# 1

# Tokenization and Morphological Analysis for Malagasy

Mary Dalrymple, Maria Liakata, and Lisa Mackie

## 1.1 Introduction

We present a tokenizer and finite-state morphological analyzer (Beesley and Karttunen, 2003) for Malagasy, based primarily on the discussion of Malagasy morphology in Keenan and Polinsky (2001) and Randriamasimanana (1986). Words in Malagasy are built from roots by means of a variety of morphological operations such as compounding, affixation, and reduplication. We analyze productive patterns of nominal and verbal morphology, describing genitive compounding and suffixation for nouns, and various derivational processes involving compounding and affixation for verbs. Our work offers a computational analysis of Malagasy morphology, and forms the basis of our computational grammar and lexicon of Malagasy within the framework of the PARGRAM project.

## 1.2 Malagasy in the PARGRAM Project

Malagasy is an Austronesian language spoken by about seventeen million people on the island of Madagascar (Lewis et al., 2015). It was one of the target languages in the Verb-Initial Grammars subproject[1], a collaborative project within the PARGRAM initiative to develop computational lexicons and grammars of the verb-initial languages Malagasy and Welsh within the shared linguistic framework of Lexical Functional

[1]http://users.ox.ac.uk/~cpgl0015/pargram/

*A LaTeX Package for CSLI Collections.*
Edie Tor and Ed Itor (eds.).

Grammar (Butt et al., 2002). As with the other grammar development projects within the PARGRAM initiative, our grammar is implemented within the XLE grammar development environment[2] (Crouch et al., 2008), which implements LFG theory. The project was a partnership between the University of Oxford and the University of Essex, and ran from 2004 to 2007.

The objective of the PARGRAM initiative is to develop parallel grammars for a range of different languages[3] using a shared linguistic framework and shared grammar writing techniques and technology. All members of the PARGRAM initiative strive for accurate, linguistically well-motivated grammatical analyses according to a common set of linguistic principles and assumptions: the grammars are "parallel" in the sense that differences among grammars reflect empirically demonstrable differences between languages, not theoretical disagreements or arbitrary decisions on the part of individual grammar writers. Within the space of these common goals, the projects within the PARGRAM umbrella have a range of practical motivations. For example, the English, German, and Japanese grammars aim for very broad and robust coverage for industrial applications. In contrast, although all of the PARGRAM grammars have the capacity to feed into large practical applications, the primary focus of attention in the Urdu and Hungarian projects is on theoretical linguistic issues: whether a coherent large-scale grammar and lexicon of these languages can be written in conformance with the linguistic assumptions common to all the PARGRAM grammars. The Welsh and Malagasy grammars fall at this end of the PARGRAM spectrum; the focus of the project was to produce coherent, internally consistent, linguistically well-motivated large-scale grammars and lexicons of these languages, following the common PARGRAM assumptions and using the common tools. As our grammar development work for Welsh and Malagasy progressed, we found that our analyses of these languages as exemplars of the verb-initial type share a good deal of commonality at the phrase structure level of analysis, creating theoretical synergy within the Verb-Initial Grammars subproject. However, the languages also differ in interesting ways: most importantly, Welsh is a VSO language, whereas Malagasy is VOS. Exploring differences between these languages has enhanced our understanding of the range of variation possible within the verb-initial type. A web-based interface to a number of PARGRAM grammars, including the Malagasy grammar, is available from the Norwegian INESS project webpages.[4]

[2] `http://www2.parc.com/isl/groups/nltt/xle/`

[3] For details see the main PARGRAM webpages, `http://pargram.b.uib.no/`.

[4] `http://clarino.uib.no/iness/xle-web`

Like all grammars within the PARGRAM project, our Malagasy grammar relies heavily on a computational component for morphological analysis. The need for automatic morphological analysis for Malagasy is acute, since entering into the lexicon each of the many hundreds of surface forms associated with a single verb, noun, or adjective root would miss important linguistic generalizations and impede progress in grammar development. For most of the other PARGRAM grammar development efforts, the task of building a morphological analyzer does not arise, since large-scale morphological analyzers already exist for many of the PARGRAM languages. For those grammars, the task is instead to incorporate these already existing morphological analyzers, which had often been originally created for shallow grammatical analysis or information retrieval applications, and which therefore tend not to contain detailed grammatical information and, indeed, to contain unnoticed mistakes, gaps, and inconsistencies. As discussed by Kaplan and Newman (1997), the challenge for PARGRAM projects incorporating a pre-existing morphological analyser is to overcome the problems arising from the lack of detailed grammatical information that these analyzers make available; Kaplan and Newman discuss ways of alleviating these problems by combining pre-existing analyzers with additional transducers to correct and augment their results.

The Malagasy grammar shares with a few of the other Pargram grammars (Arabic, Turkish, Urdu, Welsh) the opportunities and difficulties that arise when a morphological analyzer is developed in tandem with a syntactic grammar and lexicon. The advantages are that the morphological analyzer can be tuned to provide exactly the syntactic information that the grammar writer expects, and the division of labor between the morphology and syntax can be made in a well-motivated manner, rather than being imposed on the grammar writer by a pre-existing morphology component. The disadvantages are that the grammar development effort tends to be delayed if any problems arise in developing the morphological analyzer, and any changes to the architecture of the morphological analyzer can necessitate overhauling the syntactic lexicon and grammar to restore compatibility between the two. In related work, Çetinoğlu and Oflazer (2006) explore some issues in developing a morphological analyzer for Turkish, an agglutinative, morphologically complex language, in the context of the PARGRAM project.

