Non-concatinative Finite-State Morphotactics of Amharic Simple Verbs

Saba Amsalu
&
Girma A. Demeke

Abstract

In this paper we have described non-concatenative finite-state morphotactics of Amharic simple verbs. A morphological analyzer (transducer) that analyses simple Amharic verbal stems into their roots and feature tags is developed. The transducer also functions as a morphological synthesizer. It has an interface that works for Amharic text written in Unicode encoded Fidel script. We used Xerox Finite-State Tool (XFST) and Lexicon Compiler (LEXC) to construct the finite-state lexical transducer.

1. Introduction

Morphological analysis is the basis for most natural language processing tasks. Automatic procedures of synthesizing and analyzing word forms enable the creation, verification and update of broad-coverage lexica which reflect evolving usage and are less subject to lexical gaps. In light of this, we took advantage of the superior properties of finite-state tools in describing morphological processes and the bidirectional (synthesis/ analysis) functionality they provide.

There are previous studies to develop Amharic morphological analyzer using finite-state machines (cf. Amsalu and Gibbon 2005a&b and Fissaha and Haller 2003). These studies have made substantial contribution in describing the morphotactics of Amharic. However, they are far from being used for practical purposes. Overgeneration and incompleteness are the major drawbacks observed in them. Overgenerations resulted due to poor handling of constraints which is itself partly due to the generality of the theoretical studies made on the area. The dependence on incomplete lexicon has also limited these studies from having good coverage of the language. Therefore, we came up with a strategy of addressing a subset of the language thoroughly and progressively developing the full morphological analyzer.

In Section 2, the morphology of Amharic verbal stems is discussed with emphasis on non-concatinative processes. A separate treatment of irregular verbs is also made in this section. We give a detailed account of the finite-state morphological analyzer developed in Section 3. A brief clarification of the interface and internal representation of the morphological analyzer is provided in Section 4. Finally, concluding remarks and planned activities are forwarded in Section 5.

2. Morphology of Amharic simple verbs

Most natural languages construct words by linearly concatenating morphemes. Semitic languages, however, have unique non-concatinative properties in addition to the conventional concatenative modifications. In Amharic, non-concatenative operations of vocalic intercalation, reduplication accompanied by vowel insertion and radical reduction are the main components of word formation processes.

Verbs are the most complex category of words in Amharic. They are created mostly from triradical consonantal roots which are inflected by a process of merging with vocalic components based on various templates.

The verbal stems generated from the roots do not exist as free forms.[1] Instead, they undergo further inflections via the conventional concatenation of affixes (prefixes, suffixes, circumfixes).

In simple verbs intercalation, radical reduction, conditioning due to occurrence of flat or sharp consonants and gemination are the elements of word formation. Verb derivations are numerous and productive. For now we focus only on simple verbs.

2.1 Vowel intercalation

Simple verb stems are formed from consonantal roots with the infixation of vocalization. The vocalic elements often mark a grammatical category, either aspect (perfective or imperfective) or mood (imperative-jussive) (cf. Table 1). Any verbal stem in Amharic is, therefore, marked for either of the above grammatical categories, with the exception of the infinitive.

The infinitival stem, though categorized as a verb, is not marked for either of the above grammatical categories. Unlike other verbs, this form does not serve as a base for the inflection of other (verbal) grammatical categories, such as tenses, aspectual and modal categories (i.e.

[1] The imperative form seems an exception to this generalization, as it can stand as a free form. However, when this stem stands alone, i.e. without taking any visible additional morpheme, it is always interpreted as being marked to second person masculine singular. Consider the following examples:

hid ‘you (masculine. Singular) go!’
hid-i ↦ (hij-i) ‘you (feminine, singular) go!’
hid-u ‘you (plural) go!’.

