*Table 3: Main inflectional morphemes of adjectives*

| | Affix | Example | Gloss |
|---|---|---|---|
| Prefixes of tense (T) | *m-* | *madio* | clean (present) |
| | *n-* | *nadio* | clean (past) |
| | *h-* | *hadio* | clean (future) |

## 5.5 Derivational affixes of adjectives

Table 4 shows different affixes of adjectives generally considered as derivational. The words containing those affixes are not numerous. In the resource, we analyse them as derivatives and encode them as adjectival lemmas. Since our database does not attempt to link them to their bases (cf. Section 3), these affixes have no formal existence as such in the database: the affix/base segmentation is not encoded.

In some denominal adjectival lemmas in *m-*, the initial *m-* is not present in the base noun, as in *màmy* "sweet" from *hàmy* "sweetness", *màty* "dead" from *fàty* "death", but it does not take the *m:n:h* tense alternation. In these entries, we analyse *m-* as a derivational adjectivizing prefix, just like those of Table 4, and we do not encode the affix/base segmentation.

*Table 4: Derivational affixes of adjectives*

| Affix | Example | Gloss |
|---|---|---|

| | | |
|---|---|---|
| *ba-* | *bakaka* | badly plaited, as mats |
| *do-* | *dorehitra* | very red, scarlet |
| *fo-* | *forehitra* | consumed |
| *faha-* | *faharoa* | second |
| *fa-* | *farofy* | sickly |
| *ka-* | *kaozatra* | very lean |
| *ki-* | *kiboribory* | round |
| *ki-…-ina* | *kibotaina* | plump |
| *ko-* | *kosesy* | frequent |
| *sa-* | *saresaka* | talkative |
| *so-* | *somatroka* | drab |
| *ta-* | *takariva* | about dusk |
| *tan-…-ana* | *tandemena* | faint |
| *to-* | *tolantsika* | arqued |
| *tsa-* | *tsatselika* | agile |
| *tsi-* | *tsilotidotika* | dirty |
| *va-* | *varozaka* | weak, exhausted |
| *-om-* | *somariaka* | glad |
| *-il-* | *kilitika* | extremely small |
| *-ir-* | *kiritika* | extremely small |

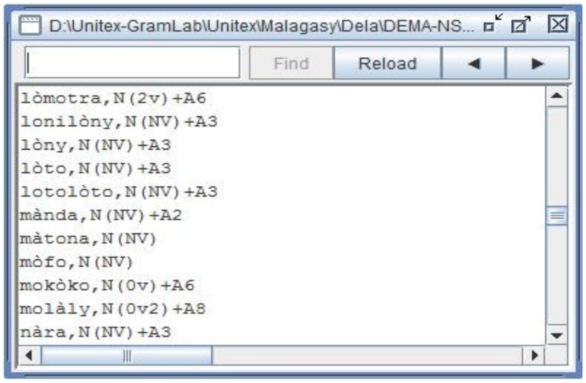
## 6 Adjectival forms of nouns

In our lexical database, adjectival inflected forms of nouns as *vintánina* "who has a destiny", from *vìntana* "destiny", or *mazòto* "diligent, industrious", from *zòto* "diligence", are encoded in a dictionary of nouns, in the form of resources that allow for segmenting and generating these adjectival forms. For example, the noun *lòto* "filth, dirtiness" receives the inflectional prefix forming adjectives *a-*, and the tense prefixes *m-*, *n-* or *h-*, hence *malòto* "is filthy, dirty", *nalòto* "was filthy, dirty", *halòto* "will be filthy, dirty".

We report in this section on the resources for these forms: lexical entries and graphs of affixation, and we list the relevant inflectional morphemes.

