Application of corpus-based techniques to Amharic texts

Sisay Fissaha and Johann Haller

Institute for Applied Information Sciences– University of Saarland
Martin-Luther-Str.14, D-66111, Saarbrücken, Germany
Tel +49-681-3895126, Fax +49-681-3895140
{sisay, hans}@iai.uni-sb.de
http://www.iai.uni-sb.de

Abstract

A number of corpus-based techniques have been used in the development of natural language processing application. One area in which these techniques have extensively been applied is lexical development. The current work is being undertaken in the context of a machine translation project in which lexical development activities constitute a significant portion of the overall task. In the first part, we applied corpus-based techniques to the extraction of collocations from Amharic text corpus. Analysis of the output reveals important collocations that can usefully be incorporated in the lexicon. This is especially true for the extraction of idiomatic expressions. The patterns of idiom formation which are observed in a small manually collected data enabled extraction of large set of idioms which otherwise may be difficult or impossible to recognize. Furthermore, preliminary results of other corpus-based techniques, that is, clustering and classification, that are currently being under investigation are presented. The results show that clustering performed no better than the frequency base line whereas classification showed a clear performance improvement over the frequency base line. This in turn suggests the need to carry out further experiments using large sets of data and more contextual information.

1 Introduction

Manual development of lexical resources is a tedious and error prone activity. The increasing availability of huge amount of electronic text has made the development of different automatic and semi-automatic techniques possible. Among the wide range of applications developed in this area are monolingual lexicon development (such as terminology extraction), bilingual translation lexicon, and grammar induction (Charniak ,1996; Hong, et. al. 2001).

Amharic, which belongs to the Semitic family of languages, is one of the most widely spoken languages in Ethiopia. However, it is among the least researched or supported languages in the world. It is only recently that research in the development of different Natural Language Processing (NLP) tools for analysing Amharic text has begun. Among the rare attempts made in this direction are the development of an Amharic word-processing program (Daniel and Yonas, 1994), a stemming algorithm (Alemayehu, 1999), a word parser (Bayou, 2000), and a parts-of-speech tagger (Getachew, 2001). All these works are at an experimental stage and the outputs are prototypes which are limited in scope (Alemu et. al., 2003).

The current study is part of a wider project that attempts to integrate Amharic into the CAT2 machine translation system. CAT2 is a transfer-based MT system which was initially developed in 1987 as an alternative prototype to the Eurotra MT system. Since then considerable improvements have been made to it, and it has been applied to a number of languages. One subtask of the project is the development of an English-Amharic bilingual transfer lexicon. This paper explores some of the corpus-based techniques which help to speed up the task. In order to enable better understanding of the rest of the discussion, a brief description of Amharic morphology is provided in the next section. Sections two and three describe the extraction of collocations using a technique that assumes minimal linguistic knowledge. The last section discusses other corpus-based techniques that are currently being investigated.

1.1 Amharic Morphology

Amharic, like other Semitic languages such as Arabic, exhibits the root-pattern morphological phenomenon. This is especially true of Amharic verbs. Amharic verbs consist of stems and affixes. A stem again consists of root consonants (typically three) and vowel patterns. Stems encode different types of morphosyntactic information by changing the arrangement of the root consonants and vowel patterns as shown in the sample derivational paradigm of the verb *sbr* ('break') in Table 1.

Simple derived verb forms		
Categories	Pattern	Word
Perfect	CVCXVC[1]	*säbbär*
Imperfect	CVCC	*säbr*
Gerund	CVCC	*säbr*
Imperative	CCVC	*sbär*
Complex derived verb forms – Perfect aspect		
Causative	*as*-CVCXVC	*assäbbär*
Passive	*tä*-CVCXVC	*täsäbbär*
Reduplicative	CVCVCXVC	*säbabbär*

Table 1. Simple verb forms of *sbr* ('break')

The derivational processes in the complex derived forms are either internal in which CV patterns are changed, or external where derivational affixes (such as *as-,tä-*) are attached to the simple derived forms. Derivational processes may also involve a combination of internal and external changes. Some of the derivations in this class serve to express adverbial functions as the language has limited lexicalized adverbs. The stem forms take various affixes. For example, arguments of a verb are indicated on the verb using suffix pronouns (see Table 2). The subject is marked on the verb using subject suffix pronouns which agree with the subject in number, gender, and person. The direct object and some prepositional phrase complements are optionally marked on the verb. Functional elements like negation maker, conjunctives (some of them), some auxiliary verbs and relative markers are also bound morphemes and are attached to the verb.

