A Two-level Morphology of Malagasy

Mary Dalrymple, Maria Liakata, and Lisa Mackie
Centre for Linguistics and Philology
University of Oxford
Walton Street, Oxford OX1 2HG UK
mary.dalrymple@ling-phil.ox.ac.uk
mal@aber.ac.uk
lisa.mackie@ling-phil.ox.ac.uk

Abstract

We present a two-level model of Malagasy nominal and verbal morphology (Beesley and Karttunen, 2003), based primarily on the discussion of Malagasy morphology in Keenan and Polinsky (1998) and Randriamasimanana (1986). Words in Malagasy are built from roots by means of a variety of morphological operations such as affixation and reduplication. The present paper analyzes productive patterns of nominal and verbal morphology, describing genitive compounding and suffixation for nouns, and various derivational processes involving compounding and affixation for verbs.

1 Overview of Malagasy Morphology

Malagasy is an Austronesian language spoken by about six million people on the island of Madagascar. With Welsh, it is a focus of the Verb-Initial Grammars subproject (users.ox.ac.uk/~cpgl0015/pargram/) within the PARGRAM initiative, a collaborative project to develop computational lexicons and grammars within the shared linguistic framework of Lexical Functional Grammar (Butt et al., 2002). Because of the complicated and productive patterns of Malagasy verbal and nominal morphology, the development of such a grammar relies heavily on a computational component for morphological analysis. 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).

The research reported here is supported by a grant from the Economic and Social Research Council, UK (Project RES-000-23-0505).

As Keenan and Polinsky (1998) 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 (1998) analyze certain alternations in deictic forms and demonstratives as inflection, but since the forms involved form a small closed class, we treat these forms by listing them in the lexicon. The morphological analyzer described here handles many of the productive cases of nominal and verbal derivational morphology.

In the following, we describe our treatment of affixal verbal morphology and genitive compounding. These phenomena are treated within the framework of two-level morphology and implemented using the Xerox finite-state tools LEXC and XFST (Beesley and Karttunen, 2003). Malagasy also has reduplicative morphological processes, in which a new root is formed by reduplicating part or all of a basic root. It is well-known that reduplication requires special treatment in a finite-state morphological model. Although the COMPILE-REPLACE algorithm described by Beesley and Karttunen (2000; 2003) provides a means of treating these cases, we have not yet addressed the problem of reduplication in our morphological analyzer.

Malagasy roots may have one or more syllables; most roots 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, 1998). In our analysis roots are assumed to be strong unless they are either guessed trisyllabic roots ending in a weak syllable or listed in the lexicon as weak.

Our Malagasy morphological lexicon contains a large number of root forms, since, as we will see, the properties of many roots are idiosyncratic and must be individually specified. However, 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. We define Syllable (Syll) as in (1); 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:

(1)
```
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]];
```

2 Genitive compounding

Our analysis of verbal and nominal morphology closely follows the exposition of Keenan and Polinsky (1998). Nominal morphology consists mainly in the formation of genitive compounds.[1] These are of the form Head+NP_{gen}, where the Head can be any of the following: noun (in which case NP_{gen} expresses the possessor), passive verb (NP_{gen} is the agent), preposition (the NP_{gen} is the prepositional object) or adjective (the NP_{gen} is an agent or indirect cause). In such expressions, the head and the 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 NP_{gen}.

2.1 Compounding rules

The following rules handle alternations in the final and first syllables of the Head and NP_{gen} respectively. The hyphen and apostrophe are part of Malagasy orthography.

1. Head is weak, that is, ends in one of *ka*, *tra*, *na*

   (a) NP_{gen} begins with a vowel Vo:
   CV + Vo → C'Vo (remove final vowel in Head and concatenate)

   (b) 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) NP_{gen} begins with a vowel Vo:
   CV + Vo → CVn'Vo (prefix *n'* and concatenate)

[1] Other nominal morphological processes involve cases of insertion of verbs, adjectives and nouns into nouns (the latter for the formation of compound nominals). These cases are not dealt with in the present work.

(b) NP_{gen} begins with a consonant Co with corresponding stop consonant S:
CV + Co → CVn-S (S not bilabial), or
CV + Co → CVm-S (S bilabial)

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.

