INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

# Analysis of implemented part of speech tagger approaches: The case of Ethiopian languages

**Wubetu Barud Demilie**[1]*

**1** Department of Information Technology, Wachemo University, Hossana, Ethiopia, P.O. Box 667

OPEN ACCESS

**Received:** 19.10.2020

**Accepted:** 12.12.2020

**Published:** 30.12.2020

**Citation:** Demilie WB (2020) Analysis of implemented part of speech tagger approaches: The case of Ethiopian languages. Indian Journal of Science and Technology 13(48): 4661-4671. https://doi.org/10.17485/IJST/v13i48.1876

* **Corresponding author**.

wubetubarud@gmail.com, wubetuB@wcu.edu.et

**Funding:** None

**Competing Interests:** None



Published By Indian Society for Education and Environment (iSee)

**ISSN**
Print: 0974-6846
Electronic: 0974-5645

## Abstract

**Objective**: To review Part of Speech (POS) tagging works that have been done for the Ethiopian languages. **Methods**: All methods that have been implemented to develop POS tagging for the Ethiopian languages have been mentioned. **Findings**: Since all implemented POS tagging methods have been mentioned in this work, the result will be used for future natural language processing researchers to select the best methodology. **Novelty**: The work includes all implemented POS tagging research works for the Ethiopian languages.

**Keywords:** Natural language processing; Ethiopian language; part of speech; labeling

## 1 Introduction

In the real world which is becoming a single village, the information and knowledge for human languages are becoming abundant. The interaction between each human language and culture is increasing as technology is advancing[(1)]. The need to work on and improve natural language technology is becoming necessary than ever before. Natural language processing is part of artificial intelligence which is the process of developing software applications that enable computers to understand human languages. Natural language processing applications may be done at different levels including word level, phrase level, sentence level, or semantic level. Computers cannot understand human languages simply as human beings can do so. They cannot understand the syntax of words and their semantics in sentences. But, as the data of each natural language is being increased, it becomes difficult for us humans to analyze and get the necessary contents manually from it. Human beings need the help of computers to manipulate the existing large amount of data. Such a requirement of computer's help leads natural language processing to emerge as an exciting discipline of information technology and related fields. Many languages, especially on the African continent, are under-resourced in that they have very few computational linguistic tools or corpora (such as lexica, taggers, parsers, or tree-banks) available[(2)]. Developing a POS tagger application is not a simple task due to many factors. One of the factors is the absence of a single method that can solve the POS tagging problems completely for any language. This study will concentrate on the implemented POS taggers for the Ethiopian languages. Hence, this paper is set to explore the analysis of all the implemented POS

tagger approaches and to identify the best and the recommended algorithm for the Ethiopian languages.

## 2 Part of Speech (POS) tagging

POS tagging means assigning labeling implies appointing linguistic classes for example suitable POS labels to each word in normal language messages and sentences. It is additionally the identification of the transform syntactic class of each word structure utilizing lexical and relevant data [1]. Assigning of POS tag to each expression of an unannotated text by hand is very tedious, which brings about the presence of different ways to deal with robotizing the undertakings. In this way, POS labeling is a strategy to computerize the explanation cycle of lexical classes in like manner. The cycle takes a word or a sentence as information allocates a POS tag to the word or each word in the sentence and creates the labeled content as yield [3] [4]. POS labels are otherwise called word classes, morphological classes, or lexical labels. The criticalness of these is the huge measure of data they give about a word and its neighbors.

POS labeling is valuable for syntactic parsing as taggers diminish equivocalness from the parser's info sentence, which makes parsing quicker by making the computational issue more modest, and the outcome will be less questionable. It likewise settles a few ambiguities that are not tended to by the syntactic parser's language model [3] [4]. The determination of the semantic highlights from the lexical portrayals and how they are related to the POS taggers are consistently a troublesome decision. Subjective etymological decisions, the application for which labeling is done, the presentation expected of the tagger, lastly the disambiguation power offered by the current language innovation are terrifically significant factors in deciding lexical component determination.

POS tag sets typically contain many different word classes. It is also a non-trivial task. Because some words in languages are ambiguous. They can belong to more than one class; the actual class depends on the context of use. There are many publicly available POS taggers on the web for different foreign languages. For example, it is possible to see the English version of the Hidden Markov Model-based POS tagger using Stanford tagger/parser [5] [6]. Example: "**We can can the can.**" (the word "**can**" corresponds to auxiliary verb, verb, and noun respectively). It generates word class information as follows.

Input sentence: "**We can can the can.**"

And the output sentence will be:

**"We/PRP can/MD can/NN the/DT can/NN. /."**

Where, PRP=Pronoun, MD=Verb, Modal, NN =noun, singular, common, DT=determinant, and.

=sentence terminator

And another POS tagger tags this sentence differently. For example, the Real-Time POS tagger tags it as follows.

Input text= "**We can can the can.**"

Output text= "**We +PRONPERS can +VAUX can +VI the +DET can +NOUN. +SENT**"

Where, PRONPERS= Personal pronoun, VAUX=Auxiliary verb, VI=Infinitive verb, DET=

Determinant, Noun= Noun and SENT= Sentence terminator

Different POS tagger approaches tag the same word differently and performances will also vary based on the POS tagger approach that has been selected by researchers of the area [7].

## 3 Related works

Numerous analysts have utilized various ways to deal with build up a POS tagger for the Ethiopian languages. The main endeavor was by [8] who endeavored to build up a Hidden Markov Model-based POS tagger for the Amharic language. An aggregate of 25 POS label sets has been separated from 300 words on a page which was likewise utilized for preparing and testing the POS tagger. The label sets have been filled in as a reason for the label sets utilized by resulting scientists. The shortcoming of this investigation was the created POS tagger can't appoint the POS tag of obscure words.