We can see in the above examples the impetrative stem of go is ‘hid’ which is also a default second masculine singular form. As there is a clear morphological marker for feminine and plural, the second masculine singular can, therefore, be considered as marked by a zero morpheme. This means that the imperative stem cannot also be considered as a free form and, hence, there is no exception to the above generalization.

progressive aspect and mood of intention), Furthermore, all other verb stems should necessarily take agreement markers in their free form. Infinitival verbal stems, however, may not do so. On the other hand, an infinitival verbal stem must take a prefix *mä* (or *m* when a stem has initial vowel). Infinitives, i.e m(ä)- + the (infinitival) stem, are considered as nominal categories. This is because the phrase headed by an infinitive is understood as an NP and can be a subject or an object like other NPs headed by simple or derived nominal categories. However, unlike other nominal categories such as simple nouns and derived result and abstract nouns, they take arguments like verbs. Due to the dual behaviors nominal as well as verbal that they have in syntax, infinitives in Amharic are categorized as nominalized verbs. Their nominal behavior emanates from the prefix *mä-* (or *m-*)[2] and their verbal behavior, from the stem, i.e. from the non-concatinative vocalic element inserted in the root. The attachment of the prefix *mä-* (or *m-*) to the stem, i.e. their nominalization, is assumed to take place in syntax (cf. Demeke 2006 and Manahlot 1977). We have therefore considered the infinitival stem, as one of the verbal stems in this work.

In general, five types of simple verbal stems (including the infinitive) can be derived with the infixation of vocalic elements to a root (and in some cases with the lengthening of the penultimate radical of a root), Table 1 shows the derivation of these five simple verbal stems from the root $\sqrt{}$*sbr* 'to break'.

[2] This prefix is considered as a nominalizing morpheme (see Demeke. 2006, Manahlot 1977 among many others).

Table 1: Simple verbal stems of the root √*sbr*

Stem type	Stem
Perfective	säbbär
Imperfective	säbr
Imperative	sbär
Gerund	säbr
Infinitive	sbär

2.2 Verb Types

Leaving semantics aside, verbs in Amharic can be divided into various types or classes based on the pattern of gemination, the number of radicals (of the root), and the quality of vowels inserted between the radicals. Based on these criteria, Bender and Fulas (1978) classify Amharic verbs into 42 classes and/ or sub-classes. As this classification considers the phonological pattern of each verb type, we have adopted it in this work. We discuss below this classification in more detail.

2.2.1 Radical types

The radicals of a root are basically consonants. Based on such consonants, verbs in Amharic can be classified as mono-radical, biradical, triradical and multiradical. The majority roots in Amharic like other Semitic languages have three radicals. However, there are also a significant number of verbs that are multiradical and biradical. The quinqiradical and sixiradical verbs are rare and, according to Bender and Fulas (1978: 23), these types of verbs can be "best viewed as derived from triliterals". Furthermore, only a single verb is identified as a mono-radical. This mono-radical verb is highly irregular. It is only found in imperfective and infinitive forms. Therefore, we have treated this mono-radical verb along with other irregular verbs in Section 2.3. In general, from the point of view of the number of radicals, regular verb classes in Amharic can be classified into biradicals, triradicals and quadriradicals. Due to phonemic loss, we may have a verb having less radicals in the surface form. We discuss this pattern of reduction in Section 2.2.4.

2.2.2 Gemination

Based on the pattern of gemination, verbs in Amharic are categorized into three types. These are type A, type B and type C. We show this pattern of gemination for triradical roots in Table 2.

Table 2: Basic non-concatenative verbal inflectional forms of triradical roots

Verb types	**Type A**	**Type B**	**Type C**
Perfective	$C_1VC_2C_2VC_3$-	$C_1VC_2C_2VC_3$-	$C_1VC_2C_2VC_3$-
Imperfective	-$C_1VC_2C_3$(-)	-$C_1VC_2C_2C_3$(-)	-$C_1VC_2C_2C_3$(-)
Imperative	$C_1C_2VC_3$(-)	$C_1VC_2C_2C_3$(-)	$C_1VC_2C_3$(-)
Gerund	$C_1VC_2C_3$-	$C_1VC_2C_2C_3$-	$C_1VC_2C_3$-
Infinitive	-$C_1C_2VC_3$	-$C_1VC_2C_2C_3$	-$C_1VC_2C_3$

Unlike most non-Ethio-Semitic languages each root type in Amharic is lexical. It is lexical in a sense that any of the type can be transitive, intransitive or can fall into any kind of semantic category: "These types [type A, type B and type C] are not conditioned either by the nature of the consonants or by the meaning of the verb. Indeed, the verbs in any one of these types may be active, transitive, verbs of state, and so on, and consists of any kind of consonants" (Leslau 1995: 283).