For our Malagasy morphological analyzer, we use the Xerox finite-state tools LEXC and XFST (Beesley and Karttunen, 2003), which are employed by many of the PARGRAM grammars. As with any finite-state morphological transducer, our Malagasy morphological analyzer is bidirectional: it can be used in grammatical analysis to produce morphologically analyzed input to a parser, or in generation to produce a surface form from a specification of lexical properties (Beesley and Karttunen, 2003). In the following, we often describe the tokenizer and morphological component in terms of analysis as opposed to generation of a surface string, but this is only for expository purposes.

### 1.3 Malagasy Morphology: An Overview

As Keenan and Polinsky (2001) note, there is very little inflectional morphology in Malagasy: there is no verb agreement or nominal inflection for agreement features, for example. Keenan and Polinsky (2001) analyze certain alternations in deictic forms and demonstratives as inflection, but since the forms involved belong to a small closed class, we treat them by listing them in the lexicon. The morphological analyzer described here handles many of the productive cases of nominal and verbal derivational morphology, consisting primarily of affixal verbal morphology and genitive compounding.

Besides verbal affixation and genitive compounding, Malagasy also exhibits root reduplication (Keenan and Razafimamonjy, 1998, Keenan and Polinsky, 2001), usually giving a diminished, attenuated, or pejorative meaning. In reduplication, the stressed syllable of the root and the material to its right are copied, and there is often adjustment to the final syllable of the full root and/or the initial syllable of the reduplicant along the lines described for genitive compounding on page 7, including dropping of the weak endings *ka/tra/na*. Keenan and Polinsky (2001) provide the following examples:

(1)

| Root | Reduplicated root |
|---|---|
| bé 'big, numerous' | bèbé 'fairly big, numerous' |
| fótsy 'white' | fòtsifótsy 'whitish' |
| hadíno 'forget' | hadìnodíno 'forget a bit' |
| éntana 'baggage' | èntanéntana |
| fántatra 'known' | fàntapántatra |
| résaka 'conversation' | rèsadrésaka |

Reduplicated roots participate in the same morphological processes as unreduplicated roots except for prefixation with the passive prefix *tafa-*, which is possible only with non-reduplicated roots.

It is well known that reduplication requires special treatment in a finite-state morphological model, and the COMPILE-REPLACE algorithm described by Beesley and Karttunen (2000, 2003) provides a means of treating these cases. We have implemented a preliminary treatment of Malagasy reduplication using the COMPILE-REPLACE algorithm, but in the following we leave aside discussion of reduplication, concentrating instead on our treatment of nominal and verbal morphology using the standard LEXC and XFST tools.

## 1.4 Lexical Information

Malagasy roots consist of one or more syllables. Roots with more or fewer than three syllables are regular or 'strong', and have penultimate stress if they are multisyllabic. Three-syllable roots take penultimate stress unless they end in one of the 'weak syllables' *na/ny, ka, tra*, in which case they usually receive antepenultimate stress and are called 'weak roots' (Keenan and Polinsky, 2001). Weak and strong roots behave differently in the processes we treat, and are listed separately in our morphological lexicon.

Our lexicon contains 2,446 roots, including 2,033 roots which form nouns, adjectives or verbs, 288 roots which form adjectives or verbs, and 125 roots which form only verbs. Indeclinable forms, including proper names, adverbs, some prepositions, and free pronouns, are not listed in the morphological lexicon, and so are passed through the morphological analyzer unchanged and treated by the syntax as unanalyzed tokens. We also allow for guessed roots, defined in terms of permissible root patterns; these roots are marked with the tag `+Guess`, and are permitted, though dispreferred, in syntactic analysis. In our treatment of guessed forms, we define Syllable (`Syll`) as in (2); this allows the definition of weak guessed roots as consisting of two syllables followed by one of the weak endings *na, ka, tra*. Strong guessed roots are then defined as consisting of one to four syllables, and subtracting the weak root patterns.

(2) Guessed roots:

```
Syll = [((Nasal) ([t|d]) Consonant) (Vowel) Vowel];
WeakKTRoot = [Syll^2 [[[T|t] [R|r]]|[K|k]] [A|a]]];
WeakNRoot = [Syll^2 [[N|n] [A|a]]];
StrongRoot = [Syll^{1,4} – [WeakKTRoot|WeakNRoot]];
```

The definitions in (2) follow standard XFST notation, as defined by Beesley and Karttunen (2003): square brackets '[' ']' indicate grouping, parentheses '(' ')' indicate optionality, '|' indicates union or disjunction, '–' indicates subtraction of the second set of strings from the first set of strings, and '^' followed by a number or range of numbers indicates the number of times the immediately preceding string is repeated. Guessed roots participate in the same morphological processes as known roots, and interact with other components of the morphological analyzer in the same way: the LEXC and XFST components do not distinguish between known and guessed roots.

## 1.5 Genitive Compounding

Our analysis of verbal and nominal morphology closely follows the exposition of Keenan and Polinsky (2001). Nominal morphology consists mainly in the formation of genitive compounds, and the rules governing genitive compounding are quite regular and consistent. Genitive compounds are of the form Head+$\mathrm{NP}_{gen}$, where the Head can be any of the following: noun (in which case $\mathrm{NP}_{gen}$ expresses the possessor), passive verb ($\mathrm{NP}_{gen}$ is the agent), preposition ($\mathrm{NP}_{gen}$ is the prepositional object) or adjective ($\mathrm{NP}_{gen}$ is an agent or indirect cause). In such expressions, the Head and $\mathrm{NP}_{gen}$ are concatenated, and the concatenation is regulated by rules referring to properties of the final syllable in the Head and the first syllable in $\mathrm{NP}_{gen}$.

