Development of Amharic Grammar Checker Using Morphological Features of Words and N-Gram Based Probabilistic Methods

Aynadis Temesgen
Department of Computer Science
Addis Ababa University
temesgen.aynadis@gmail.com

Yaregal Assabie
Department of Computer Science
Addis Ababa University
yaregal.assabie@aau.edu.et

Abstract

Amharic is one of the most morphologically complex and under-resourced languages which effectively hinder the development of efficient natural language processing applications. Amharic words, especially verbs, are marked for a combination of several grammatical functions which makes grammar checking complex. This paper describes the design and development of statistical grammar checker for Amharic by treating its morphological features. In a given Amharic sentence, the morphologies of individual words making up the sentence are analyzed and then n-gram based probabilistic methods are used to check grammatical errors in the sentence. The system is tested with a test corpus and experimental results are reported.

1 Introduction

With the rise of electronic documents, the need of natural language processing (NLP) applications that automatically process texts has drastically increased. One of such important NLP applications is grammar checker which automatically checks grammatical errors in texts and also possibly suggests the user to choose among other alternatives. Initially, most of the grammar checkers were based on checking styles, uncommon words and sentence structures, but now they are upgraded to high capacity with the capability of analyzing complex sentence structures, not only as a part of other programs but also as easy software to be installed in many operating system (Richardson, 1997; Liddy, 2001; Mudge, 2010; Mozgovoy, 2011). Various techniques and methods have been proposed so far to build systems that could check the grammars of texts. Among the most widely used approaches to implement grammar checkers are *rule-based*, *statistical* and *hybrid* (Tsuruga and Aizu, 2011; Ehsan and Faili, 2013; Xing *et al*, 2013). Rule-based systems check grammars based on a set of manually developed rules which are used to match against the text. However, it is very difficult to understand and include all grammatical rules of languages, especially for complex sentences. On the other hand, in statistical grammar checking, part-of-speech (POS)-annotated corpus is used to automatically build the grammatical rules by identifying the patterns of POS tag sequences in which case common sequences that occur often can be considered correct and the uncommon ones are reported to be incorrect. This has lead statistical approaches to become popular methods to build efficient grammar checkers. However, it is very difficult to understand error messages suggested by such checking system as there is no specific error message. Hybrid grammar checking is then introduced to benefit from the synergy effect of both approaches (Xing *et al*, 2013). A number of grammar checkers have been developed so far for many languages around the world. Among the most notable grammar checkers are those developed over the past few years for resourceful languages such as English (Richardson, 1997; Naber, 2003), Swedish (Arppe, 2000; Domeij *et al*, 2000), German Schmidt-Wigger, (1998), and Arabic (Shaalan, 2005), etc. However, to our best knowledge, there is no commercial Amharic grammar checker or published article that presents grammar checking for Amharic.

This paper presents statistical-based Amharic grammar checker developed by treating the morphological features of the language. The organization of the remaining part of the paper is as follows. Section 2 discusses an overview of the grammatical structure of Amharic. Section 3

presents the statistical methods applied to develop the system. Experimental results are presented in Section 4. In Section 5, we present our conclusion and recommendation for future works. A list of references is provided at the end.

2 Grammatical Structure of Amharic

2.1 Amharic Morphology

Amharic is the working language of Ethiopia having a population of over 90 million at present. Even though many languages are spoken in Ethiopia, Amharic is the dominant language that is spoken as a mother tongue by a large segment of the population and it is the most commonly learned second language throughout the country (Lewis *et al*, 2013). Amharic is written using its own script which has 33 consonants (base characters) out of which six other characters representing combinations of vowels and consonants are derived for each character.

Amharic is one of the most morphologically complex languages. Amharic nouns and adjectives are marked for any combination of number, definiteness, gender and case. Morover, they are affixed with prepositions. For example, from the noun ተማሪ (*tämari*/student), the following words are generated through inflection and affixation: ተማሪዎች (*tämariwoč*/students), ተማሪው (*tämariw*/ the student {masculine}/his student), ተማሪየ (*tämariyän*/my student), ተማሪየን (*tämariyän*/my student {objective case}), ተማሪሽ (*tämariš*/your {feminine} student), ለተማሪ (*lätämari*/for student), ከተማሪ (*kätämari*/ from student), etc. Similarly, we can generate the following words from the adjective ፈጣን (*fäṭan*/fast): ፈጣኑ (*fäṭanu*/fast, {definite} {masculine} { singular}), ፈጣኖች (*fäṭanoč/fast* {plural}), ፈጣኖቹ (*fäṭanoču*/fast {definite} {plural}), etc.