As we can see from Table 2, in type A, except in the perfective form, there is no gemination at all. Type B is typically characterized by the gemination of the second radical throughout the conjugation. Type C is different from the two in that the gemination of the penultimate consonant is in the perfective and imperfective aspects only. The root *√sbr* 'break', *√flg* 'want' and *√brk* 'bless' can be examples of type A, type B and type C respectively.

Gemination is only read and not written in Amharic script. Indeed, in most Ethiopic based scripts, no special representation is used for a geminate consonant different from the non-geminated one. In Amharic writing, gemination is, therefore, only identified from the context. Apparently, gemination is not part of the implementation in this paper, as we work on corpus written in Unicode encoded Amharic script.

2.2.3 Occurrence of vowels

In principle, the only vowel intercalated in simple verbs is /ä/. This is true especially for all verb types except type C. Type C verbs take the mid central low vowel immediately following the first radical, i.e. initial radical, of the root as in *barräk-ä* 'lit. He blessed'. Therefore, when vowels other than /ä/ occur, except on the first syllable of Type c, it is, the result of either radical reduction, or assimilation or due to avoiding impermissible sequence of consonants.

The usual epenthetic vowel in Amharic is the high central vowel /ï/. This vowel is inserted when a cluster of consonants occur at the beginning of a word or when three consonants (could be a geminated with a non-geminated consonant) occur sequentially at any position. We discuss radical reduction and the change of vowel quality caused by such reduction in the following section.

2.2.4 Radical reduction

A substantial number of roots are subject to radical reduction. Reduction takes place mostly on roots with the occurrence of the glides /h/, /ʔ/, /y/, /w/ or labio-velars such as /k^w/, /q^w/, /g^w/, and /h^w/. The glottal ejective /h/ and the glottal stop /ʔ/ may be reduced to the vowel /*a*/[3]. This can occur, indeed, only if there is a vowel following such radicals, otherwise they will be totally lost without leaving any of their features.[4] The reduction of /w/ may result with making the preceding consonant flat, i.e. labialized, and /y/ semi-palatal, i.e. sharp consonant. Table 3

[3] Practically speaking, the glottal consonants /h/ and /ʔ/ may not be considered as reduced to the vowel /a/. There is more to it. These glides are rather lost. But before their loss, they force the following vowel, if there is any, to be a low vowel. Since the vowel that follows these radicals (or any radical for that matter, recall the above discussion) is /ä/ (or /ï/ in the case of impermissible consonant cluster, the result of such phonological process, i.e. lowering will, of course, be /a/. Taking the perfective verb formation of the root *√ʔwq*, we show this phonological process step-by-step as follows.

√ ʔwq – (root)
ʔäwäq- (underlying form)
ä↦ a/ʔ- (assimilation)
ʔawäq-
ʔ↦Ø/-a (loss of /ʔ/)
awäq (surface form)

[4] Note that the loss of such glides (and other radicals as well) and a change of vowel quality may not happen in all stems formed from the same root. Reduction takes place mostly with verbal categories but not nominal. Furthermore, the change of the vowel quality, i.e. assimilation, seems to occur only if the consonant in question, i.e. /h/ or /ʔ/ is to be dropped. For instance such phonological process does not happen with nominals derived from the same root. Consider the following:

(i) √ ʔwq ‘to know
ʔïwqät ‘knowledge’
(ii) √ mhr ‘to learn’
tä-marä verb.perfective
tïmhïrt ‘education’

shows the stems formed from the root *√ʔwq* ‘to know’ (with a lost radical /ʔ/).

