Computational Linguistics and Chinese Language Processing
Vol. 11, No. 4, December 2006, pp. 315-332

Tokenization and Morphological Analysis for Malagasy[1]

Mary Dalrymple*, Maria Liakata*, and Lisa Mackie*

Abstract

The authors 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 [1998] and Randriamasimanana [1986]. Words in Malagasy are built from roots by means of a variety of morphological operations such as compounding, affixation and reduplication. The authors analyze productive patterns of nominal and verbal morphology, and describe genitive compounding and suffixation for nouns and various derivational processes involving compounding and affixation for verbs. This work offers a computational analysis of Malagasy morphology, and forms the basis of a computational grammar and lexicon of Malagasy within the framework of the PARGRAM project.

Keywords: Malagasy, Austronesian, Morphological Analyzer, Fnite-State Morphology, PARGRAM

1. Malagasy in the PARGRAM Project

Malagasy is an Austronesian language spoken by about six million people on the island of Madagascar [Grimes 1999]. Along with Welsh, it is a focus of the Verb-Initial Grammars subproject (http://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].

The objective of PARGRAM is to develop parallel grammars for a range of different languages[2] using a shared linguistic framework and shared grammar writing techniques and technology. However, each project within the PARGRAM umbrella is driven by a different set of goals. For example, the English, German and Japanese grammars have been under development for a number of years; these grammars aim for very broad and robust coverage

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

* Centre for Linguistics and Philology, University of Oxford, Walton Street, Oxford OX1 2HG UK, telephone: +44 (0)1865 280 403 fax: +44 (0)1865 280 412
E-mail: mary.dalrymple@ling-phil.ox.ac.uk; mal@aber.ac.uk; l.mackie@lmh.oxon.org

[2] For details see http://www2.parc.com/isl/groups/nltt/pargram/

[Received April 29, 2006; Revised October 20, 2006; Accepted November 7, 2006]

for industrial applications. In contrast, the 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 is on producing 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 the grammar development work for Welsh and Malagasy work has progressed, the researchers have found that 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 continues to enhance understanding of the range of variation possible within the verb-initial type.

Like all grammars within the PARGRAM project, the development of the Malagasy grammar relies heavily on a computational component for morphological analysis. 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 created for shallow grammatical analysis or information retrieval applications. The challenge for these grammar development projects, then, is to overcome the problems arising from the lack of detailed grammatical information that these transducers made available.

The Malagasy grammar shares with a few of the other PARGRAM grammars (Arabic, Turkish, Urdu, Welsh) the difficulties and opportunities 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 labour between the morphology and syntax can be made in a well-motivated manner, rather than being imposed on the grammar writer. 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. Despite these disadvantages, the need for automatic morphological analysis for the Malagasy grammar project 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. 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 the Malagasy morphological analyzer, the researchers use 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, the 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, the authors 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.

2. Malagasy Morphology: An Overview

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 belong to a small closed class, the authors identify 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, the third productive type of morphological process in Malagasy is reduplication [Keenan and Razafimamonjy 1998], in which a new root is formed by reduplicating all or part of a basic root, giving a diminished, attenuated, or pejorative meaning: for example, reduplication of the root *fots*y 'white' gives *fotsifots*y 'whitish'. 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. The researchers have implemented and are currently testing a treatment of Malagasy reduplication using the COMPILE-REPLACE algorithm, but, as this has not yet been completely integrated into the full Malagasy grammar, the authors concentrate in the following on describing the treatment of nominal and verbal morphology.

3. Lexical Information

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]. Weak and strong roots behave differently in the processes that are treated here, and are listed separately in the morphological lexicon.