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

2.2 Implementation

The rules governing genitive expressions are quite regular and consistent. The morphology of such expressions is modelled by the Xerox finite-state calculus, with a lexicon written in LEXC and more general orthographic and phonological rules written in XFST (Beesley and Karttunen, 2003). 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. 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, indicating that it takes strong root suffixes listed in the Nstrong sublexicon. The Nstrong sublexicon adds the +Noun tag to the lexical/Upper side of the transducer, and specifies the StrongSuff continuation class. The StrongSuff lexicon permits the form to terminate with no suffixation, or alternatively allows genitive suffixation. For example, the transducer relates the Lower string, the unsuffixed noun *akanjo*, to the morphologically analyzed lexical/Upper string which forms the input to syntactic analysis:

(2) LEXC transducer:
Upper: akanjo +Noun
Lower: akanjo
'clothes'

The related form *akanjoko* 'my clothes' is analyzed with the StrongSuff continuation class allowing pronominal genitive suffixation, relating the suffix *ko* on the surface/Lower side to the tag +1SgGen on the lexical/Upper side:

(3) LEXC transducer:
Upper: akanjo +Noun +1SgGen
Lower: akanjoko
'my clothes'

However, the LEXC lexicon on its own is not sufficient for modelling all entries and their combinations. As illustrated above, we also need a set of XFST rules to cater for phonological and orthographic alternations induced by morphological operations. These rules apply irrespective of the individual entries to be combined, and are controlled by tags introduced by LEXC to control the alternations. These tags, which were not shown in (2) or (3), are orthographically distinguished from the lexical tags of the LEXC transducer by the use of a carat '^'. The XFST rules define an XFST transducer, which is composed with the LEXC transducer in full morphological analysis. We illustrate with the genitive compound *akanjon-olona* 'a person's clothes':

(4) Lexical: akanjo +Noun +GEN+ olona +Noun
Surface: akanjon-olona
'a person's clothes'

The XFST transducer is defined by a series of rules, the first of which recognises the hyphen '-' in the surface form *akanjon-olona* as a signal of genitive compounding and relates the hyphen in the surface string to the tag +GEN+. The next rule adds *n* to a strong root when it is followed by +GEN+ and a nonbilabial consonant:

(5) [..] → n || _ ^StrongRoot +GEN+ \[Bilabial]

XFST also defines cleanup rules to remove tags such as ^StrongRoot. Thus, the XFST transducer relates the following two strings:

(6) XFST Upper: akanjo ^StrongRoot +GEN+ olona
Surface: akanjon-olona
'a person's clothes'

When the XFST transducer is composed with the LEXC transducer, the result is:

(7) Lexical: akanjo +Noun +GEN+ olona +Noun
LEXC Lower = XFST Upper: akanjo ^StrongRoot +GEN+ olona
Surface: akanjon-olona
'a person's clothes'

Our current treatment of genitive compounding depends on the presence of the hyphen or apostrophe to signal the compound boundary; however, in a minority of cases 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 that we have introduced for forms that do not appear in the LEXC lexicon.

3 Verbal morphology

Malagasy exhibits rich and complex verbal morphology. 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 (1998), though simplifying somewhat.

First, there are a small number of root passives, that is, roots which are passive verbs. These refer more to the result than the process. The LEXC transducer encodes the schematic relation for passive roots in Figure 1, which is very similar to patterns for noun roots with optional genitive compounding. ROOT represents the form of the passive root. ROOTTYPE is one of ^StrongRoot, ^WeakKTRoot or ^WeakNRoot; this information is needed by the XFST rules to control certain morphological alternations. (GEN) represents optional genitive compounding with the agent argument of the passive verb. An example for the root passive *haino* 'be listened to' is:

(8) Lexical: haino +Verb +1SgGen
Surface: hainoko
'be listened to by me'

Passive roots

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

Suffix passives

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

Prefix passives

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

Circumstantial form

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

Active Verbs

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

Figure 1: Verbal patterns

As above, the LEXC transducer is composed with the XFST transducer, which performs necessary adjustments as the morphemes are concatenated.

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 epenthesis C.[2] They can be prefixed by a tense prefix TENSE, denoting past or future, optionally followed by a causal prefix *amp*. This form can also undergo genitive compounding or imperative suffixation.

