# Evaluation of Corpora, Resources and Tools for Amharic Information Retrieval

Tilahun Yeshambel[1(✉)], Josiane Mothe[2], and Yaregal Assabie[3]

[1] IT PhD Program, Addis Ababa University, Addis Ababa, Ethiopia
tilahun.yeshambel@uog.edu.et
[2] INSPE, Univ. de Toulouse, IRIT, UMR5505 CNRS, Toulouse, France
josiane.mothe@irit.fr
[3] Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia
yaregal.assabie@aau.edu.et

**Abstract.** Amharic is the working language of Ethiopia. It is the second-most commonly spoken Semitic language in the world next to Arabic. Amharic is morphologically complex and under-resourced, which poses tremendous challenges for natural language processing. The development of fully functional Amharic text processing applications is a non-trivial task for researchers and developers. Despite attempts to develop some applications, lack of standards in corpus collection and resource development resulted in the problem of interoperability. The aim of this paper is to present and evaluate the accessibility of Amharic corpora, resources and tools with the purpose of highlighting the status of Amharic language processing applications. We present available resources and linguistic tools, assess their usability and effectiveness, investigate the implications of the morphological complexity and put the way forward in the development of Amharic text processing applications.

**Keywords:** Amharic language · Amharic NLP tools · Amharic resources · Challenges of Amharic language processing · Morphological complexity

## 1 Introduction

Digital information is available in different formats such as text, audio, image, or video. Powerful and sophisticated applications are mandatory in order to manage all these digital information. With the huge and continuously increasing amount of information available on the Web, languages for representing data and knowledge occupy a central place in managing this tremendous quantity of data (Mihalcea and Mihalcea 2001). For this purpose, Natural Language Processing (NLP) tools and resources play great roles to improve various automatic computational tasks. The effectiveness of NLP tools, the sizes and qualities of resources have significant impacts on the performance of Information Retrieval (IR), Information Extraction (IE), Sentiment Analysis (SA), Machine Translation (MT), Question-Answering (QA), Named Entity Recognition (NER), Relation Extraction (RE), and Text Classification (TC) and other applications (Jochim 2013).


M. A. Delele et al. (Eds.): ICAST 2020, LNICST 384, pp. 470–483, 2021.
https://doi.org/10.1007/978-3-030-80621-7_34

NLP tools are used for classification of sounds of a language at phonological level, analysis of word components such as suffix, prefix and root at morphological level, word analysis such as lexical meaning and part-of-speech at lexical level, sentence analysis at syntactic level, disambiguation of word in the context at semantic level, etc. (Jurafsky and James 2000). Developing NLP tools and resources for languages is not trivial especially for morphologically complex languages.

Amharic is the working language of Ethiopia currently having a population about 110 million (countrymeters 2020). It is the second-most commonly spoken Semitic language in the world next to Arabic. The growth of Amharic digital data accelerates the demand for technologies and NLP tools for on-line data processing. However, Amharic is still considered as under-resourced language since there are few attempts made to develop Amharic NLP tools, corpora and applications. Furthermore, the complex morphology of Amharic has hindered the development of NLP applications for the language. Although the number of resources has been increasing and the performance of the systems is improving since 2000, the existing tools and resources have not been reviewed, assessed, evaluated, compared and presented in a systematic way till now.

The purpose of this paper is to present state-of-art linguistic tools and resources on Amharic. Our contribution also relies on a first evaluation and comparison of these resources so that their quality and standard is better understood, which facilitates research and development in the field of Amharic NLP. More specifically, this paper explains characteristics of Amharic language and provides an extensive overview of state-of-the-art linguistic tools and resources. We present the key characteristics of Amharic morphology as it plays a central aspect for the development of many NLP applications. Tools and resources are also practically evaluated by conducting some preliminary experiments.

The rest of this paper is organized as follows. Section 2 provides background information about Amharic language and its specificities where we examine the orthography and morphology of the language. Section 3 discusses Amharic corpora and resources available digitally along with Amharic text processing tools. In Sect. 4, we present our evaluation on accessible corpora, resources and tools. Finally, we make our conclusion in Sect. 5.

## 2 Amharic Language

Amharic is the working language of Ethiopia and serves as a common communication language among different language speakers throughout the country. A wide variety of literature including religions, fictions, poetries, plays, newspapers, magazines, businesses, etc. are produced in Amharic. Amharic is the family of Semitic language[1] and the most widely used language in Ethiopia having a population of around 110 million (countrymeters 2020). It is also the second most spoken Semitic language in the world after Arabic (Gamback et al. 2006). The language has its own 34 base characters where each of them has 7 forms/orders making them total of 238 characters. In addition, there are also dozens of labialized characters, 10 unique punctuation marks and 20 unique digits (Yaqob 1997). Each Amharic character is grouped into either consonantal or syllabary writing system and its writing system is from left to right. There are no upper and

[1] Other Semitic languages are for example: Arabic, Tigrinya, Hebrew, Geez, etc.

lower case letter variations and no conventional cursive form words in Amharic writing system.

