Data-driven Amharic-English Bilingual Lexicon Acquisition

Saba Amsalu

Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld
Kiskerstrasse 6, Bielefeld
Tel: 0049 (0)521 1063519
{saba@uni-bielefeld.de}

Abstract

This paper describes a simple approach of statistical language modelling for bilingual lexicon acquisition from Amharic-English parallel corpora. The goal is to induce a seed translation lexicon from sentence-aligned corpora. The seed translation lexicon contains matches of Amharic lexemes to weekly inflected English words. Purely statistical measures of term distribution are used as the basis for finding correlations between terms. An authentic scoring scheme is codified based on distributional properties of words. For low frequency terms a two step procedure of: first a rough alignment; and then an automatic filtering to sift the output and improve the precision is made. Given the disparity of the languages and the small size of corpora used the results demonstrate the viability of the approach.

1. Introduction

Parallel corpora have proved to be valuable resources for bilingual lexical information acquisition which can be used for multifarious computational linguistic and information retrieval tasks. However, extracting this information is a non-trivial task for several reasons which may be related to the properties of the languages considered or to the common problems that come with translated documents such as deletion, insertion, splitting, merging, etc.. The problem gets even more challenging when the languages considered are disparate. Often other tools, such as morphological analysers and taggers which may not be available for resource-poor languages are required. Amharic-English translated texts are such pair of languages that happen to belong to different language groups and apparently have different syntactic and morphophonological structures.

This paper describes a word alignment system that is designed to make comprehensive use of limited amount of Amharic-English corpora without giving any assumptions on the relative nature of the two languages. The goal of the study is to come up with efficient methods of language modelling to generate seed translation lexicon for use in a project of lexical acquisition from corpora not aligned at any level. Thus, the method takes advantage of corpus characteristics of short aligned units. Automatic filtering is used to improve the precision of the extracted material for low frequency words.

A brief account of previous studies is presented in Section 2. followed by a short examplary description of the grammatical characteristics of Amharic with relevance to corpus-based lexical acquisition in Section 3.. In Section 4., the orthography of Amharic is introduced. In Section 5., methodological aspect on how the problem is approached is discussed. Evaluation results are reported in Section 6.. Concluding remarks and problems that are open for subsequent studies are also forwarded in Section 7..

2. Previous work

There are several word-alignment strategies devised by computational linguists for major languages such as English, French and Chinese (Dagan et al., 1993; Fung and Church, 1994; Simard et al., 1992; Gale and Church, 1994; Gale and Church, 1994; Sahlgren and Karlgren, 2005; Melamed, 2000; Wu and Xia, 1995; Wu and Xia, 1994; Kay and Röscheisen, 1993). Broadly speaking the approaches used are either statistical or linguistic or a hybrid of both approaches. Statistical approaches are more commonly used on language pairs that have high similarity and also with those that have a relatively less complex morphological structure. In other cases linguistic approaches predominate for obvious reasons.

A work that deals with language pairs identical to the analysis in this project is the one made on Hebrew-English pairs (Choueka et al., 2000). The Hebrew-English alignment algorithm creates an alignment $< i, j >$ where i and j correspond to positions in source and target texts. The algorithm relies on the assumption that positions of translation words are distributed similarly throughout two texts. A word is represented by a vector whose entries are distances between successive occurrences of the word. They use lemmatizers for both languages and assert lemmatization is a must when dealing with Semitic languages. An exploration on Amharic by (Alemu et al., 2004) deals with an attempt to extract noun translations from the bible. Yet, nouns are relatively minimally inflected and not a problem to align in Amharic, specially when the bible is the data source.

In this paper a novel statistical method of bilingual lexical acquisition from Amharic-English parallel corpora that makes no use of lemmatizers and addresses words of all parts of speech is presented.

3. Morphology and syntax of Amharic

Amharic and English differ substantially in their morphology, syntax and the writing system they use. As a result various methods of alignment that work for other languages do not apply for them. Examplary description of the grammar of Amharic words and sentences that suffices the relevance to text alignment is subsequently presented.

