IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 21, Issue 4, Ser. III (Jul - Aug 2019), PP 16-29
www.iosrjournals.org

Morphological Analysis of Ethio-Semitic Languages

Yitayal Abate[1],Eskedar Yirga[1],Yaregal Assabie[2], andMesfin Abate[2]
[1]*Department of Computer Science, Debre Berhan University, Ethiopia,*
[2]*Department of Computer Science, Addis Ababa University,Ethiopia*
Corresponding Author: Yitayal Abate

Abstract: *Ethio-Semitic languages includeGe'ez, Amharic, Tigre and Tigrinya, Argobba, Harari, and Gurage. In this paper, we present the morphological analysis of Ge'ez and Amharic verbs.Ge'ez is the classical language of Ethiopia. Itceased to be spoken in the twelfth or thirteenth century,ithas remained the language of literature and of theliturgy.Currently,Amharic is an official working language which is widely spoken throughout the country as a first and a second language.Both ofthem the most morphologicallycomplex languages whicheffectively hinder the development of efficient natural language processing applications. Because of this, morphological analysis of highly inflected languages like Ge'ez and Amharic is a non-trivial task.We proposed a memory-based supervised machine learning method to develop Ge'ezand Amharic verbs morphological analyzer. It extrapolates new unseen classes based on previous examples in memory. We treat morphological analysis as a classification task which retrieves the grammatical functions and properties of morphologically inflected verbs. As the task is geared towards analyzing the vowelled inflected Ge'ezand Amharic verbs with their grammatical functions of morphemes, the morphological structure of verbs and the way how they are represented in memory-based learning is investigated. The performance of the model is evaluated using 10-fold cross-validation technique. The overall accuracy using IB2 and TRIBL2 algorithms is 93.24% and 92.31% respectively for Ge'ez verbs and for Amharic verbsusing IB1 and IGTREE algorithms is 93.59 % and 82.26%.*

Keywords:*Amharic verbs Morphology, Ge'ezverbs Morphology, Memory-Based Learning,Morphological Analyzer*

Date of Submission: 26-07-2019 Date of Acceptance: 12-08-2019

I. Introduction

Natural Language Processing (NLP) is a field of Computer Science that investigates interactions between computers and human languages, whichis used for both generating human-readable information from computer systems and converting human language into more formal structures that a computer can understand [1]. It studies the problems of automated generation and understanding of natural human languages[2]. Well known problems of NLP are morphological analysis, part of speech tagging, word sense disambiguation, and machine translation [1].Morphology deals the identification, analysis,and description of the structure of a given language's morphemes and other linguistic units, such as root words, affixes, parts of speech, etc. [3]. Morphology and its analysis play an important role in many natural language processing tasks. Morphologically complex languages are especially challenging due to the combinatorial explosion of possible morpheme structures [4].Morphological analysis is the basic process for any Natural Language Processing task. It is the first step in Natural Language Processing [5].Therefore, the role of themorphological analyzer is very significant in the field of natural language processing (NLP) applications like machine translation (MT), information extraction (IE), information retrieval (IR), spell checker, lexicography, etc. [6]. There are two broad categories of approaches in computational morphology: rule-based and corpus-based [7].Rule-based typically need to be manually constructed for each language and/or sub-domain of a language, which makes the development of such a morphological analyzer very costly and time-consuming. Corpus-based means using machine learning approaches to morphology which extracts linguistic knowledge automatically from an annotated or unannotated corpus [8].Machine learning algorithms can be supervised or unsupervised. The input and corresponding output data are used in supervised learning. In unsupervised learning, only input samples are used [9]. Machine learning approaches that use supervised learning approach includes support vector machine (SVM), inductive logic programming (ILP), hidden Markov model (HMM) and memory-based learning (MBL) [10]. These paradigms have been used to implement low-level linguistic analysis such as morphological analysis [10, 11, 12, 13]. Among various alternatives, the choice of the approach depends on the problem at hand. In this work, we employed MBL to develop amorphological analyzer for Ge'ez and Amharic verbs.Memory-Based Learning (MBL) is a simple and robust machine-learning method which has been successfully applied to a wide range of

NLP tasks. Some of the tasks in which memory-based learning has been applied are Part-Of-Speech Tagging, Phoneme-To-Grapheme Conversion, Disfluency Detection in Spoken Language, and Semantic Role Labeling [14, 15, 16, 17] due to its capabilities of incremental learning from examples. Among the MBL algorithms, IBl and IGTREE for Amharic and IB2 and TRIBL2for Ge'ez are used in this study**.** All algorithms rely on the k nearest neighbor classifier which uses some distance metric to measure the distance between each neighbor of features [8, 18, 19, 20].

The remaining part of the paper is organized as follows: Section 2 presents the characteristics of Ethio-Semitic languages with special emphasis on its morphology of verbs; Section 3 presents the proposed system for morphological analysis. Section 4 describes experimental results, and in Section 5, conclusion and future works are presented. List of Referencesis provided at the end.

II. Characteristics of Ethio-Semitic Languages

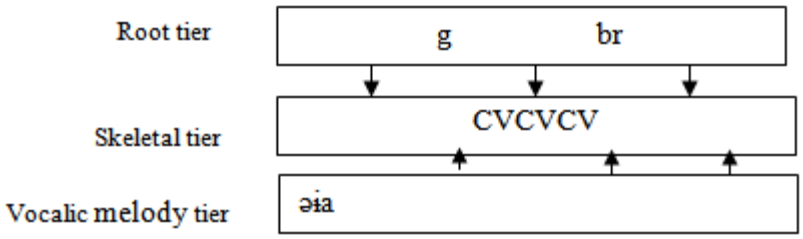
Ge'ez and Amharic languagesare the Semitic branches of the supper family of Afroasiatic which are related to Hebrew, Arabic,and Syrian languages. Ge'ez is the classical language of Ethiopia.It is still used as the liturgical language of Ethiopian Orthodox Tewahido Church. Many ancient kinds of literature were written in Ge'ez. The religious texts, ancient philosophy, tradition, history,and knowledge of Ethiopia were being written in Ge'ez. Although Ge'ez ceased to be spoken in the twelfth or thirteenth century, it has remained the language of literature and of liturgy[21,26]. A study of the Ge'ez writing systems is essential to understanding the history of Ethiopia and the evolution and modern usage of the language. It is the only language in Ethiopia which has its own alphabets [22, 23,28]. An alphabet or *fidəl* of a language represents its sound. They are also called hohyat /ሆህያት፡in Amharic/[24,28]. Geʼez has essentially 29 main syllographs or alphabets, all consonants, and each with six more derivations while the rest are essentially those with additional strokes and modifications added on to the main forms to indicate a vowel sound associated with it or to make aural adjustments in the basic consonant sound [25,28]. Hence, it has a total of 203(29x7) syllographs.

Amharic is an official working language in Ethiopia which is widely spoken throughout the country as a first and a second language after Ge'ez ceased to be spoken.Amharic is the second most widelyspoken Semitic language, next to Arabic. It uses a unique script which has originated from the Ge'ez alphabet. It has 34 basic consonant forms and 7 vowels, giving 238(34x7) syllable patterns. Semitic languages like Ge'ez, Amharic and Arabic have complex morphology in which words consist of discontinuous consonantalroots with vowel intercalation [26,27,28].

Morphology of Ge'ez and Amharic Verbs

Verbs are the most important part of both languages which serves as a basis for almost all other Part-Of-Speech (POS) categories. To illustrate the importance of the verb some scholars [24, 26,28] say that the verb is the language. The main verbs in both languages are perfect and imperfect. Perfect is usually past or completed action. It includes past perfect, past continuous, past participle. The imperfect one is usually apresent, continuous and future action. The end of all perfect verbs is the first order while all imperfect verbs have in the end the six orders. This is the pronoun ውእቱ/እሱ(he). The other main verbs are subjunctive, imperative, infinitive and gerundive [22, 26, 28].As described in [23,28], a verb in Ge'ez and Amharic must pass through three stages of formation. These are Seed formation, Tense-mood formation,and person-gender-number formation. Semitic language verbs are produced as the result of the intercalation of vowels and roots in a certain template. This fact is valid for Ge'ez and Amharic verbs too. Such intercalation of vowels and consonants in a given template produces what is called aseed. The produced seed grows up to a more natural verb form by attaching itself to tense -mood and person-gender-number marker affixes [24,26, 27].

