Morphologically Annotated Amharic Text Corpora

Tilahun Yeshambel
IT PhD Program
Addis Ababa University
Addis Ababa, Ethiopia
tilahun.yeshambel@uog.edu.et

Josiane Mothe
INSPE, UT2J, IRIT CNRS
Univ. de Toulouse
Toulouse, France
josiane.mothe@irit.fr

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

ABSTRACT

In information retrieval (IR), documents that match the query are retrieved. Search engines usually conflate word variants into a common stem when indexing documents because queries and documents do not need to use exactly the same word variant for the documents to be relevant. Stemmers are known to be effective in many languages for IR. However, there are still languages where stemmers or morphological analyzers are missing; this is the case for Amharic which is the working language of Ethiopia. Morphological analysis is the key to derive stems, roots (primary lexical units) and grammatical markers of words such as person, tense and negation markers. This paper presents morphologically annotated Amharic lexicons as well as stem-based and root-based morphologically annotated corpora which could be used by the research community as benchmark collections either to evaluate morphological analyzers or information retrieval for Amharic. Such resources are believed to foster research in Amharic IR.

CCS CONCEPTS

• Information systems ~ Information Retrieval • Information systems ~ Document representation • Information systems ~ Dictionaries

KEYWORDS

Information retrieval; Corpus; Morphological annotation; Under-resourced language; Amharic

ACM Reference format:

Tilahun Yeshambel, Josiane Mothe and Yaregal Assabie. 2021. Morphologically Annotated Amharic Text Corpora. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'21), July 11-15, virtual event, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3404835.3463237

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463237

1 Introduction

A core component of information retrieval system is query-document matching which the system uses to retrieve document for a given query. Because queries and documents do not need to use exactly the same word variant for the documents to be relevant, search engines usually apply word stemming. Stemming is the process of conflating the variants of a word (most often inflected forms) to their stem or base responsible for the word meaning. Search engines can then treat words with the same stem as synonyms in the query-document matching process. For example, the English word variants developer, develops, developed, undeveloped, etc. may all be considered as synonyms to the stem or root "develop". Other natural language processing applications such as text classification and categorization require word stemming and rooting as pre-treatments [1].

Automatic morphological analyzers for word stemming and rooting often rely on rule-based algorithms, such as Porter's stemmer, which is the reference for English [2]. Other rule-based stemmers have been developed for different languages such as Portuguese [3] or Arabic [4]. Rule-based algorithms are, however, difficult to implement in the case of morphologically complex languages [5]. Corpus-based stemming and rooting are alternatives to rule-based stemming; they rely on word similarity, n-gram frequencies and dictionaries or lexicons [6] [7] but need resources and references to optimize their hyperparameters.

The development and evaluation of such algorithms need textual corpora and references [8] [9] [10] [11]. In the case of well studied languages such as English, evaluation forums like TREC[1] [12], CLEF[2] [13], and NTCIR[3] [14] are used to develop and evaluate these algorithms on different tasks. For digitally under-resourced languages, tools and reference corpora are usually not available, which is the case for Amharic. In such cases, annotated lexicons are possible alternatives to serve as references for automatic morphological analyzers [15]. Although Amharic is

[1] Text REtrieval Conference (http://trec.nist.gov)
[2] Conference and Labs of the Evaluation Forum (http://www.clef-initiative.eu)
[3] NII Testbeds and Community for Information access Research (http://research.nii.ac.jp/ntcir)

the working language of Ethiopia, it is still an under-resourced language characterized by lack of standard digital resources required for natural language processing and IR.

In this paper, we present a collection which consists in two lexicons of 170,000 morphologically annotated Amharic terms where both stems and roots are annotated, as well as corpora of texts where documents have been re-written using these lexicons. These texts are part of the 2AIRTC, the Amharic Adhoc Information Retrieval Test Collection where documents, queries and query relevance are provided [16].

The rest of this paper is organized as follows. Section 2 gives background information about Amharic language and its morphology. Section 3 discusses related work. In Section 4, we present the morphological annotation process, the structure of the morphological tags and the details of our resources. Section 5 discusses about the implications of using the resources in Amharic IR. Finally, we draw conclusions and highlight future directions in Section 6.

2 Amharic language

Amharic is the working language of the government of Ethiopia, currently having a population of over 110 million, and serves as the *lingua franca* of the country. It is widely used as a medium of communication in governmental, religious, educational, social and business institutions of the country. Amharic is a Semitic language and uses Ethiopic alphabet for writing. The writing system has 34 base characters which change their shapes into seven different forms due to vowels.

Owing to its Semitic characteristics, Amharic is known to have a rich and complex morphological structure. According to Assabie [17], thousands of surface words can be generated from a base form. Amharic words are inflected for person (first, second, third), gender (feminine, masculine), number (singular, plural), case (subjective, objective, possessive), definiteness (definite, indefinite), tense (past, present, future), aspect (perfective, imperfective), politeness (impolite, polite), etc. [18][31]. This is achieved by adding prefixes, infixes and suffixes.