### 1.5.1 Target Phenomena

The following informal rules follow the treatment of Keenan and Polinsky (2001), and represent alternations in the final and first syllables of Head and $\mathrm{NP}_{gen}$ respectively. The hyphen, which sometimes alternates with apostrophe, is part of Malagasy orthography. The expressions 'C', '$\mathrm{C}'$' stand for consonants and 'V' and '$\mathrm{V}'$' stand for vowels, while 'S' denotes a stop consonant. The lowercase characters denote the corresponding letters. The expression to the left of the '+' sign stands for the final portion of the head word, consisting of some strong consonant 'C' and some vowel 'V' in the case of (1a) and (2) and some vowel 'V' followed by one of the weak endings *ka*, *tra*, *na* in (1b). The expression to the right of the '+' sign stands for the initial character of the $\mathrm{NP}_{gen}$.

1. Head is weak, that is, ends in one of *ka*, *tra*, *na*
   (a) $\mathrm{NP}_{gen}$ begins with a vowel V:
   CV + V → C-V (remove final vowel in Head and concatenate)
   (b) $\mathrm{NP}_{gen}$ begins with a consonant C with corresponding stop consonant S:
      i. Head ends in *na*:
      Vna + C → Vn-S (S not bilabial), or
      Vna + C → Vm-S (S bilabial)
      ii. Head ends in *ka* or *tra*:
      V[ka|tra] + C → V-S
2. Head is not weak:
   (a) $\mathrm{NP}_{gen}$ begins with a vowel V′:
   CV + V′ → CVn-V′ (prefix *n-* and concatenate)
   (b) $\mathrm{NP}_{gen}$ begins with a consonant C′ with stop consonant S:
   CV + C′ → CVn-S (S not bilabial), or
   CV + C′ → CVm-S (S bilabial)

Keenan and Polinsky (2001) provide the following examples of genitive compounding with full nominal genitives:

(3)

| Head | $\mathrm{NP}_{gen}$ | Compound | rule |
|---|---|---|---|
| tóngotra<br>'foot' | akóho<br>'chicken' | tòngotr'akóho<br>'chicken's foot' | (1a) |
| órona<br>'nose' | sáka<br>'cat' | òron-tsáka<br>'a cat's nose' | (1b.i) |
| táolana<br>'bones' | bíby<br>'animal' | tàolam-bíby<br>'bones of animals' | (1b.i) |
| sóroka<br>'shoulder' | záza<br>'child' | sòro-jáza<br>'a child's shoulder' | (1b.ii) |
| tóngotra<br>'foot' | fàrafára<br>'bed' | tòngo-pàrafára<br>'foot of a bed' | (1b.ii) |
| tráno<br>'house' | andríana<br>'noble' | trànon'andríana<br>'a noble's house' | (2a) |
| tráno<br>'house' | soavály<br>'horse' | trànon-tsoavály<br>'house for horses' | (2b) |
| tráno<br>'house' | bíby<br>'animal' | trànom-bíby<br>'house for animals' | (2b) |

Similar to noun genitive expressions are pronominal suffixed genitives. If the Head ends in a non-weak syllable or *na*, then the GEN1 suffixes are attached to the Head. Otherwise, the GEN2 suffixes are attached.

(4)

| | GEN1 suffix | GEN2 suffix |
|---|---|---|
| 1sg | ko | o |
| 2sg | nao | ao |
| 3sg or pl | ny | ny |
| 1pl.incl | ntsika | tsika |
| 1pl.excl | nay | ay |
| 2pl | nareo | areo |

Keenan and Polinsky (2001) provide the following examples of pronominal genitive compounding:[5]

(5)

| | GEN1<br>tráno 'house' | GEN2<br>sóroka 'shoulder' | GEN2<br>tóngotra 'foot' |
|---|---|---|---|
| 1sg | tránoko | sóroko | tóngotro |
| 2sg | tranonáo | sorokáo | tongotráo |
| 3sg or pl | tránony | sórony | tóngony |
| 1pl.incl | tranontsíka | sorotsíka | tongotsíka |
| 1pl.excl | tranonáy | sorokáy | tongotráy |
| 2pl | | sorokaréo | tongotraréo |

### 1.5.2 Implementation

The XLE grammar development environment requires a tokenizer and morphological analyzer for the language being analyzed, and allows the specification of a sequence of alternative morphological analyzers to be used when analysis with the first alternative fails. Our implementation relies on the Xerox finite-state calculus, with a tokenizer written in xfst, a lexicon written in lexc, and more general orthographic and phonological rules written in xfst (Beesley and Karttunen, 2003). The output of the morphological analyzer is a lemma (generally the root form) and a set of morphological tags, which is the input to syntactic analysis; this obviates the need for listing each surface form separately in the syntactic lexicon. Instead, the syntactic component contains information about the syntactic content and behavior of each lemma and morphological tag, as analyzed by the morphological component. We discuss the interface between XLE and the morphological component and display some example parses in Section 1.7.

#### Tokenization

XLE expects a string as input, which is first tokenized according to the rules of the tokenizer for the language being analyzed. Each token is then individually passed to the morphological analyzer for finite-

[5] Keenan and Polinsky (2001) do not list the second-person plural possessor form for *tráno* 'house'.

state morphological analysis. Most grammars within the PARGRAM initiative employ at least two morphological analyzers: an analyzer for known forms, and a guesser for forms that fail to be analyzed by the known-form analyzer. Following this paradigm, our known-form transducer and guesser are extracted separately, and applied to the output of the tokenizer in sequence; only forms that fail to obtain an analysis with the known-form analyzer are passed to the guesser.

In most cases, tokenization in Malagasy is straightforward, with tokens usually delimited by whitespace. For the sentence *Hanketo izy.* 'he will come here', the tokenizer produces the following result, where TB indicates a token boundary:

(6) hanketo TB izy TB .

The tokenizer (optionally) decapitalizes the first word of the sentence, inserts token boundaries at spaces, and separates punctuation by a token boundary. Each token is then passed separately to the morphological analyzer for analysis.