In non-concatenative morphology, other formal means are involved in the creation of new morphological forms. In the case of internal modification, a stem with a different form is created,for instance, by replacing a vowel pattern or a consonant pattern (or both) with another one. In Semitic languages, particularly Ge'ez and Amharic verbal roots may appear in a number of different templates with specific patterns of consonants(C) and vowels(V), sometimes in combination with a prefix [35,28]. McCarthy hypothesized in [36], that the verb in Arabic has elements arranged on three independent tiers at the underlying level of representation in the lexicon, the three tiers being the consonantal tier, also called the root tier, the skeletal tier, and the vocalic melody tier. For example,in Ge'ez,the verb ገበራ /gəbɨra/ will have the structure shown in Fig.1.



In Fig.1, the three tiers give the Ge'ez verb ገብሬ/*gəbɨra/* (she did). This is also true for Amharic nounሰባራ /*səbara/* These three tiers are linked together by association lines. In addition to such non-concatenative morphological features, Ge'ez and Amhariclanguages use different affixes to create inflectional and derivational morpheme.

Affixation in Ge'ez and Amharic Verbs

Affixation realizes the two-next stage of verb formation: person-gender-number and tense-mood formation. Therefore, verbs attain their maximum growth level through affixation and are verbs of this stage which are the most common and natural verbs of Ge'ez and Amharic [24, 27]. It has all types of the affixes: infixes, prefixes, suffixes,and circumfixes. Infixation is realized that the internal modification of the seed, which means forming seeds from root via intercalating of the vocalic patterns with the consonants [26, 27].Prefixes of Ge'ez and Amharic verbs can be categorized into two as negation marker and affirmative marker prefixes. The affirmative marker prefixes can further be divided into two as verbal-stem-marker and person-marker prefixes. As a result of such type of features, tens of thousands of verbs (in surface forms) are generated from a single verbal root. For example, the maximum possible inflections of a given Ge'ez verb, has around 1388 possible surface verb forms. As verbs are marked for various grammatical units, a single verb can form a complete sentence. For instance, ቀተልክዎሙ /*ḱətəlikɨwomu/* in Ge'ez or ገደልኳቸው /*gədəlik^{w}ć́əwu*/in Amharic, means*I killed them*in English, are analyzed as follows:

verbal root: ቅትል/*ḱtl/*
verbal stem: ቀተል/*ḱətəl/*
subject: ክ/*-ku/* (I)
object: ዎሙ-*womu/* (them)

verbal root: ግድል/*gdl/*
verbal stem: ገደል/*gədəl/*
subject: ኳ/*-k^{w}/* (I)
object: ቸው-*ć́əwu/* (them)

Table 1: Prefixes of Indicative, Subjunctive, and Jussive Ge'ez Verbs

Indicative	Subjunctive	Jussive	Person referred
አ- /ʔə-/	አ –/ʔə-/	አ- /ʔə-/	አነ /ʔənə/ (I, 1psn[1])
ነ - /nə-/	ነ –/nə-/	ነ - /nə-/	ንሕነ /nihinə/ (We, 1ppn[2])
ተ- /tə-/	–	ተ- /tə-/	አንተ /ʔənitə/ (You, 2psm)
ተ- /tə-/	–	ተ- /tə-/	አንትሙ/ʔənitimu/ (You, 2ppm)
ተ- /tə-/	–	ተ- /tə-/	አንቲ /ʔəniti/ (You, 2psf)
ተ- /tə-/	–	ተ- /tə-/	አንትን /ʔənitin/ (You, 2ppf)
ተ- /tə-/	ተ- /tə-/	ተ- /tə-/	ይእቲ /yiʔti/ (She, 3psf)
የ - /yə-/	የ - /yə-/	የ - /yə-/	ውእቱ /wiʔtu/ (He, 3psm)
የ - /yə-/	የ - /yə-/	የ - /yə-/	ውእቶሙ/wiʔtomu/ (They, 3ppm)
የ - /yə-/	የ - /yə-/	የ - /yə-/	ውእቶን /wiʔton/ (They, 3ppf)

For example, the Ge'ez verb ቀተለ /*ḱətələ/* (he killed) has its indicative, subjunctive and jussive forms as ይቀትል/*yiḱətɨl/* (he will kill), ይቅትል /*yiḱɨtɨl/* (he must kill) andይቅትል /*yiḱɨtɨl/*(for him to kill).The possible prefixes of Ge'ez and Amharic verbs along with their grammatical function are summarized as shown in Table 2.

[1]First person singular neutral.
[2]First person plural neutral.

Table 2: List of Ge'ez and Amharic verbs prefixes along with their syntactical functions

Prefixes		Syntactical function
Ge'ez	**Amharic**	
አ- /ʔə-/	አ–/ʔə-/	Causative Stem Marker
እ- /ʔ-/	እ–/ʔ-/	Indicative, Subjunctive and Jussive Moods Marker
አስተ- /ʔəsitə-/	አስ–/ʔəsi-/	Causative-Reciprocal Stem Marker
አ- /ʔi-/	አል–/ʔəli-/	Negation Marker
ና- /na-/	–	Causative Stem Marker
ናስተ- /nasitə-/	–	Causative-Reciprocal Stem Marker
ን- /ni-/	–	Indicative, Subjunctive and Jussive Moods Marker
ንት- /nit-/	–	Reciprocal and Reflexive Stems Marker
ታ- /ta-/	ታ–/ta-/	Causative Stem Marker
ታስተ- /tasitə-/	–	Causative-Reciprocal Stem Marker
ተ- /tə-/	ተ–/tə-/	Reflexive and Reciprocal Stems Marker
ት- /ti-/	ት–/ti-/	Indicative, Subjunctive and Jussive Moods Marker
ትት- /tit-/	–	Reciprocal and Reflexive Stems Marker
ያ- /ya-/	–	Causative Stem Marker
ያስተ- /yasitə-/	–	Causative-Reciprocal Stem Marker
ይ- /yi-/	ይ– /yi-/	Indicative, Subjunctive and Jussive Moods Marker
ይት- /yit-/	–	Reciprocal and Reflexive Stems Marker

For instance, the base-stem Ge'ez verb ቀተለ /ḱətələ/ (he killed) has አቅተለ /ʔəḱitələ/ (he caused somebody to killed), አስተቃተለ/ʔəsitəḱatələ/ (he caused others to be killed each other), ተቀትለ /təḱətilə/ (he is killed by), ተቃተለ /təḱatələ/ (he gets killed with somebody) and as its causative, causative-reciprocal, reflexive and reciprocal stem forms respectively. The subjects of the perfective, imperative and gerundive verbs are indicated through suffixation. In both languages, the main verbal suffixes are two types. These are 'subject marker'[3] and 'object marker'[4] suffixes [24,26, 27]. For example, if we take the Ge'ez seed ቀተል /ḱətəl/,and Amharic seedገደል/gədəl/ certain suffixes can be attached to them to produce various inflected surface verbs as shown in Table3.