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. 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

[2]The root may have a passive meaning, but this is not necessarily the case.

meaning. *voa/tafa* passives may not be prefixed by a tense prefix, while *a* passive does take a tense prefix.

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 secondary active affixes *ank(a), 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.

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*. 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 *a* imperative with root final *a*. Active verbs can be marked for tense via a tense prefix TENSE (distinguishing past, present, and future). They can also receive a prefix for causality and reciprocation. The active verb roots may be null prefix or they can be prefixed by the active prefixes *ank-/amp*.

3.1 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, suffixed passives 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:

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

The lexicon VPassRoot represents the passive verbs: passive roots, verbs derived from other listed roots, or guessed passive verbs ending in either a strong or weak syllable:

```
LEXICON VPassRoot
                        PassiveRoot;
                        OtherRoot;
<StrongRoot +Guess:0>   VPassStrong;
<WeakKTRoot +Guess:0>   VPassKTWeak;
<WeakNRoot  +Guess:0>   VPassNWeak;
```

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

```
LEXICON PassiveRoot
haino         StrongSuff; ! be heard

LEXICON OtherRoot
fantatra      TR2RWeak; !be known
```

In this example *haino* is a passive root; its continuation class indicates that it is a member of the class of strong roots. In contrast, *fantatra* is a weak 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^Ftr2r  PassSuff;
+Verb:^WeakKTRoot^Ftr2r  PassImpSuff;
+Verb:^WeakKTRoot^Ftr2r  ActImpSuff;
```

The feature ^Ftr2r is referred to by the XFST rule in (9), which transforms *tra* to *r* if the *tra* syllable is followed by the feature ^Ftr2r.

(9) [t r a] → r || _ ^Ftr2r

This rule applies after removal of the tag ^WeakKTRoot, which separates the root from the ^Ftr2r tag in the Lower 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 modelling local morphological dependencies and alternations. However, in the morphology of Malagasy verbs there are long distance dependencies which cannot be modelled 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 recognised 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 lexical preference 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 bellow.

```
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 (9) 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 in the interpretation phase, when a certain phrase is being analysed or generated. We use flag diacritics to store root-specific information, and therefore they are entered together with the lexical entry for the root. However, 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 associates the feature PASS with the value I for this root, and ensures that the root *fantatra* takes a passive in *ina* and not in *ana*. This is coupled with 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 defined as specifying the value I for the feature PASS, while the suffix *ana* specifies 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 P-type and R-type flag diacritics to model long distance dependencies between verbal and nominal affixes. In Malagasy, verbs may be formed from roots which are lexically nouns or adjectives. For example, the noun root *halatra* 'theft' is specified with the continuation class NTR2RWeak. At this point, nominal or verbal suffixes may be added. To avoid incorrect combinations, we require a previous flag NOUN or VERB to have been set to positive.

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 modelling verb morphology non-trivial. For instance, a root may not accept a certain prefix or suffix, or the transformation it undergoes during affixation may not correspond perfectly to its overall continuation class. These exceptions can be handled by more sophisticated flag diacritics, and we anticipate that in the final version of the system, each entry may be accompanied by several flag diacritics. Our current model overgenerates: for example, roots need to be marked for their behaviour during active *an* prefixation. A particular instance would be that of the root *voly*, whose initial consonant undergoes b>v transformation to become present tense *mamboly* (m+an+voly).

As noted by Beesley and Karttunen (2003), more general cases can be ruled out by means of filters, sets of rules that apply on the lexical level – that is, on the Upper side of the LEXC transducer. Such filters can be used to exclude groups of continuation classes from combining with a certain affix or can 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. Encoding such interactions in the morphological analyzer provides important constraints for syntactic analysis.

4 Acknowledgments

We are grateful to Charles Randriamasimanana for help in the development of the system.

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)*, pp. 1–12.

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

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar Project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation.*

Keenan, Edward L. and Maria Polinsky. 1998. Malagasy. In Andrew Spencer and Arnold Zwicky (editors), *The Handbook of Morphology*. Oxford: Blackwell Publishers.

Randriamasimanana, Charles. 1986. *The Causatives of Malagasy*. Honolulu: University of Hawaii Press.