Amharic consonants which mostly carry the semantic core of a word can form the root of the word. The Amharic root is a series of base character while a word is a collection of phonemes. Word can be a single morpheme or contain several of them. Different types of prefixes and suffixes can be attached to the stem to form the inflected words. Words are formed by modifying the root itself internally and not simply by the concatenation of affixes to word roots (Shashirekha and Gashaw 2016). Amharic has a complex morphology. For instance, according to the experimental report of Assabie (2017), tens of thousands of words can be generated from a given Amharic verbal root. Phonemes, morphemes, roots, stems, and surface words are word units of Amharic. Its words formation involves affixation, reduplication, changing the form of character in a stem (Argaw *et al* 2004). Surface word can be formed by changing stem form or by adding affixes that are usually used for changing the tense, number, gender and case (subjective, objective, possessive) of a word. Amharic word classes undergo complex derivations and inflections. From a given word class, we can generate many words that have similar or different word classes. Assabie (2017) reported that Amharic nouns can be derived from verbal roots by in-fixing vowels between consonants, adjectives by suffixing bound morphemes, stems by prefixing or suffixing bound morphemes, stem-like verbs by suffixing bound morphemes, and nouns by suffixing bound morphemes and compound words. Amharic nouns can also be inflected with number, definiteness, gender, objective case, and possessive case. Amharic adjectives can be derived from verbal roots by in-fixing vowels, nouns by suffixing bound morphemes, stems by suffixing bound morphemes and compound words of nouns and adjectives by affixing vowels. Amharic adjectives can be inflected with number, definiteness, gender, and cases (objective, possessive, etc.). Amharic verbs have even more complex morphology than other word classes. They can be derived from verbal roots, verbal stems, compound words of stem and verbs, sub words and verbs. They can also be inflected for person, gender, number, case, tense, and mood. From a given Amharic root, it is possible to generate a number of stems out of which a large number of surface words can be derived. Furthermore, preposition and negative marker can be attached as affixes on different word classes to generate surface words (Assabie 2017).

## 3 Amharic Corpora, Resources and Tools

### 3.1 Corpora

Amharic corpora and resources are required for the evaluation purpose in different research fields such as MT, IR and text analysis in general. Some of the Amharic text corpora which are available digitally and utilized for the development of Amharic NLP tools, IR or other text-centered tasks by different researchers are presented below.

**Walta Information Center (WIC) Corpus.** this corpus is prepared by linguists at Addis Ababa University with the financial support of Walta Information Center. It is available both in Amharic characters and transcribed form called System for Ethiopic Representation in ASCII (SERA). It contains 1,065 Amharic news articles which have

200,863 words (33,408 unique words) organized in 4,035 sentences. Since the corpus contains news items, the domain of the corpus is much diversified. It includes topics like politics, economics, science, sport, religion, business, etc.

**Ethiopian Language Research Center (ELRC) Corpus.** this is the annotated version of WIC corpus. It has been annotated with Part-Of-Speech (POS) tags manually by the Ethiopian Language Research Center (ELRC) at Addis Ababa University (AAU). The corpus is tagged with 30 different POS tags (Demeke and Getachew 2006).

**Amharic Bible Corpus.** this corpus is prepared by students at AAU. It has 39 categories which contain different number of sections. It contains a total of 13,300 sentences organized into 924 documents (Bruck and Tilahun 2015).

**AAU NLP Task Force Corpus.** this corpus is prepared by language technology staff members from IT Doctorial Program at Addis Ababa University for Amharic, Afaan Oromo and Tigrigna languages with diverse content (Abate et al. 2018). The project is still on-going and the corpus is continuously being updated.

**Amharic Wikipedia Corpus.** the Amharic Wikipedia was established in 2004. Currently, it contains 13,657 articles and 15 main categories about various topics such as science, language, history, mathematics, engineering, philosophy, Ethiopia, geography, etc.

**Amharic Corpus for Machine Learning (ACML).** this corpus has been prepared by Gamback (2012). The data set consists of free texts collected from Ethiopian News Headlines (ENH), Walta Information Center (WIC) and Amharic fiction "Fikir Iske Meqabir" (FIM). It is a set of 10,000 ENH articles for a total of 3.1 million words, 1,503 words from the WIC corpus, and 470 words from FIM book. Some machine learning-based tagging experiments have been carried out on this corpus (Alemu 2006). The purpose of this corpus was to collect word frequency, prefix and suffix. This corpus is not publicly accessible.

**Amharic Adhoc Information Retrieval Test Collection (2AIRTC).** Yeshambel *et al.* (2020b) created Amharic IR test collection based on CranField and Text REtrieval Conference (TREC) format. The test collection has 12,538 domain independent documents collected from various sources, a topic set with 240 user information need, and the associated relevance *judgment*. The corpus is created to enhance researches in Amharic IR and is made publicly accessible to research community.