Table 3: List of Ge'ez and Amharic Verbal Subject Marker Suffixes along with Indicated Persons

Seed		Subject Marker Suffix		Inflected Verbs		Subject Indicated
		Ge'ez	**Amharic**	**Ge'ez**	**Amharic**	
ቀተል /ḱətəl/ from Ge'ez and ገደል/gədəl/ from Amharic	+	-ኩ /-ku/	-ኩ /-ku/	ቀተልኩ /ḱətəliku/	ገደልኩ/gədəliku/	I
		-ነ/-nə/	-ን/-ni/	ቀተልነ /ḱətəlinə/	ገደልን/gədəlin/	We
		-ክ/-kə/	-ክ/-ki/	ቀተልክ /ḱətəlikə/	ገደልክ/gədəlik/	You (2psm)
		-ክሙ/-kimu/	-ኣችሁ/-aćihu/	ቀተልክሙ/ḱətəlikimu/	ገደላችሁ/gədəlaćihu/	You (2ppm)
		-ኪ /-ki/	-ሺ/ši/	ቀተልኪ /ḱətəliki/	ገደልሺ/gədəliš/	You (2psf)
		-ክን/-kin/	-ኣችሁ/-aćihu/	ቀተልክን /ḱətəlikin/	ገደላችሁ/gədəlaćihu/	You (2ppf)
		-ኣ/-ə/	-ኧ/-ə/	ቀተለ /ḱətələ/	ገደለ/gədələ/	He
		-ኡ /-u/	-ኡ/-u/	ቀተሉ /ḱətəlu/	ገደሉ/gədəlu/	They (3ppm)
		-ኣት/-ət/	-ኧች/-əćə/	ቀተለት /ḱətələt/	ገደለች/gədələć/	She
		-ኣ/-a/	-ኡ/-u/	ቀተላ /ḱətəla/	ገደሉ/gədəlu/	They (3ppf)
		-የ/-yə/	-የ/-yə/	ቀቲልየ /ḱətiliyə/	ገድየ/gədiyə/	I
		-ኦ /-o/	-ኦ /-o/	ቀቲሎ /ḱətilo/	ገድሎ/gədilo/	He
		-ሙ /-mu/	-ኡ /-u/	ቀቲሎሙ /ḱətilomu/	ገደሉ/gədəlu/	They (3ppm)
		-ን /-n/	-ኡ/-u/	ቀቲሎን /ḱətilon/	ገደሉ/gədəlu/	They (3ppf)
		-ኣ/-i/	-ክ/-ki/	ቅትል /ḱitil/	ገደልክ/gədəlik/	You (2psm)
		-ኡ/-u/	-ኣችሁ/-aćihu/	ቅትሉ /ḱitilu/	ገደላችሁ/gədəlaćihu/	You (2ppm)
		-ኢ/-i/	-ሺ/ši/	ቅትሊ /ḱitili/	ገደልሺ/gədəliš/	You (2psf)

Two central rules of suffixation that govern the concatenation process of morphemes to produce surface verbs are:

1. Seed + subject-marker = surface verb (only with SMS)
2. Seed + subject-marker + object- indicator = surface verb (with both SMS and OMS)

For example: 1.ቀተለ /ḱətələ/ + ኩ /ku/ = ቀተልኩ/ḱətəliku/ (I killed).

2.ገ ደለ /gədələ/ + ኩ /ku/ = ገ ደልኩ/ḱətəliku/ (I killed).

Table 4 shows the concatenation process of object markers to a verb to produce much more inflected verbs by taking the verb ቀተለ /ḱətələ/ and ገ ደለ /gədələ/ from Ge'ez and Amharic respectively [24,26, 27].

[3]It is a morpheme that indicates a subject of a verb.
[4]It is a morpheme that indicates an object of a verb.

Table 4: Example of Ge'ez and Amharic Verbal Object Marker Suffixes along with Indicated Persons

Verb	Object Marker Suffix		Object Indicated
	Ge'ez	**Amharic**	
ቀተለ /*ḱətələ*/ from Ge'ez and ገደለ/*gədələ*/ from Amharic /*ḱətələ*/	-ኒ /-*ki*/	-ኝ /-ñi/	Me
	-ነ /-*nə*/	-ን /-ni/	Us
	-ከ /-*kə*/	-ህ /-hi/	You (2psm)
	-ክሙ /-*kimu*/	-አችሁ /-aćihu/	You (2ppm)
	-ኪ /-*ki*/	-ሽ /ši/	You (2psf)
	-ክን /-*kin*/	-አችሁ/-aćihu/	You (2ppf)
	-ክዎ/-*kiwo*/, -ሁ/-*hu*/, -ኦ/-*o*/, -ዎ/-*wo*/, -ዮ/-*yo*/	-ው /-wu/	Him
	-ዎሙ/-*womu*/, -ክሙ/-*kimu*/, -ሙ/-*mu*/, -ሆሙ/-*homu*/,-ዮሙ /-*yomu*/	-አቸው /-aćəwu/	Them (3ppm)
	-ዋ /-*wa*/,-ሃ /-*ha*/, -ኣ/-*a*/,-ያ/-*ya*/	-አት /-ati/	Her
	-ዎን/-*won*/, -ሆን /*hon*/, -ን /-*n*/, -ዮን/-*yon*/	-አቸው/-aćəwu/	Them (3ppf)

In both languages, circumfixes are the subject markers of indicative, subjunctive and jussive seeds. For such type of verbs too, the object markers are attached immediately after the circumfix of the seed. Table 5 illustrates the circumfixes of ቀደሰ /*ḱədəsə*/ and መጣ/mət'a/ category.

Table 5: List of Ge'ez and Amharic Verbal Subject Marker Circumfixes along with Their Moods

Circumfixes		Subject indicated	Used in moods
Ge'ez	**Amharic**		
እ-እ /*ʔ-ʔ*/	እ-አለሁ/ʔ-aləhu/	I	Indicative, subjunctive and jussive
ት-እ /*t-ʔ*/	ት-አለህ/ti-aləh/	You (2psm) & She (3psf) (for Ge'ez only)	Indicative, subjunctive and jussive
ት-ኢ /*ti-ʔi*/	ት-አለሽ/ti-aləš/	You (2psf)	Indicative and jussive
ት-ኡ /*ti-ʔu*/	ት-አላችሁ/ti-alaćihu/	You (2ppm)	Indicative and jussive
ት-ኣ /*ti-ʔa*/	ት-አላችሁ/ti-alaćihu/	You (2ppf)	Indicative and jussive
ን-እ /*ni-ʔ*/	እን-አለን/ʔni-alən/	We (1pp)	Indicative, subjunctive and jussive
ይ-እ /*yi-ʔ*/	ይ-አል/yi-al/	He (3psm)	Indicative, subjunctive and jussive
ይ-ኡ /*yi-ʔu*/	ይ-አሉ/yi-alu/	You (3ppm)	Indicative, subjunctive and jussive
ይ-ኣ /*yi-ʔa*/	ይ-አሉ/yi-alu/	You (3ppf)	Indicative, subjunctive and jussive
-	ት-አለች/ti-aləć/	She (3psf)	Indicative, subjunctive and jussive

In general, anexample of the type of morphemes and the way they concatenate to produce surface verbs can be summarized in Table 6.

Table 6: Possible Type of Morphemes Concatenated to Form a Verb in Ge'ez and Amharic

No	Possible morphemes of a verb	Example	
		Ge'ez	**Amharic**
1	[NegPref][PosPre][5][Seed][SMS][OMS]	[ኢ][አስተ][ፋቀድ][ከም][ዎሙ]	[አል][አስ][ገደል][ኩ][ት]
2	[NegPref][PosPre][Seed][SMS]	[ኢ][አስተ][ፋቀድ][ከሙ]	[አል][አስ][ገደል][ኩም]
3	[PosPre][Seed][SMS][OMS]	[አስተ][ፋቀድ][ከም][ዎሙ]	[አስ][ገደል][ኩ][ት]
4	[NegPref][Seed][SMS][OMS]	[ኢ][ፈቀድ][ከም][ዎሙ]	[አል][ገደል][ኩ][ትም]
5	[NegPref][Seed][SMS][OMS]	[አስተ][ፋቀድ][ከሙ]	[አስ][ገደል][ኩ]
6	[NegPref][Seed][SMS]	[ኢ][ፈቀድ][ከሙ]	[አል][ገደል][ኩም]
7	[NegPref][PreCirc][6][Seed][SufCirc][7][OMS]	[ኢ][ት][ፍቅድ][ዎሙ]	[አል][ተ][ገደል][ኡም]
8	[PreCirc][Seed][SufCirc][OMS]	[ት][ፍቅድ][ዎሙ]	[ት][ግደል][ው]
9	[NegPref][PreCirc][Seed][SufCirc]	[ኢ][ት][ፍቅዱ]	[አት][ግደሉ]

The Conjugation Patterns of Verbs

Conjugation patterns are the basic templates through which the surface verbs of thecategory are formulated. The templates are effectively used during the declarationprocess carried out to find the inflected surface forms of the verb. Table 7depicts the basicconjugation patterns and their corresponding templates by taking the Ge'ez verbቀተለ/*ḱətələ*/ and Amharic verbገደለ/*gədələ*/ (he killed) as an example. The conjugation patterns (column-IV) are produced after the intercalation of the root with vowels of varies patterns (column-V)

[5]NegPref = Negation Prefix, PosPre = Positive Prefix.
[6]PreCirc = Prefix Circumfix (i.e., the prefix part of the circumfix).
[7]SufCirc = Suffix Circumfix (i.e., the suffix part of the circumfix).

in the templates given in column-III of the table. Almost all other verbs share the template, conjugation pattern and vocalic patterns depicted in this Table.

