# An Amharic Stemmer : Reducing Words to their Citation Forms

**Atelach Alemu Argaw**
Department of Computer and Systems Sciences
Stockholm University/KTH, Sweden
atelach@dsv.su.se

**Lars Asker**
Department of Computer and Systems Sciences
Stockholm University/KTH, Sweden
asker@dsv.su.se

## Abstract

Stemming is an important analysis step in a number of areas such as natural language processing (NLP), information retrieval (IR), machine translation(MT) and text classification. In this paper we present the development of a stemmer for Amharic that reduces words to their citation forms. Amharic is a Semitic language with rich and complex morphology. The application of such a stemmer is in dictionary based cross language IR, where there is a need in the translation step, to look up terms in a machine readable dictionary (MRD). We apply a rule based approach supplemented by occurrence statistics of words in a MRD and in a 3.1M words news corpus. The main purpose of the statistical supplements is to resolve ambiguity between alternative segmentations. The stemmer is evaluated on Amharic text from two domains, news articles and a classic fiction text. It is shown to have an accuracy of 60% for the old fashioned fiction text and 75% for the news articles.

## 1 Introduction

Stemming is the process of reducing morphological variants of a word into a common form. For morphologically less complex languages like English or Swedish, this usually involves removal of suffixes. For languages like Amharic or Arabic, that have a much richer morphology, this process also involves dealing with prefixes, infixes and derivatives in addition to the suffixes. Stemming is widely used in IR, with the assumption that morphological variants represent similar meaning. It is applied during indexing and is used to reduce the vocabulary size, and it is used during query processing in order to ensure similar representation as that of the document collection. In cross language information retrieval (CLIR) where a query is typically posed in one language, and the document collection from where the documents are retrieved is in another language, some form of translation is required. For low resource languages such as Amharic, machine readable dictionaries (MRDs) play a crucial role by enabling look up of translations of query terms. In most cases such MRDs have all entries represented only by their citation form. Thus in CLIR applications, it is of outmost importance that query terms in the source language are reduced to the exact corresponding citation form as presented in the MRD. In this paper we address this particular problem of stemming Amharic words and reducing them to their citation forms for CLIR applications.

The remainder of the paper is organized as follows. Section 2 provides a background information about the Amharic language, followed by related work in Section 3 and a brief description of Amharic morphology in Section 4. Section 5 presents the resources utilized, while Section 6 deals with a detailed description of the stemmer. In section 7 we describe experiments conducted to evaluate the performance of the stemmer and discuss the obtained results. We give concluding remarks in Section 8.



*Proceedings of the 5th Workshop on Important Unresolved Matters*, pages 104–110,
Prague, Czech Republic, June 2007. 

## 2 The Amharic Language

Amharic is the official working language of the federal government of the Federal Democratic Republic of Ethiopia and is estimated to be spoken by well over 20 million people as a first or second language. Amharic is the second most spoken Semitic language in the world (after Arabic). It is today probably the second largest language in Ethiopia (after Oromo, a Cushitic language) and possibly one of the five largest languages on the African continent. Following the Constitution drafted in 1993, Ethiopia is divided into nine fairly independent regions, each with it's own nationality language. However, Amharic is the language for country-wide communication and was also for a long period the principal literal language and medium of instruction in primary and secondary schools of the country, while higher education is carried out in English. Despite it's wide speaker population, computational linguistic resources for Amharic, as most 'low resource' languages, are very limitted and almost non existent.

Written Amharic uses a unique script which has originated from the Ge'ez alphabet (the liturgical language of the Ethiopian Orthodox Church). Written Ge'ez can be traced back to at least the 4th century A.D. The first versions of the language included consonants only, while the characters in later versions represent consonant-vowel (CV) phoneme pairs. In the modern Ethiopic script each syllable pattern comes in seven different forms (called orders), reflecting the seven vowel sounds. The first order is the basic form; the other orders are derived from it by more or less regular modifications indicating the different vowels. There are 33 basic forms, giving 7*33 syllable patterns (syllographs), or fidels. Two of the base forms represent vowels in isolation, but the rest are for consonants (or semi-vowels classed as consonants) and thus correspond to CV pairs, with the first order being the base symbol with no explicit vowel indicator. The writing system also includes four (incomplete, five-character) orders of labialised velars and 24 additional labialised consonants. In total, there are 275 fidels, but not all the letters of the Amharic script are strictly necessary for the pronunciation patterns of the spoken language; some were simply inherited from Ge'ez without having any semantic or phonetic distinction in modern Amharic. There are many cases where numerous symbols are used to denote a single phoneme, as well as words that have extremely different orthographic form and slightly distinct phonetics, but with the same meaning. So are, for example, most labialised consonants basically redundant, and there are actually only 39 context-independent phonemes (monophones): of the 275 symbols of the script, only about 233 remain if the redundant ones are removed. The script also has a unique set of punctuation marks and digits. Unlike Arabic or Hebrew, the language is written from left to right.

