Finite state morphology of Amharic

Dafydd Gibbon
Bielefeld University

Finite State Morphology of Amharic

Saba Amsalu and **Dafydd Gibbon**
Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld
Universität strasse 25
D-33501, Germany
{saba,gibbon}@uni-bielefeld.de

Abstract

For several computational linguistic tasks we require a morphological decomposition strategy. This paper describes non–linear morphology, modelled with finite–state (FS) techniques and implemented in a well–known FS toolset. We present a complete analysis of Amharic words of all categories. Analyses display the root, pattern and feature tags indicating part of speech, person, number, gender, mood, tense, etc.

1 Introduction

Amharic is a Semitic language, the official language of Ethiopia. Document production in Amharic is increasing rapidly, with conventional printing and word–processing, but little has been done to exploit these documents as a valuable resource for use in automatic language processing. Experimental computational work on specific aspects of Amharic is in progress at Addis Ababa University and elsewhere; e.g. (Alemayehu & Willett, 2002), (Fissaha & Haller, 2003a), (Fissaha & Haller, 2003b) and (Alemu, Asker & Getachew, 2003). We report here on the first complete account of finite–state Amharic morphology for all parts of speech, which was designed as a front–end for parallel corpus alignment, and implemented using the Xerox Finite State Tools.

2 Objectives

The goal of this work is to construct a generic morphological analyser for applications such as machine translation, sense disambiguation, lexicography, and terminology extraction. We aim to construct a tool that will analyse Amharic words from a natural language text transliterated into phonemic ASCII respresentation (SERA)[1]. The system has to produce accurate component roots/stems and feature tags that indicate part of speech, person, number, gender, mood, tense, etc. ; and it also has to give correct surface forms when run in the reverse direction.

3 Amharic Morphology

Amharic verbs exhibit the typical Semitic non–linear word formation with intercalation (interdigitation) of consonantal roots with vocalic patterns. This also applies to deverbal nouns and adjectives. We use the term 'root' for lexical morphemes consisting of consonants, 'radical' for consonant constituents of roots; and 'stem' for intercalated forms.

3.1 Verbs

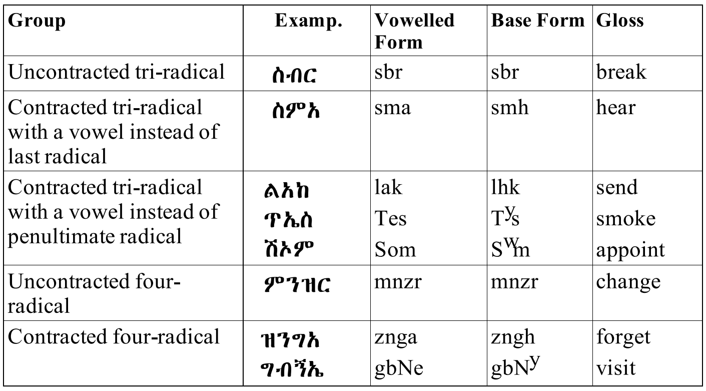
Verbs are morphologically the most complex POS in Amharic, with many inflectional forms; numerous words with other POS are derived primarily from verbs. Roots mainly consist of three radicals. It is controversial whether non–triradical roots are derived from triradicals; see (Dawkins, 1960); cf. (Bender & Fulas, 1978); (Yimam, 1999). Dawkins' classification is shown in Table 1. Simple verbs have five verbal stems that are formed by intercalation of vowels with skeleton patterns of the types CVCVC, CVCC etc.; see (Dawkins, 1960) (Bender & Fulas, 1978). These stems are: Perfective, Contingent, Jussive, Gerundive and Infinitive.

Aspect	Pattern	Stem	Description
Perfect	CVC̲VC	säb̲är	broke
Contingent	CVC̄C	säbr	break, will break
Jussive	CCVC	sbär	break! let sb. break!
Gerund	CVCC	säbr	breaking
Infinitive	CCVC	sbär	to break

Table 2:
Conjugation of a typical triradical Type A verb root *sbr*.