[1] The symbols C, V and X in the template forms indicate consonant, vowel, and gemination of the previous consonant respectively.

Person	Pronoun	Subject	Object	Posses.
Singular				
1st	*əne*	*-ku/hu*	*-ñ*	*-e*
2nd m	*antä*	*-k/h*	*-h*	*-h*
Pol.	*ərswo*	*-u*	*-wot*	*-ačč̈äw*
f.	*anči*	*-š*	*-š*	*-š*
3rd m.	*əssu*	*-ä*	*-w*	*-u*
f.	*əsswa*	*-äč*	*-at*	*-wa*
Pol.	*ərsacäw*	*-u*	*ačč̈äw*	*ačč̈äw*
Plural				
1st	*əññə*	*-n*	*-ačč̈əhu*	*-ačč̈ən*
2nd	*ənnantä*	*-ačč̈əhu*	*-ačč̈əhu*	*-ačč̈əhu*
3rd	*ənnässu*	*-u*	*-ačč̈äw*	*-ačč̈äw*

Table 2. Independent and suffix pronouns of Amharic.

All arguments of a verb are optional and may only be indicated by suffix pronouns, that is, a verb may stand alone as a sentence. The typical word order of Amharic main and subordinate clauses is SOV.

Amharic nouns inflect for gender, number and case. However, gender in Amharic is mostly natural. Pre-nominal modifiers like adjectives, relative clauses take a similar set of inflectional morphemes as nouns. The head noun appears at the end of a noun phrase. Amharic has both preposition and postposition. Some of the prepositions are bound morpheme and are prefixed to nouns. Amharic also uses circumpositions to indicate positional relations. Table 3 shows sample prepositions and postposition.

Preposit.	Meaning	Postposit.	Meaning
bä-	at, in	*lay*	Top
lä-	for	*wəst*	Inside
kä-	from	*fit*	Front
səlä-	about	*h^{w}ala*	back
əndä-	like		
wädä	to		

Table 3. Sample prepositions and postpositions in Amharic

The definite article in Amharic is a bound morpheme and is attached to a noun or to the first inflected element in a noun phrase whereas other determiners like demonstrative pronouns are independent morphemes.

In general, although the above description of Amharic morphology is far from complete, we hope it gives some idea of its complexity. The distribution of the different affixal elements across different part of speech categories is varied. Some are specific to a particular part of speech category while others are applied across several parts of speech category.

2 Collocation Extraction

"Collocation" may be defined in different ways depending on the type and scope of the application one envisages. As this study is intended to extract as much lexical knowledge as possible from Amharic text corpus, we use a definition of collocation which includes fixed expressions, idioms, and terminologies as described in (Kita et al., 1994). In the past, different approaches have been suggested for purposes of automatic identification of collocations (Church et al., 1990; Kita et al., 1994; Smadja, 1991). Most of these approaches rely basically on statistical methods for collocation identifications. Others, especially those involved in the identification of technical terms make extensive use of linguistic knowledge in order to characterise terms (Daille 2001; Hong et al. 2001).

In this section, we extend the experience gained through the development of a terminology extraction program, known as AUTOTERM (Hong et. al, 2001), for European languages (German, English, French, Italian, and Spanish) for the extraction of collocation from an Amharic text corpus. AUTOTERM combines different filtering techniques (linguistic and statistical) for identification of candidate terms that denote concepts in the specific domain described by the input text. AUTOTERM is mainly concerned with identification of terms. It makes use of significant linguistic knowledge provided by a parser. However, due to lack of broad coverage and a robust Amharic morphological analyser and parser, we have adopted a statistical approach to the identification of collocations. In particular, our emphasis is on the extraction of collocations containing a contiguous sequence of words using minimal linguistic knowledge.