**Amharic Morphologically Annotated Corpora (AMAC).** Yeshambel *et al.* (2020c) created stem-based and root-based morphologically annotated corpora semi-automatically. The annotation segments surface words into affixes, stems and roots. The number of annotated documents is 6,069 full text documents which contain 72,814 sentences or 1,592,351 morphologically annotated words.

### 3.2 Resources

**Dictionary.** the Amharic machine readable dictionaries (MRDs) entries are represented by their citation forms. Some of the commonly used Amharic dictionaries are Aklilu (1987), Berhanu (2004) and Birhan (1993). Aklilu (1987) developed the Amharic–English dictionary which contains 15,000 Amharic words along with their English translations. Berhanu (2004) built the Amharic-French dictionary that contains 12,000 Amharic entries. Birhan (1993) developed Amharic-Amharic dictionary which contains 56,000 entries. This dictionary is used to handle synonyms.

**Stopword List.** Yeshambel *et al*. (2020a) built morpheme-based Amharic stopwords based on semantics and corpus statistics (frequency, mean, variance and entropy). The stopword list is created by considering the characteristics of the language from a large morphologically annotated corpus. As the root form conflates all variants of a term to a common representation, the stopword list contains roots of words. The applicability of the created stopword list was evaluated using Amharic IR system. Prior to this work, few stopword lists were created manually for a specific purpose. For example, Mindaye et al. (2010) built a stopword list of 77 entries while Eyassu and Gambäck (2005) created a stopword list with 745 entries for the purpose of stopword removal in IR system. However, the entries were selected without considering morphological characteristics and statistical information of the terms. Thus, their applicability is limited to the specific systems for which they are created.

### 3.3 Tools

NLP is used to analyze, understand, alter, or generate natural language. This can be achieved by using automated NLP tools. Few numbers of Amharic NLP tools were developed using rule-based and machine learning approaches for various purposes. Some of the important Amharic NLP tools are text preprocessing tools, stemmers, POS taggers, morphological analyzers, morphological synthesizers and parsers. Preprocessing is the primary task used to make the data ready or clean for further processing. Amharic words need to be preprocessed for further Amharic NLP application development. Text preprocessing is used to enhance IR and other text-based tasks. A stemmer aims at reducing inflectional and derivational written words form to their stem or base form. It removes prefixes, suffixes, in fixes, circumfixes and redundant character from a given word. POS taggers aim at marking words in a given text with relevant parts-of-speech and are used to understand a text in natural language. Morphological analysis helps to find the minimal units of word which holds linguistic information for further processing whereas morphological synthesis is the process of returning surface forms from a sequence of underlying (lexicon) forms. Parsing is the process of analyzing a given sentence by identifying its constituents. It identifies the subjects and objects of a sentence, different phrases such as noun phrases, adverbial phrases, and adjective phrases.

**Text Preprocessing Tools.** Relevant Amharic text preprocessing tools include tokenizer, normalizer, and transliterator. Tokenization is splitting a text into sentences

and then into words. Normalization is the process of transforming texts into a single canonical form which then commonly used in various text-based tasks. Text normalization requires being aware of what type of text is to be normalized and how it is to be processed after wards. In Amharic writing system there are characters with the same pronunciation but different symbols which are called homophones. Amharic base characters having such property are {አ /ʔə/ and ዐ /ʔə/}, {ሠ /sə/ and ሰ /sə/}, {ሀ /hə/, ሐ /hə/, ኀ /hə/ and ኸ /hə/}, and {ጸ /ts'ə/ and ፀ /ts'ə/}. Amharic character normalization involves changing Amharic character with the same phone into one canonical orthographic form. Transliteration is required as Amharic texts are published using Ethiopic script and a variety of fonts. Some of which may not be Unicode compliant. In order to simplify the analysis and to have a unified representation of texts, Amharic texts are transliterated into American Standard Code for Information Interchange (ASCII) representation. Amharic Unicode representations such as UTF-8 become common and easy to analyze text. However, for NLP tasks, transliterated texts are preferable. In Amharic fonts, except Unicode supported ones, characters which have diacritic markings need more than one byte in their internal representation so as to use one byte for the basic character and additional one byte for the diacritic marking. For example, ሉ /lu/needs two bytes to represent internally, one byte for ለ /lə/and another byte for "-". This creates difficulty to use Amharic character/Fidel that has diacritic markings by considering it as a single unit. For this reason, first the Amharic texts need to be changed in to Unicode representation and then transliterated it into ASCII letters (Yaqob 1997). As for Latin languages, tokenization can be done by considering Amharic punctuations and space between words. Normalization is achieved my mapping each homophone character to a single orthographic canonical form. Due to such simplicity, researchers develop their own tokenizer and normalizer when a need arises. However, a standard canonical form is not yet set by linguists. As a result, researchers randomly choose one of the orthographic forms. Transliteration can also be done by mapping the Amharic characters into ASCII letters using a mapping table. ASCII representation of consonant and vowels for transliteration is not standardized and researchers develop their own transliteration rules. One of the most commonly used transliteration rule is SERA (Gasser 2011).