The Amharic writing system uses multitudes of ways to denote compound words and there is no agreed upon spelling standard for compounds. As a result of this - and of the size of the country leading to vast dialectal dispersion - lexical variation and homophony is very common.

## 3 Related work

Pioneering the work on morphological analysis of Amharic verbs, Abiyot (Bayou, 2000) designed and implemented a prototype word parser for Amharic verbs and their derivation. He designed a knowledge-based system that parses verbs, and nouns derived from verbs. He used root pattern and affixes to determine the lexical and inflectional category of the words. He tested his system on a limited number of words (200 verbs and 200 nouns) and the result showed that 86% of the verbs and 84% of the nouns were recognized correctly. Another prototype morphological analyzer for Amharic was developed by Tesfaye Bayu (Bayu, 2002) where he used an unsupervised learning approach based on probabilistic models to extract morphemic components (prefix, stem and suffix) to construct a morphological dictionary. He also investigated an approach whereby he applied the principle of Auto segmental Phonology to identify morphemic component of a stem such as consonantal root, vocalic melodies and CV-templates. The first system was able to parse successfully 87% of words of the test data (433 of 500 words). This result corresponds to a precision of 95% and a recall of 90%. Tested with 255 stems, the second system identified the morphemic componentes of 241 (or 94% of the) stems correctly.

Fissaha and Haller (Fissaha and Haller, 2003) discuss the morphology of Amharic verbs in the context of Machine Translation and present an implemetation of a morphological analyser for Amharic using Xerox Finite State Tools (XFST). The different classification shemes for Amharic verbs that have been forwarded are discussed followed by the implication such classifications have on the implementation strategy. They claim that morphological analysis for Amharic with XFST can handle most of the morphologcal phenomena except some derivation processes which involve simultaneous application of both stem interdigitation and reduplication. Saba and Gibbon (Amsalu and Gibbon, 2005) extend the XFST implementation of Amharic morpholgy to include all word categories. Testing with 1620 words text from an Amharic bible, they report recall levels of 94% for verbs, 85% for nouns, and 88% for adjectives while they report precisions of 94% for nouns, 81% for adjectives, 91% for adverbs, and 54% for verbs, at the above specified recall levels.

A more recent work that applies Conditional Random Fields to segment and part of speech tag Amharic words is done by Fissaha (Adafre, 2005). He reports an accuracy of 84% for the word segmentation. The work deals with bound morphemes of prepositions, conjunctions, relative markers, auxiliary verbs, nagation marker and coordinate conjunction, but leaves out other bound morphemes such as definite article, agreement features such as gender and number, case markers, etc, and considers them to be part of the word. The best result (84%) is obtained by using character, morphological and lexical features.

There has been a work done by Alemayehu and Willet (Alemayehu and Willett, 2002) which investigates the effectiveness of stemming in information retrieval for Amharic. They compare performance of word-based, stem-based, and root-based retrieval of 40 Amharic queries against 548 Amharic documents, and show better recall levels for stem and root based retrieval over word based, but they don't provide information on the precision of these experiments.

All the above mentioned works attempt to address the need to develop a morphological analyser for Amharic, and show that there has been a great deal of effort put in the design and implementation of each system. Although that is the case, none of them are publicly available, and/or are limitted in some way. For our current task of stemming for the purpose of CLIR dictionary lookup, full fledged morphological analysis is most likely an overkill since we only need citation forms of words, and precision plays a very important role.