Amharic verb inflections and derivations are even more complex than those of nouns and adjectives. Several verbs in surface forms are derived from a single verbal stem, and several stems in turn are derived from a single verbal root. For example, from the verbal root ውስድ (*wsd*/to take), we can derive verbal stems such as *wäsd*, *wäsäd*, *wasd*, *wäsasd*, *täwäsasd*, etc. From each of these verbal stems we can derive many verbs in their surface forms. For example, from the stem *wäsäd* the following verbs can be derived:

ወሰደ (*wäsädä*/he took)
ወሰደች (*wäsädäč*/she broke)
ወሰድኩ (*wäsädku*/I broke)
ወሰድኩት (*alwäsädkutĭm*/I took [it/him])
አልወሰድኩም (*alwäsädkum*/I didn't take)
አልወሰደችም (*alwäsädäčĭm*/she didn't take)
አልወሰደም (*alwäsädäm*/he didn't take)
አልወሰደኝም (*alwäsädäñĭm*/he didn't take me)
አስወሰደ (*aswäsädä*/he let [someone] to take)
ተወሰደ (*täwäsädä*/[it/he] was taken)
ስለተወሰደ (*sĭlätäwäsädä*/as [it/he] was taken)
ከተወሰደ (*kätäwäsädä*/ if [it/he] is taken)
እስኪወሰድ (*ĭskiwäsäd*/until [it/he] is taken)
ሲወሰድ (*sĭwäsäd*/when [it/he] is taken)
⋮
etc.

Amharic verbs are marked for any combination of person, gender, number, case, tense/aspect, and mood resulting in thousands of words from a single verbal root. As a result, a single word may represent a complete sentence cosutructed with subject, verb and object. For example, ይወስደኛል (*yĭwäsdäñal*/[he/it] will take me) is a sentence where the verbal stem ወስድ (*wäsd*/ will take) is marked for various grammatical functions as shown in Figure 1.

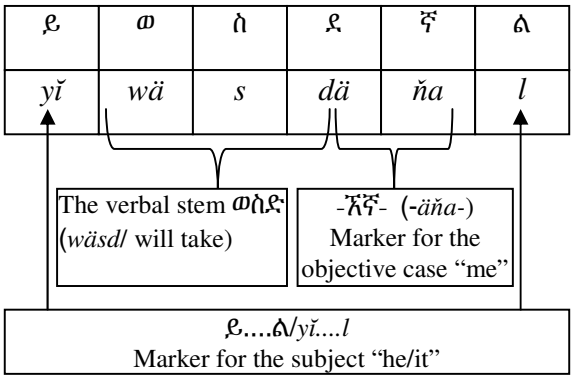

Figure 1. Morphology of the word ይወስደኛል.

2.2 Grammatical Rules of Amharic

Common for most languages, if not for all, grammar checking starts with checking the validity of the sequence of words in the sentence. This is also true for Amharic. In addition, since Amharic is morphologically complex language where verbs, nouns and adjectives are marked for various grammatical functions, the following agreements are required to be checked: adjective-noun, adjec-

tive-verb, subject-verb, object-verb, and adverb-verb (Yimam, 2000; Amare, 2010).

Word Sequence: Amharic language follows subject-object-verb (SOV) grammatical pattern as opposed to, for example, English language which has SVO sequence of words. For instance, the Amharic equivalent of sentence "John ate bread" is written as "ጆን (*jon*/*John*) ዳቦ (*dabo*/bread) በላ (*bäla*/ate)". Here, the part-of-speech (POS) tags of individual words are used as inputs to check the validity of grammatical patterns.

Adjective-Noun Agreement: Amharic nouns are required to agree for number of modifying adjectives. For example, ረጃጅም ልጆች (*räjajĭm lĭjoč*/ tall {plural} children) is a valid noun phrase whereas ረጃጅም ልጅ (*räjajĭm lĭj*/ tall {plural} child) is an invalid noun phrase construction.