This lexicon currently 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. Guessed roots are also allowed for and are defined in terms of permissible root patterns; these roots are marked with the tag +Guess, and are permitted, though dispreferred, in syntactic analysis. In treatment of guessed forms, the authors 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]];
```

The definitions in (1) 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 amount of times the immediately preceding string is repeated. Note that the use of '^' here is different than in the definition of the continuation classes and orthographic rules. In the latter case, '^' designates a feature to be interpreted by the XFST rules.

4. Genitive Compounding

This 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. These are of the form Head+NP*gen*, where the Head can be any of the following: noun (in which case NP*ge*n 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 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*.

4.1 Target Phenomena

The following informal rules follow the treatment of Keenan and Polinsky [1998], and represent alternations in the final and first syllables of Head and NP*ge*n respectively. The hyphen, which sometimes alternates with apostrophe, is part of Malagasy orthography. The expressions ‘C’, ‘Co’ stand for consonants and ‘V’ and ‘Vo’ stand for vowels, whereas ‘S’ denotes a stop consonant. The lowercase characters denote the corresponding letters. The expression to the left of the ‘+’ sign stands for the final syllable 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 NP*gen*.

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

   (a) NP*ge*n begins with a vowel Vo: CV + Vo → C-Vo (remove final vowel in Head and concatenate)

   (b) NP*ge*n 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 *k*a or *tra*:
      V{ka|tra} +C → V-S

2. Head is not weak:

   (a) NP*ge*n begins with a vowel Vo:
   CV + Vo → CVn-Vo (prefix *n*-and concatenate)

   (b) NP*ge*n begins with a consonant Co with corresponding stop consonant S:
   CV + Co → CVn-S (S not bilabial), or
   CV + Co → CVm-S(Sbilabial)

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.

(2)

person	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

4.2 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 tokenizer written in XFST, a lexicon written in LEXC, and more general orthographic and phonological rules written in XFST [Beesley and Karttunen 2003].

As with the other grammar development projects within the PARGRAM initiative, the grammar is implemented within the XLE grammar development environment ([Crouch *et al.* 2006], see also http://www2.parc.com/isl/groups/nltt/xle/). The XLE 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. The output of the morphological analyzer is the input to syntactic analysis, obviating the need for listing each surface form separately in the syntactic lexicon. Instead, the syntactic component contains information about the syntactic content and behaviour of each root and affix combination as analyzed by the morphological component.

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-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, the 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 *Hanket*o *izy*. 'he will come here', the tokenizer produces the following result, where TB indicates a token boundary:

(3)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 nonpronominal genitive compounding, however, the situation is more complex. The compound form *akanjon-olona* 'a person's clothes' consists of two noun roots *akanj*o 'clothes' and *olona* 'person', and has the following structure:

(4)akanjon-olona
akanjo (epenthetic N; cf. 2a above) (compound boundary) olona

In the original treatment of nonpronominal 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 the researchers to revise the treatment of nonpronominal genitive compounds, moving most of the work to the tokenizer.