The grammatical relations like subject, object, and syntactic information might be indicated morphologically at word level. For instance, the word ይሰጣታል /*jɨsəṭatalɨ* 'he will give her'/ is composed of the subject marker for imperfect tense ይ...አል /*jɨ...ʔalɨ*/, imperfect verbal stem ሰጥ /*səṭɨ*/ and the object marker አት /*ʔatɨ*/.

Verbs exist in perfective, imperfective, jussive, gerund, and infinitive forms. Each form has its own stem template. Stems of verbs can be classified as basic and derived stems. The basic stems of verbs are modified either internally (by inserting infix) or externally (by attaching prefix) to form derived stems. The derived stems include passive, causative, infinitive, and reduplicative. Passive stems are formed by attaching the prefix 'ተ /*tə*/' on basic stems, causative stems are derived variably by attaching the prefix 'አ /*ʔə*/', 'አስ /*ʔəsɨ*/', or 'አት /*ʔətɨ*/', and infinitive stems are formed by adding the prefix 'መ /*mə*/' on basic stems. Moreover, derived stems can be formed by reduplicating a character of basic stem. Both types of stems may be preceded by many prefixes and followed by many suffixes.

The orthography of Amharic language can be the combination of one or more functional words and inflectional morphemes. Amharic morphemes play significant roles both in morphology and syntax. Most Amharic words are composed of a basic form and many attached affixes. Prefixes can be the prepositions (ከ /*kə*/, በ /*bə*/, etc.) or genitive (የ /*jə*/), negations (አል /*ʔəlɨ*/), and conjunctions (እንደ /*ʔɨnɨdə*/, etc.) while suffixes include plural marker (ኦች /*ʔotʃɨ*/ or ዎች /*wotʃɨ*/), possessive (ኤ /*ʔe*/, ኡ /*ʔu*/, አችን /*ʔətʃɨnɨ*/, ሽ /*ʃɨ*/, etc.), or a definite marker (ኡ /*ʔu*/, ው /*wɨ*/, ዋ /*wa*/), and connectors (ና /*na*/). Gender, number, case, and definite markers can be suffixed to the stem of nouns. The sets {ኢት /*ʔitɨ*/, ዋ /*wa*/, ኡ /*ʔu*/, ው /*wɨ*/}, {ኦች /*ʔotʃɨ*/, ዎች /*wotʃɨ*/, አን /*ʔənɨ*/, እየ /*ʔijə*/, አት /*ʔətɨ*/}, {ን /*nɨ*/, ዬ /*je*/, ኤ /*ʔe*/, ዎ /*wo*/, ህ /*hɨ*/, ሽ /*ʃɨ*/, ኡ /*ʔu*/, ዋ /*wa*/, አችሁ /*ʔətʃihu*/, አችን /*ʔətʃɨnɨ*/, አቸው /*ʔətʃəwɨ*/}, and {ኡ /*ʔu*/, ዋ /*wa*/, ው /*wɨ*/, ኢቱ /*ʔitu*/, ይቱ /*jɨtu*/} are gender, number, case, and definite markers, respectively. The most common suffixes to derive adjectives are አማ /*ʔəma*/ and አዊ /*ʔəwi*/ from nouns, and ኢ /*ʔi*/ from verbs. The language has a number of lexical variations and clitics [19]. Sometimes, there is no clear demarcation between clitics and content word in the orthography. The clitics such as prepositions and conjunctions, which have syntactic roles, indicate grammatical relations with the content words. An Amharic content word can represent a phrase, a clause or a sentence [20] [21].

From a computational point of view, segmenting a word into its morphemes is very crucial in many Amharic IR and NLP applications. For instance, Amharic IR systems require words in documents and queries to be segmented correctly into their stems, roots and affixes. However, separating morphemes from surface words is a challenging task. This problem has a negative impact on the performance of different applications as it results in vocabulary mismatch problem for words generated from the same root form. Therefore, Amharic raw text needs to be morphologically analyzed to get the desired results.

3 Related work

Information retrieval and natural language processing have benefited a lot from the advancement of automatic tools relying on corpora. This has triggered a renewed interest to create annotated corpora that can be used for training and measuring the validity, accuracy and effectiveness of IR and NLP components and systems as described below.

Grubenmann *et al.* [22] built annotated Swiss German text corpus for sentiment analysis manually. The corpus consists of more than 200,000 phrases from Facebook comments and online chats. Similarly, Maher *et al.* [23] annotated an Arabic corpus for sentiment analysis manually. The corpus was annotated with five labels (positive, negative, dual, neutral, and spam). Dukes and Habash [24] built the first publicly accessible annotated Quranic Arabic corpus. The annotation involves morphological segmentation, part of speech (POS) tagging, and syntactic analysis using dependency grammar. It was carried out by automatic morphological tagging using diacritic edit-distance

followed by manual verification in online collaboration. Marcus *et al.* [25] constructed a large Treebank corpus consisting of 4.5 million English words and made it accessible to the research community. The annotation was done automatically and then corrected by human annotators. Zinkevičius *et al.* [26] built a morphologically annotated Lithuanian corpus automatically with an analyzer. The annotated corpus contains a set of morphological features such as lemmas, prefixes, and suffixes for 1 million words.