Table 3: Verb conjugations of the root *√ʔwq*[5]

Stem type	Perfective	Imperfective	Imperative	Gerund	Infinitive
Pattern	CVCVC	CVCC	CCVC	CVCC	CCVC
Stem	awäq	awq	ïwäq	awq	awäq

The vowel /o/ occurs alternatively in dialects in cases where flat consonants such as /k^w/, /q^w/ and /g^w/ appear followed by the vowel /ä/ as in *k^wä, q^wä,* and *g^wä*. In such cases two phonological processes will take place: (1) the vowel /ä/ will be changed to the vowel /o/ and, (2) the flat consonant will be reduced to its non-flat counterpart. Hence forms such as *k^wä, q^wä,* and *g^wä* in some dialects will be *ko, qo,* and *go* respectively.

When the vowel following the flat consonant is a high central vowel, then in the dialects where the above changes are observed, it is converted to its counterpart high back vowel /u/. See Table 4 for these alternative forms for the same verb.

[5] Given the morphological pattern of Amharic, is /ï/ it is the fact that the expected initial vowel in both the infinitival and imperative conjugations of the root *√ʔwq* However, as we can see in Table 3, this vowel is changed to /a/ in the infinitive but not in the imperative. The explanation is that the radical /ʔ/ is lost before the insertion of the vowel /ï/ in the imperative but not in the infinitive. This means that, if the loss of such radicals happens before the insertion of vowels, then we do not see a change of a vowel quality as expected.

Table 4: Conjugation of the root *√q^wt’r*

Stem type	Perfective	Imperfective	Imperative	Gerund	Infinitive
Pattern	CVCCVC	CVCC	CCVC	CVCC	CCVC
Stem (variant 1)	*q^w*ätt’ä r	*q^w*ät’r	*q^w*ït’är	*q^w*ät’r	*q^w*t’ä r
Stem (variant 2)	qot’är	qot’r	qut’är	qot’r	qut’är

Although the variation of the above two stem forms is dialectal, we note that the difference of vowel qualities in these two forms is a result of a phonological process as pointed out above, where the stem in variant 1 is a basic form and variant 2 is a derived one.

The vowel /*e/* also refers to an underlying sharp consonant such as c^y and t’^{y} followed by /ä/ as in c^y*ä* and *t’^{y}ä,* making *ce* and *te* respectively. When the vowel following the flat consonant is a high central vowel, i.e. /ï/, then in the dialects where the above changes are observed, it is converted to /i/, a high front vowel. Table 5 describes the conjugation of the root *t’^{y}s.*[6]

[6] As mentioned above, the existence of sharp consonants such as t’^{y} and c^y may better be seen as a result of a phonological process caused by the radical /y/ on the preceding radical prior to its being dropped. However, going to such detail is not relevant to our present work, hence; we take these types of consonants as being part of the root.

Table 5: Verb conjugations of the root √$t'^{y}s$

Conjugation	Perfective	Imperfective	Imperative	Gerund	Infinitive
Pattern	CVC	CVC	CVC	CVC	CVC
Stem	t’es	t’es	t’is	t’es	t’es

2.3 Irregular verbs

Irregular verbs are very few in Amharic. The most common ones are: *näw* (is), *all*ä (to exist, be present, available), *näbbärä* (was), *al*ä (say), *t*ä*w*ä (to leave), *yïša* (to want).[7] Among these *näbbärä, allä, alä*, and *täwä* actually have affinity to one or other basic classes. For example, the first three are basically triradical roots having the following radicals as in √*nbr*, √*hlw* and √*bhl* respectively. We take *näbbärä* and *allã* as irregular because they are found only in the perfective form. Besides, the negative form of *allä* is irregular which is *yälläm*. On the other hand, we take *alä* (or the root √*bhl*) as irregular due to its inconsistent loss of radicals. As can be seen in Table 6 below, the first two radicals of this root are lost in the perfective, imperfective and infinitive. On the other hand, in other two conjugations, only the glide /h/ is lost.

