TALN 2005
Developments Towards an Electronic Amharic Corpus

Daniel Yacob
Ge'ez Frontier Foundation
7802 Solomon Seal Dr, Springfield, VA 22152, USA
yacob@geez.org

Abstract

The state of Amharic natural language processing was aptly assessed at TALN 2003 by Atelach, Asker and Mesfin. A public Amharic corpus and a comprehensive lexicon were two of the most needed items in absence for Amharic language researchers. Since the 2003 assessment some progress has been made in these two areas and researchers have begun informal collaboration to address the common goal of developing these public resources. In this same period Ethiopia's legal system has changed to cloud the issue over what the legal status of an Amharic corpus would be. While a promising start is underway, corpus developers and researchers alike will have to familiarize themselves with the new legislature in Ethiopia and reexamine the status of their holding to avoid potential unintended violations.

1 Introduction

Amharic is the most studied and best understood language of Ethiopia, it also serves as the country's lingua franca. Researchers today, both inside and outside of Ethiopia, are increasingly interested in computational investigations of the Amharic language. The lack of a freely available electronic corpora, lexicon, and transcription standard, coupled with the complexities of Amharic orthography are a significant barrier to would be researchers.

Amharic, along with its ten sibling Ethio-Semitic languages found in Ethiopia and neighboring Eritrea, is written in the Ethiopic syllabary. Amharic has been a written language for roughly 600 years and has as rich legacy of both typeset and calligraphic literature. Significant amounts of electronic corpora in Amharic, however, did not exist prior to the 1990s. The rise of desktop and internet publishing over the last decade has helped accumulate a body of data but of limited value. This paper will review the available materials, their usefulness, complications in working with Amharic orthography, and the new legal ramifications that face corpus development in the research community.

2 Electronic Corpus

Prior to the boom in desktop publishing in the late 1980s electronic text in Amharic existed as the product of experimental explorations into computer environs and was unappreciable in quantity. As personal computers came down in price in the late

1980s and early 1990s and word processing software became more practical and extensible the stage was set for Amharic publishing to get underway in a significant manner. However, due to the political status at that time the vast majority of Amharic desktop publishing would occur outside of Ethiopia amongst the Ethiopian Diaspora.

The largest bodies of work developed would then be in periodical literature of the Diaspora which were news publications and largely of political content. Very little of this content likely survives in the present day and may no longer be accessible under modern computer systems. Ethiopic script lacked a standard for representing the letters of the syllabary electronically until Amendment 10 to the ISO-10646 standard in 1997 which later becoming a part of a formal Unicode standard in 2000 when major software vendors would begin to support Ethiopic script. In this time more than seventy computer encoding systems were devised for Ethiopic script, none compatible with the other, and supported presently by only a handful of surviving software companies. Unlike the well established western languages, having an electronic document in Amharic does not equate readily to being able to read the document.

Following the change of governments in Ethiopia in 1991, the liberalization of press laws and downward cost of personal computer becoming affordable to Ethiopian businesses, the stage was then set by the mid 1990s for the sustained generation of electronic materials. Periodical publications such as weekly newspapers and magazines would lead this production of electronic content. Unfortunately hard disk capacities were still limited and there was little to no appreciation of the electronic record. A newspaper team producing a publication one week would delete the files on the following week without giving it a second thought.

Book authors and many magazine producers did not publish their own materials. Rather, they would take handwritten manuscripts to a publishing house that would typeset the text electronically and produce the publication for a fee. The electronic version was generally not preserved for the future except in the case of books. Authors would not automatically be given an electronic copy however since the typing and typesetting was considered a service with the cost absorbed by the publishing house. An electronic copy could, reluctantly, be made available to the author for a significant larger fee.

Perception would begin to change with the arrival of the internet and exposure to online newspapers. Electronic information exchange was nearly non-existent in Ethiopia prior to email service and given the encoding difficulties there was little expectation that a document written on one computer should be readable on another.