With the growing need of automatic text processing, there is also an increased interest in building Amharic corpora. A POS tagged corpus that contains 1,065 text documents having 202,671 words was built manually by the Ethiopian Languages Center at Addis Ababa University [19]. This corpus exists in both Ethiopic and Romanized version known as SERA forms. It has been used to develop a chunker [27] and a machine learning POS tagger [28]. Seyoum *et al.* [20] created the morpho-syntactically annotated Amharic Treebank using a semi-automatic approach to develop a text parser. The corpus contains 5,000 annotated sentences, out of which 1,000 sentences were manually annotated with POS tags, morphological information, and syntactic relations of words. These sentences have been used for training a machine learning system to annotate the other sentences automatically. Fifty six POS tags were compiled based on the morpho-syntactic properties of words. The universal dependence approach is applied and clitics are separated from content words manually. Gezmu *et al.* [21] built a POS tagged corpus consisting of 25,199 documents using syntactic information of words. Each word in the corpus was tagged automatically using HornMorpho analyzer [29] and manual intervention was made to correct erroneous results. The morphological analyzer generates the derived stems of non-verbal words rather than basic stems. For verbs, it generates only roots rather than stems producing incorrect representations. This leads to the generation of similar representations for many semantically non-related words. On the other hand, since the collection is simply set of annotated documents without topic set and relevance judgment, the whole corpus is not suitable for IR experiments.

Although morphological properties have been used to create POS tagged corpora, there is no morphologically annotated large Amharic corpus to date. In this work, we construct two lexicons with 170,000 words annotated with their composition based on stems and roots, as well as two corpora annotated with these morphological features.

4 Construction of morphologically annotated Amharic lexicons and corpora

The main problems facing the development of Amharic IR and NLP applications are the morphological complexity of the language and lack of resources. Thus, this work aims at addressing these important issues. In this section, we present how we created both the stem-based and root-based morphologically annotated lexicons and corpora.

4.1 Methodology

4.1.1 Data sources

The documents are part of the 2AIRTC ad hoc IR collection [16]. Recently released and made open access in the IR domain, where query and document relevance are provided along with the document collection from the 2AIRTC collection, we re-used the documents only. The 2AIRTC documents were collected from news agencies (Walta[4], Fana[5], AMM[6]), Amharic Wikipedia[7], blogs[8], web, and individuals. Documents are full text and of various genres and topics such as sport, health, religion, history, social issues, culture and economy and thus encompass a large variety of words.

4.1.2 Annotation and evaluation method

We morphologically annotated parts of the 2AIRTC documents semi-automatically. We focus on stem and root extractions since these are two core text operations in many text-based applications such as text mining, translation, IR, and question answering. An Amharic native annotator morphologically segmented unique words from the corpus into their affixes and basic stems or roots. The accuracy of the annotations was then evaluated and corrected by a linguist in the case of inappropriate annotations. This process generates the lexicons. Then, each occurrence of surface words in the corpus was replaced by its corresponding morphologically segmented word to produce the corpora. This process is detailed hereafter.

4.2 Morphological annotation process

Morphological annotation is made by extracting unique words from preprocessed Amharic documents. Documents were preprocessed by tokenizing and removing non-Amharic words such as tags for further processing. After preprocessing is made, a total of 170,000 unique words were identified from all documents. Here, the aim was to minimize errors and the time required for annotation as each word would be annotated only once. The annotation process of the stem-based and root-based corpora is shown in Figure 1.

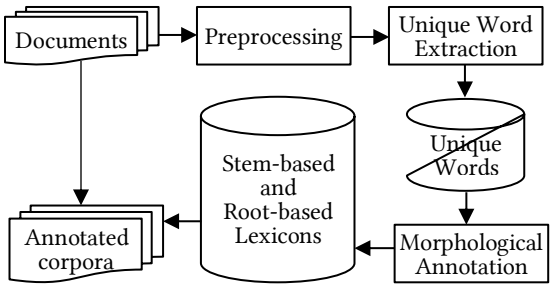

Figure 1. Morphological annotation process

[4] http://www.waltainfo.com/
[5] https://www.fanabc.com/
[6] https://www.facebook.com/AmharaMassMediaAgencyAMMA/
[7] https://am.wikipedia.org/wiki/
[8] http://www.danielkibret.com/

The stem-based and root-based morphological forms of each word were created manually by an annotator. Many of the Amharic surface words are constructed from more than one morphological segment called morphemes. Morphemes are divided into prefix, suffix, infix, stem and root. Therefore, the morphological annotation was performed by segmenting each word into its morphemes. The general structure of a morphologically annotated word *W* is:

$$[\ p\_\ ]^*\ w\ [\ \_s\ ]^*$$

where *p* is a prefix morpheme, "_" is a morphological segment marker, *w* is the root or stem of *W*, *s* is a suffix morpheme, "[…]" denotes optionality, and "*" denotes the possibility of multiple occurrences. The number of prefixes and suffixes varies from one to five [30]. For example, the word ካልተስማማናቸውም /*kalitəsimamanatʃəwimi* 'if we are not comfortable for them'/ is annotated as follows.