Another examination endeavor was made by [9]. He applied the Conditional Random Fields way to deal with creating Amharic language POS labeling and word division utilizing a little clarified corpus of 1000 words. The POS labels utilized by the specialist were gotten by consolidating a portion of the classifications proposed by [8]. Inside the given size of the information and an enormous number of obscure words in the test corpus (80%), a precision of 74% for POS labeling and 84% for Amharic language word division was gotten. The accomplished outcomes were acceptable particularly when they had seen from the outcomes accomplished in obscure word acknowledgment techniques for POS labeling tests. A few highlights were analyzed for division and POS labeling.

Character highlights and word reference-based highlights were discovered to be helpful for the division assignments while morphological and lexical highlights fundamentally improve the consequences of the POS labeling task. The outcomes could be accomplished since the Conditional Random Field approach permits character highlights for the division errands while

coordinating a few covering highlights, for example, morphological and lexical highlights for POS labeling along these lines empowering ideal usage of the accessible data. As needs are, Conditional Random Fields were relevant for morphologically rich and complex dialects like Amharic. As a rule, the scientist managed confined parts of the morphological investigation of Amharic language, which was Amharic language word division and POS labeling. Besides, these undertakings were completed generally and freely because of the scarceness of assets.

At last, the scientist suggested that future work ought to investigate how division and POS labeling could be incorporated into a solitary framework that considers fine-grained POS labeling of Amharic language words. The creator additionally suggested that the advancement of a standard Amharic language POS label sets and explanation of a sensibly estimated corpus ought to be given need.

The work by[(2)] applied three supervised POS taggers, for example, Hidden Markov Model, Support Vector Machine, and Maximum Entropy for the Amharic language. The creators utilized a physically explained corpora of 210,000 tokens created at the Ethiopian Language Research Center (ELRC) of Addis Ababa University for preparing and testing the POS tagger errands. They likewise utilized the decreased 10 label sets that have been utilized in[(9)], the first label sets created at ELRC (comprising 30 label sets), and the diminished label sets of the ELRC label sets (comprising 11 label sets). On predefined folds, all POS taggers got equivalent aftereffects of (92.5%-92.8%) on the diminished label sets and (85.5%-88.3%) on the full label sets. The Support Vector Machine tagger had the best presentation on obscure words yet was a bit more regrettable on known words. Trigrams 'n' Tags gave the best outcomes for realized words yet had the most exceedingly awful exhibition on obscure words. The Maximum Entropy approach gave the best precision on its folds, 90.1% on the full label sets, and equivalent consequences of (94.5% - 94.65%) on two diminished sets.

Generally speaking, Support Vector Machine was marginally in a way that is better than Trigrams 'n' Tags on the two more modest label sets and better on the enormous label sets, and to some degree better than Maximum Entropy on every one of the three label sets. At long last, to improve labeling precision, the scientist suggested that further investigations ought to be led on three fundamental ideas including unequivocal morphological handling to treat obscure words, consolidating taggers that draw on various qualities of the preparation information, and semi-directed or solo POS labeling for the Amharic language.

In[(10)] also conducted an Amharic language POS tagger developed for factored language modeling. Hidden Markov Model and Support Vector Machine based taggers have been trained using the Trigrams 'n' Tags and Support Vector Machine Tools. For this purpose, the researchers have used the same data used by[(2)]. Then, the overall accuracy of 82.99% and 85.50% have been achieved for Trigrams 'n' Tags and Support Vector Machine-based taggers respectively. Accordingly, this indicates that Support Vector Machine based taggers perform better than Trigrams 'n' Tags based taggers although Trigrams 'n' Tags based tagger was more efficient about speed and memory requirement. Therefore, the Support Vector Machine tagger was used to tag the texts for factored language models development for which the estimation of the probability for each word depends on the previous one or two words and their POS. Then, using these language models, they have improved the accuracy of Amharic speech recognition (1.32%)[(10)].

In[(11)] have developed a POS tagger for Tigrigna language by applying a hybrid (which was a combination of Brill transformation-error driven learning and Hidden Markov Model) approaches. He has collected a total of 26,000 words from Tigrigna news broadcasting agencies and annotate manually with their corresponding word classes and 75% (20,000) of the words were used for training purposes and the remaining 25% (6000) of it was used for testing purpose. In addition to this, he has identified 36 tag sets for the entire tagging process. This study finds the tag of a word from the raw text in two main steps. The first step was performed by the Hidden Markov Model tagger and it first annotates the given raw text and provides a level of confidence (threshold value) for each tag sequence. The second step was performed by comparing the confidence level of each tag sequence with the minimum confidence level that was set by the researcher using the output analyzer module. During those steps, if the confidence level is less than that of the minimum confidence level, a window size of two (bigram of the word) is given to the rule-based tagger for correction. Otherwise, it was treated as a correct tag. He conducted different experiments for the three types of taggers namely the Hidden Markov Model, rule-based, and hybrid taggers to test the performance of the tagger that he had developed. Finally, he has got an accuracy of 89.13% for the Hidden Markov Model, 91.8% for rule-based, and 95.88% for the hybrid taggers.