[7] Note that all these forms are marked to the third person singular masculine subject marker and hence, the English translations given to them are not the natural equivalent ones.

Table 6: Conjugations of the root √*bhl* 'to say'

"Stem" type	Conjugation	Gloss (literal)
Perfective + Agr	alä	'he said'
Imperfective+ Agr	yïl	'he will say/ says'
Gerund+ Agr	bïlo	'(he) having said'
Imperative (+Agr)	bäl	'say!'
Infinitive + *m–*	malät	'to say'

Whether *täwä* is a biradical or triradical is difficult to determine. Furthermore, the loss of the final radical /w/ is inconsistent in its conjugations as can be seen in the following table. Hence, we have treated this verb as irregular.

Table 7: Conjugation of the irregular verb *täwä*

"Stem" type	conjugation	Gloss (literal)
Perfective + Agr	täwä	'he left'
Imperfective+ Agr	yïtäw	'he will leave'
Gerund+ Agr	tïto	'(he) having left'
Imperative (+ Agr)	täw	'leave!'
Infinitive + *mä–*	mätäw	'to leave'

As we can see in Table 7 the verb *täwä* only takes the consonant /t/ in the gerund and, in fact, with the loss of one of its radical, i.e. /w/.[8]

The copula *näw* 'he is' is highly irregular for the following three basic reasons: First, it is found only in the present tense;. Second, the negative form of this copula is a supplative form which is *aydäläm* 'he is not'; Third, unlike other verbs it takes an object agreement marker for the identification of a subject (cf. Table 8).

Table 8: The conjugation of the copula *näw*

Conjugation	**Gloss**
näň	'It is me', 'lit. I am'
näh	'It is you', 'lit. You (masc., sing.) are'
näš	'It is you', 'Lit. You (fem., sing.) are'
näwot	'It is you', 'Lit. You (respect, sing.) are'
näw	'It is he', 'Lit. He is'
nat	'It is she', 'Lit. She is'
nän	'It is we', 'Lit. We are'
naccïhu	'It is you', 'Lit. You (pl.) are'
naccäw	'It is you', 'Lit. They are'

[8] Triradical roots that lose the final radical usually add the consonant /t/ in the lost radical position in their infinitive and gerund conjugations. Furthermore, when there is only a single radical in either the infinitival or gerund or both conjugations, the addition of /t/ will occur in the final syllable (cf. Table 6, infinitive).

As pointed out in Section 2, the only mono-radical verb in Amharic is among the highly irregular verbs. It is only found in the imperfective and infinitival forms. Hence, all the above discussed irregular forms, including the negative forms of *allä* 'there is, he exists' and the copula *näw* 'lit. He is, it is' are handled separately.

In general, in this work we have considered nearly 1300 roots of regular and irregular verbs obtained from various sources, mainly Bender and Fulas (1978) and Dawkins (1960). From such roots we were able to generate a total of about 6400 simple verb stems. In subsequent sections, we discuss in brief the system we developed and the procedure we used to generate such simple verbal stems.

3. Finite-state representation

Finite-state machines (FSMs) have proved to be straightforward and powerful tools in describing morphological processes. Finite-state machines are compiled from regular expressions with which most morphological phenomena can be described. Finite-state transducers (FSMs with two tapes) are bidirectional. This gives them additional power to serve both as analyzers and generators. In our project we make use of Xerox Finite-state Tool (XFST) and Lexicon Compiler (LEXC) (cf. Beesely & Katunen 2003) to generate a set of stems for each root. These tools are well tested and have been used on other Semitic languages such as Arabic and Hebrew with successful results (see Sholomo & Wintner 1998 and Beesely 1996, 1998 & 2001 among others).

Using the above tools we have constructed lexicon automata and rule transducers. The lexicon automaton accepts the roots in the language. The rule transducer, on the other hand, consists of the rules by which the root lexicon is modified to form stems. These two components are combined to form a single transducer which has two tapes. One of the tapes, the lower side, consists of stems and the other tape, upper side, consists of roots and feature tags.