The corpus used consists of 720,000 words taken from an Amharic newspaper text. The text has been encoded using SERA transliteration scheme which represents Amharic script using Latin alphabets (Yacob, 1993). As in the original text, gemination is not indicated, which is one of the shortcomings of the Amharic script. Tools have been developed to convert the transliterated text into the original script and vice versa. For the current work, the transliteration has been found adequate and no significant loss of linguistic information has been observed.

Using this text corpus, bi-gram counts have been generated and measures of associations have been computed using Log-likelihood ratio. The most significant associations are marked in the text and the process is repeated again until no more significant associations are found. Table 4 and 5 show the top 11 bi- and tri-grams respectively.

No.	Collocation	Meaning
1	*yäadis abäba*	'PREP-Addis Ababa'
2	*adis abäba*	'Addis Ababa'
3	*əgr k^{w}as*	'football'
4	*qädäm sil*	'before'
5	*bäadis abäba*	'PREP-Addis Ababa'
6	*mkr bet*	'Council', 'Parliament'
7	*lämawäq täcl^{w}al*	'it was possible to known'
8	*bäahunu wäqt*	'this time'
9	*bälela bäkul*	'on the other hand'
10	*qalä mləls*	'interview'
11	*betä krstiyan*	'church'

Table 4. Top 11 bi-gram collocation

No.	Collocation	Meaning
1	*bäityop'yana bäertra mäkakäl*	'between Ethiopia and Eritrea'
2	*ngd mkr bet*	'Chamber of Commerce'
3	*yäafrika andnät drjt*	'Organization of African Unity'
4	*əgr k^{w}as fedärešn*	'football federation'
5	*bälela bäkul dägmo*	'on the other hand'
6	*yäadis abäba ngd*	'Addis Ababa Chamber …'
7	*ţäqlay ministr mäläs*	'Prime Minister Meles'
8	*käqrb gize wädih*	'recently'
9	*kägize wädä gize*	'from time to time'
10	*miliyon br bälay*	'more than…million Birr'
11	*wädä adis abäba*	'to addis ababa'

Table 5. Top 11 tri-gram collocations

The results show different kinds of interesting collocation candidates in terms of syntactic patterns and semantic compositionality. The top ranking collocations exhibit different syntactic patterns such as noun-noun, adjective-noun, verb-verb, verb-noun, etc. An example of noun-noun collocation is *əgr k^{w}as* ('football') which is compositional in meaning. However, syntactically it behaves as a single unit taking prefixes and suffixes only at the peripheries. Multi-words like *adis abäba* (capital city of Ethiopia, lit. 'new flower') are names and hence they should be treated as a single unit. *qalä mləls* ('interview', lit. 'exchange of words') consists of a noun *qalä* ('word') and an adjective *mləls* ('exchange') which is not syntactically and semantically compositional. A verb-verb collocation *qädäm sil* ('before') has an idiosyncratic meaning and function that are different from any of the constituting elements. *lämawäq täcl^{w}al* ('it was possible to know') also reveals another two-verb collocation which is typical for newspaper text.

Tri-gram word sequences also reveal important collocations. Some of the tri-gram collocations appearing at the top of the list are the complete forms or extensions of the corresponding top ranking bi-grams. Semantically, however, some exhibit idiosyncratic relations. *mkr bet* ('Council', 'Parliament') in the context of *ngd* ('trade') undergoes a slight semantic transform and gives rise to a different meaning (*ngd mkr bet*-'Chamber of commerce'). Although we considered only bi- and tri-gram word sequences, higher-level n-grams are also expected to reveal useful collocations. An attempt should also be made to extract discontinuous collocations. A small experiment to extract a discontinuous collocation within a window of 5 words revealed the following collocations candidate, *mäskäräm* ('September')… *qän* ('day') which is a natural consequence of the Amharic date format ('month day' *qän* 'year'). Enlarging the window size may also help discovering other meaningful collocations.