Eight months following the arrival of public internet service to Ethiopia the Ethiopian News Headlines (ENH) service was launched that featured a selection of articles from the capital city's newspapers. These articles were retyped from the newspapers and published in the Unicode character set amongst others. To date the news service has generated over 10,000 articles from more than 110 newspapers that are available in its archives and in zipped archived bundles for each month and year that may be downloaded freely and used under "fair use" rules governing literature. This corpus served as the basis of the collocation extraction research of Sisay and Haller in 2003 as well as the speech recognition research of Solomon Abate among others.

While the ENH offers a large collection of newspaper articles, the typical article is under 500 words in length, is largely of political content, and as retyped content is subject to having more typographic errors than the original document.

Other online publications in Amharic with large numbers of documents include the government news organ the Ethiopian News Agency (ENA) and the private, yet government affiliated, Walta Information Service. The ENA is noted for offering good quality Amharic language but has not yet begun publishing in Unicode leaving researchers to find a way to convert the ENA materials into a usable form. Walta on the other hand has begun partial publishing in Unicode since the middle of 2003 and has nearly made a full transition to Unicode (archives remain in a legacy encoding). The archives of each are available on a per-article basis. Like the ENH articles they are relatively short but the content characteristic of these services is broader, offering more than political subjects.

Research into Amharic machine translation is underway in 2004 at Stockholm University under the direction of Dr. Lars Asker where work on a parallel lexicon is being developed. Parallel corpus is even more limited for Amharic and again relies heavily on news services, in this case the bilingual Walta. Translation of Open Source software made in recent years offers a significant amount of translated phrases (over 40,000 phrases). However, the phrases are usually of 3 words or less, perhaps 10% representing one more sentences, having 20% redundancy, are of technical content, and reflecting the mixed quality of untrained volunteer translators.

A small collection of books have been typed up and made available to the research community. Under arrangement with the authors, the researcher must first sign a contract assuring that he or she will use the content for research purposes only and not commercially. This library made available by the Ge'ez Frontier Foundation offers much lengthier content under broader subject matter. The typographic correctness of the materials remains uncertain at this time.

3 Lexicon & Spelling

Atelach, et.al. (2003) identified the need for an Amharic lexicon and spelling checker as essential to the research community. The Ge'ez Frontier Foundation has been working actively in 2004 and continuing into 2005 to provide a basic word list. The lexicons given in the dictionaries of Amsalu Aklilu (1979 EC), Desta Tekle Wold (1962 EC) and Tesema Habte (1951 EC) are being extracted to form the basis for a comprehensive public lexicon. A rough version of the Amsalu lexicon comes with the Aspell version 0.60.2 open source spelling checker and has some initial tagging for affix rules.

A more refined version of the Amsalu lexicon will be available in the coming months and affix tagging will be an ongoing effort for some years to come. Amharic is a highly inflected language where the affix rules are largely governed by the presence of a midfix and many 10s of thousands of derived forms of some nouns and verbs become possible. The tagging of the lexicon as per their derivational classes will be essential to detecting proper word formation.

The comprehensive lexicon effort should produce its first unified lexicon in 2005. The next stage in refining the lexicon will be to resolve differences in spelling

that may emerge. The three dictionaries that form the basis of the lexicon are widely considered to be of the highest quality so discrepancies are expected to be minimal.

A complexity that enters into Amharic spelling are the presence of Ge'ez loan words and words derived from a Ge'ez root. Ge'ez is the ancient language of Ethiopia that is analogous in the role that Latin played for the Romance language of Europe. Ge'ez had a richer phonemic inventory and required additional letters for its orthography. In Amharic orthography these additional letters from Ge'ez would take on the phonemic value of its nearest neighbor. The result being two syllabic series for 's' ('ሰ' and 'ሠ'), two series for 'ts' ('ጸ' and 'ፀ'), two for 'a' ('አ' and 'ዐ') and 4 for 'h' ('ሀ', 'ሐ', 'ኀ' and 'ኸ'). This redundancy in Amharic becomes a source of confusion and the letters are treated as interchangeable by the lay person. Common Amharic spelling then becomes highly flexible and "correctness" is not a matter of precision but one of acceptable proximity. For example the fourth month of the year, canonically "ታኅሣሥ" ("Tahsas") may have any of the logical, phonetically equivalent, forms:

ታኅሣሥ	ታሕሣሥ	ታህሣሥ	ታኽሣሥ
ታኅሣስ	ታሕሣስ	ታህሣስ	ታኽሣስ
ታኅሳሥ	ታሕሳሥ	ታህሳሥ	ታኽሳሥ
ታኅሳስ	ታሕሳስ	ታህሳስ	ታኽሳስ

Table 1: Logical Amharic Renderings of the month "Tahsas".