In[(12)] also tried to develop supervised POS tagging for the Amharic language. This work was different from previous works because of its degree of cleaning the corpus, good feature selection, and parameter values in the selected and implemented approaches. Besides, the features used in other machine learning-based tagging methods, the researcher included two other unique features, the vowel patterns, and the radicals. These additional features reduced the impact of the data sparsity problem to some degree. All these factors had a significant impact on the final performance of the achieved result. The data set used for this study consists of 207,000 tokens (186,000 for training and 21,000 for testing). The original tag sets developed at ELRC (consisting of 31 tags) were used.[(12)] the experimental result shows; the highest POS tagging accuracies have been achieved in

both Conditional Random Fields and Support Vector Machine, followed by Brill tagger and Trigrams 'n' Tags. The Conditional Random Fields tagger achieved an average accuracy of 90.95% on 10 fold cross-validation while under the same circumstance, the Support Vector Machine achieved an average of 90.43%. Brill tagger and Trigrams 'n' Tags achieved comparable results. Even though the results obtained in this experiment were higher than the previous results, it was still far behind Arabic and English languages, where accuracies were above 97%. Therefore, as a recommendation, the researcher stated that to achieve the required accuracy using stochastic methods, there should be a cleaned corpus.

In[13] have built up a POS tagger for the Afaan Oromo language by utilizing the Hidden Markov Model. In this work, they have utilized the Hidden Markov Model methodology for building up the tagger and they have gathered 159 sentences (with an aggregate of 1621 words for both preparing and testing purposes) from various sources to make the corpus adjusted, and they have utilized 17 label sets.

For the labeling cycle, they have utilized two stages to appoint word classes to a given Afaan Oromo text. The primary period of the tagger trains on the preparation information to register and store the lexical and momentary probabilities of the preparation information by utilizing unigram and bigram models of the Viterbi calculation by taking the put away data and the second period of the tagger acknowledges untagged Afaan Oromo messages and tokenized into words. After this, the tagger relegates the right POS tag for each of the tokenized words. The presentation of the tagger has tried utilizing a ten times cross-approval component and they got an exactness of 87.58% and 91.97% for unigram and bigram models individually. At long last, they have prescribed different analysts to build up a POS tagger for other neighborhood dialects by utilizing a similar methodology.

In[14] conducted POS labeling trials to distinguish the best strategy for under-resourced and morphologically rich languages like Amharic utilizing various sorts of approaches) and preparing information sizes (25%, half, 75%, and 100% of the preparation set). The POS label sets and the corpus used to prepare and test the taggers utilized were the ones created by ELRC. The creators had the option to show then Memory-Based Tagger was a decent labeling system for under-resourced and morphologically rich dialects, for example, Amharic with little size informational indexes contrasted and different strategies, especially Trigrams 'n' Tags. Besides, dividing words made out of morphemes of various POS labels and label theories mixes are additionally distinguished as they were promising headings to improve labeling execution for morphologically rich and under-resourced dialects individually. At long last, the specialists suggested that the best taggers recognized ought to be applied in programmed discourse acknowledgment just as measurable machine interpretation undertakings.

The Amharic language POS tagger, which was done by[9], was experienced utilizing a little size of preparing corpus, bringing about a word mistake pace of over 25%.

In[15] focused on checking, amending, and retagging Amharic language text corpus by partaking in the Amharic language news stories of 1065 (comprises 210,000 words) gathered at Stockholm University from an Ethiopian web news document, and afterward morphologically broke down and physically POS labeled at Addis Ababa University.

200,863 word POS labeled corpus of Amharic language news writings were made by cleaning, normalizing, and checking a public accessible physically labeled corpus. The corpus has been increased with three diverse label sets (each 30, 11, and 10 labels). The labeled corpus was utilized as the reason for testing the AI procedures and apparatuses created for the Amharic language. The labeling precision of around 90% was accomplished on the most troublesome label sets which were not extremely promising, and not valuable for the errand of labeling the rest of the corpus. Other than this,[15] improved the word blunder rate accomplished by[9] to figures underneath 10% utilizing a 200,000-word corpus. Yet, the number was still high when contrasted with better-resourced language, for which Word Error Rate of 2–4% was normal. Along these lines, the analyst suggested that further investigations ought to be directed at confirming, rectifying, and retagging Amharic language text corpus[15].

In[16] developed a POS tagger for the Kafi-noonoo language by applying a hybrid (which was a combination of Brill transformation-error driven learning and Hidden Markov Model) approaches. He has collected a total of 354 untagged sentences from two different genres and annotated them using an incremental corpus preparation approach. After assigning word class information on each word within the sentences, both Hidden Markov Model and rule-based taggers were trained on 90% of the tagged sentences to generate probabilities i.e. lexical and transitional probabilities for the statistical component of the hybrid tagger and a set of transformation rules for the rule-based component of the hybrid tagger. Both the rule-based and Hidden Markov Model taggers have been trained on 90% of the tagged sentences. In addition to this, he has identified 34 tag sets for the entire tagging process. Finally, he has got an accuracy of 77.19% for the Hidden Markov Model, 61.88% for rule-based, and 80.47% for the hybrid tagger.

In[17] also conducted an iterative automatic annotation process using the WebAnno tool and Margin Infused Relaxed Algorithm[18], an online machine learning algorithm, and produced an F1 score of 0.89 for Amharic language documents collected from the web. For this research, they have adapted the tag sets used by previous researchers (consisting of 11 tags) that were compatible with the Universal tag sets[19].

In the work of[20], he has investigated the utilization of one of the conditions of the craftsmanship probabilistic model for grouping characterization, the Adopted Transformation-based Error-driven learning approach, and has gathered 17,473 words from around 1100 sentences containing 6750 unmistakable words. At last, the adjusted Brill's Tagger indicated a precision of 80.08% though the improved Brill's Tagger result demonstrated an exactness of 95.6%.