The appearance of three morphological variants of the word Addis Ababa (i.e. *yäadis abäba*, *bäadis abäba, adis abäba*) among the top 20 collocation candidates reveals the morphological complexity of the Amharic language. Amharic is a highly inflected language. This in turn results in sparse data problem. A shallow stripping of the most common prefixes and suffixes has been done with the hope of improving the result. Some improvements have been observed among the top ranking bi-grams. However, for the majority of the bi-grams only minor changes have been observed. This is because, in addition to prefix and suffix morphemes, internal change is the main mechanism of encoding grammatical information in Amharic. This in turn indicates the importance full-fledged morphological analysis.

3 Amharic idioms

The next experiment attempts to exploit the facts about the nature of Amharic idioms which are explicated in Amharic grammar literature (Amsalu, 1988). Amsalu classifies Amharic idioms based on their syntactic patterns and also the type of the most frequently occurring words in them. Table 6 shows some example of Amharic idioms taken from Amsalu (1988).

Idiom	Meaning
bäl bäl alä	'was instigated', 'incited'
yäsäw säw	'a perfect man'
hullu agärš	'adaptable'
lbä däandana	'stout-hearted'
qalä sätṭe	'promised'

Table 6: Sample Amharic idioms taken from Amsalu (1988)

Some typical syntactic patterns of Amharic idioms are V V, N N, N V, and Adj N. However, these patterns are the standard ways of forming noun phrases and verb phrases in Amharic, and therefore might not be very useful for the identification of idioms. One observation we made of the list provided in Amsalu (1988) is that total or partial reduplications of words seems to be a common phenomenon in Amharic idioms, e.g. *bäl bäl alä, yäsäw säw*. Hence duplicate sequences may give some hint as to the formation of idiomatic expressions. He also indicates that among the most commonly used words in building Amharic idioms are names of parts of the body, e.g. *lb* ('heart'), *anjät* ('intestine'), *joro* ('ear') etc, and common nouns and verbs like *lg* ('child'), *qal* ('word'), *mola* ('became full'), *qärrä* ('remained') etc. Therefore duplicate sequences and the above word types have been used as criteria in extracting candidate idioms. As duplicate sequences can be morphological variants of each other (and hence

may not be an exact copy of each other), a string similarity measure has been used and those string pairs that fall above the threshold of 0.7 have been extracted. Tables 7 and 8 shows the top 19 collocations extracted using this method.

1	*lyu lyu*	'different'
2	*Ĉärqa Ĉärq*	'textile'
3	*alfo alfo*	'occasionally'
4	*däräja bädäräja*	'step by step'
5	*wana wana*	'main'
6	*ahun ahun*	'recently'
7	*hulät hulät*	'two two'
8	*biyans biyans*	'at least'
9	*mäto metr*	'hundred meter'
10	*yäərs bäərs*	'each other'
11	*zoro zoro*	'finally'

Table 7: idioms containing duplicate sequences

1	*əgr kwas*	'football'
2	*mkr bet*	'council'
3	*qalä mləls*	'interview'
4	*betä krstiyan*	'church'
5	*mulu lämulu*	'completely'
6	*wädä ityop'ya*	'to Ethiopia'
7	*qal aqäbay*	'correspondent'
8	*frd bet*	'court'

Table 8: idioms containing the most commonly used idiom forming words

Some of the collocations share a common semantic space with the corresponding single word whereas others carry idiosyncratic semantic values not predictable from the corresponding single word unit. For example, at abstract level

- *lyu* and *lyu lyu* have close semantic relations. *lyu* ('different', 'distinct', 'special', 'specific', 'extraordinary', 'recluse' etc.) is used to express wide varieties of meaning and is usually used to single out an object or a set of homogenous objects whereas *lyu lyu* ('various', 'miscellaneous', 'distinct', 'separate') has a very restrictive meaning which is used to characterize a group consisting of heterogeneous elements.
- *alfo alfo* ('occasionally', 'now and again', 'from time to time') does not seem to have any semantic relation with the corresponding single word *alfo* (the gerund form of the verb *aläfä* 'he passed').