In Amharic verbs, the only vowel which is genuinely intercalated is *ä*. (cf. Table 2) shows the conjugation of the root *sbr*–typical triradical, type A (penultimate gemination in perfective stem only). When vowels other than the usual *ä*

[1] SERA (System for Ethiopic Representation in ASCII) is widely used for transliteration between Ethiopic syllables and ASCII



Group	Examp.	Vowelled Form	Base Form	Gloss
Uncontracted tri-radical	ስብር	sbr	sbr	break
Contracted tri-radical with a vowel instead of last radical	ስምአ	sma	smh	hear
Contracted tri-radical with a vowel instead of penultimate radical	ልአክ ጥኤስ ሽኦም	lak Tes Som	lhk T^ys S^wm	send smoke appoint
Uncontracted four-radical	ምንዝር	mnzr	mnzr	change
Contracted four-radical	ዝንግአ ግብኜ	znga gbNe	zngh gbN^y	forget visit

Table 1: Dawkins' classification of roots.

occur in stems, it is the result of historical consonantal reduction, or to conditioning by sharp or flat consonants. The vowel *a* occurs due to the reduction of the glide *h* in the root. The vowel *o* alternatively occurs in dialects in cases where flat consonants such as k^w*ä*, q^w*ä*, g^w*ä* etc. occur to create the forms *ko, qo, go* etc. When the vowel is short it is converted to *u* instead of *o*. The vowel *e* also refers to an underlining sharp consonant such as C^y*ä*, T^y*ä*, making *Ce, Te.*

The stems have the patterns of gemination, commonly referred to as Types A, B and C (the Fidel script does not distinguish between geminate consonants; they are read but not written):

- *Type A*: penultimate consonant geminates in Perfect only
- *Type B*: penultimate consonant geminates throughout the conjugation
- *Type C*: penultimate consonant geminates in Perfect and Contingent.

Several linguists have categorised Amharic verbs formally on the basis of root and stem structure; cf. (Bender, 1968), (Bender & Fulas, 1978), (Dawkins, 1960), (Markos, 1994). A detailed study of verb morphology is given by (Bender, 1968) and (Bender & Fulas, 1978): 42 verb classes based on three main morphotactic criteria which provide input to phonological rules:

1. consonantal skeleton (one or more radicals);
2. gemination pattern (Types A, B, C);
3. occurrence of vowels other than *ä* (i.e. *e, o, a*).

Amharic verbs are not derived from other POS but from other verbs, mainly by affixation, penultimate consonant reduplication and vowel insertion; cf. (Amare, 1989), (Yimam, 1995). Except for the second person masculine jussive, the stem is always minimally inflected with a subject marker. The verb may be inflected for Person, Gender, Number, Mood and Tense. The verb is also inflected for beneficative, malfactive, causative, transitive, passive, dative, negative (Berhane, 1992).

3.2 Nouns

Amharic nouns are either simplex (e.g. *bEt* 'house', *merEt* 'earth' and *Isat* 'fire') or derived. The latter are derived from verb roots, adjectives or other nouns (e.g. *TyaqE* 'question' from *Tyq* 'to ask' , *degnet* 'generosity' from *deg* 'generous',*xumet* 'post, title' from xum 'an appointed person').

Deverbal nouns are derived from verb roots by intercalating different vowels between the radicals, by adding suffixes to the root without vowel intercalation, or by consonant reduction; cf. (Dawkins, 1960), (Amare, 1989), (Yimam, 1995). Affixation is the major process when deriving them from adjectives and other nouns. Nouns

Singular	Plural	(Alternative)	Gloss
mezgeb	mezagbt	mezgeboc	archive(s)
anbessa	anabst	anbessoc	lion(s)

Geez pl.noun	Amharic pl.
Mekuannt	mekuanntoc
Liqawnt	liqawntoc

Table 3: Treatment of Geez singular and plural borrowings.

are inflected for Number, Gender, Case and Definiteness. Most plural nouns are formed by adding a plural marker affix (*–oc* or *–woc* — their distribution is determined phonologically) to the singular form, although when referring to groups belonging to a certain tribe or country *–yan* is affixed. Nouns from the liturgical Geez language do not necessarily have these plural suffixes. Often, another operation in addition to plural marker affixation occurs. Table 3 lists noun borrowings from Geez: some Geez plural nouns are incorporated into Amharic as singulars and get an additional plural marker. Some collective nouns are, however, formed by full reduplication of the singular noun with insertion of a linking vowel *a*.