ከ	አል	ተ	ስማም	አን	አቸው	ም
p	*p*	*p*	*w*	*s*	*s*	*s*

In this example, the Amharic word is split into 7 morphological morphemes (3 prefixes, a stem and 3 suffixes) which are preposition (ከ /*kə* 'from'/), negation (አል /*ʔəli* 'not'/), passive form (ተ /*tə*/), the verbal stem (ስማም /*simami* 'comfort'/), subject pronoun (አን /*ʔəni* 'we'/), object pronoun (አቸው /*ʔətʃəwi* 'they'/), and the negation marker (ም /*mi* 'not'/). Similarly, definite markers and conjunctions can be suffixed to a word while genitive can be attached as prefix. Therefore, a word can be segmented into a base (stem or root) and affixes like number, gender, tense, aspect, etc.

The stem and the root of each word are stored in two lexicons. Each of the stem-based and the root-based morphological lexicons thus contain 170,000 unique entries. The two lexicons are organized into two columns: the surface word and the morphologically annotated form. Therefore, the stem-based morphological lexicon contains unique words and their stem-based annotated forms (morphologically segmented basic stems and affixes) whereas the root-based morphological lexicon contains unique words and their root-based annotated forms (morphologically segmented roots and affixes), the details of which are discussed below. As a final resource, the stem-based and the root-based corpora are created based on the stem-based and root-based lexicons, respectively.

4.3 Stem-Based Annotation

The stem-based morphological annotation segments surface words into basic stems and affixes. The focus of this work was to segment basic stems from the rest of morphemes. The stem extraction is performed by removing characters, changing the shapes of characters, or simply segmenting stem from the rest of the word. Generally, Amharic words can be classified as derived or non-derived words. The stem-based annotation process is presented below.

Amharic verbs are marked for person, gender, number, tense, subject, object, and negation by attaching a series of affixes. Therefore, their annotation was performed by segmenting a word into its affixes and stem. The following example shows the morphological annotations of the verbs ስላላመጣን /*silalamət'ani* 'since we did not bring'/ and አልተሰበረም /*ʔəlitəsəbərəmi* 'it was not broken'/.

ስላላመጣን	አልተሰበረም
ስለ_አል_አ_መጥ_አን	አል_ተ_ሰበር_ኧ_ም
silə_ʔali_ʔa_mət'i_ʔani	*ʔəli_tə_səbəri_ʔə_mi*
[*pre*]-[*neg*]-[*cau*]-[*stem*]-[*1,p*] [9]	[*neg*]-[*pas*]-[*stem*]-[*3,s,m*]-[*neg*]

In the case of stems derived from basic stems, annotation was made based on basic stems as they are better for conflating more variants than derived stems. Therefore, in this work verbs are annotated morphologically into their basic stems and affixes. Non-derived nouns, adjectives, adverbs and functional words are annotated into their morphemes by segmenting them into their prefixes, stem, and suffixes. For example, the noun ከየቤቱና /*kəjəbetuna* 'from each house and'/ and the adjective የደንቹን /*jədəgotʃuni* 'of the (many) generous-[*acc*]'/ are annotated as shown below.

ከየቤቱና	የደንቹን
ከ_እየ_ቤት_ኡ_እና	የ_ደግ_አች_ኡ_ን
kə_ʔijə_beti_ʔu_ʔina	*jə_dəgi_ʔotʃi_ʔu_ni*
[*pre*]-[*dis*]-[*stem*]-[*def*]-[*conj*]	[*gen*]-[*stem*]-[*p*]-[*def*]-[*acc*]

Similarly, the adverbial word የነጋችን /*jənəgatʃini* 'of tomorrow-[*1,p,pos*]'/ and the functional word እንደዚያም /*ʔinidəzijami* 'like that-[*foc*]'/ are annotated as follows.

የነጋችን	እንደዚያም
የ_ነገ_አችን	እንደ_እዚያ_ም
jə_nəgə_ʔətʃini	*ʔinidə_ʔizija_mi*
[*gen*]-[*stem*]-[*1,p,pos*]	[*tran*]-[*stem*]-[*foc*]

Many Amharic words are derived from other word classes by attaching affixes. For example, nouns can be derived by attaching affixes to the base of adjectives, nouns, verbs, stems and roots [32]. Accordingly, the verbal noun ለስብራታችን /*ləsibiratatʃini* 'for our state of being broken'/ and the abstract noun በሰውነቶቻችን /*bəsəwinətotʃatʃini* 'with our bodies'/ are annotated as follows.