There are also some erroneously extracted bi-grams like *mäto metr* which consist of two distinct words having different base forms. One observation concerning these bi-grams is that they differ in one or more of their consonantal radicals. In Semitic languages, consonantal radicals usually carry semantic values (*sbr* –'break' for *säbbärä*- 'he broke'). It may be useful to work with only consonantal radicals in order to extract word sequences having the same semantic base.

The discussions hitherto have been concerned with the identification of collocations from Amharic text corpora. A simple extension of the technique developed for the extraction of terms for European languages provided us with word sequences that exhibit different collocation behaviour. Currently, an attempt is being made to apply other corpus-based techniques to Amharic text corpus which will be briefly discussed in the next section.

4 Work in progress

There are a number of other knowledge free corpus-based techniques that can be used to assist in the acquisition of lexical knowledge for resource poor languages like Amharic (Goldsmith 2001). Clustering is one such exploratory technique for discovering the natural grouping underlying data. Among the many applications of this technique are improving language modelling, word sense disambiguation and text categorization (Manning and Schütze 1999). Currently, we are investigating the application of this technique for bootstrapping the development of the linguistic resources required for other knowledge intensive tasks. Specifically we attempt to group Amharic words into linguistically motivated classes using minimal linguistic knowledge. These techniques may be used to bring together morphologically related words which in turn may assist in further data preparation activities, such as compiling an initial set of parts-of-speech tags, or generating semantic classes.

We describe a preliminary experiment we conducted on clustering Amharic words using surface morphological features in the next section[2].

[2] The Weka software has been used for analysing the data.

stem	word	nopre	bä-	…	nosuf	-ä	…	CVCV	…	aa	äa	…	cat
wana	*bäwana*	0	1	…	1	0	…	1	…	1	0	…	adj
mägzat	*lämägzat*	0	0	…	1	0	…	0	…	0	1	…	n
aläf	*aläfä*	1	0	…	0	1	…	0	…	0	1	…	v

Table 9. Data set used for clustering Amharic words

Test	Prefix	Suffix	Pattern	Vowel pattern	Classification	Clustering
1	X	X	X	X	70.16%	56.80%
2	X	X			75.00%	44.38%

Table 10. Result of clustering sample Amharic words

4.1 Experiment

The data used for the experiment is manually generated and consists of 248 words taken from the newspaper texts. Only four syntactic categories N(oun), V(erbs), ADJ(ectives), and ADV(erbs) have been considered, and functional categories like prepositions, determiners which are unbound morphemes have been eliminated. Five distinct occurrences of the stem forms have been extracted from the text and have been analyzed into the constituting morpheme. The attributes used for the experiment include consonant and vowel (CV) patterns, prefixes, suffixes and vowel patterns.

The suffix and prefix morphemes, CV templates and vowel patterns, which occur more than once, have been used to form the feature sets. Each word has been assigned a vector of values on the basis of these features. The result is coded in the form shown in Table 9 above. For example, the first element in Table 9 is an adjective which takes the prepositional prefix *bä* but no suffix, and has a template form of CVCV and vowel pattern of *aa.*

The experiment used a simple k-mean clustering algorithm. The result is shown in column 7 of Table 10 above. For the purpose of comparison, the result using 10-fold cross validation as test mode for classification is also shown in column 6.

The accuracy of the classification algorithm is better than the frequency base line (which assigns all words the most frequently occurring class) of 54%. Most adjectives and adverbs are incorrectly classified as nouns. This may be due to the fact that adjectives and adverbs behave morphologically like nouns by taking a similar set of prefixes and suffixes. Clustering seems to be a difficult task as the above results show.

In general, given the complexity of the phenomenon we are trying to model and the simplicity of the information used, the above result suggests the need to carry out further experiments. However, the real applicability of such techniques is yet to be tested using large set of data and more context information.