There are two genders in amharic, masculine and feminine. For things that are not naturally male or female, the gender female tends to be used when the entity is small or adorable; the gender male is used otherwise. The feminine gender suffix (*–it* or *–yt*, phonologically conditioned) is used to mark femineneness in cases which otherwise would be masculine.

Definiteness markers are suffixes that vary depending on the gender of the noun (*–u* or *–wa* for feminine and *–u* or *–w* for masculine).

3.3 Pronouns, Adjectives, Adverbs, Prepositions, Conjunctions

Amharic pronouns can be free or bound to other POS. In the accusative and genitive, free personal pronouns take the affixes for nouns.

Adjectives are generally derived from verbs. The number of simplex adjectives is relatively small. Some simple adjectives are *qey* 'red', *deg* 'generous'. Adjectives are also derived from nouns or from verbal morphemes (Amare, 1989): cf. *brtu* 'strong', from *brth* 'be strong', *hayleNa* 'forceful', from *hayl* 'force, energy'. Like nouns, adjectives are inflected for Number, Case, Gender and Definiteness.

Adverbs in Amharic are very few, about seven common items, some derived from adjectives by suffixing *Na*; cf. (Amare, 1989) and (Yimam, 1995). Adverbial functions are often accomplished with noun phrases, prepositional phrases and subordinate clauses.

Conjunctions and prepositions have similar behaviours, and are often placed in the same class (*mestewadid*): no affixation, not used as base for derivations, syncategorematic and only occurring with other words.

3.4 Compounding

Amharic has compound verbs, nouns and adjectives. Compound verbs are created by combining the words *ale* 'said' or *aderege* 'did', with meaningless morphemes such as *qeT*: *qeT ale* 'he stood straight up', *qeT aderege* 'he made sth. straight'.

Compound nouns are formed by concatenating two nouns or a noun and an adjective with the linking vowel *–e–*: *bEtekrstiyan* 'church' = *bEt+e+krstiyan* = 'house+e+Christian'.

Compound adjectives are also formed by concatenating a noun and an adjective: *IgreqeCn* 'wanderer' = *Igr+e+qeCn* = 'leg+e+thin'.

Graphemic changes occur in word formation due to occurrence of vowels in sequence, and palatisation: *aa* → *a*, *ia* → *iya* and when a dental consonant is followed by the vowel *e* or *i* it changes to palatal *de* → *je*, *di* → *ji* or sometimes *di* → *j*.

4 The morphological analyser

The morphological analyser takes a string of morphemes as an input and gives an output of lexical forms, i.e. underlying morphemes and morphosyntactic categories.

Many basic procedures in natural language processing standardly employ FS techniques for implementation, including tokenisation, phonological and morphological analysis, shallow parsing, spelling correction and others; cf. (Karttunen, 2003). Morphological constructions can be described particularly efficiently with regular expressions; cf. (Beesley & Karttunen, 2003), (Kay, 1987), (Koskenniemi, 1984), and (Kiraz, 2000). Morphological analysis using finite state transducers (FSTs) is based on the assumption that the mapping of words to their analysis constitutes a regular relation, i.e. the underlying forms constitute a regular set, the surface forms constitute a regular set, and there is a (possibly many–to–many) regular relation between these sets. In languages whose morphotactics is morph concatenation only, FSTs are straightforward to apply. Handling non–concatenative (or partially concatenative) languages is more challenging; cf. especially (Kay, 1987), (Beesley & Karttunen, 2003), (Trost, 2003).

4.1 Formal properties of word forms

The basic morphological modelling convention for Amharic is that there is a small finite upper bound to root length (e.g. *sbr*) and to intercalated stems:

$root + vocalism + template = stem$

e.g. $sbr + ä + CVCC = säbr$

Words are constructed from stems by concatenation of prefixes and suffixes. The reversibility property of FSTs is useful: the 'generate' mode is used for generation, the 'accept' mode for analysis (cf. Figure 1).