## 4 Amharic Morphology

Amharic has a rich verb morphology which is based on triconsonantal roots with vowel variants describing modifications to, or supplementary detail and variants of the root form. A significantly large part of the vocabulary consists of verbs, which exhibit different morphosyntactic properties based on the arrangment of the consonant-vowel patterns. For example, the root *sbr*, meaning 'to break' can have the perfect form *säbbär* with the pattern CVC-CVC[1], imperfect form *säbr* with the pattern CVCC, gerund form *säbr* with the pattern CVCC, imperative form *sbär* with the pattern CCVC, causative form *assäbbär* with the pattern *as*-CVCCVC, passive form *täsäbbär* with the pattern *tä*-CVCCVC, etc. Subject, gender, number, etc are also indicated as bound morphemes on the verb, as well as objects and possesion markers, mood and tense, beneficative, malfactive, transitive, dative, negative, etc, producing a complex verb morphology.

Amharic nouns can be inflected for gender, number, definiteness, and case, although gender is usually neutral. Adjectives behave in the same way as nouns, taking similar inflections, while prepositions are mostly bound morphemes prefixed to nouns. The definte article in Amharic is also a bound morpheme, and attaches to the end of a noun. We have given a very brief description of some aspects of Amharic morphology, detailed information can be found in (Bender, 1968), (Bender and Fulas, 1978), (Yimam, 1995).

We have constructed 65 rules based on the entire Amharic morphology for the purpose of this study. The rules vary from simple affixation rules to each word category to allowed combinations of prefixes and suffixes for each word category and set of affixes.

[1] C stands for consonants and V for vowels

## 5 Resources

### 5.1 The Corpora

We have utilized three different sources of text for the development of the stemmer and the experiments. The first is a collection of news articles from an online news repository, Ethiopian News Headlines (ENH), which is available at http://www.ethiozena.net. This corpus consists of 3.1 million words of Amharic news text in a little more than 10,000 articles. This corpus was used to collect word frequency and prefix and suffix statistics i.e. the number of times an affix occurs attached to a known stem, and the occurence statistics was used to disambiguate between alternative segmantations of a given word. The second text source is another Ethiopian news agency, Walta Information Center (WIC) which can be found at http://www.waltainfo.com. We used news items downloaded from WIC to evaluate the stemmer on independent news texts from another source. The third text, which was also used for evaluation, is from the Amharic novel "Fikir Iske Meqabir" (FIM) by the renowned Ethiopian author Dr. Hadis Alemayehu. This text (FIM) was selected for the evaluation in order to see how well the stemmer would perform on a text that differed substantially in style from the news collection.

### 5.2 The Dictionaries

The simplest and most straight forward way for the stemmer to verify that a suggested segmentation is correct is to try to look up the stem in a dictionary. For this purpose we used three different dictionaries, an Amharic - English, an Amharic - French, and an Amharic - Amharic dictionary. The Amharic - English dictionary, by Dr. Amsalu Aklilu, contains 15 000 Amharic words with their English translations (Aklilu, 1981). The Amharic - French dictionary (Abebe, 2004) has 12 000 Amharic entries while the Amharic - Amharic dictionary by Kesatie Birhan has 56 000 entries (Tesema, ). All three dictionaries were made available to us in electronic form, transliterated to SERA and then merged and represented in a form suitable for the stemmer.

### 5.3 Transliteration

The dictionaries and all Amharic news texts mentioned above are published using Ethiopic script and using a variety of fonts, some of which are not Unicode compliant. In order to simplify the analysis and to have a unified representation of the texts, we transliterated all Amharic texts into SERA which is a system for ASCII representation of Ethiopic characters (Firdyiwek and Yacob, 1997).

The transliteration was done using a file conversion utility called `g2` which was made available to us by Daniel Yacob of the Ge'ez Frontier Foundation (http://www.ethiopic.org/).

## 6 The Stemmer

The stemmer first creates a list consisting of all possible segmentations of the word that is to be stemmed. In a second step, each such segmentation is then verified by matching each candidate stem against the machine readable dictionary. If no stem matches the dictionary, the stemmer will modfy the stem and redo the matching. If more than one stem matches, the most likely stem will be selected after disambiguating between the candidate stems based on statistical and other properties of the stems. In the cases when exactly one stem matches the dictionary then that segmentation will be presented as the output from the stemmer.

### 6.1 Segmentation

For each new word the stemmer first creates a list of possible segmentations by applying a list of morphological rules for allowed prefixes and suffixes. In this way, the word `Indeminorewna` would for example be segmented into the following 9 different ways:

```
(1)   Indeminorewna
(2)   Indeminorew -na
(3)   Indeminore -w -na
(4)   Inde- minorewna
(5)   Inde- minorew -na
(6)   Inde- minore -w -na
(7)   Inde- mi- norewna
(8)   Inde- mi- norew -na
(9)   Inde- mi- nore -w -na
```

For each of the 9 possible segmentations, the remaining stem is then matched against the (merged) three dictionaries. In this case, the only one that is found as entry in the dictionary is `nore`, so alternative 9 is selected as the most likely segmentation of the word.