Subject-Verb Agreement: Amharic verbs are marked for number, person and gender of subjects. For example, ልጆቹ መስኮት ሰበሩ (*lĭjoču mäskot säbäru*/the children broke a window) is a valid Amharic sentence. However, ልጅቷ መስኮት ሰበረ (*lĭjt^{w}a mäskot säbärä*/the girl broke {masculine} a window) is not a valid Amharic sentence since the subject ልጅቷ (*lĭjt^{w}a*/the girl) is feminine and the verb ሰበረ (*säbärä* /broke {masculine}) is marked for masculine.

Object-Verb Agreement: Amharic verbs are also marked for number, person and gender of objective cases. For example, in the sentence ልጆቹ መስኮቶቹን ሰበሯቸው (*lĭjoču mäskotoču säbär^{w}ačäw*/the children broke {plural} the windows), the verb ሰበሯቸው (*säbäru*/broke {plural}) is marked for the plural property of the object መስኮቶቹን (*mäskotoču*/the windows).

Adverb-Verb Agreement: Tenses of verbs are required to agree with time adverbs. For example, ትላንት ሰበሩ (*tĭlant säbäru*/ [they] broke yesterday) is a valid verb phrase construction whereas ትላንት ይሰብራሉ (*tĭlant yĭsäbralu*/ [they] will break yesterday) is an invalid construction.

3 The Proposed Grammar Checker

The proposed grammar checker for Amharic text passes through three phases:

- Checking word sequences;
- Checking adjective-noun-verb agreements;
- Checking adverb-verb agreement.

In the first two phases, we employ the n-gram based statistical method. The n-gram probabilities are computed from the linguistic properties of words in a sentence.

3.1 Representation of the Morphological Properties of Words

To check grammatical errors in an Amharic sentence, the morphological properties of words is required. The morphological property of Amharic words contains linguistic information such as number, gender, person, etc. Such linguistic information is used to check whether the linguistic properties of one word agree with that of the other words in the sentence. For this task, we used an Amharic morphological analyzer known as HornMorpho developed by Gasser (2011). After performing morphological analysis for a given word, the morphological property of the word is stored along with its POS tag using a structure with four slots as shown in Figure 2.

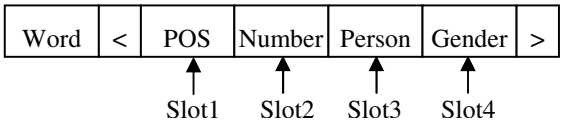

Figure 2: A structure for representing the linguistic properties of words

Slot1: This slot contains information about the POS tag of the word. The corpus we used in this work contains 31 types of POS tags, and the value for this slot is retrieved from the corpus. In addition to checking the correct POS tag sequence in a sentence, this slot is required to check agreements in number, person and gender as well.

Slot2: This slot holds number information about the word, i.e. whether the word is plural (P), singular (S), or unknown (U). In the case of nouns and adjectives, it has three values: P, S, or U. Since Amharic verbs are marked for numbers of subject and object, the value for this slot are combinations of the aforementioned values for subject and objective cases. We use the symbol "^" to represent such combinations. For example, a verb marked for plural subject and singular object is represented as SP^OS; a verb marked for singular subject and singular object is represented as SS^OS; etc.

Slot3: This slot stores person information about the word, i.e. first person (P1), second person (P2), third person (P3), or unknown (PU). The slot has four different possible values for the nouns and adjectives: P1, P2, P3 and PU. However, verbs can

have a combination of these four values for subject and object grammatical functions. Examples of slot values for verbs are the following.

SP1^OP1: verb marked for first person subject and first person object
SP2^OP1: verb marked for second person subject and first person object
SP3^OP1: verb marked for third person subject and first person object
SP2^OP3: verb marked for second person subject and third person object
⋮
etc.

Slot4: This slot holds gender information about the word, i.e. whether the word is masculine (M), feminine (F), or unknown (U). In the case of nouns and adjectives, it has three values: M, F, or U. The values of this slot for verbs are are combinations of the aforementioned values for subject and objective cases. Accordingly, a verb marked for masculine subject and feminine object is represented as SM^OF; a verb marked for feminine subject and masculine object is represented as SF^OM; etc.