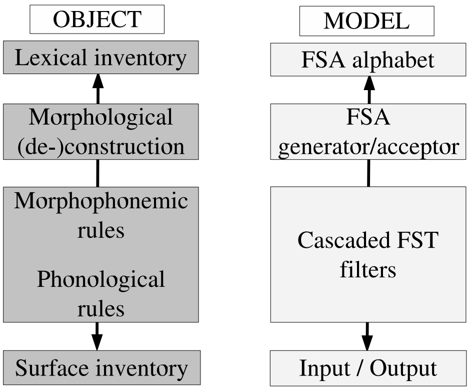


Figure 1: Modelling conventions for FSTs.

The absence of a lexicon of Amharic words in their base form is a major problem. About 1277 Amharic verb roots were compiled from (Bender & Fulas, 1978); other irregular verbs were gathered from (Dawkins, 1960). Deverbal nouns and adjectives were also obtained from these sources. Non–derived adjectives, adverbs, prepositions and conjunctions are few, and were manually collected. Simplex nouns are also hard to find. Lists of names were collected from the Bible, as well as place–names, kinship terms, body parts, local environmental terms and numbers (cardinal and ordinal), and implemented with the Xerox lexicon compiler (LEXC).

Semitic stem interdigitation has been treated several times; cf. (Kay, 1987), (Kataja & Koskenniemi, 1988), (Beesley & Karttunen, 2003). Kay designed a multitape FS technique for the interdigitation of roots, CV–templates and vocalisations in Arabic, and (Kataja & Koskenniemi, 1988) demonstrated interdigitation of Semitic roots (taking Ancient Akkadian as an example) using intersection over regular languages.

In (Beesley & Karttunen, 2003) a 'merge' operator for Arabic stems is described, a pattern filling algorithm which combines two regular languages, a CV template and fillers (root & vocalisation). The output of the merge operator is a regular expression that can be computed by the compile-replace algorithm of XFST. This algorithm works well for Amharic too. A more straightforward approach, however, would be to simply insert vocalisation between radicals. This requires accessing positions between consonant sequences. We used a novel bracketing 'diacritic' convention to locate vowel positions and right and left contexts to descriminate between different positions.

4.2 Internal changes

Derived verbs with internal changes involving penultimate consonant reduplication and vowel insertion are handled mostly by single replace rules. For example to generate *säbabär* from *säbär*, the rule used is:

$\{b\}(-->)\{bab\}jjä\_ä$ which results in säbabär, while retaining the original underived *säbär*.

4.3 Affix concatenation

The regular operation concatenation is used to concatenate affixes to the stem. When concatenating, illegal sequences of vowels are avoided by using replace rules and also impermissible affix combinations are controlled by introducing constraints:

$[P1][P2][P3][P4][P5][stem1jstem2j...]$

[[S1|S2|S3] [S4] [S5][S6|S7]] [S8]

where P1-P5 stand for prefix categories and S1-S8 are suffix categories that a verb stem can take. Prefixes, stems and suffixes have specific positions. In case of prefixes, all categories may occur together, but no more than one from each category. There are constraints on the suffixes: [S1|S2|S3] are alternatives and cannot exist together in one word. The same is true for [S6|S7]. Similar procedures of concatenation are applied for other POS as well.

4.4 Full stem reduplication

Reduplication of collective nouns is handled by using the self concatenation operation *word^2* which concatenates a word to itself with the compile-replace algorithm of (Beesley & Karttunen, 2003), and using a bracketing rule to find the mid position to insert the vowel.

A second method that also gives the same results is without using the compile-replace algorithm just with the self concatenation operator and a temporary file to deal with singleton elements in the lexicon at a time to avoid over production of unwanted results. This operation demands the use of a shell script outside the Finite State Tool we used (Xerox Finite State Tool-XFST).

4.5 Phonological processes

During affix concatenation, it is possible for vowels to occur in sequence that would result in a change of grapheme. To handle this problem simple replace rules are used. For example,

$\{aa\} - > \{a\}$, replaces the sequence aa by a.

$\{ae\} - > \{aye\}$ replaces the sequence ae by aye.

Finally, palatisation was handled by a replace rule that replaces dentals with palatals:

$\{di\}(->)\{pi\}$, maps di to pi and retains di

$\{di\} - > \{p\}$, maps the retained di to p

(the order of operation matters)

$\{de\} - > \{pe\}$ maps each de to pe

The transducers created for each class of verbs are finally merged by the union operation. This single transducer is then used whenever analysis of surfaces forms need to be made. The transducers for the different POS are not put together for evaluation purposes cf. Section 5.