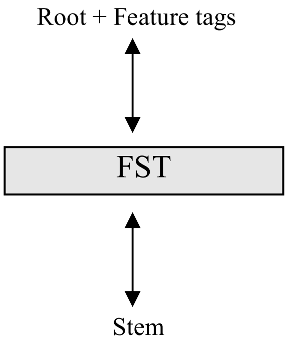

Figure 1: A finite-state stem generator/analyzer

Figure 1 describes the transducer as a black box. It takes a valid stem and produces its root and feature tags that are the causes for the inflection and in the reverse direction; it takes root plus feature tags and produces the corresponding stem.

3.1 Lexicon automata

Our lexicon automata are mainly on from the list of roots compiled by Bender and Fulas (1978) which are 1277 in numbers. As mentioned above, Bender and Fulas classify these verbs into 42 classes and/ or sub-classes based on the structure of stems that are formed from them. We also gathered irregular verbs from the study made by Dawkins (1960). We do not, however, claim that our corpus is complete, but we believe it is exhaustive enough to address most of the words found in the literature.

Xerox Lexicon Compiler (LEXC) is used for compiling the root lexicon. Any string outside the roots in the language is discarded. In Figure 2 an automata that accepts the strings *√mkr*, *√mnn*, *√mrg*, *√mrmr*, *√mskr* and *√mst'r* is depicted. In the same manner our lexicon automata stores all the roots in the language.

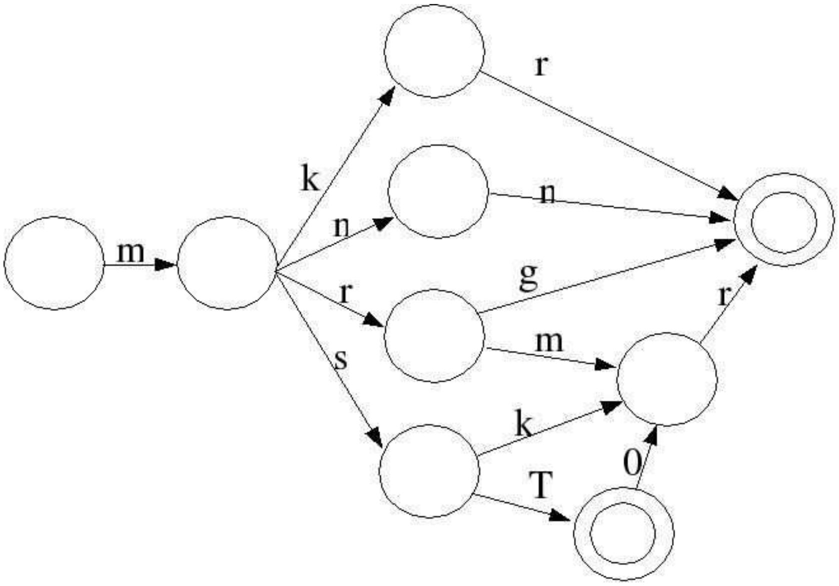

Figure 2: A finite-state automata accepting valid strings (roots)

3.2 Rule transducers

Rule transducers describe the rules of forming stems from the consonantal roots. Rules need to be carefully designed. They should be strict in that they should be able to handle constraints. Failure to define constraints will result in the formation of illicit stems. They should also be complete to provide good coverage of the language.

3.2.1 Rule one: Vowel intercalation

The intercalation problem is solved using the *merge* and *compile replace* algorithms of XFST. *Merge* is a pattern-filling operator that combines a template and fillers in one network. Figure 3 demonstrates the modification of a triradical root to form the perfective stem.