In the 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:

(5) akanjo TB +GEN+ TB olona

The root *akanj*o is a three-syllable root; the hypothetical root *akanjona* 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 1998]:

(a) Head is weak, reconstructed by the tokenizer as *volana* 'month', with Rule 1a requiring removal of the final vowel; or

(b) Head is strong, reconstructed by the tokenizer as *vola* 'money', with Rule 2a requiring insertion of -n

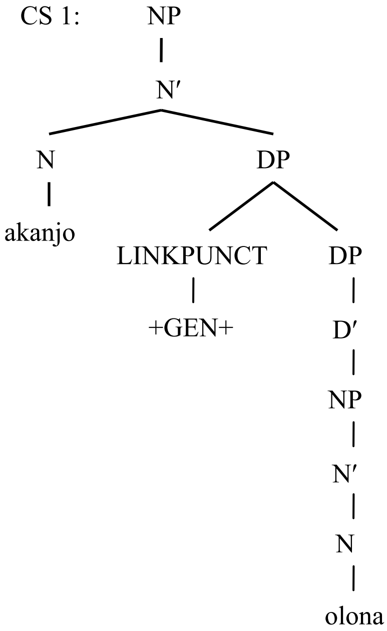

Figure 1. Phrase structure tree for akanjon-olona

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

The 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 signalled by special punctuation. The authors have left the treatment of these forms for future work, since it is unclear 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 genitive compounds, on the other hand, are best treated by the morphological analysis component, which is now described. 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/Upper side of the transducer, and permits the form to terminate with no suffixation, or alternatively allows genitive suffixation. Thus, the transducer relates the Lower string, the unsuffixed noun *akanjo*, to the morphologically analyzed lexical/Upper string which forms the input to syntactic analysis:

(6)LEXC transducer:
Upper: akanjo +Noun
Lower: 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/Lower side to the tag +1SgGen on the lexical/Upper side:

(7)LEXC transducer:
Upper: akanjo +Noun +1SgGen
Lower: 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*. The phrase structure tree that is produced for *akanjoko* is shown on the left side of Figure 2; the right side shows the root and series of tags that is output by the morphological analyzer and analyzed by the syntactic component.

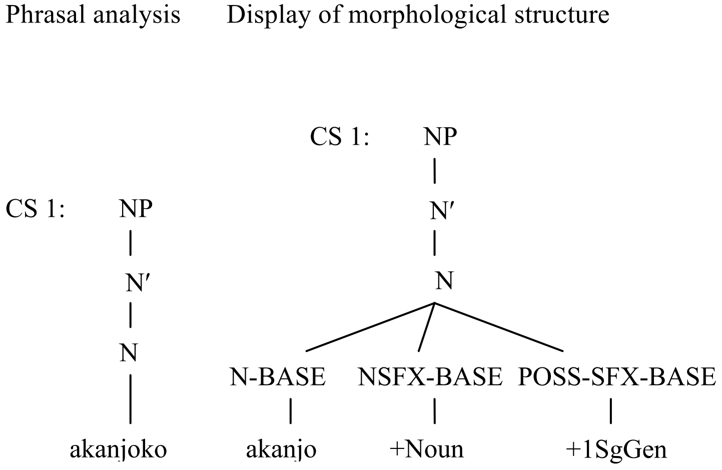

Figure 2. Analysis of akanjoko

5. Verbal Morphology

In many cases, however, a set of XFST rules is also needed to cater to 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. These tags 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. In the following, the authors describe the use of these tags in the analysis of verbal morphology.

5.1 Target Phenomena

Malagasy exhibits rich and complex verbal morphology [Randriamasimanana 1986; Keenan and Polinsky 1998]. 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 3, which is very similar to patterns for noun roots with optional pronominal genitive compounding. ROOT represents the form of the passive root. ROOTTYPE is one of ˆStrongRoot, ˆWeakKTRoot or ˆWeakNRoot; 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, using the affixes listed in (2). An example for the root passive *araka* 'be followed' is:

(8)Lexical: araka +Verb +3Gen
Surface: arany
'be followed by him/them'

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

The largest category of passive verbs is 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 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.

Passive **root**s

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

Suffix **passive**s

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

Prefix **passive**s

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

Circumstantial **for**m

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

Active **Verb**s

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

Figure 3. Verbal patterns

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 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, possibly, 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 an 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*.

5.2 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 ˆTNSˆ 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.