5 Conclusion

In this paper, we applied some of the corpus-based techniques for analyzing Amharic text corpora. The first part is concerned with the extraction of collocations. The aim is to extract as many collocations as possible using co-occurrence frequency information. Extraction of contiguous sequences of collocations showed encouraging results. Although the emphasis was on the extraction of contiguous collocations, an attempt has also been made to extract discontinuous collocations spanning several words apart. In addition to simple frequency counts, we also attempted to incorporate some language specific features for the identification of idioms, that is, duplicate sequences of words and words frequently occurring in Amharic idioms. This resulted in extraction of idiomatic expressions which would not have been possible using pure frequency counts. Due to the complexity of Amharic morphology, the attempt to normalize the surface forms by stripping off the most frequently occurring suffix and prefix morphemes brought only minimal improvements. Hence proper morphological analysis is required for improving the results. The last part of the paper presented a preliminary result obtained through the application of clustering and classification techniques. Clustering shows relatively poor performance whereas the result obtained for classification is

significantly better than the frequency base line. This suggests the need to carry out further research with large set of data.

6 Bibliographical References

Abiyot Bayou. 2000. *Developing automatic word parser for Amharic verbs and their derivation*. Master Thesis. Addis Ababa University.

Amsalu Aklilu. 1988. Characteristics of Amharic idiomatic expressions: In *Proceedings of the eigth international conference of Ethiopian studies.* pp. 571-578

Atelach Alemu, Lars Asker and Mesfin Getachew. 2003. Natural Language Processing for Amharic: Overview and Suggestions for a Way Forward. In *Traitement Automatique des Langues Naturelles*, Batz-sur-Mer, 11-14 juin 2003

Béatrice Daille. 1994. *Combined approach for terminology extraction: lexical statistics and linguistic filtering*. Université Paris VII.

Church, K. W., Gale, W., Hanks, P. and Hindle, D. 1991. Using Statistics in Lexical Analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Zernik U (ed), Lawrence Erlbaum Associates, pp. 115-164.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

Daniel Yacob. 1996. *System for Ethiopic Representation in ASCII (SERA)*. http://www.abys siniacybergateway.net/fidel/

Eugene Charniak. 1996. Tree-bank grammars. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, pp. 1031-1036, Menlo Park. AAAI Press/MIT Press.

Goldsmith, J. A. 2000. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27(2) pp. 153-198,

Kenneth Beesley. 1996. Arabic finite-state morphological analysis and generation. *Computational Linguistics*, vol 1, pp 89-94.

Kita, K., Kato, Y., Omoto, T. and Yano, Y. 1994. Automatically Extracting Collocations from Corpora for Language Learning. In *Corpora in Language Education and Research: A Selection of Papers from Talc94*, UCREL Technical Papers, Lancaster University, pp. 53-64.

Melamed, I. Dan. 1998. Empirical Methods for MT Lexicon Development. In *Proceedings of the Third Conference for Machine Translation in the Americas, AMTA'98*, Langhorne PA, Springer-Verlag, LNAI 1529, pp. 18-30.

Mesfin Getachew. 2001. *Automatic part of speech tagging for Amharic language: An experiment using stochastic HMM*. Masters Thesis. Addis Ababa University.

Munpyo Hong, Sisay Fissaha and Johann Haller. 2001. Hybrid Filtering for Extraction of Term Candidates from German Technical Texts, In *Proceedings of Terminologie et Intelligence Arteficielle, TIA'2001*

Nega Alemayehu. 1999. *Development of stemming algorithm for Amharic text retrieval*. PhD Thesis, University of Sheffield.

Pierre Isabelle and Laurent Bourbeau. 1984. TAUM-AVIATION: its technical features and some experimental results. *Computational Linguistics*, 11(1), pp. 18—27.

Smadja, F. A. and McKeown, K. R. 1990 Automatically Extracting and Representing Collocations for Language Generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp.252-259.