While logical under the rules of the syllabary, the final column however is not also probable, leaving us with only 12 renderings likely to be found. Equating of the various spellings in text processing has been accomplished through the device of equivalence classes for the Amharic syllabary whereby:

[=ሀ=] ≡ {ሀ,ሐ,ኀ,ኸ} *(all 'h' syllables)*
[=ሳ=] ≡ {ሣ,ሳ} *(all "sa" syllables)*
[=ስ=] ≡ {ሥ,ስ} *(all 's' syllables)*

The equivalence classes may then be applied to form the regular expression for the possible renderings in the expression string "ታ[=ሀ=][=ሳ=][=ስ=]". A metaphone matching approach has also been devised for Amharic where all renderings simplify into the single string "ትህስስ" ("thss"). The Perl language package, Text::Metaphone::Amharic in fact applies the Amharic character classes in its metaphone implementation also via the Perl package Regexp::Ethiopic::Amharic.

The Regexp::Ethiopic package contains classes for a few other Ethiopian languages using Ethiopic script. The same month in Tigrinya would render canonically as "ታሕሣሥ" (taħsas) but could not be matched by the Regular expression derived for Amharic as 'ሕ' ('ħ') would no longer be a member of the [=ሀ=] equivalence set (likewise 'ኸ' which becomes 'x' in Tigrinya). The correct regular expression for valid Tiginya renderings is then "ታሕ[=ሳ=][=ስ=]" matching the 2nd column in the above table and the metaphone key becomes "ትሕስስ".

Without exception the Ethio-Semitic languages of Ethiopia and Eritrea use the Ethiopic script, many but not all of the members of the other language families (Cushitic, Omotic and Nilo-Saharan) will also use Ethiopic script. Unlike the Ethio-Semitic languages that have literal histories of some length, the primary issue that has prevented recent corpora development for the less populous languages has been the lack of character encoding support for their written elements. Only since Unicode 4.1.0 (March 31. 2005) has there been a standard supporting the written syllables of Bench, Blin, Me'en, Mursi, Sebatbeit, Suri and Xamtanga.

Prior to the last change of governments the early 1990s, it was legally permissible by the central government for publications to be in only one of Amharic, Tigrinya or Afan Oromo. Needless to say, significant corpora in the remaining 75 or so languages are not to be found. Since then some language communities have elected to adapt Latin script as the basis for their orthography. Spelling problems encountered by such communities are primarily related to the representation of long and short vowels and geminated consonants, none of which could have been represented in Ethiopic. Hence "Adooleessa", the seventh month of the year in Afan Oromo, is likely to be rendered with any number of doubled letters, e.g. "Adolleesaa", "Addoolessaa", etc. Prevalence of a multitude of renderings here is due in large part to user confusion over where lengthening is needed and the inherited acceptance of lax spelling practices from Amharic. By migrating to Latin script however these languages may enjoy the benefit of the vast computational resources and methodologies developed for western languages.

The resources discussed for Ethiopic script are effective for coping with the complexities of modern orthography for tasks such as pattern matching applied in stages of text retrieval and spelling correction. Spelling correction very much depends upon a good quality lexicon. Amharic does not have an authoritative reference for spelling nor a recognized authority responsible for defining Amharic rules and vocabulary. The closest such authority would either be the Ethiopian Orthodox Church or The Amharic Language School at Addis Ababa University.

The combined lexicons of the three authors mentioned will provide a strong basis for a spelling checker as a database for canonical spellings. In a number of cases it will likely be necessary to allow for two and three acceptable renderings of a word. The faculty of the Amharic Language School at Addis Ababa University will be enlisted to refine and add to the unified lexicon as well as decide which amongst alternative spellings can be deemed acceptable.