ለስብራታችን	በሰውነቶቻችን
ለ_ስብር_አት_አችን	በ_ሰው_ነት_አች_አችን
lə_sibiri_ʔati_ʔatʃini	*bə_səwi_nəti_ʔotʃi_ʔatʃini*
[*pre*]-[*stem*]-[*nom*]-[*1,p,pos*]	[*pre*]-[*stem*]-[*nom*]-[*p*]-[*1,p,pos*]

[9] *pre*: preposition, *acc*: accusative, *cau*: causative, *pas*: passive, *nom*: nominative, *pos*: possessive, *def*: definite, *foc*: focus, *tran*: transitive, *pas*: passive, *gen*: genitive, *adj*: adjectivizer, *neg*: negative, *dis*: distributive, *conj*: conjunction, *f*: feminine, *m*: masculine, *s*: singular, *p*: plural, *1*: first person, *2*: second person, *3*: third person person

Amharic adjectives can also be derived from the stems of nouns, verbs, adverbs, and verbal roots by attaching suffixes. Similarly, derived adverbs are formed from various word classes. Derived adjectives and adverbs are annotated into the stem of their corresponding nouns, verbs, or adjectives. For example, the adjective ጥቁሮቹ /*t'ɨk'urotʃu* 'the blacks'/ and ሰማያዊ /*səmajawi* 'blue'/, and the adverbs በፍጥነት /*bəfɨt'ɨnətɨ* 'quickly'/ are annotated as follows.

ጥቁሮቹ	ሰማያዊ	በፍጥነት
ጥቁር_አች_ኡ	ሰማይ_አዊ	በ_ፍጥን_ኧት
t'ɨk'urɨ_ʔotʃɨ_ʔu	*səmajɨ_ʔawi*	*bə_fɨt'ɨnɨ_ʔətɨ*
[*stem*]-[*p*]-[*def*]	[*stem*]-[*adj*]	[*pre*]-[*stem*]-[*nom*]

Amharic has many types of functional words. This includes prepositions, conjunctions, demonstratives, etc. Unlike English, many of them undergo morphological processes. They might merge with each other and other word classes, or affixes might be added as indicated in the following examples. For example, the words ስለዚህም /*sɨləzihɨmɨ* 'therefore-[*foc*]'/ and ከሌሎች /*kəlelotʃɨ* 'from others'/ are annotated as follows.

ስለዚህም	ከሌሎች
ስለ_እዚህ_ም	ከ_ሌላ_አች
sɨlə_ʔizihɨ_mɨ	*kə_lela_ʔotʃɨ*
[*pre*]-[*stem*]-[*foc*]	[*pre*]-[*stem*]-[*p*]

The stems of some words are generated after applying regressive and progressive assimilation. The suffix -ኢ /*ʔi*/ sometimes palatalizes the last character whereas the suffix -ኢያ /*ʔija*/ usually palatalizes the final coronal consonants of a stem with the loss of -ኢ /*-ʔi*/. For instance, the word ጎጂ /*godʒi* 'harm'/ is annotated as ጎድ_ኢ /*godɨ_ʔi*/ whereas the word መጨረሻ /*mətʃ'ərəʃa* 'end'/ is annotated as መ_ጨረስ_ኢያ /*mə_tʃ'ərəsɨ_ʔija*/.

Amharic has many homonym words for which the morphological annotation is still similar. For example, the stem-based morphological annotation of the word ቤቱ /*betu* 'the house' or 'his house'/ is annotated as ቤት_ኡ where ቤት /*betɨ* 'house'/ is the stem, and the suffix ኡ /*ʔu*/ can refer to 'the' or 'his' depending on the word context. The word በጉ /*bəgu* 'the sheep' or 'a male sheep' or 'foolish'/ can be annotated as በግ_ኡ for the three cases. In such cases, only a single annotation is made for homonym words. However, the morphological annotations of some homonym words can be different due to the use of the word for multiple purposes. For example, the word ቀኑ /*k'ənu*/ may mean 'the day' or 'they become jealous' whereas the word በቀለ /*bək'ələ*/ may function as proper name or verb, requiring multiple annotations. As a result, we applied multiple annotations (or context-annotation) to avoid ambiguity for such types of words (see Table 1).

Table 1. Examples of stem-based multiple annotation.

Word	Stem-based annotation
የአበበ	የ_አበበ /*jə_ʔəbəbə* 'of Abebe'/
/*jəʔəbəbə*/	የ_አበብ_ኧ /*jə_ʔəbəbɨ_ʔə* 'of flowered-[*3,s,m*]'/
ማሩን	ማሩ_ን /*maru_nɨ* 'Maru-[*acc*]'/
/*marun*/	ማር_ኡ_ን /*marɨ_ʔu_nɨ* 'the honey-[*acc*]'/
	ማር_ኡ_ን /*marɨ_ʔu_nɨ* 'they gave us mercy'/