In the case of non-pronominal genitive compounding, the situation is more complex. The compound form *akanjon-olona* 'a person's clothes' consists of two noun roots *akanjo* 'clothes' and *olona* 'person', and has the following structure:

(7) akanjon-olona
akanjo (epenthetic 'n'; cf. 2a above) (compound boundary) olona

In our original treatment of non-pronominal genitive compounding, this form was treated as a single token, and was handled by the morphological analyzer (Dalrymple et al., 2005). However, this approach interacted badly with the standard configuration of XLE grammars, where the known-form analyzer and guesser are separate transducers, applied in sequence. If both roots are known, both can be analyzed by the known-form transducer; however, if one root is unknown, the entire compound fails to be recognized by the known-form analyzer. This means that the entire compound must be handled by the guesser, even if one of the roots is known. This undesirable result has led us to revise the treatment of non-pronominal genitive compounds, moving most of the work to the tokenizer.

In our current treatment, the tokenizer 'undoes' the effects of the compounding rules given above, proposing one or more underlying forms for analysis by the morphological component. For example, the compound form *akanjon-olona* is tokenized as follows:

(8) `akanjo TB +GEN+ TB olona`

The root *akanjo* is a three-syllable root; the hypothetical root *akan-*

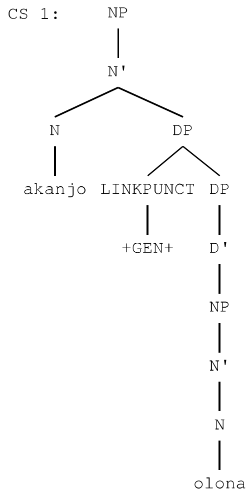


FIGURE 1 Phrase structure tree for *akanjon-olona*.

*jona* would have four syllables, which is impossible for a weak root. For this reason, rule 1 above, which requires that Head is a weak root, does not apply. Rule 2a covers the case in which the second member of the compound begins with a vowel; "undoing" rule 2a entails removing the epenthetic *n* inserted after the Head, producing *akanjo*. The compound boundary is treated as a separate token, represented by the special symbol `+GEN+`, to signal to the syntactic analysis component that genitive compounding has taken place. The phrase structure tree that is produced for *akanjon-olona* is shown in Figure 1, in which the leaves of the tree correspond to the tokens produced by the tokenizer.

Some forms can be tokenized in more than one way. For example, the compound *volon-dRabe* 'Rabe's month'/'Rabe's money' is ambiguous (Keenan and Polinsky, 2001):

1. Head is weak, reconstructed by the tokenizer as *volana* 'month', with Rule 1a requiring removal of the final vowel; or
2. Head is strong, reconstructed by the tokenizer as *vola* 'money', with Rule 2a requiring insertion of -*n*

After tokenization, the morphological analyzer is presented with all possible forms resulting from 'undoing' the compounding rules. Analysis proceeds as for the simple cases, with forms recognized by the known-form analyzer given preference in syntactic analysis over hypothetical forms analyzed by the guesser.

Our current treatment of genitive compounding relies on the presence of the hyphen or apostrophe to signal the compound boundary. In a small minority of cases, however, genitive compounding involves only concatenation of roots, and is not signaled by special punctuation. We have left the treatment of these forms for future work, since we are unsure how the treatment of such forms will interact with the guesser: almost any simple form can be given a spurious analysis as a compound composed of two hypothetical, guessed roots.

**Pronominal genitives**

Pronominal genitive compounds, on the other hand, are best treated by the morphological analysis component, which we now describe. The LEXC lexicon is a finite-state transducer which specifies a relation between an Upper 'lexical' string and a Lower 'surface' string for a form (Beesley and Karttunen, 2003). Roots and affixes are organized into sublexicons according to their phonological and prosodic properties, e.g. whether the root is weak or strong. The lexicon also specifies possibilities for transitions when a particular form is encountered.

For example, the noun root *akanjo* 'clothes' is listed in the Noun sublexicon with continuation class `Nstrong`, meaning that it takes the strong root suffixes listed in the `Nstrong` sublexicon. The `Nstrong` sublexicon adds the `+Noun` tag to the lexical side of the transducer, and permits the form to terminate with no suffixation, or alternatively allows genitive suffixation. Thus, the transducer relates the surface string, the unsuffixed noun *akanjo*, to the morphologically analyzed lexical string which forms the input to syntactic analysis:

(9) LEXC transducer:
Lexical: `akanjo +Noun`
Surface: `akanjo`
'clothes'

The related form *akanjoko* 'my clothes' involves pronominal genitive suffixation; the `NStrong` sublexicon relates the first person singular genitive suffix *ko* on the surface side to the tag `+1SgGen` on the lexical side:

(10) LEXC transducer:
Lexical: `akanjo +Noun +1SgGen`
Surface: `akanjoko`
'my clothes'

Here, the LEXC lexicon on its own is sufficient for analysis of the combination of the root *akanjo* and the pronominal genitive suffix *-ko*.

## 1.6 Verbal Morphology

In many cases, we also need a set of XFST rules to cater for phonological and orthographic alternations induced by morphological operations. The XFST rules define an XFST transducer, which is composed with the LEXC transducer in full morphological analysis (Karttunen et al., 1992). The XFST transducer is controlled by diacritics introduced by LEXC, orthographically distinguished by use of a carat '^'. These diacritics appear neither in the surface string nor in the lemma and morphological tags that are provided to the syntactic component, but are crucial in the definition of the XFST rules, as we now describe.

### 1.6.1 Target Phenomena

Malagasy exhibits rich and complex verbal morphology (Randriamasimanana, 1986, Keenan and Polinsky, 2001). Verbs are classified according to the case of their arguments: nominative, accusative and genitive. Verbs which take a genitive complement are non-active verbs, a category which includes passive verbs and circumstantial verbs. Passive verbs are formed in three different ways, each corresponding to different semantics. The following discussion follows Keenan and Polinsky (2001), though simplifying somewhat.