Amharic is a Semitic language that has a complex morphology which combines consonantal roots and vowel intercalation with extensive agglutination (Amsalu and Gibbon, 2005; Fissaha and Haller, 2003; Bayou, 2000), an inherent Semitic property. Articles, prepositions, conjunctions and personal pronouns are often inflectional patterns of other parts of speech and can only seldom occur as disengaged morphemes. Apparently, sentences in Amharic are often short in terms of the number words they are consisted of. For the reader to assimilate the flavour of the problem, just picking the first sentence in the bible in Amharic and English,

በመጀመሪያ እግዚአብሔር ሰማያትንና ምድርን ፈጠረ

In the begining God created the heavens and the earth

we obtain a ratio of 1:2 words.

This is a common case as far as the two languages are concerned. The texts that are used in the experiment presented in this paper have a ratio of 22179:36733, which is approximately 1 Amharic word to 1.7 English words.
But if we try to consider morphemic substratum we observe a different result. In Figure 1, a projection at nearly morpheme level is presented.

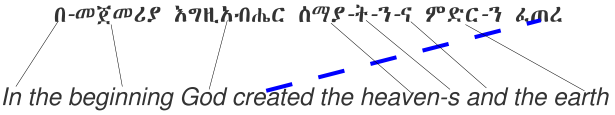
በ-መጀመሪያ እግዚአብሔር ሰማያ-ት-ን-ና ምድር-ን ፈጠረ
In the beginning God created the heaven-s and the earth

Figure 1: Morphemic alignment.

Definiteness in Amharic is not necessarily explicitly represented. It is often left to be understood contextually. When it is explicit the definite article is realized as a suffix and rarely the indefinite article is expressed with a number coming before the noun such as '*and säw*', literally it means '*one man*', parallel to the English '*a man*'. The definite article '*the*' that occurs three times in the English sentence in Figure 1 is in all cases implicit in the Amharic translation. Hence, there are floating words in the English side that are not aligned. The object marker'ን' in Amharic also does not exist in English. This paper does not give a detailed account of Amharic morphology; better treatments are given by (Yimam, 1994; Yimam, 1999; Bender and Fulas, 1978; Berhane, 1992; Dawkins, 1960; Amare, 1997; Markos, 1991; Amsalu and Gibbon, 2005).

Syntactically Amharic is an SOV language. It does not have free order as in other Semitic languages. The generalisation given by (Choueka et al., 2000) about the free word order for Semitic languages does not hold for Amharic. Taking their own example,

The boy ***ate*** *the apple* (English)

the correct representation in Amharic is:

The boy the apple ***ate***

This forbids a linear alignment of Amharic words with their English equivalents which are revealed in SVO order. The broken line in Figure 1 shows a cross-over alignment that accommodates this discord in syntax. In a two dimensional Cartesian plane of alignments between source and target texts we do not expect a linear path, rather it would be skewed at the position of inversion of the verb and object. See the chart in Figure 2 for the portray of the mapping of our example sentences.

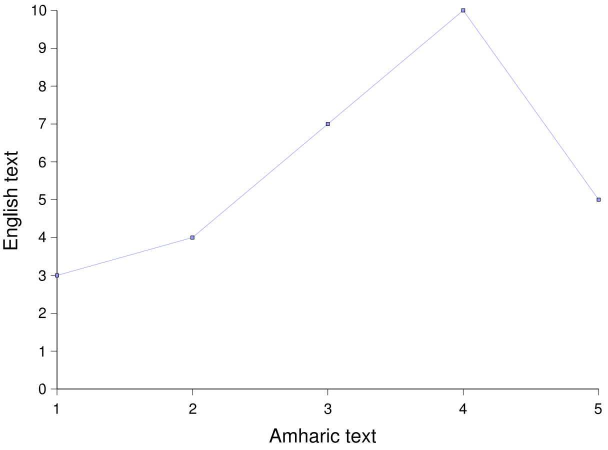

Figure 2: Non-linear alignment.