For example, the linguistic information built for the noun ፕሬዚዳንቱ (*prezidantu*/the president {masculine}) is: ፕሬዚዳንቱ <N|S|P3|M>. Likewise, the linguistic information for the verb ደረሰችበት (*däräsäčtbät*/she reached at him) is: ደረሰችበት <V|SS^OS|SP3^OP3|SF^OM>. Accordingly, the linguistic information about each word in the entire corpus is automatically constructed so as to use it for training and testing.

3.2 Word Sequences

To check the validity of POS tag sequence for a given sentence, we use n-gram probability p_t computed as:

$$p_t(w_n \mid w_1 w_2 ... w_{n-1}) = \frac{count(w_1 w_2 ... w_{n-1} w_n)}{count(w_1 w_2 ... w_{n-1})} \qquad (1)$$

where n is the number of words in a sequence and w is POS tags of words. We have calculated n-gram values for n=2 (bigram) and n=3 (trigram) where they are saved in repository and used in grammar checking process. The probabilities of sequence occurrences are determined from the corpus, which is used to train the system. The training process starts by accepting the training corpus and the n-value as inputs. For each sentence in the corpus, the sequences of POS tags of words are extracted. For each unique sequence of POS tags, the probability of the occurrence of the sequence is computed using n-gram models. The n-gram probabilities of POS tag sequences stored in the permanent repository are accessed to check grammatical errors in a given sentence. The probability p_{st} of the correctness of the POS tag sequence of words in a given sentence construction is computed as:

$$p_{st} = \prod_{i=1}^{n} p_{t_i} \qquad (2)$$

where n is the number of POS tags extracted in the sentence. Sentence with higher values of p_{st} are considered to be having a valid sequence of words whereas those with low values are regarded as having unlikely sequence of words. Finally, the decision is made based on a threshold value set by empirical method. The training process is illustrated in Figure 3.

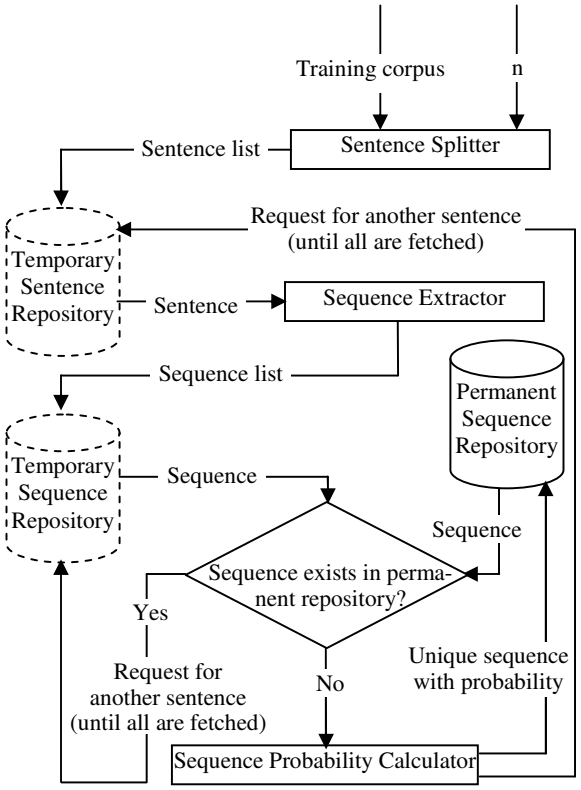

Figure 3: A flowchart of the training process for checking sequences of words.

3.3 Adjective-Noun-Verb Agreements

The agreements between words serving various grammatical functions in Amharic sentence are also checked using n-gram approach. Number, per-

son and gender agreements are checked at this phase. We perform this task by analyzing the four slots representing linguistic information about words as discussed in Section 3.1. Since the values for number, person, and gender depends on the word class, the POS tag information is required. Thus, for each word in the corpus, we extract such information as <slot1,slot2>, <slot1|slot3> and <slot1|slot4> where slot1, slot2, slot3 and slot4 represent POS tag, number, person and gender information, respectively. Given the POS tag *w* of a word, the sequence probability p_a of adjective-noun-verb agreement for a slot is computed as:

$$p_a(w_n v_n \mid w_1 v_1 ... w_{n-1} v_{n-1}) = \frac{count(w_1 v_1 ... w_n v_n)}{count(w_1 v_1 ... w_{n-1} v_{n-1})} \quad (3)$$