**Passive roots**

```
ROOT  +Verb    (GEN)
...   ROOTTYPE  ...
```

**Suffix passives**

```
TENSE (Caus) ROOT +Verb    (C)VnaPass (GEN|IMP)
...     amp   ...  ROOTTYPE (C)ina/ana  ...
```

**Prefix passives**

```
VTPass    ROOT +Verb    (GEN|IMP)
voa/tafa  ...  ROOTTYPE  ...
```

**Active Verbs**

```
(TENSE|NOM) [(Recip)(Caus)][ACTIVE PassROOT|NullPrefROOT] +Verb    (IMP)
 ...           if      amp    i/an   ...                    ROOTTYPE  ...
```

**Circumstantial form**

```
TENSE (Caus)(ACTIVE) ROOT +Verb    (C)VnaPass
...     amp   i/an    ...  ROOTTYPE (C)ina/ana
```

FIGURE 2 Verbal patterns in the LEXC transducer.

The schematic relation for verbal patterns encoded by the LEXC transducer is given in Figure 2. The first line in each category represents the lexical side of the LEXC transducer, and the second line represents the surface side, which is input to the XFST transducer. Ellipses (...) represent surface strings of varying form: for example, ROOT is realized as any root form, and GEN is realized as any of the pronominal genitive suffixes in (4). ROOTTYPE is a diacritic encoding the type of the root, one of `^StrongRoot`, `^WeakKTRoot` or `^WeakNRoot`; this information is needed by the XFST rules to control certain morphological alternations. The LEXC transducer is composed with the XFST transducer, which performs necessary adjustments as the morphemes are concatenated.

The first category defined in Figure 2 is root passives: that is, roots which are passive verbs. These refer more to the result than the process. The pattern for passive roots is very similar to patterns for noun roots with optional pronominal genitive suffixation:

(11)

| Root passive | With genitive suffix |
|---|---|
| araka 'be followed' | arany 'be followed by him/them' |

The largest category of passive verbs are suffix passives. These are formed by the suffixation of *ina* or *ana* to a root, which is usually preceded by a root-dependent consonant C epenthesis. They can appear with a past or future tense prefix, represented as TENSE, optionally followed by a causal prefix *amp*. This form can also undergo genitive compounding or imperative suffixation. Keenan and Polinsky (2001) provide the following examples of suffix passives; these examples do not contain a TENSE prefix or any additional suffixation:

(12)

| Root | Suffix passive | |
|---|---|---|
| áraka 'follow' | aráhina 'be followed' | *k*→*h, -ina* |
| rákotra 'cover' | rakófana 'be covered' | *tr*→*f, -ana* |
| tóhy 'continue' | tohízina 'be continued' | $\emptyset \rightarrow$ *z, -ina* |

A third type of passive is prefix passives. These are formed by prefixing a root with any of *a*, *voa*, *tafa*. Passives in *a* refer to the process rather than the result, and usually their subject functions as an instrument. The imperative is formed by prefixing with *a* and adding the corresponding passive imperative suffix. Passives in *voa/tafa* refer to the end result rather than the process and have a perfective meaning. *voa/tafa* passives may not be prefixed by a tense prefix, while *a* passives do take a tense prefix. Keenan and Polinsky (2001) provide the following examples of prefix passives:

(13) Prefix passive:

| Root | Prefix passive | |
|---|---|---|
| tolotra 'offer' | atolotra 'be offered' | *a* |
| arina 'put upright' | tafarina 'get up' | *tafa* |
| araka 'follow' | voaraka 'be followed' | *voa* |

There are a few active verb roots, but the majority of active verbs are derived from roots by means of the active prefixes *i, an*. Active verbs can be marked for tense via a tense prefix TENSE (distinguishing past, present, and future). Genitive suffixing is not allowed, but the formation of imperatives is possible: present tense (*m-*) actives take suffix *a*, where consonant mutation and epenthesis (*-(C)a*) apply. If no epenthetic consonant intervenes, they fuse the *-a* imperative with root final *-a*. They can also receive a prefix for causality and reciprocation. The active verb roots may also appear with causative *ank-/amp-*.

In the circumstantial form of a verb, an oblique argument or adjunct of an active verb is made the subject. The circumstantial is built from roots prefixed by primary active affixes *i, an* and, possibly, secondary causative affixes *ank-, amp-* by means of the suffixation *-Cana*, where *C* is the root-specific epenthetic consonant mentioned above in the context of suffix passives. Tense is marked in the same way as for suffix passives. Keenan and Polinsky (2001) provide the following examples of present active forms and circumstantials:

(14)

| Root | Present active | Circumstantial |
|---|---|---|
| omé | manomé 'gives' | anomézana |
| sóratra | manóratra 'writes' | anorátana |
| íla | míla 'needs' | ilána |
| ánatra | miánatra 'studies' | ianárana |
| tóhy | mitóhy 'continues' | itohízanax |

### 1.6.2 Implementation

As discussed above, the LEXC lexicons contain information about subclasses of individual roots as well as more general structural information regarding verb forms. For example, verbs are formed on the basis of a tense prefix sublexicon which contains separate past, present and future prefixes, including a `^TNS^` tag to control morphological alternations with overt tense prefixes. In the following, 0 represents the empty string.

```
LEXICON Tense
PresentTense+:0         Secondary;
PastTense+:no^TNS^      Secondary;
FutureTense+:ho^TNS^    Secondary;
PresentTense+:0         Vroot;
```

```
PastTense+:no^TNS^      Vroot;
FutureTense+:ho^TNS^    Vroot;
```