4. Amharic orthography

Amharic uses a syllabary script called Fidel, with graphemes denoting consonants with an inherent following vowel, which are consistently modified to indicate other vowels or, in some cases, the lack of a vowel. There are no different representations of upper and lower cases, hence there is no special marking of the beginning of a sentence and first letter of names or acronyms. Words are separated by white space. The streams of characters, however, are written left-to-right deviating from its relatives Hebrew and Arabic.
Differences in writing system reflects on attempts to align cognates. Amharic and English do not share many words such as, say, English and German do, but scientific words, technical words and names of places and people or objects are often either inherited from English or both take them from some other language. Phonetically, cognates sound the same. For example the word *'police'* is also *'pəli:s'* in Amharic phonologically decoded, but when written in Fidel it is ፖሊስ. In effect, it does not have any relation whatsoever to its English complementary.

5. Parallelizing words

Statistical methods of modelling relations of translation words have a limitation in that they require a large amount of corpora to align a relatively smaller size of lexicon in comparison to the total number of words in the texts. The size of corpora needs to be even bigger when highly inflected languages are used, because all variants of a given word are considered different which will have tremendous effect in altering the frequencies of occurrences. The dearth of large amounts of corpora is, on this account, a bottleneck for many languages. On the other hand linguistic approaches require computational linguistic tools which in the case of Amharic operational systems are not immanent. There are only prototype level systems for morphological analysis (Bayu, 2002; Bayou, 2000; Amsalu and Gibbon, 2005) and POS taggers (Getachew, 2001; Adafre, 2005).

Therefore, in this paper a statistical method that tries to make optimal use of bounded amount of corpora without causing too much of degradation in the outputs is proposed. Attempts to align words with attenuated distributional similarity are also made. For that cause, a filtering system that filters outputs obtained from the first alignment is developed.

The assumption in taking distributional properties of words as the measure for their equivalence emanates from the belief that equivalent terms are distributed similarly throughout the texts. Hence, the distribution of each term in the source language is compared to the distribution of every term in the target language.

The final aim is to parallelize Amharic lexemes to weakly inflected English words. The alignment algorithm does not exclude function words from computation rather the scoring scheme which is discussed in Section 5.3.. distills them by keeping their scores low. From the Amharic side a significant proportion of the words have a high probability of being included in the lexicon, while in the English side there will be floating words which would in many cases be function words. A demonstration on our examplary bitext segment is presented in Figure 3

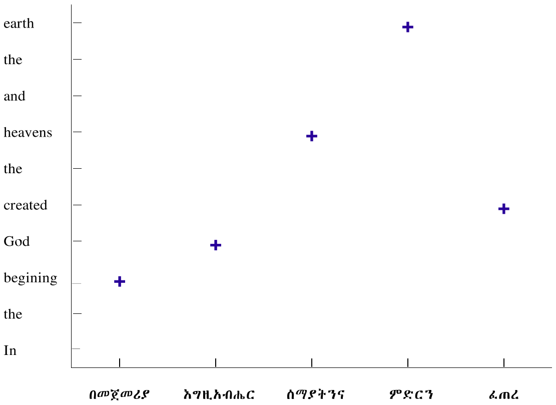

Figure 3: Aligned Words.

Gaps for non-aligned words and crossing alignments to overcome syntactic differences are extant. Details of the alignment heuristics are discussed in subsequent subsections.

5.1. Data preparation

The data sources used for testing the canonization and the systems and subsystems developed thereof are the books of Matthew and Mark in the bible. Preliminary processing steps fundamentally consisted of:

- text segmentation,
- tokenization and
- splitting and merging.

All operations with the exception of splitting and merging are machine based. At the same time distribution values of tokens were extracted. Translation memory which is not the major part of this work was also produced as a by-product.

5.2. Term distribution measures

The distribution of a term is simply a measure of how frequently and where in the document it occurs. Texts are often divided into smaller segments inorder to decrease the amount of search space and consequently have limited options. In the case of this paper the small segments are sentences. Three parameters are used to describe the distribution of each term:

1. *Global-frequency*: Frequency of occurrence in the corpus;
2. *Local-frequency*: Frequency of occurrence in a segment; and
3. *Placement*: Position of occurrence in the corpus.

5.3. Scoring Scheme

The scoring scheme formularized is an original novel scheme that gives categorical scores for each distinct pair of distributions and favours those that are distributed similarly. The scoring scheme also handles function words robustly.