```
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 *fantatra* is *fantarina*.

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*, one 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, the authors 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:ˆWeakKTRoot          WeakKTEnding ;
+Verb:ˆWeakKTRootˆFtr2r    Suffixes ;
```

The feature ˆFtr2r is referenced by the XFST rule in (9), which transforms *tra* to *r* if the *tra* syllable is followed by the feature ˆFtr2r.

(9)[t r a] → r || __ ˆFtr2r

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 ˆFtr2r. This rule applies after removal of the tag ˆWeakKTRoot, which separates the root from the ˆFtr2r tag in the Lower string of the LEXC transducer. Directly after the application of this rule, the rule to remove the tag ˆFtr2r applies, preventing its appearance in the surface string and its interference with the application of other rules. Similar rules cater to 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 *fantari̲na* and not *fantara̲na* 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 one 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 ˆFpassi 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:ˆFpassi>          TR2RWeak;
```

However, this means that when the features ˆWeakKTRootˆFtr2r are added by the continuation class TR2RWeak, they are not immediately next to the root, but rather ˆFpassi 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 at runtime, when a certain phrase is being analyzed or generated. The authors use flag diacritics to store root-specific information, and, therefore, they are entered together with the lexical entry for the root. As 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, 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*.

The researchers 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, the lexicon does not list every root form in both the noun lexicon and the verb lexicon; nevertheless, one 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 but not nominal suffixes.

This is handled 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. Then one must 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 the 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 affix; this can be handled by more sophisticated flag diacritics which govern the permissible affixes that each root can accept. The current 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 Upper side of the LEXC transducer. Such filters are 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 one has 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. The current treatment incorporates 5 filters disallowing certain affix combinations for all roots. Encoding such interactions in the morphological analyzer provides important constraints for syntactic analysis.

6. Conclusion

The authors have presented a computational implementation of the derivational morphology of Malagasy, concentrating on the treatment of genitive compounding and affixal verb morphology. This approach closely follows the analysis of Keenan and Polinsky [1998] 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.

The morphological analyzer provides a solid basis for continuing work on the syntactic lexicon and grammar of Malagasy. Currently, the syntactic lexicon has been populated with the root and affix forms generated by the morphological analyzer and accepts these forms as input for syntactic analysis. The authors have encoded the syntactic contributions of each affix as well as default syntactic contributions for large classes of verb, noun, and adjective roots, and in current work are refining the lexicon to account for subclasses of roots with exceptional, non-default behaviour. The 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. With the morphological analysis component in place, the researchers anticipate now being able to make rapid progress in expanding the coverage of the Malagasy grammar and syntactic lexicon.

Acknowledgments

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

References

Beesley, K. R., and L. Karttunen, "Finite-state non-concatenative morphotactics," in *Proceedings of the Fifth Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON-2000)*, 2000., pp. 1-12.

Beesley, K. R., and L. Karttunen, *Finite-Stat*e *Morphology*. CSLI Publications, Stanford, 2003.

Butt, M., H. Dyvik, T. H. King, H. Masuichi, and C. Rohrer, "The Parallel Grammar Project," in *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation,* Taipei, 2002.

Çetinoğlu, Ö., and K. Oflazer, "Morphology-syntax interface for Turkish LFG," in *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, 2006.

Crouch, D., M. Dalrymple, R. Kaplan, T. King, J. Maxwell, and P. Newman, XLE documentation. Technical report, Palo Alto Research Center, Palo Alto, CA. www2.parc.com/istl/groups/nltt/xle/doc/xle_toc.html, 2006.

Dalrymple, M., M. Liakata, and L. Mackie, "A two-level morphology of Malagasy," in *Proceedings of PACLIC 19, the 19th Asia-Pacific Conference on Language, Information, and Computation.* Taipei: Academia Sinica, 2005.

Grimes, B. F, *ETHNOLOGUE: Languages of the World.* SIL International, 1999. URL www.sil.org/ethnologue/.

Keenan, E. L., and M. Polinsky, "Malagasy," in Andrew Spencer and Arnold Zwicky (editors), *The Handbook of Morphology*. Blackwell Publishers, Oxford, 1998.

Keenan, E. L., and J. P. Razafimamonjy, "Reduplication in Malagasy," in *The Structure of Malagasy III: UCLA Working Papers in Syntax and Semantics*. Los Angeles: UCLA Linguistics Department, 1998.

Randriamasimanana, C., *The Causatives of Malagasy*. Honolulu, 1986.