### 6.1 Lexical entries

The resource contains about 400 entries of nouns (Fig. 5), 115 of which have adjectival inflected forms. The code "N" is for 'noun' and the codes in parentheses identify the graphs of allomorphy. The graphs of allomorphy are those described in Subsection 4.2. The codes "A6", "A2" or "A3" are for the graphs of affixation. For example, nouns with the code A3, such as *lòto,* receive the *m-*: *n-*: *h-* morphological alternation of tense and the prefix of formation of adjectives *a-*, hence *malòto*, which is thus segmented as *m-a-lòto*, where the lemma is represented by the nominal stem *lòto*. The 'A3' graph encodes this segmentation in three morphemes and formally represents *a-* as an inflectional prefix.



```
lòmotra,N(2v)+A6
lonilòny,N(NV)+A3
lòny,N(NV)+A3
lòto,N(NV)+A3
lotolòto,N(NV)+A3
mànda,N(NV)+A2
màtona,N(NV)
mòfo,N(NV)
mokòko,N(0v)+A6
molàly,N(0v2)+A8
nàra,N(NV)+A3
```

*Figure 5: Resource for adjectival inflected forms of nouns.*

This contrasts with *marènina* "deaf", analysed as *m-arènina* in Subsection 5.2: as *marènina* is an adjectival lemma, it does not contain any inflectional prefix of formation of adjectives, and the first *a* is part of the stem. Thus, graphs of affixation for adjectival forms of nouns are different from those for adjectival entries.

### 6.2 Graphs of affixation

We inventory thirteen graphs for adjectival forms of nouns.

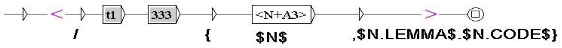


*Figure 6: Graph of affixation 'A3'.*

The 'A3' graph (Fig. 6) describes the combinations of affixes that make up adjectival forms such as *m-a-dìty* "gummy, resinous", *m-a-sìra* "salty" and their own inflected forms. The '**t1**' box represents the *m-:n-:h-* morphological alternation, the '**333**' box contains the prefix *a-* forming adjectives (PAdj) and the **<N+A3>** box indicates to which nouns the graph is applicable, here *dìty* "gum, resin", *sìra* "salt", among others.

### 6.3 Grammatical codes

The additional codes of parts of speech used in this part of the database are:

- N for 'noun'
- PAdj for 'prefix forming adjectives'
- SAdj for 'suffix forming adjectives'.

### 6.4 Inflectional morphemes of adjectives

About 65 of the nominal entries that we studied produce adjectival forms that take the *m-:n-:h-* alternation. Table 5 contains the additional inflectional morphemes used in this part of the database.

*Table 5: Inflectional morphemes of adjectives*

| | Affix | Example | Gloss |
|---|---|---|---|
| Adjectivizing prefixes (PAdj) | *m-* | *màizina* | dark |
| | *i-* | *mimànda* | having defences |
| | *a-* | *matànjaka* | strong |
| | *an-* | *mangàtsiaka* | cold |
| | *am-* | *mamirifiry* | cold |
| Adjectivizing suffixes (SAdj) | *-ina* | *fasehina* | sandy |
| | *-ana* | *nofosana* | fleshy |
| | *-na* | *faikána* | dreggy |

The inflectional adjectivizing prefix *m-* occurs in words as *màizina* "dark" from *àizina* "darkness", *mànitra* "fragrant" from *hànitra* "fragrance": in these forms, it does not take the role of tense prefix. They form a syntactic future with *ho*.

## 7 Tests

In a novel by Clarisse Ratsifandrihamanana, a well-known Malagasy writer, which has 2,400 sentences and about 55,600 tokens, we recognize with our database about 380 unique forms of adjectives. The representation of the result with Unitex is different for base or derived adjectives and for adjectival forms of nouns.

Two experiments are presented in this section: segmentation and generation.

### 7.1 Segmentation

In this experiment, the Unitex platform used the lexical database and the graphs to segment text. The segmentation of *mareforèfo* "a bit fragile", a derived adjective that accepts the *m-:n-:h-* morphological alternation, is presented in Fig. 7.