### 6.2 Disambiguation

If more than one of the candidate stems are matched in the dictionary, those segmentations that have a stem that matches an entry in the dictionary are ranked according to length and frequency of the stem. The longest stem that have a match in the dictionary is selected and if more than one stem of equal length matches the dictionary then the stem that is more frequent is preferred before the less frequent. The frequency score is based on how often the stem occurs in the ENH corpus described above. The word `beteyazew` would for example be segmented in the following ways:

```
(1)  beteyazew
(2)  beteyaze -w
(3)  beteyaz -e -w
(4)  be- teyazew
(5)  be- teyaze -w
(6)  be- teyaz -e -w
(7)  be- te- yazew
(8)  be- te- yaze -w
(9)  be- te- yaz -e -w
```

In this case the three stems `teyaze` (5), `yaze` (8) and `yaz` (9) all have matching entries in the dictionary but `teyaze` is selected as the most likely stem since it is the longest.

### 6.3 Modification

For approximately 30% of the words, the stem does not match the dictionary. In these cases, the stem will be slightly modifed and a second attempt to match the entries in the dictionary will be done. For example the word `IndegeleSut` should correctly be segmented into `Inde- geleSe -u -t`. With the approach described so far, the segmentation based on prefixes and suffixes would yield the stem `geleS` which will not have a match in the dictionary. Instead, for the dictionary lookup to succeed, we first need to add the vowel `e` at the end of the stem. For the word `astawqWal` which should correctly segment into `astaweqe -W -al` we will first have to insert `e` both between `w` and `q` and again after `q` to reach the correct form of the stem. This process of modifying the stem by adding vowels, is applied to the candidate stems if no matches by the unmodifed stems are made in the dictionary. For the current implementation of the stemmer, this is done by inserting one of the vowels 'e' or 'a' between the consonants if the unmatched stem contains two consecutive consonants, or after the last consonant if the stem ends in a consonant. If exactly one of the modifed stems will match the dictionary, then that segmentation will be ranked as the most likely. If more than one modifed stem matches, then the longest will be selected. For the words where this modification of the stem is done, approximately 30% will successfully match their correct entry in the dicionary while 20% make an incorrect match and the remaining 50% will not match the dictionary at all.

### 6.4 Out-of-dictionary terms

Finally, the approximately 15% of the words that do not have any stem that matches entries in the dictionary (even after the modifiaction) will be ranked according to the length of the stem and the number of times that the stem occurs in the ENH corpus. In this case, it is the shorter stems that are preferred. For example the word `bekomixnu` will have four possible segmentations, none of which occurs in the dictionary.

```
(1) bekomixnu
(2) bekomixn -u
(3) be- komixnu
(4) be- komixn -u
```

In this case, alternative 4, `komixn` is the shortest stem that occurs as a unique word in the reference corpus and is therefor selected as the most likely segmentation before either one of the alternative stems `bekomixnu`, `bekomixn` or `komixnu`.

## 7 Experimental Evaluation

In order to evaluate the performance of the stemmer, we selected the first 1503 words (= 1000 unique words) from the WIC corpus described above. We also selected a 470 words long text from the book ”Fikir Iske Meqabir” to get a text with 300 unique words.

On the WIC data the stemmer had an overall accuray of 76.9 %. For 48 % of the words, the stemmer found exactly one segmentation with a stem that was matching the dictionary, and for these words it had an accuracy of 83.75 %. For 36.3 % of the words, the stemmer found more than one segmentation that matched the dictionary and therefor needed to do additional disambiguation between alternative segmentations. For these words, the stemmer had an accuracy of 69.1 %. For the remaining 15.7 % of the words, the stemmer found no match in the dictionary for any of the possible segmentations. For these words the stemmer had an accuracy of 73.9 %. In the cases when there is only one match in the dictionary, the extra sources for error that are introduced by having to disambiguate between alternative segmentations are avoided and hence the stemmer has best accuracy for those words that have exactly one segmentation with a stem that will match the dictionary.