In[1] developed a POS tagger for the Amharic language by using an unsupervised approach. The research raised three different and important research questions to answer and how these research questions have been answered within the study.

The first question was "How to prepare a huge amount of corpora for the study". Based on this question, 929, 526 sentences were collected for the study.

The second question was "How to modify Amharic language tag sets for POS tagging activities". The question was answered by reviewing previously conducted research works on Amharic, and Tigrinya language tag sets and exploring the specific properties of the languages and finally modifying Amharic language tag sets.

The third question was "How to apply unsupervised POS tagger on Amharic language text documents". Here, the question was answered by preparing training data sets in a way that was appropriate for the study and it was prepared by removing non-Amharic characters, segmenting sentences per line, tokenizing words and normalizing the data sets, and then applying them for the Amharic language which was already prepared data sets in different remote machines. 37 sentences of test data sets have been prepared in WebAnno with an evaluation accuracy of 66.98% for eleven-word categories. The performance achieved was less than the work of[21] unsupervised POS tagger result because the tagger was not trained very well on the test data sets which was used for[1] research work so it cannot be capable of assigning POS tag of the test data accurately.[1] have used test data sets with trained tagger and it was possible to achieve better performance. So, the evaluation result using additional seven sentences, and the accuracy was improved to 70.25%.

In[21] developed unsupervised POS tagger for the Amharic language. The training data set was constructed from the Walta Information Center corpus that contains more than 210,000 tokens. Besides, the morphological, syntactic, positional information, and frequency features were used to represent each word. In the development of the tagger, the research had followed the following procedures. Firstly, the unlabeled data were divided into 10-folds and segmented. The raw text was divided into sentences and tokenized into words. Secondly, features such as distributional, syntactic, and morphological features were extracted. Clustering was performed in the third phase and the k-means clustering algorithm, which forms groups of similar lexicons, has been selected and implemented. The last phase was mapping, which deals with looking at each cluster carefully and the most common tag was assigned for a group. Based on the experiments conducted using different features, the performance of the system shows that it achieves a maximum of 81% accuracy.[21] considered only five POS tags. Since the k-means algorithm was used, the number of clusters (k) given by the user restricts words in the corpus to be clustered in one of those clusters. Therefore, words that have other word categories were not considered. Different word categories that share similar features were also assigned together. This indicates that the features selected were not enough. In addition to that, the training data of small size (consists of 210,000 tokens) was used. This in turn maximizes the rate of unknown words. Therefore, as a recommendation, the researcher stated that future work should be conducted on hierarchical clustering by incorporating semantic features. Besides this, building a large amount of raw corpus was also recommended undertaking extensive experimentation.

Another work for the Amharic language POS tagger has been developed by using Machine Learning Approaches[22]. The work aimed to improve POS tagging performance for the Amharic language, which was never above 91%.

The data sets used in this study were categorized into three main categories, the Ethiopian Language Research Center annotated corpus that contains 210,000 words, the extended re-tagged corpus of the Ethiopian Language Research Center, and the newly annotated corpus of the Amharic language translation of the Quran and Bible. The overall average accuracy of 86.44, 95.87, and 92.27 for Ethiopian Language Research Center, ELEC-Extended, and ELRCQB tag sets respectively.

In[23] developed POS tagger using Neural Word Embedding as Features for the Amharic language. The experiments were conducted on some classifiers on the Weka environment and others developed using deep learning algorithms. In this research work, two basic tasks having a positive contribution to the Amharic language POS tagger were done. The first task was segmenting prepositions and conjunctions attached to the other POS tagger. The second task was tried to simplify the design of features by generating them automatically using the Word2Vec tool. Finally, the study was concluded within an accuracy of 88.88% for MLP, 92.8% for LSTM, and 93.7% for Bi-LSTM. The F-measure values for these networks are 88.81%, 92.75%, and 93.67% respectively.

In the work of[24], Machine Learning Approach-based Amharic language POS tagger has been developed. The researchers tried to collect a huge amount of compiled corpora from two sources. The first source was from Ethiopian Language Research Center which had around 210,000 tokens and was manually tagged with 31 tags and the second corpus was from a religious corpus containing 116,000 tokens which were manually tagged with 62 tags. All the collected corpora have been cleaned by

using different preprocessing mechanisms and the total corpus had become 16451 sentences (around 321,109 tokens). They have shown a comparison among statistical-based taggers including Conditional Random Fields, Hidden Markov Model-based Trigrams 'n' Tags, and Naive Bays based taggers. They have checked and compare the performances of all taggers with similar sizes of training and testing data set. The result of the experiment showed that the Conditional Random Fields approach was a super tagging strategy for Amharic languages, as the accuracy of the tagger was less affected, after it reaches at some point, as the amount of training data increases compared with other methods. Finally, the best accuracy obtained from their experiments using Conditional Random Fields was 94.08%. Other research works have been done for the Amharic language POS tagging which includes [25]

The Table 1 summarizes all POS tagger researches for the Ethiopian languages that have been done by different researchers in the area.