The lexicon VPassRoot represents the passive verbs: inherently passive roots, or guessed passive verbs ending in either a strong or weak syllable. In the compilation step, the transducer for guessed roots is extracted separately from the transducer for known roots, since XLE requires application of the known root transducer and the guessed root transducer in sequence. However, since known and guessed roots exhibit the same morphological patterns, they are defined together in the LEXC lexicons.

```
LEXICON VPassRoot
PassiveRoot ;
<StrongRoot %+Guess:0>  StrongSuff ; ! guessed Strong root
<WeakKTRoot %+Guess:0>  KTWeak ; ! guessed Weak KT root
<WeakNRoot %+Guess:0>   NWeak ; ! guessed Weak N root
```

Roots are listed in the lexicon with information about the continuation classes of their suffixes:

```
LEXICON PassiveRoot
araka         WeakSuff; ! be followed
fantatra      TR2RWeak; ! be known
```

In this example *araka* is a passive root; its continuation class indicates that it is a member of the class of morphologically weak roots. *fantatra* is a weak passive root with final syllable *tra*, where the `TR2RWeak` continuation class indicates that the *tra* suffix for this root is replaced with *r* during passive suffixation or the formation of imperatives. Thus, the passive form corresponding to *fanta<u>tra</u>* is *fanta<u>r</u>ina*.

As above, the XFST rules deal with surface phenomena such as syllable deletion and consonant and vowel epenthesis, which take place during affixation. In the previous example, the continuation class `TR2RWeak` is used with roots where the weak final root syllable *tra* is converted to *r* during passive suffixation or the formation of imperatives. Other weak roots convert *tra* to one of a number of other consonants which must be lexically specified for each root. One way of handling these alternations would be to have a continuation class for each of the possible combinations of suffixes and final syllables of roots. Thus, even though there are only two passive suffixes *ina/ana*, we would need separate continuation classes for the formation of passives for weak roots ending in *tra* where *tra* is transformed to *r*, *f*, *t*, or other consonants.

However, this would result in an over-sized, untidy lexicon. Instead, we keep a small number of continuation classes corresponding to possible suffixes, and signal the final syllable root transformations by means of tags referenced by rules of the XFST transducer. These tags provide the context for the application of XFST rules for the various cases of epenthesis, deletion and transformation. For instance, the `TR2RWeak` continuation class is defined in the following way:

```
LEXICON TR2RWeak
+Verb:^WeakKTRoot        WeakKTEnding ;
+Verb:^WeakKTRoot^Ftr2r  Suffixes ;
```

The feature `^Ftr2r` is referenced by the XFST rule in (15), which transforms *tra* to *r* if the *tra* syllable is followed by the feature `^Ftr2r`.

(15) `[t r a]` $\rightarrow$ `r` $\|$ _ `^Ftr2r`

This XFST rule applies to the underlying grammatical form, the output of LEXC. Formally, it resembles a standard context-sensitive phrase structure rule: the expression to the left of the arrow is replaced by the expression to the right of the arrow in a certain context. The context of application is separated from the rule by double bars, '‖'. Here, the sequence *'tra'* is replaced by *'r'* in the context immediately preceding the tag `^Ftr2r`. This rule applies after removal of the tag `^WeakKTRoot`, which separates the root from the `^Ftr2r` tag in the surface string of the LEXC transducer. Directly after the application of this rule, the rule to remove the tag `^Ftr2r` applies, preventing its appearance in the surface string and its interference with the application of other rules. Similar rules cater for alternations with prefixation, passive and imperative formation.

Features are an efficient way of modeling local morphological dependencies and alternations. However, in the morphology of Malagasy verbs there are long distance dependencies which cannot be modeled by standard FST techniques. For instance, there are roots which can form the passive in either *ana* or *ina* but not both. Thus, we want *fantarina* and not *fantarana* to be recognized as the correct passive form of *fantatra*. This is a problem both in recognition and generation as we do not want our rules to accept or generate incorrect forms. A tag could be added to the root *fantatra* to exclude the passive formation in *ana*, but the tag may be separated from the position of *ina/ana* by other morphemes and tags during passive formation.

For example, if we decided to implement the requirement for passive suffixation in *-ina* rather than *-ana* as a feature, the lexical entry for *fantatra* in the lexicon would be accompanied by a feature `^Fpassi` on the surface level as below. In the following hypothetical lexicon, the root and its associated tags are grouped together by angled brackets

'<' and '>' as is standard in LEXC:

```
LEXICON OtherRoot
<{fantatra} 0:^Fpassi> TR2RWeak;
```

However, this means that when the features ^WeakKTRoot^Ftr2r are added by the continuation class TR2RWeak, they are not immediately next to the root, but rather ^Fpassi stands in the way. As a result the rule (15) above for the transformation of the weak syllable *-tra* to *-r*, which precedes passive suffixation, cannot apply.

Fortunately, XFST allows for the treatment of such dependencies by the use of flag diacritics, non-FST handles which can store information that is not compiled into the FST. This information is used at runtime, when a certain phrase is being analyzed or generated. We use two sorts of flag diacritics: U-type (where 'U' stands for "unification"), and R-type (where 'R' stands for "requires"). Importantly, the use of flag diacritics does not change the formal power of the system (Karttunen, 2006). In our Malagasy morphological analyzer, we use flag diacritics to store root-specific information: therefore, they are entered together with the lexical entry for the root. Because they do not take effect until the interpretation phase, they do not interfere with the XFST rules. Thus, the lexical entry for *fantatra* in the previous section becomes:

```
LEXICON OtherRoot
<{fantatra} @U.PASS.I@> TR2RWeak;
```

This information uses a U-type flag diacritic, here represented as @U.PASS.I@, to associate the feature PASS with the value I for this root. This feature ensures that the root *fantatra* takes a passive in *ina* and not in *ana*. This is coupled with matching flag diacritics for the passive suffixes:

```
LEXICON PassaSuff
<+Passa:a 0:n 0:a @U.PASS.A@>   #;

LEXICON PassiSuff
<+Passi:i 0:n 0:a @U.PASS.I@>  #;
```

The passive suffix *ina* is associated with the flag diacritic @U.PASS.I@, which is defined as above as specifying the value I for the feature PASS, while the suffix *ana* is associated with the flag diacritic @U.PASS.A@ specifying the value A for the same feature. Whenever flag diacritics meet, they must match; therefore the form *fantarana* is not accepted, as the flag diacritics of *ana* do not match the flag diacritics of *fantatra*.