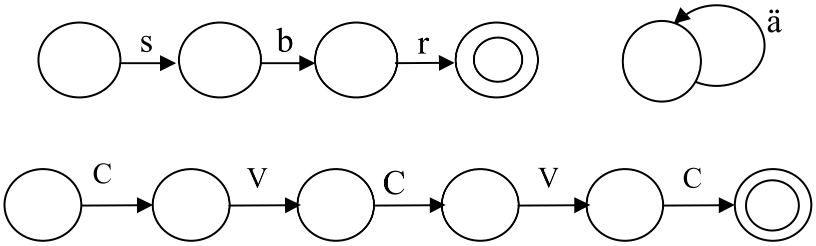

Figure 3: Template and fillers (root & vocalization)

Three networks: root (*√sbr*), template (*CVCVC*) and vocalization (ä) are the inputs for the merge operation. The template is a pattern for the perfective form of the root *√sbr*. The root and the vocalization fill into the template to form the single transducer as in Figure 4. The merge operation represented by *.m > . & . < m.* creates a single network that has an upper side *[sbr+Perf]* and a lower side ˆ[*{sbr}.m > .{CV CV C}. < m.*[ä *]ˆ]. The compile-replace algorithm, afterwards, compiles the lower side of the network by substituting the Cs of the template by the consonants in the root in the order they are written and substitutes all Vs with the vocalization. It finally replaces the original lower side created by merging by the newly compiled form. The end result would, therefore, be the transducer in Figure 4.

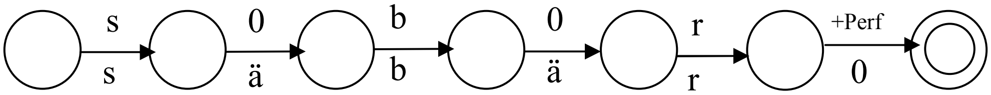

Figure 4: Lexical and surface forms

3.2.2 Rule two: Radical reduction

Recall that in some roots reduction of radicals takes place in stem formation. The reduction can be handled before or after the intercalation of a vowel depending on whether the vowel intercalated is affected or not by the reduced radical. The reduction of the glide /ʔ/ in the root *ʔwq* described in Table 3 is accomplished by replacing it by /a/ before or after intercalation of the infix in the four stems (i.e. excluding the imperative): *ʔwq –> awq*, *ʔ– > a*, and by /ï/ in the case of imperative *ʔwq –> ïwq*, *ʔ– > ï*, (but, recall the discussion in footnotes 2 and 3).

The contraction of flat consonants is made optionally to accommodate the different variants in dialects. The order of execution of intercalation and contraction is of importance in this case as the vowel plays a role in the process. Hence, first the vowel is intercalated with the operation described in the previous section: √q^{W}äl*l*ä*f –>* q^{W}äl*l*ä*f*. Then, the string /wä/ is optionally replaced by *o* while still keeping the original form: q^{W}*älläf –> qoll*ä*f* (q^{W}*äll*ä*f*) (note the optional replace operator (– >)), {wä} (– >) o.

Sharp consonants are also handled in the same way. The vowel is intercalated first: $t'^{y}s$ –> $t'^{y}äs$ and then, the replace operation is executed as follows:

t’^{y}äs –>t’es,
{yä}– >e

So far, we have described Amharic text in Latin script. But Amharic is not written in Latin script, as mentioned earlier. For reasons of reducing complexities in processing on the one hand, and the need to make the system consume text as it exists naturally on the other hand, we designed the interface to work in Amharic script while the internal operation is in Roman alphabet.

4. Text encoding

Amharic uses a syllabary script called *Fidel*. In Fidel graphemes denote consonant-vowel (CV) combinations. Consider for instance the following individual symbols:

ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
tä	tu	ti	ta	te	tï	to

If a vowel occurs at the beginning of a string, it is written independently. In other words, there are symbols for individual vowels. When a vowel, however, comes in contact with a consonant on its left side, perhaps during affixation, it is not realized with an independent vowel grapheme but along with its immediate preceding consonant by modifying the grapheme in question.