Table 7: Basic Conjugation Patterns of a Root with Their Templates and Vocalic Patterns

No	Root	Vocalic Pattern (VP)		CV-Template		Conjugation Patterns		Tense-mood
		Ge'ez	Amharic	Ge'ez	Amharic	Ge'ez	Amharic	
1	ቅ-ት-ል/*ḱ-t-l/ and* ግ-ድ-ል /*g-d-l/*	111	111	C_1əC_2əC_3ə	C_1əC_2əC_3ə	ቀተለ /*ḱətələ/*	ገደለ/gədələ/	Perfective
2		166	116	yC_1əC_2iC_3ɨ	yC_1əC_2əC_3ɨ	ይቀትል /*yiḱətil/*	ይገድል/yɨgədəl/	Indicative
3		666	616	yC_1ɨC_2iC_3ɨ	yC_1iC_2əC_3ɨ	ይቅትል/*yiḱitil/*	ይግደል/yɨgɨdəl/	Subjunctive
4		666	616	C_1iC_2iC_3ɨ	C_1iC_2əC_3ɨ	ቅትል/*ḱitil/*	ግደል/gɨdəl/	Imperative
5		666	666	yC1iC_2iC_3ɨ	C1iC_2iC_3ɨ	ይቅትል /*yiḱitil/*	ግድል/gɨdɨl/	Jussive
6		137	167	C_1əC_2iC_3o	C_1əC_2iC_3o	ቀቲሎ /*ḱətilo/*	ገድሎ/gədɨlo/	Gerundive
7		136	14	C_1əC_2iC_3ɨ	C1əC_2ay	ቀቲል /*ḱətil/*	ገዳይ/gəday/	Infinitive
		1376	1676	C_1əC_2iC_3oC_4ɨ	C_1əC_2iC_3oC_4ɨ	ቀቲሎት/*ḱətilot/*	ገድሎት/gədɨlot/	

III. The Proposed Morphological Analyzer

3.1 System Architecture

The morphological analyzer (Fig. 3) has two phases. A training phase which consists of morpheme annotation to manually annotate inflected Ge'ez and Amharic verbs, and feature extraction to create instances in a fixed length of windows and the memory-based learning to train the dataset. On the other hand, the morphological analysis phase contains the feature extraction (instance making) to deconstruct a given text, morpheme identification to classify and extrapolate, stem and root extraction to label segmented inflected words with their morpheme functions.

3.2 Training Phase

Ge'ez and Amharic verb morphemes are mostly expressed by internal phonological changes in the root. Because of this, the internal irregular changes of phonemes make the morphological analysis more complex. Even if it is a difficult task to find the roots of verbs, we investigated the morphology of Ge'ez and Amharic languages,particularly for verbs. Even though Ge'ez and Amharic verbs have more than 4 prefixes and 3 suffixes, the morphological structure is somehow straightforward to be learned by the system. After we identify the common property of all morphological formations of both languageverbs and grammatical features of all the morphemes, we built amorphological database. It is difficult to find a single representation or patterns of verbs as they are different in types due to a number of morphological and phonological processes.

Morpheme Annotation

In this study, we tried to create a memory based morphological analysis for Ge'ez and Amharic which handles all morphological productive word-classes,particularly for verbs. To do this there is a need for preparing morphologically annotated word lists manually. We were not able to get anannotated database of Ge'ez and Amharic morphology.This makes us consider building amorphologically analyzed database. Therefore, we are forced to prepare manually annotated word. In doing so we identified the different affixations of verbs. During preparing theannotated dataset for experimentation purpose the following tasks were identified and performed in the order listed: Identifying inflected words; segmenting the word into prefix, stem,and suffix; putting boundary marker between each and describing the representation of each marker. As described in [26, 27,28],verbs in Ge'ez have three segments for prefixes before the stem and three segments for suffixes after the stem. Where NegPref is for Negation Prefix; PosPref is for Positive Prefix; PreCirc is for Prefix Circumfix; SufCirc is for suffix Circumfix; SMS is for subject marker suffix, and OMS is for object marker suffix. Amharic verbs also have four segments for prefixes and four segments for suffixes.The positions of the affixes are shown as follows, where prep is for preposition; conj is for conjunction; rel is for relativation; neg is for negation; subj is for subject; appl is for applicative; obj is for objective; def is for definiteness; and acc is for accusative.

(NegPref)(PosPref)(PreCirc) + Stem + (SufCirc)(SMS)(OMS) **(a)**

(prep|conj)(rel)(neg) (subj) + Stem + (subj)(appl)(obj|def)(neg|aux|acc)(conj) **(b)**