5 Evaluation and conclusion

A preliminary evaluation of the system was made by analysing words from Amharic corpus (The Book of Matthew in the bible, Chapters 1–5). The evaluation hypothesis was that for each word class the words in it should be analysed correctly. A total number of 1620 words which contain words of all parts of speech were input into the transducers of each class. The results showed that among 468 verbs in the corpus 94% were analysed in total but taking the first 100 of analysed verbs 32% consisted also wrong analysis together with the correct ones. Among 650 nouns that exist in the corpus 85% were correctly analysed,with only a few about 7 that contain wrong analysis. For adjectives of 76 a recall of 88% with less than 1% wrong plus correct analysis was obtained. Other parts of speech were all correctly recognised. Since the input consisted of all classes of words, there were false positives. The precision levels in cases of nouns, adjectives and adverbs were 94%, 81%, and 91% respectively; while that of verbs was down to 54%. An attempt to improve the precision for verbs increased it to 65% but with an adverse effect on the recall. The low results in the precision of verb analysis are primarily a result of rules that are not inclusive for all members in a class. In addition, there is no standard spelling, creating flexibility in spelling the same words one way or another.

The results show that even without more contextual information for purposes of disambiguation, the basic recall result is already very useful. The next stage of development is to incorporate the output of the analyser into a syntax–aware tagging utility; we predict that this will increase the precision result drastically.

References

Nega Alemayu and Peter Willett. 2002. Stemming of Amharic Words for Information Retrieval. Literary and Linguistic computing 17 (1), p. 1–17.

Atelach Alemu, Lars Asker, and Mesfin Getachew. 2003. Natural Language Processing for Amharic: Overview and Suggestions for a Way Forward. In Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories, Växjö University, Sweden, November.

Getahun Amare. 1989. Amarenja Souasou Bek'elal Ak'erareb (Amharic Grammar Presented in an Easy Way). Addis Abbaba: Business Printing Press.

Kenneth R. Beesley and Lauri Karttunen. 2003. Finite State Morphology. Stanford: CSLI.

M. Lionel Bender. 1968. Amharic Verb Morphology: A Generative Approach. PhD. Dissertaion, Graduate School of Texas.

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

Girmay Berhane. 1992 Word Formation in Amharic. Journal of Ethiopian Languages and Literature. No. 2. p. 50–75

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

Sisay Fissaha and Johann Haller. 2003. Amharic Verb Lexicon in the Context of Machine Translation. TALN, p. 183–192.

Sisay Fissaha and Johann Haller. 2003. Application of Corpus-based Techniques to Amharic Texts. In Proceedings of the 10th Conference on Traitement Automatique des Langues Naturelles, volume 2, p. 173-182, Batz-sur-Mer, France, June.

Lauri Karttunen. 2003. Finite–State Technology. In: The Oxford Handbook of Computational linguistics, p. 339–357. Oxford University Press.

Laura Kataja and Kimmo Koskenniemi. 1988. Finite State Description of Semitic Morphology: a case study of Ancient Akkadian. Proceedings of 12th Conference on Computational Linguistics, I, p. 313–315.

Martin Kay. 1987. Nonconcatenative Finite–State Morphology. EACL 1987, p. 2–10.

George Anton Kiraz. 2000. Multitiered Nonlinear Morphology Using Multitape Finite Automation: A Case Study on Syriac and Arabic. Computational Linguistics, Volume 26 Issue 1, p. 178–181.

Kimmo Koskenniemi. 1984. A General Computational Model for Word–Form Recognition and Production. Proceedings of the 22nd Conference of the ACL, p. 178–181, California.

Habte Mariam Markos. 1991–1994. Towards the Identification of the Morphemic Components of the Conjugational Forms of Amharic. Proceedings of the Eleventh International Conference of Ethiopian Studies. Addis Ababa, vol. 1, p. 465–479.

Harald Trost. 2003. Morphology. In: The Oxford Handbook of Computational linguistics, p. 25–47. Oxford University Press.

Baye Yimam. 1999. Root Reductions and Extensions in Amharic. Ethiopian Journal of Languages and Literature, No 9, p. 56–88.

Baye Yimam. 1995. yamargnasewasew (Amharic Grammar). Addis Ababa: EMPDA.