**Table 1.** Summary of implemented POS tagger approaches for the Ethiopian languages

| S.No. | Author (s) | The objective of the Study | The methodology of the Study | Key Findings of the Study | Remarks |
|---|---|---|---|---|---|
| 1. | [8] | • To develop a POS tagger for the Amharic language | • Hidden Markov Model<br>• 300 tokens<br>• 25 POS tag sets | • He developed a tagging prototype using the Hidden Markov model | • The tagger can't appoint the POS tag of obscure words |
| 2. | [9] | • To show the applicability of Conditional Random Fields in POS tagging for a morphologically complex language like Amharic | • Conditional Random Fields<br>• 1000 tokens have been prepared<br>• 10 POS tag sets | • An accuracy of 74% is obtained for POS tagging and 84% for Amharic word segmentation | • The word division and POS labeling were completed generally freely because of scant assets |
| 3. | [2] | • To conduct a supervised POS tagging for the Amharic language | • Hidden Markov Model, Support Vector Machine, and Maximum Entropy<br>• 210,000 tokens<br>• 10, 30, and 11 POS tag sets | • 92.6-92.8% on the selected tag sets, • 92.5-92.8% on the reduced sets<br>• 85.5-88.3% on the full tag set | • here was no unequivocal morphological handling to treat obscure words and no other information source utilized than a labeled preparing corpus |
| 4. | [10] | • To conduct Amharic language part of speech taggers developed for factored language modeling | • Hidden Markov Model and Support Vector Machine<br>• 210, 000 tokens<br>• 30 POS tags | • Accuracies such as 82.99% for Trigrams 'n' Tags and 85.50% for Support Vector Machine is obtained | • The tagger has been less performance in assigning the POS tag of unknown words |
| 5. | [12] | • To conduct POS tagger for the Amharic language which performs better | • Conditional Random Fields, Support Vector Machine, Brill, and Trigrams 'n' Tags.<br>• 207,000 tokens<br>• 30 POS tag sets | • Average accuracies of 90.95% for Conditional Random Fields and 90.43% for Support Vector Machine are achieved. Brill and Trigrams 'n' Tags achieved comparable results | • To achieve a better performance, the the corpus should be cleaner |
| 6. | [26] | • To develop POS tagger for Afaan Oromo language | • Transformational Error driven Learning approach<br>• 223 sentences (1708 words)<br>• 18 tag sets | • The experiment was relatively good, 80.08% of the total word was correctly tagged | • The standardized and readily available corpus was very important for natural language processing application |

*Continued on next page*

*Table 1 continued*

| | | | | | |
|---|---|---|---|---|---|
| 7. | (11) | • To build up a POS tagger model for the Tigrigna language and examine the exhibition | • Hidden Markov Model, rule-based and hybrid<br>• 26,000 words<br>• 36 broad tag sets | • An exactness of 89.13%, 91.8%, and 95.88% for Hidden Markov Model, rule-based, and crossbreed taggers separately | • Preparing in huge corpus and utilizing huge label sets that can distinguish sex, number, tense, and so forth with various capabilities will improve exhibitions |
| 8. | (27) | • To direct POS labeling investigations to distinguish the best strategy for under-resourced and morphologically rich dialects | • Memory-Based Tagger, Trigrams 'n' Tags, Support Vector Machine, Conditional Random Fields tagging strategies<br>• 210,000 tokens<br>• 30 POS tag sets | • Before segmentation: (83.4-86.3%) and after segmentation: (91.6-93.5%) are achieved. Memory-Based Tagger was a decent labeling procedure for under-resourced dialects as the exactness of the tagger | • The tagger (particularly Trigrams 'n' Tags) can't allocate the POS tag of obscure words |
| 9. | (13) | • To develop POS tagger for Afaan Oromo language | • Hidden Markov Model<br>• 59 sentences (with a total of 1621 words)<br>• 17 tag sets | • The accuracy was 87.58% and 91.97% for unigram and bigram models separately | • The accuracy and effective processing of natural language processing applications that need annotated data sets were dependent upon standardized and sufficient amounts of the corpus. |
| 10. | (15) | • verifying, correcting and retagging Amharic text corpus | • Manual and Automatic way | • An accuracy of 90% was achieved | • The word error rate of the corpus prepared was still high when compared to better-resourced languages |
| 11. | (16) | • Design and develop POS tagger for Kafinoonoo language | • Hidden Markov Model, rule-based, and hybrid<br>• 354 untagged Kafinoonoo sentences | • Accuracy of 77.19%, 61.88%, and 80.47% for Hidden Markov Model, rule-based and hybrid taggers respectively | • Arrangement of a reasonable corpus that contains messages which speak to various types like papers, fiction, course books, parliamentary reports, and so on. Would have a great role in the performances |
| 12. | (28) | • To investigate the possibility of developing POS tagger for Wolaita text using small manually tagged text | • Conditional Random Fields and Hidden Markov Model<br>• 200 sentences<br>• 22 tag sets | • An accuracy of 83.58% and 74.63% using reduced tag set for supervised Hidden Markov Model and Conditional Random fields based taggers respectively | • Most of the Ethiopian languages are under-resourced and do not have large size POS tagger annotated corpus, they can benefit from a semi-supervised approach |
| 13. | (17) | • To facilitate the annotation process | • Margin Infused Relaxed Algorithm<br>• 504 tokens<br>• 11 POS tags | • F1 score of 0.89 for Amharic language documents collected from the web | • Supports annotation suggestions |