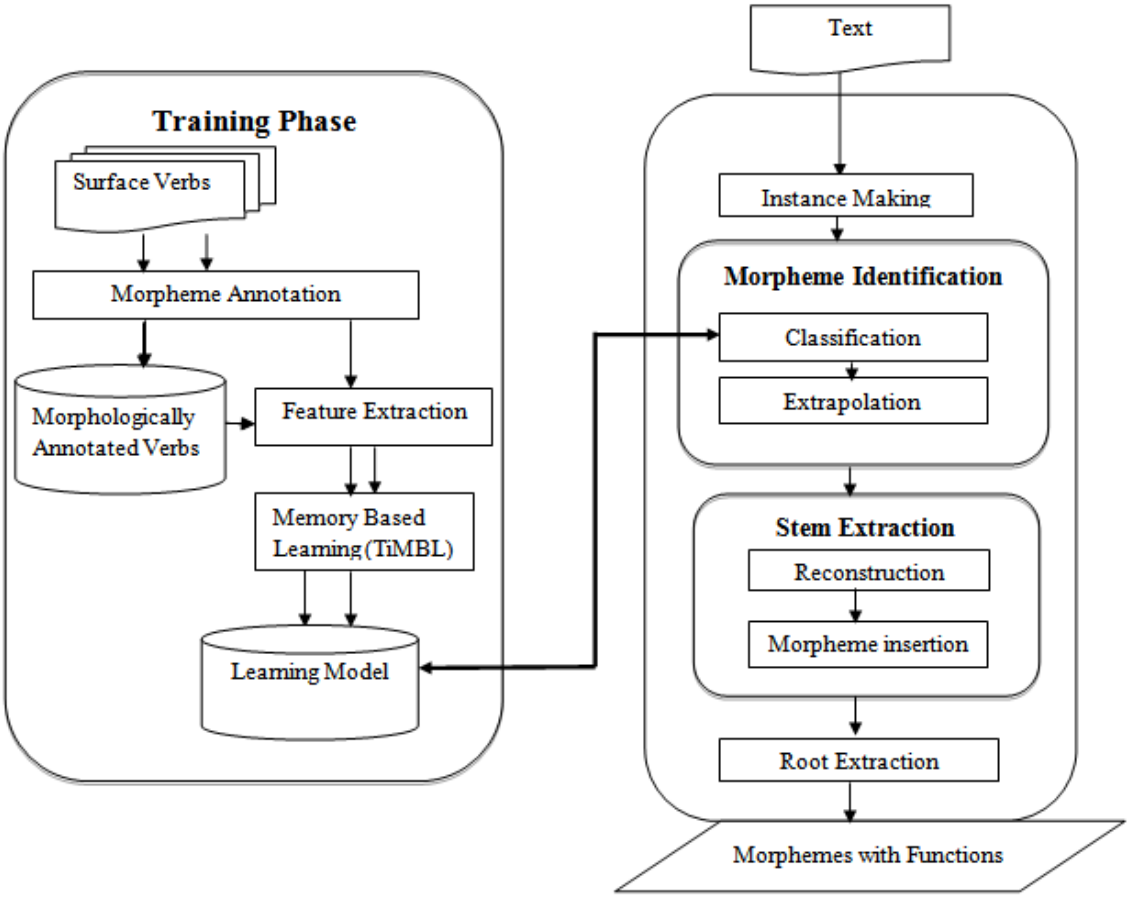


Figure 3: Architecture of the Proposed Morphological Analyzer

Analyzing all these affixes, the root template pattern of Ge'ez and Amharic verbs makes morphological analysis complex [26, 27,28]. It is a challenging task representing the features into suitable memory-based learning approach.verb stems are divided into verb roots and grammatical template. The stem is the lexical part of the verb and also the source of most of its complexity. Instances are created from the manually annotated morphological database for the purpose of learning [26]. The bracket ([ ]) can be filled with the appropriate grammatical features for each segmentation where 3, 4, S, J, A, Q, V, C, G, Z, M,K,N, U, D, indicate end of Perfective, indicative, jussive and gerundive causative reciprocal stem; Perfective, subjunctive and gerundive reflexive and reciprocal stem; Stem; First person singular object; Third person singular masculine subject; Third person singular masculine object; Third person plural masculine object; Third person singular feminine object; Third person plural feminine object; First person plural object; Second person singular masculine object markers; preposition; Negative marker; Second person plural subject marker (this is for Amharic verbs only. Ge'ez verbs have separate affixes for second and third person plural feminine and masculine.); andThird person singular feminine subject marker respectively. Lexicons were prepared manually in such a way to be suitable for extraction purpose. To study all morphologically inflected verb types, we need a morphologically annotated word list with its possible inflection forms. Then, the tokens are manually annotated like prefix, stem and suffix pattern as shown in Table 8.

Feature Extraction

After the annotated verbs are stored in a database, features are extracted automatically from the manually created morphological database to make instances based on the concept of windowing a methodfor a fixed length of the left and right context which is the average word length in the database.

Table 8: Manually Annotated Sample Verbs

Ge'ez verbs							
ʔəsitə	[3]	faḱəd	[S]	a	[F]	hu	[Q]
ʔəsitə	[3]	faḱəd	[S]	a	[F]	homu	[V]
ʔəsitə	[3]	faḱəd	[S]	a	[F]	ha	[C]
ʔəsitə	[3]	faḱəd	[S]	a	[F]	hon	[G]
tə	[4]	fəḱəd	[S]	ɨku	[J]	-	-
tə	[4]	gəbir	[S]	ə	[A]	nə	[Z]
tə	[4]	gəbir	[S]	ə	[A]	kə	[M]
-	-	noləw	[S]	ə	[A]	-	-
Amharic verbs							
		ləməd	[S]				
		ləməd	[S]	ə	[A]		
		ləməd	[S]	əćɨhu	[U]		
ʔəs	[3]	ləməd	[S]	əć	[D]		
sɨlə	[K]	ləməd	[S]	əć	[D]		
		səbər	[S]	ə	[A]	n	[Z]
ʔəs	[3]	səbər	[S]	ə	[A]	ət	[C]
ʔəl	[N]	səbər	[S]	əć	[D]	m	[N]

As described in [29], windowing method is dividing the windows where the instances are placed in the left and right context to hold afixed-length string of features, which describe the linguistic context of the token to be classified. Each instance is associated with a class. The class represents the morphological category in which the given word possesses. Instances consisting of a fixed number of features are created by windowing methods-by sliding each character down as a focus letter for each character in the left and right context. Each example focuses on one letter and includes a fixed number of left and right neighbor letters, in this case using a 10–1-10-1 window which yields twenty-two features. The input character in focus, plus the ten preceding and ten following characters are placed in the windows. As stated in [14, 30] the complex morphological analysis is placed at the rightmost part as a class.

For example, the character-based representation of the Ge'ez word **'ḱədəsomu' "ቀደሰሙ"** is as follows, in Table 9.

Table 9: Character Based Representation of the Ge'ez Word 'ḱədəsomu' "ቀደሰሙ"

No	Left context										Focus	Right context										Class
1	=	=	=	=	=	=	=	=	=	=	**ḱ**	ə	d	ə	s	o	m	u	=	=	=	0
2	=	=	=	=	=	=	=	=	=	ḱ	**ə**	d	ə	s	o	m	u	=	=	=	=	0
3	=	=	=	=	=	=	=	=	ḱ	ə	**d**	ə	s	o	m	u	=	=	=	=	=	0
4	=	=	=	=	=	=	=	ḱ	ə	d	**ə**	s	o	m	u	=	=	=	=	=	=	0
5	=	=	=	=	=	=	ḱ	ə	d	ə	**s**	o	m	u	=	=	=	=	=	=	=	S
6	=	=	=	=	=	ḱ	ə	d	ə	s	**o**	m	u	=	=	=	=	=	=	=	=	Q
7	=	=	=	=	ḱ	ə	d	ə	s	o	**m**	u	=	=	=	=	=	=	=	=	=	0
8	=	=	=	ḱ	ə	d	ə	s	o	m	**u**	=	=	=	=	=	=	=	=	=	=	V

The equality mark (=) is used as a filler symbol for positions beyond the beginning or end of the word. Characters that do not mark the end of a morpheme are classed with the default category **0** (zero). Table 9 shows the structure of features to make seven (7) instances associated with their classes, which are derived from the word **'ḱədəsomu' "ቀደሰሙ"**. For example, the class of the sixth instance is '**Q**', which indicates the morpheme ending in '**o**' is a suffix which represents Third-person singular masculine object marker/ accusative'-o' (**ኦ**).This shows the character based representation of words transcribes their deep structure of phonological process and segments each letter one at a time.

Using syllables rather than characters as features do involve an extra pre-processing step, namely syllabification [29]. This process is very precise, while mistakes are made on inflected verbs. For example, in the syllabification process for the word**'ḱədəsomu' "ቀደሰሙ"** the syllable ***so*** is never considered as a syllable within the representation, as it is split by a morpheme boundary that yields the desired bound stem/root morpheme.

ḱədəsomu =>ḱədəs[S]o[Q]mu[V]

As illustrated in Table 10, the features refer to syllables instead of single characters. This indicates that syllable-based representation of words difficult to study their deep structure of phonological processes. Due to this, character-based representation is selected for this study.

Table 10: Syllable-basedrepresentation of the Ge'ez Word 'ḱədəsomu' "ቀደሰሙ"

No	Left context										Focus		Right context									Class
1	=	=	=	=	=	=	=	=	=	=	**ḱə**	**də**	**s**	**o**	**mu**	=	=	=	=	=	=	0
2	=	=	=	=	=	=	=	=	=	**ḱə**	**də**	**s**	**o**	**mu**	=	=	=	=	=	=	=	0
3	=	=	=	=	=	=	=	=	**ḱə**	**də**	**s**	o	**mu**	=	=	=	=	=	=	=	=	S
4	=	=	=	=	=	=	=	**ḱə**	**də**	**s**	**o**	**mu**	=	=	=	=	=	=	=	=	=	Q
5	=	=	=	=	=	=	**ḱə**	**də**	**s**	**o**	**mu**	=	=	=	=	=	=	=	=	=	=	V

Memory-Based Learning

As described in section 2.2, one can be known how much the Ge'ez and Amharic languages are complex, i.e. the morphological complexity of Ge'ez and Amharic verbs. Memory-based learning (MBL) is an approach to NLP based on a symbolic machine learning method. It has been the primary machine learning method used in the present work. It also has the following advantages in contrast to most other machine learning algorithms [14]: It presupposes no more linguistic knowledge than explicitly present in the corpus used for training, i.e., it avoids a knowledge-acquisition bottleneck; Learning is automatic and fast; Processing is deterministic, non-recurrent (i.e., it does not retry analysis generation) and fast, and is only linearly related to the length of the word from being processed [29]. Because of these advantages, it is selected to test the morphological analysis of Ge'ez and Amharic verbs.