**Stemmers.** Alemayehu and Willett (2002) built a rule-based Amharic stemmer and evaluated the effectiveness of stemming for Amharic information retrieval (Alemayehu and Willett 2003). The prefixes and suffixes were removed and took into account letter inconsistency and reiterative verb forms. The stemmer was tested on 1,221 words and performs 95.9% accuracy. The compression rates of stem and root processing are 50.3% and 58.6%, respectively. The performance of word-based, stem-based, and root-based retrieval was compared using 40 Amharic queries against 548 Amharic documents, and it was reported better recall levels for stem and root-based retrieval over word-based. Alemu and Lars (2007) developed a rule-based stemmer. The aim was to solve the problem of stemming Amharic words and reducing them to their base/stem forms for Cross Lingual Information Retrieval (CLIR) applications. A total of 65 rules were constructed on the entire Amharic morphology. The experiment was conducted on ACML data set. The correctness of the stemmer was tested through Amharic-English, Amharic-French, and Amharic-Amharic dictionaries. An accuracy of 60% for the old fashioned fiction text and 75% for the news articles were achieved. Mengistu (2009) developed an Amharic

stemmer using peak and plateau, entropy and complete word methods. A corpus containing 6, 270 words was prepared for training and testing the methods. The corpus was divided into 80:20 for training and testing, respectively. The experimental results indicate that the peak and plateau method performed 71.8% accuracy and the entropy and complete word methods performed 63.95% and 57.99% accuracy, respectively. Tekalign (2014) developed an Amharic text stemmer using the rule-based Longest-Match Method. The stemmer can remove affixes using Amharic Nyala font directly without transliteration. It removes prefix, infix and suffix from a given word to get the stem of each word. It was evaluated manually using stemmed 1,500 words, 310 prefixes and 611 suffixes and performed 85% accuracy.

**POS Taggers.** Getachew (2001) developed an Amharic POS tagger using Hidden Markov Model. The tag set was designed with words as the smallest units for tagging without splitting a word into its morphemes (prefixes, a stem and suffixes). The tagger was tested with one page of text using a set of 25 tags and scores accuracy at about 90%. Sisay (2005) used 10 of the tag sets along with Aklilu (1987) dictionary for verification purpose. The experiment was done on five news articles that had been annotated manually to evaluate the stochastic model based on conditional random fields using 5-fold cross-validation. The accuracy of the system was 74%. Gamback et al. (2009) developed a tagger using TnT, SVMTool and Mallet approaches with three different tag sets. Detailed experiment was carried out on ELRC dataset. Test results reported the accuracies of 85.56%, 88.30%, and 87.87% for TnT, SVM and MaxEnt, respectively. Gebrekidan (2010) developed a tagger using knowledge of Amharic morphology and machine learning algorithms with a set of 31 tags. The objective was to improve the performances of taggers developed by Getachew (2001), Sisay (2005), Demeke and Getachew (2006), and Gamback et al. (2009). CRF++ was used for segmenting data whereas LIBSVM was used for classification and regression purpose. The experiment was done by CRF, SVM, and TnT and Brill machine learning algorithms and 10-fold cross validation using ELRC corpus which provided average accuracy of 90.95%, 90.43%, 87.41% and 87.09%, respectively. The tagging was improved compared previous works due to better preprocessing tasks and the use of vowel patterns and root as features, machine learning algorithms and parameter tuning. The error analysis of CRF and SVM was between 39% and 45% using confusion matrices. Yifru et al. (2011) conducted experimental research to investigate the best method for under-resourced and morphologically rich languages. Disambig, Moses, CRF++, SVMTool, MBT, and TnTwere the tagging strategies implemented in this research. The experiment was conducted using 31 tags on ELRC corpus which has been divided into training, development and evaluation test sets in the proportion of 90:5:5. Accuracies of 75.1% and 74.4% were obtained with SVM and Disambig, respectively. High gain is observed on TnT strategies. It was reported that HPR and hybrid combination methods were promising to improve POS tagging performance for under-resourced languages.

**Morphological Analyzers.** Tesfaye (2002) used unsupervised learning approach based on probabilistic models to extract stems, prefixes, and suffixes for building a morphological dictionary. On top of that, a modified version of Harris's algorithm of successor frequency was applied to detect plausible word break points. An experiment was carried