4.4 Root-Based Annotation

Amharic root is the base of stem formation whereas stem is the base of surface word formation. Verbal root is represented as a sequence of radicals (consonants). Morphology provides the templates for the combination of the root consonants with the theme vowels to derive basic stems, which may be actual or potential verbs [18]. For example, the stem ሰባር /*səbarɨ*/ is derived from the root ሰ-ብ-ር /*s-b-r 'break'*/ using a ሰ-ኧ-ብ-ኣ-ር /*s-ə-b-a-r-ɨ*/ pattern. The radicals are ሰ /*s*/, ብ /*b*/, and ር /*r*/ whereas ኧ /*ə*/ and ኣ /*a*/ are vowels. The vowels change the shape of the preceding radical in a given root during stem or word formation. Some words can be formed directly from roots. For example, the noun ልብስ /*lɨbɨsɨ* 'cloth'/ and the adjective ደካማ /*dəkama* 'weak'/ are derived from the roots ል-ብ-ስ /*l-b-s*/ and ድ-ክ-ም /*d-k-m*/, respectively. Affixation applies to the outputs of the patterns to derive verbs with functions such as passive, causative, infinitive, etc. We performed morphological annotation to produce our root-based lexicon that contains roots of words and their affixes. Therefore, surface words are segmented into prefixes, root, and suffixes. For instance, the verb ስለገደሏቸው /*sɨləgədələtʃatʃəwɨ* 'since she killed them'/ is morphologically segmented with root form as ስለ_ግ-ድ-ል_ኧች_አቸው /*sɨlə_g-d-l_ʔətʃ_ʔətʃəwɨ*/. In this case, ስለ /*sɨlə* 'since'/ is prefix and ግ-ድ-ል /*g-d-l* 'kill'/ is the root whereas ኧች /*ʔətʃɨ*/ and አቸው /*ʔtʃəwɨ*/ are suffixes. Amharic has many words of various word classes that are derived from verbal roots. Thus, the root representation of such words is the verbal root. For example, ዝ-ን-ብ /*z-n-b*/ and ድ-ክ-ም /*d-k-m*/ are the verbal roots for the derived noun ዝናብ /*zɨnabɨ* 'rain'/ and the derived adjective ደካማ /*dəkama* 'weak'/, respectively. Accordingly, for instance, the noun ለስብራታችን /*ləsɨbɨratatʃɨnɨ* 'for our state of being broken'/ and the adverb በፍጥነት /*bəfɨt'ɨnətɨ* 'quickly'/, which are derived from verbal roots, are annotated as follows.

ለስብራታችን	በፍጥነት
ለ_ስ-ብ-ር_አት_አችን	በ_ፍ-ጥ-ን_ኧት
lə_s-b-r_ʔat_ʔatʃɨnɨ	*bə_f-t'-n_ʔətɨ*
[*pre*]-[*root*]-[*nom*]-[*1,p*]	[*pre*]-[*root*]-[*nom*]

Some words are derived from verbal roots containing weak consonants like ው /w/. In such cases, the weak consonants are included in the root forms as shown in the following annotations of ይሞታል /*jɨmotalɨ* 'he will die'/ and ለኑሮ /*lənuro* 'for life'/.

ይሞታል	ለኑሮ
ይ_ም-ው-ት_አል	ለ_ን-ው-ር_ኦ
jɨ_m-w-t_ʔalɨ	*lə_n-w-r_ʔo*
[*3,s,m*]-[root]-[*future*]	[*pre*]-[*root*]-[*nom*]

Amharic verbal roots are usually extracted by removing vowels from basic stems directly. However, the roots of some words are formed by palatalizing one or more characters. For example, the root of ቀጭን /*kʼətʃʼɨnɨ* 'thin'/ is ቅ-ጥ-ን /*q-tʼ-n*/, and the root of ረዥም /*rəʒɨmɨ* 'long'/ is ር-ዝ-ም /*r-z-m*/.

On the other hand, the root forms of words that are not derived from verbal roots have similar representations as their basic stems. For example, the non-derived noun በቤቶች /*bəbetotʃɨ* 'with houses'/, the adjective የደንቁ /*jədəgotʃu* 'of the (many) generous'/, and the word ከሌሎች /*kəlelotʃɨ* 'from others'/ are annotated as follows.

በቤቶች	የደንቁ	ከሌሎች
በ_ቤት_አች	የ_ደግ_አች_ኡ	ከ_ሌላ_አች
bə_betɨ_ʔotʃɨ	*jə_dəgɨ_ʔotʃɨ_ʔu*	*kə_lela_ʔotʃɨ*
[*pre*]-[*root*]-[*p*]	[*gen*]-[*root*]-[*p*]-[*def*]	[*pre*]-[*root*]-[*p*]

Like the case of stem-based annotations, the root-based annotations of many ambiguous words can be considered the same way. For example, the annotation of the word ልጁ /*lɨdʒu* 'the child' or 'his child'/ is ልጅ_ኡ /*lɨdʒɨ_ʔu*/. However, some ambiguous words may have multiple annotations as presented in Table 2.