where *v* is the value of the slot. The n-gram probability values for each unique pattern was computed and stored in a permanent repository which would be later accessed to adjective-noun-verb agreements in a given sentence. The probability p_{sa} of the correctness of the adjective-noun-verb agreements in a given sentence is then computed as:

$$p_{sa} = \prod_{i=1}^{n} p_{a_i} \quad (4)$$

3.4 Adverb-Verb Agreement

Amharic adverbs usually come before the verb they modify. When adverb appears in the sentence it usually modifies the next verb that comes after it. There could be a number of other words in between the adverb and the verb, but the modified verb appears next to the modifier before any other verb in the sentence. As Amharic adverbs are few in number, adverb-verb agreement was not checked in the previous phases. To check time agreement between the adverb and the verb, the tense for the verb that the adverb modifies should be identified. In this work, we considered four different types of tenses: perfective, imperfective, jussive/ imperative and gerundive. The pattern of time adverbs associated with each tense type was extracted from the corpus and stored in repository. Whenever these time adverbs are found in the sentence to be checked, the tense type of the next verb is extracted by using morphological analysis. If the tense type extracted from the given sentence matches with an adverb-tense pattern in the repository, the adverb and the verb are considered to have correct agreement. Otherwise, it is reported as grammatically incorrect sentence.

4 Experiment

4.1 The Corpus

We used Walta Information Center (WIC) news corpus which contains 8067 sentences where words are annotated with POS tags. We used 7964 sentences for training and the remaining for testing. In addition, to test the performance of the system with grammatically wrong sentences, we also used manually prepared sentences which are grammatically incorrect.

4.2 Test Results

In order to test the performance of the grammar checker, we are required to compute the number of actual errors in the test set, number of errors reported by the system and the number of false positives generated by the system. These numbers were then used to calculate the precision and recall of the system as follows.

$$precision = \frac{number\ of\ correctly\ flagged\ errors}{total\ number\ of\ flagged\ errors} * 100\% \quad (5)$$

$$recall = \frac{number\ of\ correctly\ flagged\ errors}{total\ number\ of\ grammatical\ errors} * 100\% \quad (6)$$

Accordingly, we tested the system with simple and complex sentences where we obtained experimental results as shown in Table1.

Type of Sentence	n-gram model	Precision	Recall
Simple	Bigram	59.72%	82.69%
	Trigram	67.14%	90.38%
Complex	Bigram	57.82%	65.38%
	Trigram	63.76%	67.69%

Table 1: Experimental results.

Experimental results were also analyzed to evaluate the performance of the system with regard to identifying various types of grammatical errors. The detection rate of the various grammatical error types is shown in Table 2.

Error type	Detection rate (%)
Incorrect word order	73
Number disagreement	80
Person disagreement	52
Gender disagreement	60
Adjective-noun disagreement	55
Adverb-verb disagreement	90

Table 2: Detection rate by error types.

4.3 Discussion

A complete system that checks Amharic grammatical errors is developed. To train and test our system, we used WIC corpus which is manually annotated with POS tags. However, we have observed that a number of words are tagged with wrong POS and many of them are also misspelled. Since Amharic is one of the less-resourced languages, to our best knowledge, there is no tool that checks and corrects the spelling of Amharic words. Although attempts have been made to correct some of the erroneously tagged words in the corpus, were were unable to manually correct all wrongly tagged words. POS tag errors cause the wrong tag patterns to be interpreted as correct ones during the training process which would ultimately affect the performance of the system. Thus, the performance of the system can be maximized if the system is trained with error-free corpus. Moreover, since the corpus is collected from news items, most of the sentences contain words which refer to third person. For this reason, occurrence of first and second person in the corpus is very small. This has affected the system while checking person disagreement. This is evidenced by the low accuracy obtained while the system detects number disagreement (see Table 2).

To our best knowledge, HornMorpho is the only tool at present publicly available to morphologically analyze Amharic words. However, the tool analyses only some specific types of verbs and nouns. Adjectives analyzed as nouns and adverbs are not analyzed at all. Since Amharic is morphologically very complex language where combinations of various linguistic information are encoded in a single word, the effectiveness of grammar checking is hugely compromised if words are not properly analyzed. Thus, the performance of the system can be greatly enhanced by using a more effective Amharic morphological analyzer.