out on the transcribed first 50 pages of Amharic text containing 5,736 words. A precision of 0.95 and a recall of 0.90 were obtained. Sisay and Haller (2003) investigated the morphology of Amharic verbs in the context of machine translation and presented the implementation of morphological analyzer for Amharic using Xerox Finite State Tools (XFST) method. The different classification schemes for Amharic verbs that have been forwarded are discussed followed by the implication. It was stated that morphological analysis for Amharic with XFST can handle most of the morphological phenomena except some derivation processes which involve simultaneous application of both stem inter-digitations and reduplication. The experiment was carried out on 3.1 million words of Amharic news text. They achieved 0.88–0.94 recall and 0.54–0.94 precision depending on the word-class. Sisay (2005) developed a morphological analyzer and tagger using conditional random. The work dealt with bound morphemes of prepositions, conjunctions, relative markers, auxiliary verbs, negation markers and coordinate conjunctions, but leaves out other bound morphemes such as definite articles, agreement features such as gender and numbers, case markers (objective, possessive), etc., and considered them to be part of the word. The best analyzing result (84%) was obtained by using character, morphological and lexical features. Amsalu and Gibbon (2006) developed Amharic analyzer using XFST. The experiment was conducted on a text of 1,620 words from an Amharic Bible corpus. The system provided precisions of 0.54, 0.94 and 0.81 for recall levels of 0.94, 0.85 and 0.88 for verbs, nouns and adjectives, respectively. Gasser (2011) developed a morphological analyzer and synthesizer using a rule-based approach and Akililu's (1987) dictionary. The analyzer marks-up the stems, roots, and POS. It uses lexical Finite State Transducer (FST). Prefix and suffix FSTs were concatenated onto the stem FST to create the full verb morph tactic FST. The remaining FSTs implemented alternation rules that apply to the word as a whole, including allomorphic rules and general phonological or orthographic rules. The analyzer was tested on 200 Amharic verbs, nouns and adjectives that were selected randomly. On this data set the analyzer performed 99% accuracy for Amharic verbs and 95.5% accuracy for Amharic nouns and adjectives. Mulugeta and Gasser (2012) developed a morphological analyzer for Amharic verbs using supervised machine learning approach. The training data used to formulate morphological rules was prepared manually from 216 romanized verbs. After training, 108 rules for affix extraction, 18 rules for root template extraction and 3 rules for internal stem alternation were used. CLOG learns rules as a first order predicate decision list. The approach was tested on 1,784 Amharic verbs and its accuracy is 86.99%. Abate and Assabie (2014) developed morphological analyzer using memory-based supervised machine learning approach. The goal was to extract stems from nouns and adjectives, and roots from verbal stems. Morphological analysis was used to classify the grammatical functions of morphemes and morphological structure of morphologically inflected words. The morphological analysis component comprises the feature extraction to deconstruct a given word, morpheme identification to split and extrapolate, stem and root extraction to label segmented inflected words with their morpheme functions. TiMBL was used as a learning tool. The experiment was conducted on 181 romanized verbs and 841 nouns. 8,075 instances were extracted from nouns and adjectives. Leave-one-out (LOO) cross-validation for IB1 algorithm and 10-fold cross-validation technique was used to test the performance of the system with IB1 and IGtree classifier engines

algorithms. An average accuracy of 96.40% was achieved with IB1 algorithm on LOO method. On the other hand, the system provided 93.59% and 82.26% accuracy using IB1 and IGtree algorithms, respectively using 10-fold method on previously unseen verbs and nouns.

**Parsers.** Abiyot (2000) developed an Amharic parser for verbs and their derivations. A knowledge-based system that parses verbs, and nouns derived from verbs was designed. Root pattern and affixes were used to determine the lexical and inflectional category of the words. The parser was tested on 200 verbs and nouns; the accuracy was 86% for the verbs and 84% for the nouns.

**Synthesizers.** Lisanu (2002) developed word synthesizer for Amharic perfect verbs. A combination of rule based and neural network approaches were used. The experiment was conducted using 14 roots and some affixes which were selected by domain experts on type A, type B and type C perfective verb forms. The rule-based approach performed near to 100% accuracy whereas the accuracy of the neural network approach is 81.48%. The accuracy of the neural network system in predicting type A, B and C perfective verb forms is 80%, 25% and 100%, respectively. On the other hand, Gasser (2011) developed Amharic word synthesizer. FST and rule-based approaches were used for implementation. An experiment was run on major verb root classes to generate 10 to 25 verbs. The system performed 100% accuracy on 330 test data.

## 4 Evaluation of Corpora, Resources and Tools

We reviewed the aforementioned corpora, resources and tools where it was understood that there was no standard that governs the format and general characteristics that the resource should have. Each researcher was found to construct resources in line with the specific needs of the NLP tasks. Moreover, the reported performances of NLP tools were based solely on the non-standard resources. Accordingly, it was difficult to make fair comparisons among similar NLP tools. On the other hand, most of the developed NLP tools are not publicly available which hinders the evaluation process. Taking this scenario into consideration, we conducted the evaluation first by qualitative analysis and then using quantitative means. Thus, accessible tools are evaluated based on their performance, accessibility, usability, utility (plug into other system), portability and maintainability. These parameters are selected since they are some of the most common techniques to evaluate computational resources and tools.

### 4.1 Qualitative Evaluations

**Corpora.** the size of most of the existing corpora is not sufficient for applications such as IR, NLP tools and machine learning. Acceptable and reputable IR experiments are based on ten thousands and million number of various documents. On top of that, the corpora are not preprocessed. They are mainly collections of documents from which formatting tags are removed. Hence, both the quantity and the quality of many corpora are not sufficient for the development of NLP tools and applications. On the other hand, most of the corpora are not publicly accessible.