Table 2. Examples of root-based multiple annotations

Words	Root-based annotation
የአበበ /*jəʔəbəbə*/	የ_አበበ /*jə_ʔəbəbə* 'of Abebe'/
	የ_እ-በ-በ_ኧ /*jə_ʔ-b-b_ə* 'of flowered-[*3,s,m*]'/
ማሩን /*marun*/	ማሩ_ን /*maru_nɨ*/ 'Maru-[*acc*]'/
	ማር_ኡ_ን /*marɨ_ʔu_nɨ* 'the honey-[*acc*]'/
	ም-ሕ-ር_ኡ_ን /*m-h-r_ʔu_nɨ* 'they gave us mercy'/

4.5 The annotated corpora

Two annotated corpora are created from the initial corpus using the two lexicons (stem-based and root-based), where each word in the initial corpus is replaced by its annotation from the appropriate lexicon. These corpora are monolingual, coded in Unicode-8 and International Phonetic Association (IPA) text file forms. Each corpus consists of 6,069 full text documents consisting of 72,814 sentences constructed from 1,592,351 morphologically annotated words. Each annotated word form contains its base (stem or root) and a full set of morphological features indicating inflection and derivations of words. The grammatical features are separated from each other and from the stem or root by underscore ('_'); radicals of a root are separated from each other by hyphen ('-'); and multiple annotations of a word are enclosed in square brackets ('[ ]') without leaving free space between them. A Python script was used to produce the final morphologically annotated corpora automatically using documents and morphologically analyzed lexicons.

5 Discussion

The lexicons we built cover a variety of domains and styles. Accordingly, they can serve as important resources in the development of Amharic IR and NLP applications. Many IR and NLP applications need stem or root extraction prior to other processes. We conducted a preliminary analysis on the usefulness of stem-based and root-based retrieval using the corpora we built, the 2AIRTC [16] and the Amharic stopword list [33] which are all available at https://www.irit.fr/AmharicResources/. We found that root-based approach is better for retrieving more number of relevant documents. The retrieval effectiveness of root-based and stem-based approaches are 0.70 and 0.57, respectively [34]. This is due to the fact that root could represent word variants using a single common form. For example, the stopwords ነበር /*nəbərɨ* 'has existed'/ and ነባር /*nəbarɨ* 'something known to be in existence for a long time'/ are variants having a common root ን-ብ-ር /*n-b-r*/. However, they do not have a common stem leading to degradation of retrieval performance in comparison to the use of roots for document representation.

6 Conclusion

Amharic is a morphologically complex and under-resourced language. In this paper, a considerable effort has been made to create stem-based and root-based morphologically annotated lexicons and corpora. The morphological annotation segments words to their constituent morphemes. The lexicons and annotated corpora could be used by different researchers to extract knowledge in different applications, to model language phenomena, and to train and test algorithms. Future work is directed at adding more vocabulary to the lexicons and extending the corpus by annotating with the contextual meaning of words in a sentence.

REFERENCES

[1] Tanja Gaustad and Gosse Bouma, 2002. Accurate stemming of Dutch for text classfication. *Language and Computers*, vol. 45, no. 1, 104-117.

[2] Martin Porter, 1980. An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, vol. 14, no. 3, 130-137.

[3] Viviane Orengo and Christian Huyck, 2001. A stemming algorithm for the Portuguese language. *In Proceedings the 8th Symposium on String Processing and Information Retrieval*, 186-193, Laguna de San Rafael, Chile.

[4] Mohammed Aljlayl and Ophir Frieder, 2002. On Arabic search: improving the retrieval effectiveness via a light stemming approach. *In Proceedings of the 11th International Conference on Information and knowledge Management*, 340–347, McLean Virginia, USA.

[5] Eduard Hovy and Jduard Lavid, 2010. Towards a science of corpus annotation: A new methodological challenge for corpus linguistics, *International journal of Translation*, vol. 22, no. 1, 13-36.

[6] Prasenjit Majumder, Mandar Mitra, Swapan Parui, Gobinda Kole, Pabitra Mitra and Kalyankumar Datta, 2007. YASS: Yet another suffix stripper. *ACM Transactions on Information Systems (TOIS)*, vol. 25, no. 4.

[7] Jasmeet Singh and Vishal Gupta, 2019. A novel unsupervised corpus-based stemming technique using lexicon and corpus statistics. *Knowledge-Based Systems, vol.* 180 , no. 2019, 147-162.

[8] Jiaul H. Paik and Swapan K. Parui, 2011. A fast corpus-based stemmer. *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 10, no. 2, 1-16.

[9] Massimo Melucci and Nicola Orio, 2003. A novel method for stemmer generation based on hidden Markov models. *In Proceedings of the 12th CIMK*, 131-138, New Orleans, USA.

[10] Alireza Mokhtaripour and Saber Jahanpour, 2006. Introduction to a new Farsi stemmer. *In Proceedings of the 15th ACM International Conference on Information and Knowledge management*, Arlington Virginia,USA.

[11] Ali Daud, Wahab Khan and Dunrene Che, 2017. Urdu language processing: a survey. *Artificial Intelligence Review*, vol. 47, no. 3, 279-311.

[12] Donna Harman, 1995. Overview of the second text retrieval conference (TREC-2). *Information Processing and Management*, vol. 31, no. 3, 271-289.