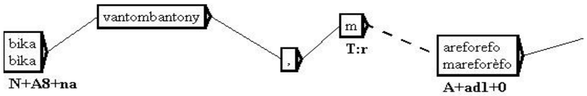


*Figure 7: Representation of a derived adjective with a tense marker.*

The two boxes connected by a broken line represent 1) *m-*, recognized as a tense marker (T) of present tense (r), and 2) *areforefo*, labeled **A+ad1+0**, which means that it is an adjective and lists two features; this form is attached to the lemma *mareforèfo.*

An adjectival inflected form of a noun is segmented as shown in Fig. 8.

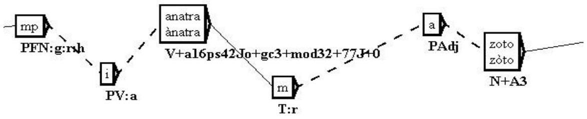
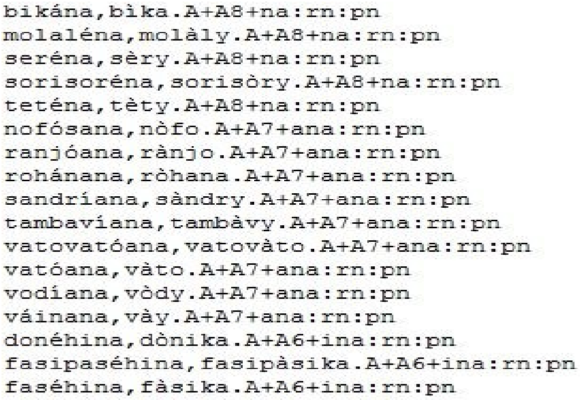


*Figure 8: Representation of an adjectival form of a noun.*

In *mpianatra mazòto* "industrious student", *mazòto* "diligent, industrious" is an adjectival inflected form of *zòto* "diligence". Unitex shows the morpheme of tense *m-* (present), the morpheme forming adjectives *a-* (Padj) and the nominal stem (N+A3).

### 7.2 Generation of adjectives

Unitex has a generator of words that lists the adjectival inflected forms specified by the database (Fig. 9). Each semicolon begins a new bundle of inflectional features. The tense codes **:r**, **:p** and **:f** are mutually exclusive because a word cannot be in the same interpretation at two different grammatical tenses, but the other code **n** qualifies the same interpretation as the tense preceding them. Thus, **A+A8+na:rn:pn** means that *bikána* "well formed" can be in the present or in the past tense, but in both cases it is in the indicative.



```
bikána,bìka.A+A8+na:rn:pn
molaléna,molàly.A+A8+na:rn:pn
seréna,sèry.A+A8+na:rn:pn
sorisoréna,sorisòry.A+A8+na:rn:pn
teténa,tèty.A+A8+na:rn:pn
nofósana,nòfo.A+A7+ana:rn:pn
ranjóana,rànjo.A+A7+ana:rn:pn
rohánana,ròhana.A+A7+ana:rn:pn
sandríana,sàndry.A+A7+ana:rn:pn
tambavíana,tambàvy.A+A7+ana:rn:pn
vatovatóana,vatovàto.A+A7+ana:rn:pn
vatóana,vàto.A+A7+ana:rn:pn
vodíana,vòdy.A+A7+ana:rn:pn
váinana,vày.A+A7+ana:rn:pn
donéhina,dònika.A+A6+ina:rn:pn
fasipaséhina,fasipàsika.A+A6+ina:rn:pn
faséhina,fàsika.A+A6+ina:rn:pn
```

*Figure 9: Generation of adjectives*

## 8 Conclusion

We described a lexical database of Malagasy adjectives under construction. The content of the resource is based on linguistic analysis, bearing in mind relevant methodological safeguards. We make a distinction between adjectives in the usual sense and adjectival forms of nouns. The database can be used by language-processing software. The resource allows for morphological analysis and generation of adjectives. It is not limited to the adjectival forms occurring in a corpus of texts, but takes into account our competence as a native speaker and grammatical tradition.

This research has several potential applications. Its results can be used in grammar textbooks to describe the inflection of Malagasy adjectives. Linguists can use them to launch queries for grammatical configurations containing adjectives, e.g. noun phrases. Finally, lexical databases remove obstacles to the construction of computer applications to process languages.