**Resources.** Amharic dictionaries are very limited in size compared to the richness of the language and also when compared to dictionaries in other languages. Amharic dictionaries are not sufficient to verify lemmatization, or to handle synonyms. Their usability is very limited as they are in the format of spreadsheet files. So far, they are used by very few researchers. Alemu and Asker (2006) used Berhanu's (2004) dictionary to evaluate the output of Amharic stemmers and for CLIR applications to reduce query terms in the source language to the exact corresponding citation form in the MRD. Aklilu (1987) dictionary has been used by Gasser (2011) and Sisay (2005) to develop Amharic morphological analyzer, POS tagger, stemmer and machine translator. Majority of the existing Amharic stopwords are not collected and organized scientifically (Mindaye et al. 2010; Eyassu and Gambäck (2005)). Researchers collected stopwords lists based on their language knowledge without validation by linguists, which is an important step considering the morphological complexity of Amharic.

**Tools.** the functionalities of the existing Amharic NLP tools are limited. Hence, it is very difficult to use them as a component in other applications like IR, question answering or text summarization. The existing tools are very often limited to certain types of words. For example, Mulugeta and Gasser (2012) developed morphological analyzer for Amharic verbs. On the other hand, Lisanu (2002) developed a morphological synthesizer for Amharic perfective verbs. Gasser's (2011) morphological analyzer is available freely on the web and in executable format. It has a command line interface, can be configured and run easily on different environment. However, this analyzer usually has not been much utilized in Amharic applications such as IR. The output of this analyzer is not suitable to integrate with IR system. The output contains more than one stems for variants. Furthermore, root and POS are represented by English while the stems are in Amharic. This by itself needs complicated preprocessing task. Taggers developed by Yifru et al. (2011) and Gebrekidan (2010) have command line interfaces. Furthermore, they are not publicly accessible and are not executable file. They are used by very few researchers (Sintayehu 2013). In general, most of the tools are not accessible which limits both the capability of evaluating them and their utility for other researchers.

Many of the existing corpora, resources and tools are not suitable for Amharic IR. IR experiments require large scientifically built corpora and resources. However, except Yeshambel *et al*. (2020b), all existing corpora are simply collections of small documents without topic set and relevant judgments. They are not used to test retrieval system automatically. As a result, in IR previous experiments (Mindaye et al. 2010; Wordofa 2013; Asefa 2013; Bruck and Tilahun 2015) the respective researchers created few user query topics relevant to the corpus they used for. However, acceptable IR experiments should be carried out by running 50+ queries on a large corpus. On the other hand, some stopword lists contain variants of stopwords rather than basic stems (Mindaye et al. 2010; Eyassu and Gambäck (2005)). However, there are many variants for a given Amharic stopword. Therefore, Amharic IR system is not able to remove stopwords from index terms and query terms using many of the existing lists. Furthermore, the existing Amharic dictionaries are designed based on citation form. The citation form of an Amharic word is different form its stem or root. Therefore, the existing dictionaries are not useful in IR such as CLIR system.

### 4.2 Quantitative Evaluations

Tools have been evaluated by the researchers who developed them but neither on the same collection nor the same task. Indeed, there is no reference collection as it is the case in other languages such as English or Arabic for example. Therefore, we started to develop an evaluation framework to be able to compare the various tools and resources. To evaluate preprocessing and NLP tools, we started with small dataset consisting of few words (verbs, nouns and adjective) with prefix, suffix and circumfix. Then, we continuously increased the dataset to evaluate how the system functions different scenarios. We observed that Gasser's (2011) morphological analyzer performs well if the word is available in Aklilu (1987) dictionary. However, if the word is not in the dictionary the performance decreases. Preliminary experiments on taggers developed by Yifru et al. (2011) and Gebrekidan (2010) were conducted using ELRC corpus as training and testing data set. They performed well on the test data set. However, when using any other dataset, their performances decrease drastically. This may be due to the size of the training corpus. The stemmer developed by Alemayehu and Willett (2002) is cited in some researches (Munye and Atnafu 2012; Wordofa 2013). However, the fully functioning system is not currently accessible. Only parts of the stemmer which are prefix and suffix removal are available. Table 1 shows a summary of evaluations on accessible tools. Since it is very difficult to make evaluation automatically and repeatable, the evaluations were made manually by building a small data set.