[13] Nicola Ferro, 2014. CLEF 15th birthday: past, present, and future. *ACM SIGIR Forum*, vol. 48, no. 2, 31–55.

[14] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato and Jun Adachi, 1999. The NTCIR workshop: the 1st evaluation workshop on Japanese text retrieval and cross-lingual information retrieval. In *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages, INV-1-INV-7.*, 1-7, Tokyo,Japan.

[15] Tasnim Chaudhury, Abdul Matin, M. S. Hossain, Asie Uzzaman and Md Masum, 2017. Annotated Bangla news corpus and lexicon development with POS tagging and stemming. *Global Journal of Research in Engineering*, vol. 17, no. 1.

[16] Tilahun Yeshambel, Josiane Mothe and Yaregal Assabie, 2020. 2AIRTC: The Amharic adhoc information retrieval test collection. In *Proceedings of CLEF 2020*, 55-66,Thessaloniki, Greece.

[17] Yaregal Assabie, 2017. Development of Amharic morphological analyzer, T*echnical Report, Ethiopian Ministry of Communication and Information Technology*, Addis Ababa.

[18] Wolf Leslau, 1995. *Reference Grammar of Amharic* (1st ed.). Otto Harrassowitz, Wiesbaden, Germany.

[19] Girma Demeke and Mesfin Getachew, 2006. Manual annotation of Amharic news items with part-of-speech tags and its challenges. *Ethiopian Languages Research Center Working Papers*, vol. 2, no. 1, 1-16.

[20] Biniyam Epherem, Yusuke Miyao and Baye Yimam, 2016. Morpho-syntactically annotated Amharic treebank. In *Proceedings of CLiF Corpus Linguistics Fest, 48-57, Blooming, IN, USA.*

[21] Andargachew Mekonnen, Biniyam Epherem, Michael Gasser and Andreas Nürnberger, 2018. Contemporary Amharic corpus: Automatically morpho-syntactically tagged Amharic corpus. *In Proceedings of the 1st Workshop on Linguistic Resources for Natural Language Processing*, 65-70, Santa Fe, USA.

[22] Ralf Grubenmann, Don Tuggener, Pius Däniken, Mark Deriu and Cieliebak, 2019. SB-Ch: A Swiss German corpus with sentiment annotations. *In Proceedings of the 11th International Conference on Language Resources and Evaluation, LRE, MC*, 2349-2353, Miyazaki, Japan.

[23] Maher Itani, Chris Roast and Samir Al-Khayatt, 2017. Corpora for sentiment analysis of Arabic text in social media. *In Proceding of* the *8th International Conference on Information and Communication Systems (ICICS)*, 64-69, Irbid, Jordan.

[24] Kais Dukes and Nizar Habash, 2010. Morphological annotation of quranic Arabic. *In Proceedings of the 7th International Conference on Language Resources and Evaluation*, 2530-2536, Valletta, Malta.

[25] Mitchell Marcus, Beatrice Santorini and Mary Marcinkiewicz, 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, vol. 19, no. 2, 313-330.

[26] Erika Rimkutė,Vidas Daudaravicius and Anrius Utka, 2007. Morphological annotation of the Lithuanian corpus. *In Proceedings of the Workshop on Balto-Slavonic natural language processing*, 94-99, Czech Republic.

[27] Abeba Ibrahim and Yaregal Assabie, 2014. Amharic sentence parsing using phrase chunking. In *Gelbukh A.(eds) Computiational Linguistics and Intelligent Text Processing* (*CICLing*), 297-306, Berlin, Heidelberg.

[28] Martha Yifiru, Solomon Teferra and Laurent Besacier, 2011. Part-of-speech tagging for under-resourced and morphologically rich languages: the case of Amharic. In *Proceedings of Conference on Human Language Technology for Development*, 50-55, Alexandria, Egypt.

[29] Michael Gasser, 2011. HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. In *Preceedings of Conference on Human Language Technology for Development*, 94-99, Alexandria, Egypt.

[30] Wondwossen Mulugeta and Michael Gasser, 2012. Learning morphological rules for Amharic verbs using inductive logic programming. In *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*, 7-12, Istanbul, Turkey.

[31] Baye Yimam, 2000. የአማርኛ ሰዋስው /jəʔamariɲa səwasɨw 'Amharic Grammar'/ (2nd ed.), CASE, Addis Ababa, Ethiopia

[32] Nega Alemayehu and Peter Willett, 2002. Stemming of Amharic words for information retrieval. *Literary and Linguistic Computing*, vol. 17, no. 1, 1-17.

[33] Tilahun Yeshambel, Josiane Mothe and Yaregal Assabie, 2020. Construction of morpheme-based Amharic stopword list for information retrieval system. In *Proceedings of the 8th EAI International Conference on Advancements of Science and Technology*, Bahir Dar, Ethiopia.

[34] Tilahun Yeshambel, Josiane Mothe and Yaregal Assabie, 2020. Amharic document representation for adhoc retrieval. In *Proceedings of the 12th International Conference on knowledge discovery and information retrieval*, online conference, Hungary,124-134.