*Continued on next page*

*Table 1 continued*

| | | | | | |
|---|---|---|---|---|---|
| 14. | (1) | • To develop POS tagger for the Amharic language by using an unsupervised approach | • 37 sentences of test data sets have been prepared in WebAnno | • 66.98% for eleven-word categories and<br>• 70.25% seven additional sentences and the accuracy was improved | • To improve the exhibition of the unaided aspect of the tagger, there is a need to assemble an enormous measure of the crude corpus |
| 15. | (20) | • To improve Brill's tagger lexically and change the rule for Afaan Oromo POS labeling with an adequately huge preparing corpus | • Adopted Transformation-based Error driven learning approach<br>• 17,473 words from around 1100 sentences containing 6750 distinct words<br>• 26 broad tag sets | • An adjusted Brill's Tagger indicated a precision of 80.08% while the improved Brill's Tagger result demonstrated an exactness of 95.6% | • Utilizing a morphologically examined corpus for the preparation of Brill's tagger's to think about the inflectional properties of the language. |
| 16. | (29) | • To building another POS labeled corpus and build up a POS tagger for the Tigrinya language | • A Supervised Learning approach dependent on Conditional Random Fields and Support Vector Machines<br>• 72,080 tokens<br>• 73 tag sets | • For reduced label sets of 20, the general accuracy of 90.89% was acquired on a defined 10-overlap cross-approval | • Data from different genres and styles to achieve a more representative corpus |
| 17. | (30) | • To propose the sentence level POS tagger utilizing a half and half methodology that improves the exhibition of Tigrigna language tagger | • Hybrid approach<br>• 3100 sentences<br>• 22 tag sets | • An accuracy of 94.8%, 95.5%, and 96.3% for rule-based, averaged perceptron, and hybrid taggers respectively | • Associating additional word-class data about corpora substance of the type of word-class marker is a helpful errand in both the semantic and language innovation field |
| 18. | (31) | • To investigate the use of hybrid (rule-based and statistical hidden Markov models) approaches to the development of POS tagging for the Afaan Oromo language | • Hidden Markov Model, rule-based and hybrid tagger<br>• 1517 sentences<br>• 35 tag sets | • An accuracy of 91.9%, 96.4%, and 98.3% for Hidden Markov Model, rule-based and hybrid taggers respectively | • To increase the performance of the tagger wide coverage/domain area of training data and morphologically segmented words were recommended |
| 19. | (32) | • To analyze word embedding's and improve par of speech tagger of Tigrinya | • Conditional Random Fields<br>• 72,000 words | • The performance of Conditional Random Field-transform was 91% in general and 80% on obscure words | • Performance can be enlarged by increasing the size of the content corpus and tuning boundaries |
| 20. | (33) | • To develop POS tagger for Ge'ez language | • Adopt Trigrams 'n' Tags tagger to the hybrid tagger | • Accuracy of 77.87%, 82.23%, and 94.32% performances for Trigrams 'n' Tags tagger, Trigrams 'n' Tags tagger with Regex tagger, and Hybrid taggers respectively<br>• 15,154 words from around 1,305 sentences<br>• 26 broad tag sets | • Hybrid tagger performs better than the Trigrams 'n' Tags and Trigrams 'n' Tags with Regex tagger used individually |

*Continued on next page*

*Table 1 continued*

| | | | | | |
|---|---|---|---|---|---|
| 21. | (21) | • To develop an unsupervised POS tagger for Amharic language | • k-means clustering<br>• 210,000 tokens<br>• 5 POS tag sets | • A maximum of 81% accuracy was achieved | • Other word categories were not considered.<br>• The features selected were not enough.<br>• The data used for this study was not enough |
| 22. | (22) | • To improve POS labeling execution for the Amharic language, which was rarely above 91% | • Brill, Trigrams 'n' Tags, and CRFSuit<br>• Annotated corpus of 210,000 words | • The overall performance of 86.44, 95.87, and 92.27 for Ethiopian Language Research Center, ELEC-Extended, and ELRCQB tag sets respectively | • Because of an enormous number of labels that might not have great permeability, they demonstrated just the main top 20 label sets of a disarray lattice for the best score |
| 23. | (34) | • To develop POS tagger for Hadiyyisa language | • Rule-based and Trigrams 'n' Tags approaches<br>• 1280 manually tagged corpus of Hadiyyisa sentences<br>• 32 identified tag sets and 10 basic tag sets | • Accuracy of 66.64%, 64.65%, 72.54%, and 73.06% with 32 determined label sets labeled corpus and 80.34%, 72.67%, 87.25%, and 89.03% for Trigram Tagger, Bigram Tagger, Unigram Tagger, and Affix Tagger respectively | • Good performance was obtained by the Trigrams 'n' Tags with unknown word handling with the back off in the sequences |
| 24. | (23) | • To examine the impact of segmentation and neural word embedding features on Amharic language POS tagger | • Support Vector Machine<br>• J-48 (Decision Tree)<br>• 220,260 instances including punctuation marks | • Accuracy of 88.88% for MLP, 92.8% for LSTM and 93.7% for Bi-LSTM. The F-measure values for these networks were 88.81%, 92.75%, and 93.67% respectively | • Improving the quality of the existing corpus and increasing its size can be a significantly important task |
| 25. | (7) | • To develop POS tagger for Awngi language using | • Hidden Markov Model<br>• 94,000 sentences (with total word of 188,760 both for training and testing sets)<br>• 23 tag sets | • Accuracy of 93.64% and 94.77% for both unigram and bigram taggers respectively | • The standardized and readily available corpus was very important for natural language processing applications |
| 26. | (35) | • To develop POS tagger for Guragigna Language | • Hidden Markov model, a hybrid approach which was a combination of rule-based and Hidden Markov Model-based and Conditional Random Fields<br>• 6,745 words<br>• 17 tag sets | • Performance analyses of the taggers were 66.56, 74.46, and 78.42 for Conditional Random Fields, Hidden Markov model tagger, and Hybrid tagger respectively | • Adding of rule-based tagger performs better result than HMM tagger alone |
| 27. | (24) | • To conducted a comparison between Conditional Random Field, Trigrams 'n' Tags, Naïve Bays taggers | • Conditional Random Field, Trigrams 'n' Tags, Naïve Bays<br>• 16451 sentences (around 321,109 tokens) | • The F-measure performance was better than the performance achieved by Trigrams 'n' Tags 87.39 %<br>• F-measure and • Naïve Bays 81.25 % F-measure.<br>• However, the best accuracy obtained from their experiment using Conditional Random Fields was 94.08% | • There was a clear correlation between the training data size and the performance of Machine Learning approaches • The larger the training data size, the better the performance |