In set-theoretic terms, we have a set of distributions and a set of terms. Let D_a be the set of distributions in the Amharic text and D_e be the set of distributions in the English text. And let T_a be the set of terms in Amharic text and T_e be the set of terms in English text. Then if an Amharic Term $Term_j \in T_a$ has a distribution $D_j \in D_a$ and an English term $Term_k \in T_e$ has a distribution $D_k \in D_e$, then the score of the translation candidates $Term_j$ and $Term_k$ is a measure of the degree of similarity between the distributions D_j and D_k.

Hence, we have an nxs and an mxs matrices; where n and m are the number of unique terms in Amharic and English respectively and s is the number of segments in either of the texts. The values in the matrix are local frequencies. Therefore, each word is a weighted vector of its distribution; where the weight is its local frequency in the respective segment. If for example $Term_j$, is an Amharic term vector, with the values: $Term_j = (0, 2, 1, 0, 0)$ and suppose we have a term in the English document, $Term_k = (0, 1, 1, 0, 1)$. Then,

$$Score_{(j,k)} = \frac{2 \cdot \Sigma(Term_j \wedge Term_k)_i}{\Sigma(Term_j + Term_k)_i} = \frac{2 \cdot 2}{6} \approx 0.67$$

where $_i$ denotes the i^{th} entry of a vector, and

$$Term_j \wedge Term_k = (0, 2, 1, 0, 0) \wedge (0, 1, 1, 0, 1)$$

$$= (0,1,1,0,0),$$

$$Term_j + Term_k = (0, 2, 1, 0, 0) + (0, 1, 1, 0, 1)$$

$$= (0,3,2,0,1)$$

If instead we have pairs of $Term_j = (0,1,1,0,0)$ and $Term_k = (0,1,1,0,1)$,

$$Term_j \wedge Term_k = (0,1,1,0,0),$$

$$Term_j + Term_k = (0,2,2,0,1)$$

$$Score_{(j,k)} = \frac{2 \cdot \Sigma(Term_j \wedge Term_k)_i}{\Sigma(Term_j + Term_k)_i} = \frac{2 \cdot 2}{5} = 0.8$$

again, for $Term_j = (0,2,1,0,0)$ and $Term_k = (0,2,1,0,1)$,

$$Term_j \wedge Term_k = (0,2,1,0,0),$$

$$Term_j + Term_k = (0,4,2,0,1)$$

$$Score_{(j,k)} = \frac{2 \cdot \Sigma(Term_j \wedge Term_k)_i}{\Sigma(Term_j + Term_k)_i} = \frac{2 \cdot 3}{7} \approx 0.86$$

The constant 2 in the numerator is algebraized to normalise the scores to range between 0.0 (for disjoint vectors) and 1.0 (for identical vectors), which otherwise would have been in the range of 0.0 to 0.5.

5.4. Thresholds

Obviously candidates with low score are bad candidates. But the question is, what values of score are low? To determine this cutting point different thresholds of score above which candidates could be true translation were tested on the corpus and the one that gives reasonably good translation pairs is selected. But again not all candidates with high score are true translations. In fact for a small size of corpus many of the candidates with a score of 1.0 are low frequency words. Hence, to control this a second threshold for frequencies is set.

5.5. Filtering mechanism

In statistical methods of alignment, the words that can most likely be correctly aligned are high frequency words. This is because there are many instances of these words that enable them to survive from accidental collisions with false translations. But for low frequency words, it is highly likely that just by chance they could co-occur with words that are not their equivalents. Specially when the test is made on a small size of corpora, low frequency words are too many and often coincide with several other low frequency words.

One commonly used method of avoiding such coincidences is to amputate low frequency words from evaluation set. Other methods of filtering are looking into knowledge sources such as the parts of speech of aligned texts, machine readable dictionaries, cognate heuristics, etc. (Melamed, 1995). In this paper a simple operation of annihilating those words that are aligned with equal score to different words is made.