We used TiMBL as a learning tool for our task [12].It is implemented in C and C++. It has also a Python version. We used both versions interchangeably because the python version has the capability to classify individual instances with simple scripts.TiMBL implements several memory-based learning algorithms (IB1, IB2, IGTREE, TRIBL,and TRIBL2). The main differences among the algorithms incorporated in TiMBL lie in the definition of similarity, the way the instances are stored in memory, and the way the search through memory is conducted [31]. There are also different optimized parameters to be tuned in TiMBL. Therefore, to get an optimal accuracy of the model we used the default settings and also tuned some of the parameters. These are the MVDM (modified value difference metric) from distance metrics, IG (information gain) from weighting metrics, ID (inverse distance) from class voting weights, and k= 3 and 5 from the nearest neighbor. The classifier engines used here are IB1 and IGTREE for Amharic verbs and TRIBL2 and IB2 for Ge'ez verbs which construct a database of instances in memory during learning process.In those classifiers a training instances are represented by tree structures. The structure starts by creating a root node. From the root node all values of a chosen feature are grown as an arc ending in a new node. The arcs are labeled with the feature value. Information about the number of training instances leading up to the node and class-information is stored in the nodes.From the nodes on the first level of the tree, branches are grown to all the values of the second feature that occur in the training set in combination with the value of the previous node. TRIBL2 is designed as a combination between IB1 and IGTREE with the aim to exploit the trade-off between the search speed of IGTREE and the generalization accuracy of' IBl. In other words, it is built to capture the best of both algorithms. During the classification of an instance, it continues to use IGTREE as long as it finds matching feature values in the weighting-governed feature ordering. It only reverts to IB1 classification when a mismatch is found. The reasoning behind this mismatch-based switching is that the switch to IB1 is only invoked when mismatching occurs, being the typical point in which IB1 can improve on decision-tree-style classification. The operation of IB2 is identical to IBl, except that IB2 only saves misclassified instances in the training data [20]. It starts with an instance base containing only a small portion of the available training instances. This initial number of training instances in memory can be set by the user. IB2 then adds instances into memory only when they are misclassified by the k-NN algorithm, on the basis of the instances already in memory at that point [18, 20].

3.3 Morphological Analysis

The training phase is the base to implement the morphological analysis phase. In this phase the instance making to make the input words to be suitable for memory-based learning classification, the morpheme identification to classify and extrapolate the class of new instances, the stem extraction to reconstruct and insert identified morphemes, and finally the root extraction to get root forms of verb stem with their grammatical functions are described.

Instance Making

The instance is a sequence of features (characters) separated by acomma. When anunknown word is given to be analyzed by the system, it accepts and deconstructs[8] as instances to make similar representation with the stored instance in memory. Feature extraction in this section is different from that described in the training phase. It states how to make instance to be analyzed here. The word is deconstructed in a fixed length of

[8] The verbs are extracted as instances by making class question mark (?) to show themorphological representation of the instances (class).

instances without identifying the class labels at the last index. For example, when a new previously unseen word needs to be segmented, the words are similarly deconstructed and represented as instances using the same information. For instance, if we want to segment the Ge'ez word 'təḱədəsomu' "ተቀደሱሙ" the system accepts the word and extract its feature as shown in Table 11.

Table 11: Instance to be classified

10	9	8	7	6	5	4	3	2	1	**F**	1	2	3	4	5	6	7	8	9	10	class
=,	=,	=,	t,	ə,	ḱ,	ə,	d,	ə,	s,	o,	m,	u,	=,	=,	=,	=,	=,	=,	=,	=,	?

Morpheme Identification

Morpheme identification is a process of identifying each morpheme of a given verb. A new inflected word is deconstructed as instances and given to the system to be classified by the classifier. In **classification** process, if there is an exact match in the memory, the classifier returns the class of that instance to the new instance. If it is misclassified by the algorithm, on the basis of the instances already in memory at that point, it will be added to memory. **Extrapolation** is performed to assign the most likely neighborhood class with its morphemes based on their boundaries. This will be done based on the similarity metric applied to the training data.

For example, to find the class of the new instance in Table 10, the instance is compared to each and every instance in the memory-based learner. The classifier tries to find that training instance in memory that most closely resembles it. So, this might be instance six in Table 9, as they share almost all features **(L5, L4, L3, L2, L1, F, R1,R2)** except **L6** and **L7**. The memory-based learner then extrapolates the **Q** class of this training instance and predicts it to be the class of the new instance. Instances which are only misclassified by the classifier will be stored in the memory-based model for farther analysis [32].

Stem Extraction

In stem extraction process **reconstruction** of individual instances into a meaningful (their original word forms) and **morpheme insertions** in their segmentation point are performed. The steps of this post-processing phase (the recompilation of words and insertion of the predicted classes in their proper morpheme boundaries) are performed. For instance, the reconstruction and morpheme insertion of the whole instances of the Ge'ez word ተቀደሱሙ'təḱədəsomu' could be as shown in Fig. 4.

Root Extraction

To extract the root of the verb, reprocess will be taken. For example, consider the verb 'ḱədəsə' in Figure 4: reprocess was taken in order to remove the vowels and to get the root 'ḱds'. In oursystem we assumed that the root of Ge'ez

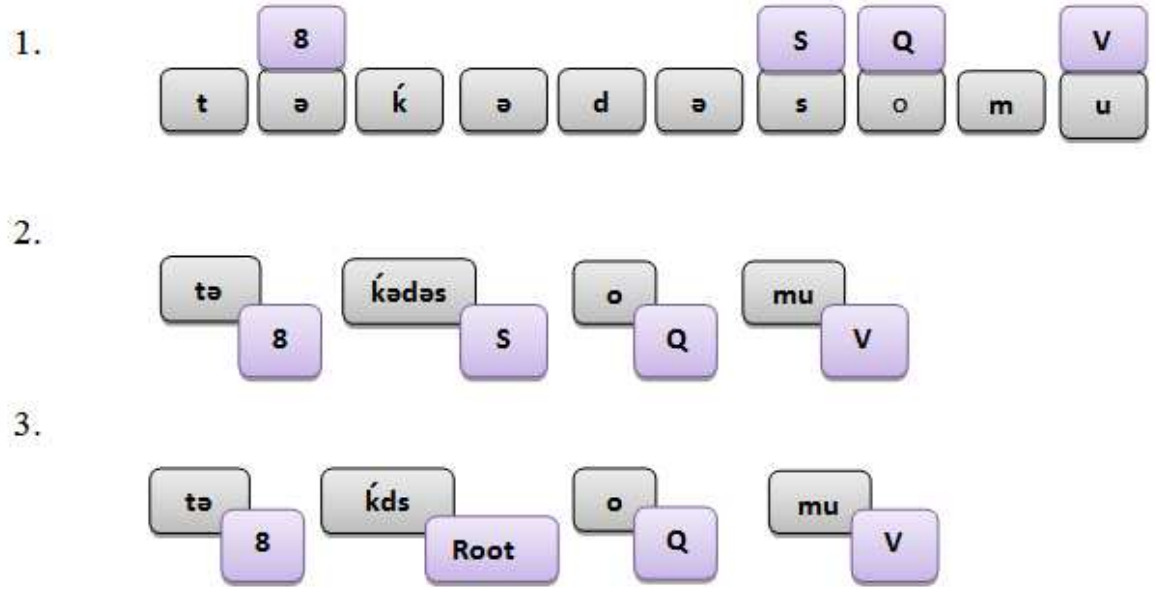


Figure 4: Reconstruction of the Morphological Analysis of the Ge'ez Word ተቀደሱሙ 'təḱədəsomu'

verbs can be deduced from the stems by removing the vowels.Because like Amharic language root in Ge'ez is represented by only with the sixth characters of consonants which we called in Ge'ez 'sads' except some words

that begin with avowel. There is an exception that we must consider. Whenever the system finds some verbs that start with avowel, it doesn't remove the vowel immediately rather it ignores and passes it as analyzed root. Moreover, mono and bi-radical verb types should also be considered not to remove their vowels since those verb types have valid meaning when they exit with vowels. For example, we can consider the four inflected words as shown in Table 12.