*Continued on next page*

*Table 1 continued*

| | | | | |
|---|---|---|---|---|
| 28. | [36] | • To build POS tagger for under-resourced language: the instance of Somali language | • Hidden Markov Model, Conditional Random Fields, and neural network<br>• 14,369 tagged tokens<br>• 19 tag sets | • An accuracy of 87.51% on a tenfold cross-approval | • Considering relative investigation on various methodologies, for example, Support Vector Machine, rule-based, and profound learning-based taggers with additionally preparing and testing information were normal |

## 4 Analysis of experimental results

As revealed by table 1, no one can produce 100% accurate results for all Ethiopian languages. Hence, all the implemented POS tagger approaches are useful in any natural language processing applications.

As the related works from the summarized table indicate, before developing any kind of POS tagger for the languages by using and of the approaches, the accuracy depends on the structure and the grammatical rules that should be identified and it needs a linguistic expert. Additionally, a detailed analysis of the morphology of the language words shows that all Ethiopian languages are morphologically rich. The types of affixation such as suffixes, infixes, reduplication, blending, compounding, and concatenation of suffixes in the language contribute a lot in generating rich morphological variants and make the word-formation process complicated. Therefore, attempting to conflate each language word manually is very tedious and extremely difficult. For this reason, applying automated conflation procedures such as the POS tagger is very important for the languages. To improve the performance of the taggers, it should be tested within a large number of corpora to prove its real performance since natural language processing applications need standard and balanced corpus (from different sources and genres) preparation. Hence, preparing the standard corpus for all Ethiopian languages could also be another research opportunity in this field. Accordingly, to enhance the performance of the tagger in all approaches of the POS tagger that have been implemented for Ethiopian languages, there is a need to build a large number of raw corpora. Hence, incorporating all necessary elements, the POS tagger can also be used as a component for developing other computational tools like morphological analyzer, parser, spell checker, thesaurus, text stemmer, word frequency counting, information retrieval, and the like of the language under consideration. Finally, evaluating the POS taggers on text collection of large size collected from different sources that can represent the characteristics of the language more than a small size sample will improve the accuracy of the POS taggers for Ethiopian languages

## 5 Conclusion

This study summarizes the works which have been done on Part of Speech Tagger (POS) for Ethiopian languages. Part of Speech (POS) taggers are otherwise called word classes, morphological classes, or lexical labels. The significance of it is the immense measure of data they give about a word and its neighbors. POS taggers are helpful for syntactic parsing as taggers decrease vagueness from the parser's information sentence, which makes parsing quicker by making the computational issue more modest, and the outcome will be less equivocal. Finally, this study can be used for future natural language processing researchers as a reference since natural language processing researches depend on POS tagger results.

### Acknowledgment

The author wants to acknowledge all researchers of the area who have been contributed a lot regardless of POS tagging research work for Ethiopian languages.

## References

1) Getinet Y. By yemisrach getinet . 2015. Available from: http://etd.aau.edu.et/bitstream/handle/123456789/6042/Yemisrach%20Getu.pdf?sequence=1&isAllowed=y.
2) Gambäck B, Olsson F, Argaw AA, Asker L. Methods for Amharic part-of-speech tagging. 2009. Available from: https://doi.org/10.3115/1564508.1564527.
3) Hasan FM. Comparison of different pos tagging techniques for some south asian languages. 2006. Available from: https://core.ac.uk/download/pdf/61799956.pdf.