6. Evaluation

The statistical language model developed is evaluated on a dataset of 20,347 Amharic and 36,537 English words, which encompass 6867 and 2613 unique words in Amharic and English respectively. The first attempt to screen the candidates with higher score and high frequency is presented in Table 1.

For score $>= 0.7$ and $\Sigma(Term_j + Term_k)_i > 5$, among the 38 errors, 30 of them are due to candidates with $\Sigma(Term_j + Term_k)_i$ between $6 - 9$. Hence, the threshold for frequency is set to > 9. Again, keeping the frequency threshold fixed the score is lowered until 0.55. For scores below 0.55 the accuracies went below 80%.

To exploit the low frequency words, a two step analysis is assembled. First a higher threshold is set for them, second a filtering algorithm is designed to screen those words with multiple equal score translations. For $\Sigma(Term_j + Term_k)_i$ between $6 - 9$ with score $>= 0.8$, 64.71% has been obtained before filtering and 82.35% after filtering.

The filter for one and two frequency words, selects all words that match with a score of 1.0 with one and only one word. After filtering, accuracies of $51.61\%$and 43.55% for two and one frequency words (i.e. $\Sigma(Term_j + Term_k)_i$ equal to 4 and 2) respectively is achieved.

6.1. Analysis of the results

The score threshold level for which a good percentage was found before filtering is 0.55. This means the distributions of translation candidates need only overlap in almost 50% of the case. This is an advantage for inflectional variants of Amharic that fail to align quite well with their counterpart. Surprisingly enough our method works well even with low frequency words. The translation pairs need to have a frequency sum > 9. This means that each word on average needs to appear in the text only 4.5 times. This is without filtering. With filtering words of frequency 3 also give good results. Most other existing systems use a higher frequency threshold (Sahlgren and Karlgren, 2005).

The weakness of this system lies on the inability to handle multiword compounds. Verbal compounds as well as many nominal compounds are written as two separate words (Amsalu and Gibbon, 2005). Split compound alignments are reckoned as wrong matches. Excluding them from the result set, the accuracy of our experimentation increases to 87.76%.

Lets give an examplary explanation of the case, for more clarity of the facts. The analogue for the word *'disciple'* in Amharic is *'däǩä mäzmur'*. The constituent words always come together. Nevertheless, a statistical alignment system knows them to be two separate words. Yet, since they always appear as a unit, each one of them are likely to match with every word in the English text with equal score. To exemplify it, suppose we have,

Score <disciple , däǩä>= 0.7 and
Score <disciple , mäzmur>=0.7

It is easy to excavate them from the result set by simply setting a conditional rule that if a word is aligned with a value which is its best score with two terms, then accouple the two terms as strings of a compound and align the single word to them, i.e.,

Score	$\Sigma(Term_j + Term_k)_i$	Correct	Compounds	Wrong	Total	% Correct
$>= 0.7$	> 5	123	16	38	177	**69.49%**
$>= 0.6$	> 9	134	8	20	162	**82.72%**
$>= 0.55$	> 9	172	9	24	205	**83.90%**

Table 1: Candidates of high score and high frequency.

Score <disciple , däk̆ä mäzmur>= 0.7

Corpus data can be used to find which string comes first.

But there are two problems that block us from using their score as a measure of their association. First, compounds could be inflected. Inflection may alter either or both of the elements. If the compound takes a prefix, the first element will be affected. If the compound takes a suffix the second element will be changed. This will mess up the scores. The second problem arises for the reason that in most cases, the second part of the compound can exist unbound. And when it occurs independently it has alltogether another meaning. In our example multiword compound, the second part *'mäzmur'* means *'song'*.

The best plausible solution would possibly be to mark compounds as one word right from the beginning. That way, even if they are inflected they will only be affected like any other word would. In an attempt to excerpt compounds from the corpus, bigram distributions of words were generated. Perhaps because the document size was small there were many non-compound bigrams that occurred as frequently as the compounds.

7. Conclusions and future work

The work described in this paper demonstrates that alignment of disparate languages using statistical methods is viable. It is also possible to gain good translation matches even for low frequency words with the assistance of simple filtering measures.