**Table 1.** Summary of evaluations on Amharic NLP tools

| Tools | Developers | Accuracy (%) | Dataset (words) |
|---|---|---|---|
| Taggers | Gasser (2011) | 64.91 | 200 |
| | Yifru *et al.* (2011) | 86.75 | 200,863 |
| | Gebrekidan (2010) | 87.30 | 200,863 |
| Stemmer | Alemayehu and Willett (2002) | 41.40 | 75 |
| | Tekalign (2014) | 25.40 | 75 |
| Synthesizer | Lisanu (2002) | 82.72 | 375 |

Morphological processing is one of the fundamental tasks of retrieval systems. The aim is to conflate variants to their common form. The retrieval effectiveness of IR system depends largely on the accuracy of linguistic processing tools such as stemmer and analyzer. However, as shown in Table 1, the performances of stemmers are low. Some of them over-stem (Tekalign 2014) and the others under-stem (Alemayehu and Willett 2002). In case of over-stemming, many non-related words are conflated to a common stem. On the contrary, many variants cannot be conflated to a common stem in case of under-stemming. These will significantly affect the precision and recall of Amharic IR system. The retrieval performance of some IR system based on the existing tools is low. For example, using the stemmer developed by Alemayehu and Willett

(2002), Amharic-English CLIR system (Munye and Atnafu 2012) achieve a retrieval performance of 0.74 whereas a retrieval system developed by Alemayehu and Willett (2003) performed recalls of 0.21, 0.35 and 0.64 at a cut of 5, 10 and 20, respectively. Moreover, identification of nouns in documents and user query is important in Amharic IR. However, the performance of the existing taggers on a corpus other than training data set is poor.

## 5 Conclusion

The morphological complexity of Amharic has brought challenges in developing useful IR resources, corpora, and tools. This paper presents efforts that have been attempted so far along with evaluations of the contributions. In comparison with other resourceful languages, few resources and tools have been developed for Amharic. However, even most of the existing resources and tools are not publicly accessible. Only few of the existing Amharic tools and resources are used in academia for research purposes. The major obstacles that hinder the progress on the development of Amharic IR are the complex morphology of the language, lack of sufficiently annotated corpus and lack of standards in resource construction and application development. Empirical evaluations conducted on accessible tools show that most of the tools are generally domain dependent and their performance drastically decreases when tested on out-of-domain scenarios.

## References

Abate, M., Assabie, Y.: The development of Amharic morphological analyzer using memory based learning. In: Ethiopia Information Communication Technology Annual Conference 2014, pp. 11–18 (2014)

Abate, S.T., et al.: Parallel corpora for Bi-Lingual English-Ethiopian languages statistical machine translation. In: Proceedings of the 27th International Conference on Computational Linguistics, New Mexico, USA, pp. 3102–3111 (2018)

Abiyot, B.: Developing automatic word parser for Amharic verbs and their derivation. Master's thesis, Addis Ababa University, Addis Ababa (2000)

Aklilu, A.: Amharic-English dictionary, 1st edn. Kuraz Printing Press, Addis Ababa (1987)

Alemayehu, N., Willett, P.: Stemming of Amharic words for information retrieval. Lit. Linguist. Comput. **17**(1), 1–17 (2002)

Alemayehu, N., Willett, P.: The effectiveness of stemming for information retrieval in Amharic. Program: Electron. Libr. Inf. Syst. **37**(4), 254–259 (2003)

Alemu, A., Asker, L.: Amharic-English information retrieval. In: Peters, C. et al. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 43–50. Springer, Heidelberg (2006). https://doi.org/10.1007/978-3-540-74999-8_5

Alemu, A., Asker, L.: An Amharic stemmer: Reducing words to their citation forms. In: Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, pp. 104–110. Association for Computational Linguistics (2007)

Amsalu, S., Gibbon, D.: Finite state morphology of Amharic. In: 5th Recent Advances in Natural Language Processing, pp. 47–51 (2006)

Argaw, A.A., Asker, L., Cöster, R., Karlgren, J.: Dictionary-based Amharic – English information retrieval. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 143–149. Springer, Heidelberg (2005). https://doi.org/10.1007/11519645_14