Table 12: Example of Stem and root extraction of verbs

	Word	Stem	Root	
1	fəḱədiku	[fəḱəd]$_{perfective}$ [iku]$_{1ps}$	[fḱd]$_{perfective}$ [iku]$_{1ps}$	Ge'ez verbs
2	śeməni	[śemə]$_{perfective}$ [ni]$_{object}$		
3	geśə	[geśə]$_{perfective}$		
4	səbərəć	[səbər]$_{perfective}$[əć]$_{3fs}$	[sbr]$_{perfective}$[əć]$_{3fs}$	Amharic verbs
5	mət'an	[mət'a]$_{perfective}$[n]$_{object}$		

The morphological analysis process of some bi-radical verbs will end in stem extraction. For example, as shown in Table 12,verbs which are listed in number 2,3 and 5 ends in stem extraction. Because when we remove vowels 'e-ə' from ሣመ /śemə/ andገሠ /geśə/ gives another meaning ሥም /śm/ ግሥ /gś/ which have different meanings respectively. Therefore, to get the roots of such type of verbs first we have to change into three radicals like ሣመ /śemə/ and ገሠ /geśə/ will be changed to ሠየመand ገየሠrespectively.Vowels in some verbal stems serve as consonants. For example, when the verb ባ(he is present)changed intoባአወ; አ is serving as aconsonant.Therefore, such exceptions should be considered during implementation process.

IV. Experiments

4.1 The Corpus

In order to train a classifier that can predict the class of new, previously unseen words correctly, a set of training examples that are manually prepared with the correct input format are needed. Because the basis of classification in TiMBL is the storage of all training examples in memory, a test of the classifier's accuracy must be done on a separate test set [8]. Therefore, we split each datasets into training and testing. The total of our corpus contains 1105 Ge'ez words and 1022 Amharic words which count 12135 and 8075instances respectively. Within these instances, 31 Ge'ez and 26 Amharic class labels occur. Therefor, we have two datasets (Amharic and Ge'ez) which are trained and tested separately with different algorithms.

4.2 Test Results

The memory-based classier has been trained and tested by the more common method called 10-fold cross-validation technique. In it, each data set was split into 10 parts of equal size. Each of these parts was used as test set once and the remaining parts were concatenated to be the training set. In the 10-fold cross-validation experiments, the system is trained on approximately 90% of the corpus and then tested on the remaining 10%. This is repeated 10 times (i.e., for 10 folds), with a different part of the corpus being used as a test corpus each time. All instances in the corpus are entirely included in either the test corpus or the training corpus. To get an optimal accuracy, this testing has been applied in two batches, the first one using TiMBL's default parameter settings and the second one with a limited form of parameter optimization. Based on these parameters of each algorithm (IB1, IGTREE, IB2 and TRIB2) Table 13 and Table 14 shows that the overall accuracy of the default and optimized parametersrespectively.Notice that we have implemented the two sets (Ge'ez and Amharic) using the four algorithms exchangeable. During this we have got nearly similar results. But here we presented Amharic (using IB1 and IGTREE) and Ge'ez (using IB2 and TRIB2).

Table 13: Average Performance of 10-fold CV Experiment with Default Parameter Setting

Evaluation method	*Algorithm*	*Compression (%)*	*Time taken (seconds)*	*Size of instances base (byte)*	*Accuracy (%)*	*Language*
10-fold CV	IB2	56.48	0.177	1898204	91.72	Ge'ez
	TRIBL2	55.98	0.0725	3761076	91.19	
	IB1	50.84	0.87831	1197760	92.02	Amharic
	IGTREE	97.63	0.07158	51866	76.27	

Many classifiers are parameterized and their parameters can be tuned to achieve the best result with a particular dataset. In most cases, it is easy to learn the proper value for a parameter from the available data [33]. After tuning a number of parameters along with the algorithms, we identified the parameters with high performance as optimized parameters. These are nearest neighbor: k=5, distance metric: modified value difference metric (-mM), feature weighting: information gain (-w2) and distance-weighted class voting: inverse distance (-d ID).

Table 14: Average Performance of 10-fold CV Experiment with Optimized Parameter Setting

Evaluation method	*Algorithm*	*Compression (%)*	*Time taken (seconds)*	*Size of instances base (byte)*	*Accuracy (%)*	*Language*
10-fold CV	IB2	56.66	0.263	1882188	93.24	Ge'ez
	TRIBL2	52.4	0.082	4066300	92.31	
	IB1	50.93	0.821	1213656	93.59	Amharic
	IGTREE	99.1	0.037	1136582	82.26	

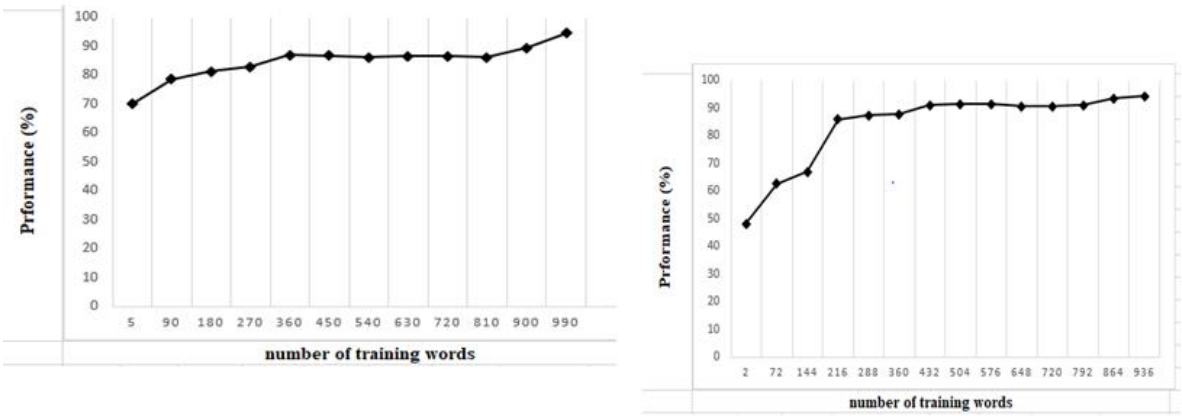
As shown in Tables 13 and 14, IB1 and IB2 show better general performance than IGTREE and TRIBL2 in both default and optimized parameter settings. It also performs better in compression the instance trees and uses less memory than TRIBL2. On the other hand, TRIBL2 processes within short seconds than IB2 on the same number of instances. It is also true for IB1 and IGTREE. IGTREE processes with in short seconds than IB1 on the same number of instances.

Aside from the percentage of correctly classified test instances, we used some more evaluation metrics that have become common in information retrieval and machine learning namely precision, recall, and F-score (f-measure). The precision, recall,and F-score for unknown- words with default and optimized parameters described in Table 15.

Table 15: 10 folds CV Results of Average Precision, Recall and F-score with Default and Optimized Parameter Settings

	With default parameters			**With optimized parameters**			
Algorithm	***Precision***	***Recall***	***F-score***	**Precision**	**Recall**	**F-score**	**Language**
IB2	*52.9*	*52.1*	*52.49*	55.6	56.3	59.95	Ge'ez
TRIBL2	*55.4*	*56.6*	*55.99*	58.8	60.3	59.54	
IB1	*81.26*	*80.80*	*81.04*	85.96	89.04	87.47	Amharic
IGTREE	*59.40*	*65.74*	*62.41*	73.04	74.12	73.58	

As shown in Table 15, optimizing some of the parameters achieved abetter result than default parameter settings on both algorithms. In memory-basedlearning, the minimum size of the training set to begin with is not yet specified.However, as stated in many kinds of literature the size of the training data matters the learning performance of the algorithm. Hence, it is crucial to draw learning curves in addition to reporting the experimental results [34]. We perform series of experiments on systematically increased amounts of training material up to the currently available total dataset in both cases. In most cases to draw a learning curve, the learning can be measured by fixing the test set against which the increased model is systematically tested. The learning curve of our system is shown in Fig. 5.