Test results have shown that trigram models perform better than bigram models. In Amharic, head words in verb phrases, noun phrases, adjective phrases are located at the end of the phrases (Yimam, 2000). This means that, for verb phrases, the nouns and adjectives for which verbs are marked come immediately before the head word (which is a verb). Likewise, sequences of adjectives modifying nouns in noun phrases come immediately before the head word (which is a noun). Thus, sequences of multiple words in phrases are better captured by trigrams than bigrams. We have also seen that grammatical errors in simple sentences are detected more accurately than in complex sentences. The reason is that complex sentences have complex phrasal structures which could not be directly treated by trigram and bigram models. However, the performance of the system can be improved by using a parser that generates phrasal structures hierarchically at different levels. We can then systematically check grammatical errors at various levels in line with the parse result.

5 Conclusion and Future Works

Amharic is one of the most morphologically complex languages. Furthermore, it is considered to be less-resourced language. Despite its importance, these circumstances lead to unavailability of efficient NLP tools that automatically process Amharic texts at present. This work is aimed at contributing to the ever-increasing need of developing Amharic NLP tools. Accordingly, the development of Amharic grammar checker using morphological features and n-gram probabilities is presented. In this work, we have systematically treated the morphological features of the language where we represented grammar dependency rules extracted from the morphological structures of words.

However, lack of error-free corpus and effective morphological analyzer are observed to be affecting the performance of the developed grammar checker. Thus, future works are recommended to be directed at improving linguistic resources and developing effective NLP tools such morphological analyzer, parser, spell checker, etc. for Amharic. The efficiency of these components is crucial not only for Amharic grammar checking but also for many Amharic NLP applications.

References

Amare, G. (2010). ዘመናዊ የአማርኛ ሰዋስው በቀላል አቀራረብ (Modern Amharic Grammar in a Simple Approach). Addis Ababa, Ethiopia.

Arppe, A. (2000). "Developing a Grammar Checker for Swedish"; In *Proceeding of the 12th Nordic conference on computational linguistic*. pp 9 – 27.

Domeij, R., Knutsson, O., Carlberger, J. and Kann, V. (2000). "Granska: An efficient hybrid system for Swedish grammar checking"; In *Proceedings of the 12th Nordic conference on computational linguistics*, Nodalida- 99.

Ehsan, N. and Faili, H. (2013), "Grammatical and context-sensitive error correction using a statistical machine translation framework"; *Softw: Pract. Exper*., 43: 187–206. doi: 10.1002/spe.2110.

Gasser, M. (2011). "HornMorpho: a system for morphological analysis and generation of Amharic, Oromo, and Tigrinya words"' In *Proceedings of the Conference on Human Language Technology for Development*.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (2013); *Ethnologue: Languages of the World, Seventeenth edition.* Dallas, Texas: SIL International.

Liddy, E. D. (2001) "Natural language processing", In *Encyclopedia of Library and Information Science*, 2nd Ed. Marcel Decker, Inc.

Mozgovoy, M. (2011). "Dependency-based rules for grammar checking with LanguageTool", *Federated Conference on Computer Science and Information Systems (FedCSIS*) Sept. 18-21, 2011, pp. 209-212, Szczecin, Poland.

Mudge, R. (2010). "The design of a proofreading software service"; In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*. pp 24-32, Stroudsburg, PA, USA.

Naber, D. (2003). "A Rule-Based Style and Grammar Checker"; PhD Thesis, Bielefeld University, Germany.

Richardson, S. (1997). "Microsoft Natuaral language Understanding System and Grammar checker"; *Microsoft*, USA,.

Schmidt-Wigger, A. (1998). "Grammar And Style Checking in German"; In *Proceedings of CLAW*. Vol 98.

Shaalan, K. (2005). "Arabic Gramcheck: A Grammar Checker for Arabic"; The British University in Dubai, United Arab Emirates.

Tsuruga, I. and Aizu W. (2011). "Dependency-Based Rules for Grammar Checking with LanguageTool". Maxim Mozgovoy. University of Aizu. IEEE. Japan, 2011.

Xing, J., Wang, L., Wong, D. F., Chao, S., and Zeng, X. (2013). "UM-Checker: A Hybrid System for English Grammatical Error Correction"; In *Proceedings of CoNLL-2013*, vol. 34.

Yimam, B. (2000). የአማርኛ ሰዋስው (Amharic Grammar). Addis Ababa, Ethiopia.