Asefa, G.: Ontology-based semantic indexing for Amharic text in football domain. Master's thesis, Addis Ababa University, College of Natural Science, Ethiopia (2013)
Assabie, Y.: Development of Amharic morphological analyzer. Technical report, Ethiopia Ministry of Communication and Information Technology, Addis Ababa, Ethiopia (2017)
Berhanu, A.: Amharic-Français Dictionnaire, 1st edn. Shama Books, Addis Ababa (2004)
Birhan, K.: YeAmarinja Mezgebe Qalat. Ethiopian Languages Study and Research Center. 1st edn. Artistic publisher, Addis Ababa (1993)
Bruck, A., Tilahun, T.: Bi-gram based query expansion technique for Amharic information retrieval system. Int. J. Inf. Eng. Electron. Bus. **7**(6), 1 (2015)
Countrymeters: Ethiopian Population. http://countrymeters.info/en/ethiopia. Accessed 11 Sept 2020
Demeke, G., Getachew, M.: Manual annotation of Amharic news items with part-of-speech tags and its challenges. In: Ethiopian Languages Research Center Working Papers, vol. 2, pp.1–16 (2006)
Eyassu, S., Gambäck, B.: Classifying Amharic news text using self-organizing maps. In: Association for Computational Linguistics, pp. 71–78. Ann Arbor, Michigan (2005)
Gamback, B.: Tagging and verifying an Amharic news corpus. Lang. Technol. Norm. Less-Resour. Lang. **79**, 79–84 (2012)
Gamback, B., Olsson, F., Atelach, A., Asker, L.: Methods for Amharic part-of-speech tagging. In: Proceedings of the first Workshop on Language Technologies for African languages, pp. 104–111. Association for Computational Linguistics (2009)
Gamback, B., Sahlgren, M., Atelach, A., Lars, A.: Applying machine learning to Amharic text classification. In: WOCAL 5: 5th World Congress of African Linguistics. Citeseer (2006)
Gasser, M.: Hornmorpho: A System for morphological processing of Amharic, Afaan Oromo, and Tigrinya. In: Conference on Human Language Technology for Development, Alexandria, Egypt (2011)
Gebrekidan, B.: Part of speech tagging for Amharic. Master's thesis, University of Wolverhampton, School of Law, Social Science and Communication, United Kingdom (2010)
Getachew, M.: Automatic part of speech tagging for Amharic: an experiment using Stochastic Hidden Markov Model (HMM) approach. Master's thesis, Addis Ababa University, Addis Ababa (2001)
Jochim, C.: Natural language processing and information retrieval methods for Intellectual Property Analysis. Ph.D. thesis, University of Stuttgart, Germany (2013)
Jurafsky, D., James, H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech, 1st edn. Prentice Hall, USA (2000)
Lisanu, K.: Design and development of automatic morphological synthesizer for Amharic perfective verb forms. Master's thesis, school of Information Studies for Africa, Addis Ababa University, Addis Ababa (2002)
Mengistu, G.: Automatic stemming for Amharic text: an experiment using successor variety approach. Master's thesis, Addis Ababa University, Addis Ababa (2009)
Mihalcea, F., Mihalcea, I.: Word semantics for information retrieval: moving one step closer to the semantic web. In: Tools with Artificial Intelligence, Proceedings of the 13th International Conference on Tools with Artificial Intelligence, pp. 280–287. IEEE (2001)
Mindaye, T., Atnafu, S., Redwan, H.: Searching the web for Amharic content. Int. J. Multimed. Process. Technol. (JMPT) **1**(1), 318–325 (2010)
Mulugeta, W., Gasser, M.: Learning morphological rules for Amharic verbs using inductive logic programming. In: Workshop on Language Technology for Normalization of Less-Resourced Languages (SALTMIL8/AfLaT2012), pp.7–12 (2012)
Munye, M., Atnafu, S.: Amharic-English bilingual Web search engine. In: Proceedings of the International Conference on Management of Emergent Digital EcoSystems, MEDES 2012, pp. 32–39 (2012)

Shashirekha, H., Ibrahim, G.: Dictionary based Amharic-Arabic cross language information retrieval. ASCTY, pp.49–60. NATP (2016)

Sintayehu, H.: Designing an information extraction system for Amharic vacancy announcement text. Master's thesis, Informatics faculty, Addis Ababa University, Addis Ababa (2013)

Sisay, F.: Part-of-speech tagging for Amharic using conditional random fields. In: Proceedings of ACL-2005 Workshop on Computational Approaches to Semitic Languages, Ann Arbor, Mich, pp. 47–54 (2005).

Sisay, F., Haller, J.: Application of corpus-based techniques to Amharic texts. In: Proceedings of MT Summit IX Workshop on Machine Translation for Semitic Languages (2003)

Tekalign, B.: Developing stemmer for Amharic text using longest-match method. Master's thesis, Arba Minch University, Arba Minch (2014)

Tesfaye, B.: Automatic morphological analyzer for Amharic an experiment employing unsupervised learning and auto segmental analysis approaches. Master's thesis, informatics faculty, Addis Ababa University, Addis Ababa (2002)

Wordofa, M.: Semantic indexing and document clustering for Amharic information retrieval. Master's thesis, Informatics faculty, Addis Ababa University, Addis Ababa (2013)

Yaqob, D.: Transliteration on the internet: the case of Ethiopic. In: Proceedings of the International Symposium on Multilingual Information Processing, Tsukuba, Japan (1997)

Yeshambel, T., Mothe, J., Assabie, Y.: Construction of morpheme-based Amharic stopword list for information retrieval system. In: the 8th EAI International Conference on Advancements of Science and Technology, Bahir Dar (2020a)

Yeshambel, T., Mothe, J., Assabie, Y.: 2AIRTC: the amharic adhoc information retrieval test collection. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 55–66. Springer, Cham (2020b). https://doi.org/10.1007/978-3-030-58219-7_5

Yeshambel, T., Mothe, J., Assabie, Y.: Morphologically annotated Amharic text corpora. In: Proceedings of 44th ACM SIGIR Conference on Research and Development in Information Retrieval, Canada (2020c). https://doi.org/10.1145/3404835.3463237

Yifru, M., Tefera, S., Besacier, L.: Part-of-speech tagging for under-resourced and morphologically rich languages: the case of Amharic. In: HLTD 2011, pp. 50–55 (2011)