4 Ethiopian Intellectual Property Laws

Until very recently Ethiopia was without copyright or intellectual property laws. Foreign and domestic works could be republished with impunity. Researchers could enjoy the luxury of using textual materials without any concern for lawsuits but will now have to reexamine their holdings.

Wishing to join the World Trade Organization (WTO) Ethiopia has begun bosltering its legal system to protect intellectual properties. In 2004 Ethiopia passed a highly progressive copyright proclamation which covers a wide range of media from books, pamphlets and speeches to music, photographs, software and databases. Copyright is protected for the life of the author plus fifty years. The copyright

directorate is under the Ethiopian Intellectual Property Office established in 2003 (Wondwossen 2004). Ethiopia has become a member of the World Intellectual Property Organization (WIPO) in 1998.

The new laws are not yet well understood by the public nor the judicial system and may be difficult to apply as both go through a learning period. Lawsuits have already been brought against suspected violators by copyright holders and in some cases dismissed when the suite was initiated out of context. When the copyright law applies will, in a practical matter, be learnt by the society through judicial trial and error.

5 Status & Conclusion

The informal collaboration of researchers toward the development of an Amharic corpus has advanced passed the recognition of the problem and participants are presently reviewing applicable standards, tools, and related initiatives to launch a formalized effort. Oxford University has generously come forward and offered to serve as the corpus repository under its Open Archives Initiative which requires that the Text Encoding Initiative guidelines be followed. The home for the Amharic corpus will most likely be at Oxford but the requirements are still being studied. The formal phase of the Amharic corpus initiative is expected to get underway in early June following the resolution of these matters.

Internet newspaper archives, the few available books to researchers, and software translations provide a sizable corpus of Amharic text in electronic form but do not yet represent a balanced corpus as prescribed in Atelach, et.al. (2003). The unified lexicon in development while a promising start, represents no more than a collection of raw materials and may not be of a widely usable quality for several years. The critical mass of resources required for natural language processing of Amharic to "take off" is likewise some years away though we can now say that important steps are underway. While the resource collection is building, corpus providers and Amharic researchers are advised to familiarize themselves on Ethiopia's new copyright and intellectual property laws avoid the unintended missteps.

References

Amsalu Aklilu (1979 EC), **አማርኛ-እንግሊዝኛ መዝገበ ቃላት** *Amharic-English Dictionary*, Kuraz Publishing Agency.

Atelach Alemu, Lars Asker, and Mesfin Getachew (2003). *Natural Language Processing for Amharic: Overview and Suggestions for a Way Forward*, In Proceedings of TALN 2003 Workshop on Natural Language Processing of Minority Languages and Small Languages, Batz-sur-Mer, France, June, 2003.

Atelach Alemu, Lars Asker, and Gunnar Eriksson (2004). *Building an Amharic Lexicon from Parallel Texts*, In Proceedings of First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, a Workshop at LREC2004.

Atkisonson, Kevin (2004), *GNU Aspell,* http://aspell.net/, GNU, v0.60.2, December 27, 2004.

Daniel Yacob (2004a), Regexp::Ethiopic, http://search.cpan.org/~dyacob/Regexp-Ethiopic/, CPAN, Version 0.14.

Daniel Yacob (2004b), Text::Metaphone::Amharic, http://search.cpan.org/~dyacob/Text-Metaphone-Amharic/, CPAN, v0.11.

Desta Tekle Wold (1962 EC), **አዲስ ያማርኛ መዝገበ ቃላት** *Addis Yamarña Mäzgba Qalat*, Artistic Printers, Addis Ababa.

Ethiopian News Headlines (1989-1997 EC), Newspaper Archives, ftp://archives.news.com.et/pub/ENH/, Addis Ababa.

Sisay Fissaha and Johann Haller (2003). *Application of Corpus-based Techniques to Amharic Texts*, In *Proceedings of the MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches*, New Orleans, Louisana, September 2003.

Tesema Habte Mikael Gisew (1951 EC), **የአማርኛ መዝገበ ቃላት** *(yä'ämärñä mäzgäbä qalat),* Addis Ababa.

Wondwossen Belete (2004), *The Intellectual Property System in Ethiopia,* Ethiopian Intellectual Property Office, Addis Ababa, December 2004.