Research on the use of other approaches that depend on simple linguistic features of texts, such as syntactically fixed realizations of terms and expressions and alignments of above word level strings in context are on their way (Amsalu and Gibbon, 2006). Empirical methods for generating more lexical items from the original corpus, given the known translations in the corpus and maximum likelihood estimates that consider every word in the documents are also being investigated. Reusing the seed lexicon to align bigger chunks of text is worth attention. The use of bigram and trigram alignments which for the corpus used here did not produce good results may be tested on a bigger size corpora.

8. References

Sisay Fissaha Adafre. 2005. Part of speech tagging for amharic using conditional random fields. In *Proceedings ACL-2005 Workshop on Computational Approaches to Semitic Languages*, pages 47–54.

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 LREC*, Lisbon.

Getahun Amare. 1997. *Zämänawi yamarNa Säwasäw bäqälal aqäraräb*. Commercial Printing Press, Addis Ababa.

Saba Amsalu and Dafydd Gibbon. 2005. Finite state morphology of amharic. In *International Conference on Recent Advances n Natural language processing 2005*, pages 47–51, Borovets.

Saba Amsalu and Dafydd Gibbon. 2006. Methods of bilingual lexicon extraction from amharic-english parallel corpora. In *World Congress of African Linguistics*, Addis Ababa.

Abiyot Bayou. 2000. Design and development of word parser for amharic language. Master's thesis, School of Graduate Studies of Addis Ababa University, Addis Ababa.

Tesfaye Bayu. 2002. Automatic morphological analyser for amharic: An experiment employing unsuppervised learning and autosegmental analysis approaches. Master's thesis, School of Graduate Studies of Addis Ababa University, Addis Ababa.

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

Girmaye Berhane. 1992. Word formation in amharic. *Journal of Ethiopian Languages and Literature*, pages 50–74.

Yaacov Choueka, Ehud S. Conley, and Ido Dagan. 2000. A comprehensive bilingual word alignment system. application to disparate languages: Hebrew and english. In *Parallel text Processing: Alignment and use of Translation Corpora*. Kluwer Academic Publishers.

Ido Dagan, Kenneth W. Church, and William A. Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very large Corpora: Academic and Industrial Perspectives*, pages 1–8, Columbus, Ohio.

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

Sisay Fissaha and Johann Haller. 2003. Amharic verb lexicon in the context of machine translation. In *Traitement*

Automatique des Langues Naturelles, TALN2003, pages 183–192.

Pascale Fung and Kenneth W. Church. 1994. Kvec: A new approach for aligning parallel texts. In *COL–ING 9J*, pages 1096–1102, Kyoto.

William A. Gale and Kenneth W. Church. 1994. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Mesfin Getachew. 2001. Automatic part of speech tagging for amharic language: An experiment using stochastic hmm. Master's thesis, School of Graduate Studies of Addis Ababa University, Addis Ababa.

Martin Kay and Martin Röscheisen. 1993. Text–translation alignment. *Computation Linguistics*, 19:121–142.

Habte Mariam Markos. 1991. Towards the identification of the morphemic components of the conjugational forms of amharic. In *Proceedings of the Eleventh International Conference of Ethiopian Studie*, Addis Ababa.

I. Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing n–best translation lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*, Boston.

I. Dan Melamed. 2000. Pattern recognition for mapping bitext correspondance. In Jean Vèronis, editor, *Parallel Text Processing: Alignment and Use of Translation Corpora*, chapter 2, pages 25–48. Kluwer Academic Publishers.

Magnus Sahlgren and Jussi Karlgren. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(3).

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 92)*, pages 67–81, Montreal.

Dekai Wu and Xuanyin Xia. 1994. Large-scale automatic extraction of an english-chinese translation lexicon. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 206–213, Columbia,Maryland.

Dekai Wu and Xuanyin Xia. 1995. Large-scale automatic extraction of an english-chinese translation lexicon. *Machine Translation*, 9(3-4):285–313.

Baye Yimam. 1994. *YamarNa Säwasäw*. E.M.P.D.A, Addis Ababa.

Baye Yimam. 1999. Root reductions and extensions in amharic. *Ethiopian Journal of Languages and Literature*, pages 56–88.