If a vowel comes in contact with another vowel forming a vowel cluster, note that, phonological alternations take place. Such kind of modification takes place with the insertion of the semi-vowels (either /w/ or /y/, depending mostly on the quality of the vowels), substitution, i.e. substituting the vowels /i/ and /e/ with /y/ and /u/ and /o/ with /w/, or deletion of one of the sequenced vowels. In some cases the would-be deleted vowel may leave its trace in the form of leaving behind some of its feature on the preceding consonant as in bälto-al → bält^{w}al 'he has eaten':

t→t^{w}/-o
o→ Ø/t^{w}-

Apparently, processing a text in Amharic script makes the computation quite complex, as these phonological processes will cause a change of a grapheme, which is, syllabic, i.e. CV. Hence, we preferred to make the entire computation in Latin script. However, we designed the interface to be in Fidel. This allows the user not to deal with transliterated data. Simple replace rules have allowed us to map Fidel symbols into Latin and vice-versa. The System for Ethiopic Representation in ASCII (SERA)[9] is used to generate the corresponding Latin symbols. Apparently, the system accepts any string from Unicode encoded Amharic text, converts it into SERA representation, analyses it and again converts the root obtained into Fidel. The feature tags are still in Latin. It also works in the same fashion on the opposite direction.

5. Conclusion and future work

This work is just track 1 of a long way towards having a full-fledged morphological analyzer. We only dealt with simple verbs. Derived verbs need to be addressed in the future. Moreover, as Amharic stems are further inflected to form words, complex concatinative operations with long distance constraints are also ahead of us. All in all, about 6400 simple verb stems have been generated from nearly 1300 roots of regular and irregular verbs. Each one of them will further be modified to form

[9] http://www.abyssiniacybergateway.net/fidel/sera-faq.html

fully inflected forms. In subsequent tracks we will be working on derived verb forms and some aspects of inflection to form fully inflected surface forms.

References

Amsalu, Saba and Dafydd Gibbon. 2005a. A complete fs model for Amharic morphographemics. In *Proceedings of FSMNLP*. Helsinki, Finland.

Amsalu, Saba and Dafydd Gibbon. 2005b. Finite state morphology of Amharic. *In Proceedings of the International Conference on Recent Advances on Natural language processing 2005*. pp 47-51, Borovets, Bulgaria.

Beesley, Kenneth R. 1996. Arabic finite-state morphological analysis and generation. In *Proceedings of COLING-96, the 16*th *International Conference on Computational Linguistics*, volume 1, Copenhagen, Denmark.

Beesley, Kenneth R. 1998. Arabic morphology using only finite-state operations. In *Proceedings of the Workshop on Computational Approaches to Semitic languages, COLING-ACL'98*, volume 1, pages 50-57, Montreal, Canada.

Beesley, Kenneth R. 2001. Finite-state morphological analysis and generation of Arabic at xerox research: Status and plans in 2001. In *ACL Workshop on Arabic Language Processing: Status and Perspective*, volume 1, pp. 1-8. Toulouse, France.

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

Bender, Lionel M. and Hailu Fulas. 1978. *Amharic Verb Morphology*. East Lansig, Michigan: African Studies Center, Michigan State University.

Dawkins, C. H. 1960. *The Fundamentals of Amharic*. Addis Ababa: Sudan Interior Mission.

Demeke 2006. The Syntax of Infinitival Clauses in Amharic. Ms. Addis Ababa University.

Fissaha, Sisay and Johann Haller. 2003. Amharic verb lexicon in the context of machine translation. *In Proceedings of Traitement Automatique des Langues Naturelles, TALN 2003*, pp. 183-192

Leslau, Wolf. 1995. *Reference Grammar of Amharic*. Wiesbaden: Otto Harrassowitz.

Manahlot, Demissie. 1977. Nominal Clause in Amharic. Doctoral Dissertation; Georgetown University.

Yimam, Baye. 1999. Root reductions and extensions in Amharic. *Ethiopian Journal of Languages and Literature*. 9:56-88.

Yona, Shlomo and Shuly Wintner. 2005. A finite-state morphological grammar of Hebrew. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages. Pp. 9-16. Association for Computational Linguists: Ann Arbor.

Addresses of the Authers:

Saba Amsalu
Bielefeld University
saba@uni-bielefeld.de

Girma A. Demeke
Ethiopian Languages Research Center
Addis Ababa University
girmaad@gmail.com