We also make use of R-type flag diacritics to model long distance dependencies between prefixes and suffixes. R-type flag diacritics are similar in structure to U-type diacritics, and are specified in the same way; crucially, however, R-type diacritics check that a certain value of a feature has been previously set by another flag diacritic specification. For example, Malagasy verbs may be formed from roots which are lexically nouns or adjectives. Since this is a very general fact about Malagasy, we do not list every root form in both the noun lexicon and the verb lexicon. Nevertheless, we must ensure that the prefixes that appear with a root are compatible with its suffixes, since if a root appears with verbal prefixes, it allows verbal and not nominal suffixes.

We handle this by setting the flag `POS` (part of speech) to `VERB` or `NOUN` at the beginning of the word, depending on what prefixes have been encountered. We then check that the suffixes that appear with a root are compatible with its prefixes. For example, the strong noun root *halatra* 'theft' is specified with the continuation class `NStrong`, which allows either the `+Noun` tag and noun suffixes with the continuation class `NStrongEnding`, or the `+Verb` tag and verbal suffixes with the continuation class `VStrongEnding`. `NStrongEnding` uses an R-type flag diacritic to check that the flag diacritic `POS` has been set to `NOUN`, preventing noun suffixes from appearing with verb prefixes; similarly, the `VStrongEnding` lexicon checks that the flag `POS` is set to `VERB`.

The continuation classes, together with the rules and flag diacritics, give a general model for the construction of different verb forms. In our analysis of Malagasy verbal morphology, there are many exceptions to be taken into account which render the task of modeling verb morphology non-trivial. For instance, a root may not accept a certain affix; this can be handled by more sophisticated flag diacritics which govern the permissible affixes that each root can accept. Our analyzer uses flags encoding 9 features with various values for the different affixes, which can be negatively specified by particular roots to disallow particular affixes or affix combinations.

As noted by Beesley and Karttunen (2003), more general cases can be ruled out by means of filters, which are sets of rules that apply on the lexical level – that is, on the lexical side of the LEXC transducer. Such filters are used to prevent groups of continuation classes from combining with a certain affix, or to merge together morphological information. For instance, the lexical tag `+Passa` indicates that we have a passive form in *ana*, which can signal either a suffix passive or a circumstantial form. However, if it is preceded by the tag `ActiveAN+`, it is unambiguously a circumstantial form. Our treatment incorporates 5 filters disallowing certain affix combinations for all roots. Encoding such interactions in the morphological analyzer provides important constraints for syntactic analysis.

## 1.7 Grammatical outputs

In syntactic analysis, the XLE grammar development environment makes crucial use of both the lexical and surface sides of the morphological analyzer. The surface form is treated as a single leaf node of the tree in the display of the phrasal analysis. XLE also requires the specification of a set of sublexical grammar rules for analysis of the lemma and tags, each of which is associated with a functional description specifying its grammatical contribution. Example (10), repeated here, shows the lexical and surface side of the transducer for the form *akanjo* 'my clothes', composed of the noun root *akanjo* with a pronominal genitive suffix.

(16) Lexical: `akanjo +Noun +1SgGen`
Surface: `akanjoko`
'my clothes'

The standard LFG phrase structure tree for *akanjoko* is shown on the left side of Figure 3: the leaf node in this tree is the surface form *akanjoko*. The lexical side of the transducer also plays a crucial analytic role, since the lemma and morphological tags are analyzed by sublexical phrase structure rules, and information associated with them is assembled to form the syntactic and semantic constraints contributed by *akanjoko*. XLE provides a special mode for display of this sublexical structure: the right side of Figure 3 shows the sublexical structure and sublexical rules for *akanjoko*, with the lemma and tags as leaves of the sublexical tree.

Phrasal analysis Display of morphological structure

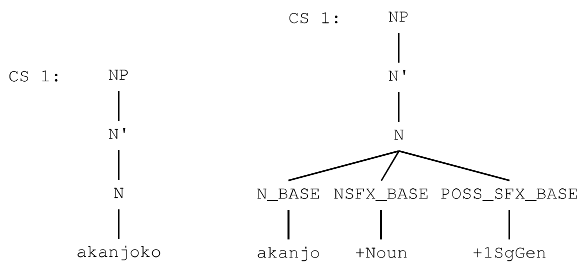


FIGURE 3 Analysis of *akanjoko*.

The phrase structure analysis of the full sentence in (17) is shown in Figure 4, and Figure 5 displays the morphological structure (lemmas and tags) for the verb *nindrana* 'borrowed' and *vola* 'money'.

| | | | | |
|---|---|---|---|---|
| (17) | nindrana | vola | i | Paoly |
| | n-indrana | vola | i | Paoly |
| | PAST-borrow | money | DET | Paul |
| | 'Paul borrowed (some) money.' | | | |

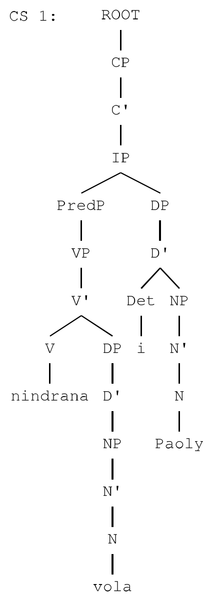


FIGURE 4 Analysis of *nindrana vola i Paoly* 'Paul borrowed some money'.