4) Bhatt PM, Ganatra A. 2019. Available from: https://www.ijrte.org/wp-content/uploads/papers/v8i1/A9254058119.pdf.
5) Jurafsky D, Martin JH. Part-of-speech tagging Speech Lang. Process. In: and others, editor. An Introduction to Natural Language Processing Comput Linguistics, and Speech Recognition.;vol. 1. 2019. Available from: https://www.semanticscholar.org/paper/CHAPTER-8-Part-of-Speech-Tagging/5691f47c1d26ebcebddf41766554baa4116c9469.
6) Stanford NLP Group, Stanford Parser. 2015. Available from: http://nlp.stanford.edu:8080/parser/.
7) Demilie WB. Parts of Speech Tagger for Awngi Language;vol. 9 of 9. and others, editor. 2019.
8) Getachew M. Addis Ababa university school of. 2001. Available from: https://www.researchgate.net/publication/343211980_Parts_of_Speech_Tagger_for_Awngi_Language.
9) Aghaei S, Aslani FS. Systemic lupus erythematosus arising in a patient with epidermodysplasia verruciformis. *Lupus*. 2006;15:47–50. Available from: https://dx.doi.org/10.1191/0961203306lu2246cr.
10) Tachbelie MY, Menzel W. Amharic Part-of-Speech Tagger for Factored Language Modeling. 2009. Available from: https://www.aclweb.org/anthology/R09-1077.pdf.
11) Abreha TG. School of Graduate Studies Part of Speech Tagger for Tigrigna Language Part of Speech Tagger for Tigrigna," p. Angeles, L., Advocacy, S., Location, O. (2002). 2010. Available from: http://etd.aau.edu.et/handle/123456789/3266.
12) Gebre BG. School of Law, social sciences and binyam gebrekidan gebre part of speech tagging for amharic supervisors. 2010. Available from: https://www.semanticscholar.org/paper/Gebrekidan-Gebre-Part-of-Speech-Tagging-for-Amharic-Binyam/62d32ce0814195e2a73be1841fc4c0b23c08e111.
13) Mamo G, Meshesha M. Parts of Speech Tagging for Afaan Oromo. *International Journal of Advanced Computer Science and Applications*. 2011;1(3):1–5. Available from: https://dx.doi.org/10.14569/specialissue.2011.010301.
14) Tachbelie MY, Abate ST, Besacier L. Part-of-Speech Tagging for Under-Resourced and Morphologically Rich Languages - The Case of Amharic. 2011. Available from: https://www.researchgate.net/publication/228964535_Part-of-Speech_Tagging_for_Under-Resourced_and_Morphologically_Rich_Languages-The_Case_of_Amharic.
15) Gamback B. Tagging and Verifying an Amharic News Corpus. 2012. Available from: http://tekstlab.uio.no/ethiopia/gamback.pdf.
16) Mekuria Z. School of Graduate Studies College of Natural Sciences Department of Computer Science Design and Development of Part-of-speech Tagger for Kafi-noonoo Language Addis Ababa University School of Graduate Studies College of Natural Sciences Department of Comp. 2013. Available from: http://etd.aau.edu.et/handle/123456789/3752?show=full.
17) Yimam SM, Biemann C, Castilho RED, Gurevych I. Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. 2015. Available from: https://doi.org/10.3115/v1/p14-5016.
18) Crammer K, Singer Y. Ultraconservative online algorithms for multiclass problems, Lect. Notes Comput. Sci. (including Subsea. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2001;2111:99–115. Available from: https://doi.org/10.1007/3-540-44581-1_7.
19) Das D, Mcdonald R. A Universal Part-of-Speech Tagset. . Available from: http://petrovi.de/data/universal.pdf.
20) Ayana AG. Improving Brill's tagger lexical and transformation rule for Afaan Oromo language. *PeerJ Prepr*. 2015;3:1–11. Available from: https://doi.org/10.7287/peer.preprints.1225.
21) Fantahun Z. Automation & robotics. 2018. Available from: https://doi.org/10.4172/2168-9695-C1-015.
22) Gashaw I. Machine Learning Approaches for Amharic Parts-of-speech Tagging. 2018.
23) Argaw M. Amharic Parts-of-Speech Tagger using Neural Word Embeddings as Features Amharic Parts-of-Speech Tagger using Neural Word Embeddings as Features. 2019. Available from: http://etd.aau.edu.et/handle/123456789/17226.
24) Hirpssa S, S G. . Available from: https://arxiv.org/ftp/arxiv/papers/2001/2001.03324.pdf.
25) Kifle N. Amharic Word Sequence Prediction - ACL Anthology. . Available from: https://www.aclweb.org/anthology/W19-3607/.
26) Abubeker MH. A thesis submitted to the School of Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the Degree of Master of Arts in Sociology. *Control*. 2010.
27) Martha YT, Solomon TA, Besacier L. Part-of-Speech tagging for under-resourced and morphologically rich languages – The case of Amharic. In: and others, editor. Conference of Human and Language Technology Development. 2011;p. 50–55. Available from: https://www.researchgate.net/publication/228964535_Part-of-Speech_Tagging_for_Under-Resourced_and_Morphologically_Rich_Languages-The_Case_of_Amharic.
28) Ganta BH. Part of speech tagging for Wolaita language. 2015. Available from: https://ijesc.org/upload/37d26e28f04568f75a74a91059e59327.Part%20of%20Speech%20Tagging%20for%20Wolaita%20Language%20using%20Transformation%20Based%20Learning%20(TBL)%20Approach.pdf.
29) Keleta Y, Yamamoto K, Marasinghe A. Tigrinya part-of-speech tagging with morphological patterns and the New Nagaoka Tigrinya corpus. *International Journal of Computer Applications*. 2016;146(14):33–41. Available from: https://dx.doi.org/10.5120/ijca2016910943.
30) Sium MA. Automatic Part of Speech Tagger for Tigrigna Language Using Hybrid Approach. *Journal of Chemical Information and Modeling*. 2016;53(9):130.
31) Emiru G. Development of part of speech tagger using hybrid approach. 2016;12(1):1–7. Available from: http://etd.aau.edu.et/handle/123456789/14045.
32) Tedla Y, Yamamoto K. Analyzing word embeddings and improving POS tagger of Tigrinya. In: and others, editor. Proc. 2017 International Conference on Asian Language Processing;vol. 2017. 2017;p. 115–118. Available from: https://doi.org/10.1109/IALP.2017.8300559.
33) Abera MK. Development of Part of Speech Tagger for Ge ' ez Language . 2017. Available from: http://etd.aau.edu.et/handle/123456789/18347.
34) Desta K, Mulugeta W. Part of Speech Tagger for Hadiyyisa Language. 2018. Available from: https://www.semanticscholar.org/paper/Part-of-Speech-Tagger-for-Hadiyyisa-Language-Desta/b14d8e291812c84dc8d1150feda9d5ea1f8e484c.
35) Deriba FG. developing parts of speech. . Available from: https://www.dictionary.com/browse/developing.
36) Mohammed S. Using machine learning to build POS tagger for under-resourced language: the case of Somali. *International Journal of Information Technology*. 2020;12(3):717–729. Available from: https://doi.org/10.1007/s41870-020-00480-2.