For the 300 unique words from Fikir Iske Meqabir, the stemmer had an overall accuracy of 60.0 % In a similar fashion as for the WIC data, the stemmer performed best on the subset of words for which there was exacly one match in the dictionary. For this group the performance was 68.8 % correct but the overall accuracy was lowered by the fact that the stemmer performed worse on the words that had either more than one match, or no match at all in the dictionary. These numbers were 54.8 % and 42.1 % respectively.

## 8 Conclusion

We have presented the design and development of an Amharic stemmer which reduces words to their citation forms for the purpose of dictionary lookup in CLIR. Given the resource constraints we have, and the specificity of the stemmer, the overall performance could be acceptable, but needs further improvment. The stemming depends highly on word entries in the three MRDs for verification purposes. These MRDs altogether consist of a limitted amount of entries, overall 83000, with a very high level of overlap, leaving 47176 unique entries. Although it is not the largest source of error, it accounts for around 15% of the words segmentation decided on corpus statistics only since they are not found in the dictionaries. We intend to use more dictionaries with the assumption that there will be a performance increase with the increasing number of citation forms to refer to. On the other hand, increasing the amount of citation forms also will increase the percentage of words that will have more than one match in the dictionaries. That would lead us to focus on the disambiguation strategy in the future. So long as the morphological rule exists, we are able to get the correct segmentation for a word in a possible segmentations list. And when we have two or more likely segmentations that are picked out since they have matching stems in dictionaries, we need to design a smarter way of disambiguation that would take into account contextual information and part of speech tags, etc, in addition to the currently used occurence frequency approach.

Although conducting a full fledged morphological analyser for Amharic is beyond the scope of this paper, we would like to note that there is a need to create a forum for collaboration and exchange among researchers involved in developing NLP resources for Amharic and other Semitic languages and organize the considerable effort that is being made individually. We also hope that some of the ideas and procedures that are described in this paper could be more generally applicable to other Semitic languages as well.

## Acknowledgements

The copyright to the two volumes of the French-Amharic and Amharic-French dictionary (”Dictionnaire Francais-Amharique” and ”Dictionnaire Amharique-Francais”) by Dr Berhanou Abebe and Eloi Fiquet is owned by the French Ministry of Foreign Affairs. We would like to thank the authors and the French embassy in Addis Ababa for allowing us to use the dictionary in this research.

The content of the “English - Amharic Dictionary” is the intellectual property of Dr Amsalu

Aklilu. We would like to thank Dr Amsalu as well as Daniel Yacob of the Geez frontier foundation for making it possible for us to use the dictionary and other resources in this work.

We would also like to thank Ato Negash of Walta Information Center for allowing us to use part of their news texts in this research.

## References

Berhanou Abebe. 2004. *Dictionnaire Amharique-Francais*. Shama Books, Addis Ababa, Ethiopia.

Sisay Fissaha Adafre. 2005. Part of speech tagging for amharic using conditional random fields. In *Proceedings of ACL-2005 Workshop on Computational Approaches to Semitic Languages*.

Amsalu Aklilu. 1981. *Amharic - English Dictionary*. Mega Publishing Enterprise, Ethiopia.

Nega Alemayehu and Peter Willett. 2002. The effectiveness of stemming for information retrieval in amharic. In *Short Communication*.

Saba Amsalu and Dafydd Gibbon. 2005. Finite state morphology of amharic. In *Proceedings of RANLP*.

Abiyot Bayou. 2000. Design and development of word parser for amharic language. Masterthesis, Addis Abeba Univeristy.

Tesfaye Bayu. 2002. Automatic morphological analyser: An experiment using unsupervised and autosegmental approach. Masterthesis, Addis Ababa University.

M. Lionel Bender and Hailu Fulas. 1978. Amharic verb morphology. In *East Lansing: Michigan State University, African Studies Center*.

M. Lionel Bender. 1968. *Amharic Verb Morphology: A Generative Approach*. Ph.D. thesis, Graduate School of Texas.

Yitna Firdyiwek and Daniel Yacob. 1997. System for ethiopic representation in ascii.

Sisay Fissaha and Johann Haller. 2003. Amharic verb lexicon in the context of machine translation. In *Actes de la 10e conference TALN, Batz-sur-Mer*.

Kesatie Birhan Tesema. *YeAmarinja Mezgebe Qalat*. Adis Abeba.

Baye Yimam. 1995. *ye amargna sewasew (Amharic Grammar)*. EMPDA.