(a) **(b)**

Figure 5: Ge'ez (a) and Amharic (b)Learning Curve with Increasing Number of Words

V. Conclusions and Future Work

For some languages, it is difficult to address and analyze all the morphological features of the language. Especially a language like Ge'ez and Amharic, it is difficult to address all features since theyare complex inflected languages. Due to this,in this research work, we addressed the morphological analysis of Ge'ez and Amharic verbs only. Therefore,the work is aimed at developing Ge'ez and Amharic verbs morphological analyzer using memory-based approach. Given the promising results, our work adds value and initiatesother researchers in the overall effort to dealing with the complex problem of developing Ge'ez and

Amharic verbs morphological analyzer. The performance of our systemcan be further enhanced by increasing the training data. Future work is recommended to be extendedinto nouns, adjectives, compound- words and complex verbs. The use of a large training data has huge importance in enhancing the performance of the system. In other words, initiation of a big project to develop an efficient full-fledged automatic morphological analyzer for both languages which works for all POS categories is required.The complex nature of Ge'ez and Amharic the system can't segment some words. So, managing spelling changes (inserting, deleting) and extracting roots are also possible toproject areas.

References

[1]. Koray Ak, OlcayTanerYıldız (2011). Unsupervised Morphological Analysis Using Tries: Dept. of Computer Science and Engineering. Isık University

[2]. JishaJayan and Rajeev Rajendran (2011). Morphological Analyser and Morphological Generator for Malayalam - Tamil Machine Translation, Tamil University, Thanjavur

[3]. http://en.wikipedia.org/wiki/Morphology_%28linguistics% last accessed on 8 May 2014 at 09:11.

[4]. Sebastian Reiner Spiegler (2011). Machine learning for the Analysis of Morphologically Complex languages: a thesis submitted to the University of Bristol in accordance with the requirements for the degree of Doctor of Philosophy, Merchant Venturers School of Engineering University of Bristol

[5]. Anand Kumar, Dhanalakshmi, and Krishnan Soman (2010). A Sequence Labeling Approach to Morphological Analyzer for Tamil Language, Department of Linguistics, Tamil University, Thanjavur, India.

[6]. Pachalam Antony, Anand Kumar and Krishnan Soman (2010). A paradigm based Morphological Analyzer for Kannada Language Using Machine Learning Approach: Vol. 3 Issue 4, p457

[7]. Michael Gasser (2011). Computational Morphology and the Teaching of Indigenous Languages,School of Informatics, Indiana University

[8]. Walter Daelemans, Jakub Zavrel, Ko van der Sloot and Antal van den Bosch (2010). TiMBL: Tilburg Memory-Based Learner version 6.3 Reference Guide: ILK Technical Report – ILK10-01http://ilk.uvt.nl/timbl/

[9]. Anand Kumar (2013). Morphology-Based Prototype Statistical Machine Translation System for English to theTamil Language, A Thesis Submitted for the Degree of Doctor of Philosophy in the School of Engineering, Amrita School of Engineering Amrita Vishwa Vidyapeetham Coimbatore-641 112, Tamilnadu, India

[10]. Samer Al Moubayed, AnanthaKrishnan and Laura Enflo (2011). Automatic Prominence Classification in Swedish Centre for Speech Technology,Royal Institute of Technology (KTH), Lindstedtsvägen 24, SE-10044, Stockholm

[11]. Sotiris Kotsiantis (2007). Supervised Machine Learning: A Review of Classification Techniques, Department of Computer Science and Technology theUniversity of Peloponnese, Greece End of Karaiskaki, 22100, Tripolis

[12]. Walter Daelemans and Antal van den Bosch (2005). Memory-Based Language Processing,University of Antwerp and Tilburg University

[13]. Phil Blunso (2004). Hidden Markov Models,

[14]. Antal van den Bosch, Erwin Marsi, and Abdelhadi Soudi (2005). Memory-based morphological analysis and part-of-speech tagging of Arabic: ILK/Dept. of Language and Information Science, Faculty of Arts, Tilburg University.

[15]. Bart Decadt, Jacques Duchateau, Walter Daelemans, and Piet Wambacq (2002). Memory-based phoneme-to-grapheme conversion. Computational Linguistics in the Netherlands 2001. Selected Papers fromthe Twelfth CLIN Meeting, Amsterdam, pp. 47-61.Rodopi.

[16]. Paul Lendvai, Antal van den Bosch, and EmielKrahmer (2003). Memory-based disfluency chunking. In Proceedings of DISS'03, Disfluency in Spontaneous Speech Workshop, Goteborg University, pp. 63 - 66.

[17]. Antal van den Bosch, Sander Canisius, Walter Daelemans, Jens Hendrickx, and Karl Sang (2004). Memory-based semantic role labeling: Optimizing features, algorithm, and output. Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004), Boston, MA.

[18]. Walter Daelemans, Jakub Zavrel, Ko van der Sloot and Antal van den Bosch (2003). TiMBL: Tilburg Memory-Based Learner version 5.0 Reference Guide: ILK Technical Report – ILK 03-10 http://ilk.uvt.nl/timbl/

[19]. Demon Hand, Metro Mannila, Padhraic Smyth (2001). Principles of Data Mining. The MIT Press.

[20]. Daha Aha, John Kibler and JolmAlbcrt (1991). Instance-based Algorithms in Machine learning vol. 6, pp. 37-66.

[21]. Wolf Leslau (1987, 1991). Comparative Dictionary of Ge'ez (Classical Ethiopic): Ge'ez-English / English-Ge'ez with an index of the Semitic roots, Wiesbaden Harrassowitz, ISBN- 1-447-02592-1

[22]. Bender Myron Lienel (1976). Language in Ethiopia: pages 23-27 and 99-106. Oxford University Press, London.

[23]. Dillmann Carl Bezold (1899). Ethiopic Grammar. London: Amsterdam Hilo Press

[24]. ዘርአዳዊትአድሐና (መምህር) (1996) ፣ መርኖሰዋስዉዘልሣነግእዝ ፣ አዲስአበባ ፣ብርሃንናሰላምማተምያድርጅት

[25]. GabriellaScelt (2001). The Comparative Origin and Usage of the Ge'ez writing system of Ethiopia, A paper submitted to Professor Pilar Quezzaire-Belle in partial fulfillment of the requirements for Arts of Africa, AH

[26]. ደሴቀሉብ(መምህር) (2002). ትንሣኤግእዝ, በኢትዮጵያኦርቶዶክስተዋሕዶቤተክርስተያንበሰንበትትምህርትቤቶችማደራጃመምሪያማኅበረቅዱሳን: አዲስአበባ

[27]. Mulugeta, W., Gasser, M (2012). Learning Morphological Rules for Amharic Verbs UsingInductive Logic Programming. In: Proc. of SALTMIL8/AfLaT

[28]. Yimam, B. (2000). Ye'amarignasewasew (Amharic Grammar). Eleni Printing Press, Addis Ababa.

[29]. Antal van den Bosch and Walter Daelemans (1999). Memory-based morphological analysis for theDutch language: computational linguistics, Tilburg University.

[30]. Walter Daelemans and Antal van den Bosch (2009). Memory-based language processing, CLiPS Research Group, University of Antwerp, ILK / Tilburg center for Creative Computing, Tilburg University, the Netherlands.

[31]. Jaime Teixeira, Carlos Oliveira and CoentraoMoutinho (2006). Machine Learning of European Portuguese Grapheme-To-Phone Conversion using a Richer Feature Set, vol. 4, no 6

[32]. Guy DePauw and Gilles-Maurice deSchryver (2008). Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes, Language Technology Group, University of Antwerp and Department of African Languages and Cultures, Ghent University, Belgium

[33]. Payam Refaeilzadeh, LeitangHuanliu (2008). Cross-validation Arizona State University.

[34]. Walter Daelemans and AntalVanden Bosch (2005). Memory-based language processing, Published in the United States of America by Cambridge University Press, New York

[35]. Geert Booij (2009). The Oxford Handbook of Grammatical Analysis Oxford: Oxford University Press pp 563-589
[36]. John McCarthy (1981). A prosodic theory of non-concatenative morphology, Linguistics Department Faculty Publication Series, University of Massachusetts Amherst

Yitayal Abate. " Morphological Analysis of Ethio-Semitic Languages" IOSR Journal of Computer Engineering (IOSR-JCE) 21.4 (2019): 16-29.