## 1.8 Conclusion

We have described a computational implementation of the productive nominal and verbal morphology of Malagasy, concentrating on our treatment of genitive compounding and affixal verb morphology. Our approach closely follows the analysis of Keenan and Polinsky (2001) and realizes the aforementioned morphological processes in terms of LEXC continuation classes, associated with groups of productive roots and general structural information, and general orthographic and phonetic rules implemented as XFST rules.

Our morphological analyzer provides a solid basis for our syntactic lexicon and grammar of Malagasy. The syntactic lexicon has been populated with the lemmas and tags generated by the morphological analyzer, and accepts these forms as input to syntactic analysis. We

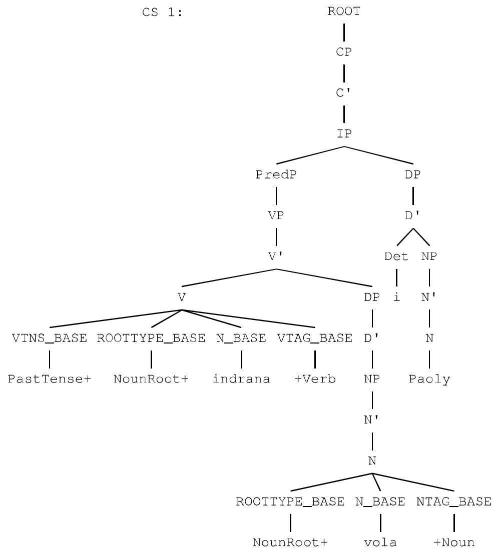


FIGURE 5 Analysis of *nindrana vola i Paoly* showing morphological analysis of *nindrana* and *vola*.

have encoded the syntactic contributions of each affix as well as default syntactic contributions for large classes of verb, noun, and adjective roots, together with refining the lexicon to account for subclasses of roots with exceptional, non-default behavior. Our Malagasy grammar comprises 22 preterminal categories appearing on the left-hand side of phrase-structure rules with regular-expression right-hand sides covering a number of possible expansions.

## 1.9 Acknowledgments

Lauri Karttunen's seminal and influential work on finite-state morphology is the basis upon which our proposals are built. We are grateful to Lauri and to the late Charles Randriamasimanana for help in the development of the system. The Verb-Initial Grammars subproject of PARGRAM initiative was supported by a grant from the Economic and Social Research Council, UK (Project RES-000-23-0505).

This paper is a revised and updated version of Dalrymple et al. (2006). We thank the editors of the *International Journal of Computational Linguistics and Chinese Language Processing* for permission to publish this paper, and we are very grateful to Ron Kaplan and an anonymous reviewer for helpful comments on this version.

## References

Beesley, Kenneth R. and Lauri Karttunen. 2000. Finite-state non-concatenative morphotactics. In *Proceedings of the Fifth Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON-2000)*, pages 1–12. Association for Computational Linguistics.

Beesley, Kenneth R. and Lauri Karttunen. 2003. *Finite-State Morphology*. Stanford: CSLI Publications.

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar Project. In *Proceedings of the International Conference on Computational Linguistics (COLING2002) Workshop on Grammar Engineering and Evaluation*, pages 1–7. International Committee on Computational Linguistics, Taipei.

Çetinoğlu, Özlem and Kemal Oflazer. 2006. Morphology-syntax interface for Turkish LFG. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*. Association for Computational Linguistics, Sydney.

Crouch, Dick, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell, III, and Paula S. Newman. 2008. *XLE Documentation*. Palo Alto Research Center, Palo Alto, CA.

Dalrymple, Mary, Maria Liakata, and Lisa Mackie. 2005. A two-level morphology of Malagasy. In *19th Pacific Asia Conference on Language, Information and Computation*. PACLIC Steering Committee, Taipei.

Dalrymple, Mary, Maria Liakata, and Lisa Mackie. 2006. Tokenization and morphological analysis for Malagasy. *International Journal of Computational Linguistics and Chinese Language Processing* 11(4):315–332.

Kaplan, Ronald M. and Paula Newman. 1997. Lexical resource reconciliation in the Xerox Linguistic Environment. In D. Estival, A. Lavelli, K. Netter, and F. Pianesi, eds., *Computational Environments For Grammar Development and Linguistic Engineering*, pages 54–61. Proceedings of a workshop sponsored by the Association for Computational Linguistics, Madrid, Spain, July 1997.

Karttunen, Lauri. 2006. Numbers and Finnish numerals. *SKY Journal of Linguistics* 19:407–421.

Karttunen, Lauri, Ronald M. Kaplan, and Annie Zaenen. 1992. Two-level morphology with composition. In *Proceedings of the International Conference on Computational Linguistics (COLING92)*. International Committee on Computational Linguistics, Nantes.

Keenan, Edward L. and Maria Polinsky. 2001. Malagasy. In A. Spencer and A. M. Zwicky, eds., *The Handbook of Morphology*, pages 563–623. Oxford: Blackwell.

Keenan, Edward L. and Jean Paulin Razafimamonjy. 1998. Reduplication in Malagasy. In *The Structure of Malagasy III*, UCLA Working Papers in Syntax and Semantics. Department of Linguistics, UCLA, Los Angeles.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig. 2015. *Ethnologue: Languages of the World*. SIL International. 18th Edition.

Randriamasimanana, Charles. 1986. *The Causatives of Malagasy*. Honolulu: University of Hawai'i Press.