## Acknowledgements

We address our special thanks to the reviewers for their relevant comments and suggestions for the clarity of our paper.

## References

Andriamise, L, Ranaivoson, JF, Rakotoalison, SF 2011, "Les locutions support en malgache. Le cas de *misy azy*," *Lexis and Grammar*, University of Cyprus, pp. 21-28.

Berlocher, I, Huh, HG, Laporte, E, Nam, JS 2006, "Morphological annotation of Korean with directly maintainable resources", *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa (Italy).

Catz, I & Catz, S 2017, *Standard Malagasy Grammar & List of 410+ Common Verb Conjugation,* edited by Kimmerling Razafindrina, Fetra Marc Humbert Rahajason and Vololona Fenohaja.

Crystal, D 2008, *A dictionary of linguistics and phonetics,* UK, Blackwell.

Dalrymple, M, Liakata, M & Mackie, L 2005, "A Two-level Morphology of Malagasy", *Proceedings of the 19th Pacific Asia Conference on Language, Information and Computation*, pages 83–94, Taipei, Taiwan, R.O.C, Institute of Linguistics, Academia Sinica.

Elia, A 1978 "Pour un lexique-grammaire de la langue italienne : les complétives objet", *Lingvisticae Investigationes* 2 ( 2), pp. 233-276.

Gross, M 1989, "La construction de dictionnaires électroniques", *Annales des Télécommunications*, vol. 44 (1-2), pp. 4-19.

Hamitramalala, R 2017, *Vers une typologie des collocations à verbe support en malgache*, PhD, Université de Montréal.

Jaozandry, M 2014, *Les prédicats nominaux du Malgache. Étude comparative avec le français*, PhD, Université Paris-Nord.

Keenan, E & Polinski, M 1998, "Malagasy (Austronesian)", *The handbook of Morphology,* ed.

Andrew Spencer and Arnold M. Zwicky, New Jersey, John Wiley & Sons.

Kennedy, Ch & McNally, L 1999, "From Event Structure to Scale Structure : Degree Modification in Deverbal Adjectives", Tanya Matthews and Devon Strolovitch (eds), *9th Semantics and Linguistic Theory Conference*, pp. 163-180.

Koskenniemi, K 1983, *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. PhD, University of Helsinki.

Laporte, E 1991, "Separating Entries in Electronic Dictionaries of French", *Sprache - Kommunikation - Informatik. Akten des 26. Linguistischen Kolloquiums, Poznan 1991*, J. Darski and Z. Vetulani eds., Tübingen: Max Niemeyer, pp.173-179.

Malzac, RP 1950, *Grammaire malgache,* Paris, Société d'éditions géographiques, maritimes et coloniales.

Neme, A & Paumier, S 2019, "Restoring Arabic vowels through omission-tolerant dictionary lookup", *Language Resources and Evaluation* 54, pp. 487-551.

Paumier, S 2003, *Unitex manual*, University of Marne-la-Vallée, Paris.

Rajaona, S 1972, *Structure du malgache¸* Fianarantsoa, Ambozontany.

Rajaonarimanana, N 1995, *Grammaire moderne de la langue malgache,* Paris, L'Asiathèque.

Rajemisa-Raolison, R 1969, *Grammaire malgache,* Fianarantsoa, Ambozontany.

Ralalaoherivony, BS 1995, *Lexique-grammaire du malgache. Constructions adjectivales,* Thèse de doctorat, Université Paris 7.

Ranaivoarison, J, Laporte, E & Ralalaoherivony, BS 2013, "Formalization of Malagasy conjugation", *Language and Technology Conference*, Poznań (Poland).

Ranaivoson, JF 1996, *La nominalisation en malgache. Étude des formes* manao N, PhD, Université d'Antananarivo.

Richardson, J 1885, *A new Malagasy-English Dictionary,* Antananarivo, The London Missionary Society.

TEI Consortium, eds. (2022) *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* 4.4.