# Designing a Rule Based Stemmer for Afaan Oromo Text

**Debela Tesfaye** dabookoo@yahoo.com
*Faculty of Informatics/Department of*
*Information Science Addis Ababa*
*University Jimma, 378, Ethiopia*

**Ermias Abebe** ermiasabe@gmail.com
*Faculty of Informatics/Department of*
*Information Science Addis Ababa*
*University Addis Abeba, Ethiopia*

## Abstract

Most natural language processing systems use stemmer as a separate module in their architecture. Specially, it is very significant for developing, machine translator, speech recognizer and search engines. In this work, a stemming system for Afan Oromo is presented. This system takes as input a word and removes its affixes according to a rule based algorithm. The result of the study is a prototype context sensitive iterative stemmer. Error counting technique was employed to evaluate the performance of this stemmer. The errors were analyzed and classified into two different categories: under stemming and over stemming errors. For testing purpose corpus which is collected from different public Afaan Oromo newspapers and bulletins is used. Newspapers, bulletins and public magazines are considered as consisting different issues of the community: social, economical, technological and political issues. This will reduce the probability of making the corpus biased toward some specific words that do not appear in everyday life. According to the evaluation of the experiments, it can be concluded that an overall accuracy of the stemmer is encouraging which shows stemming can be performed with low error rates in high inflected languages such as Afan Oromo.

**Keywords:** Afan Oromo stemmer, Rule based Stemmer, Context sensitive stemmer

## 1. INTRODUCTION

We can find a variety of internet search engines with advanced search parameters for retrieving documents in a document collection. During the development of search engines we can notice an ongoing specialization on the searching features. These engines are becoming more and more sophisticated trying to cover user's demands to access specific information.

One of the attempts to make the search engines more effective in information retrieval was the usage of word stemming. For grammatical reasons, documents are going to use different forms of a word, such as organize, organizes, and organizing. Additionally, there are families of derivationally related words with similar meanings, such as democracy, democratic, and democratization. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set. Stemming makes families of derivationally related words with similar meanings represented using single term. A stemming algorithm is a procedure that reduces all words with the same stem to a common form by stripping of its derivational and inflectional suffixes [1]. Using Stemming, many contemporary search engines associate words with prefixes and suffixes to their word stem, to make the search broader in the meaning that it can ensure that the greatest number of relevant matches is included in search results.

Afaan Oromo, one of the major languages that is widely spoken and used in Ethiopia, is morphologically very productive; derivation, reduplication and compounding are also common [2]. Obviously, these extensive inflectional and derivational features of the language are presenting various challenges for text processing and information retrieval tasks in Afaan Oromo.

This paper describes the development and evaluation of a context sensitive rule based stemmer for Afan Oromo. Most of the concepts are adopted from stemming algorithm developed by Porter [3]. We have chosen to modify the stemming algorithm developed by Porter [3] because it is well known and is frequently used in experimental IR systems.

## 2. STEMMING ALGORITHMS

A basic characteristic of stemming algorithm is whether it is context-free or context sensitive, which refers to any attribute of the remaining stem.

### 2.1 Context-free

In context-free algorithm, no restriction is placed on the removal of a suffix and thus any ending, which matches, is accepted for stripping.

### 2.1. Context sensitive

In context sensitive algorithms, however, various restrictions are placed on the usage of the suffix. Therefore, such kind of algorithm requires the construction of suffix dictionary and the formation of a set of rules defining the morphological context of the suffixes. The dictionary gives the exact suffix form, while the rules define conditions to be tested in order to apply the rules.

One of the widely used context sensitive rule based stemmer is that of Porter [3]. The Porter stemmer has five steps and applying rules within each step. Within each step, if a suffix rule matched to a word, then the conditions attached to that rule are tested on what would be the resulting stem, if that suffix was removed, in the way defined by the rule. For example such a condition may be, the number of vowel characters, which are followed be a consonant character in the stem (Measure), must be greater than one for the rule to be applied [3]. Within each phase there are various conventions to select rules, such as selecting the rule from each rule group that applies to the longest suffix.

In context sensitive stemmers, If the set of rules defining the correct morphological context for the suffix is satisfied, it is replaced by another string, either the null string (if the suffix is to be removed) or specified replacement string (for example, to create nominal forms to adjectival forms). Both dictionary and rules require careful analysis of vocabulary and language behavior, and are thus time-consuming to create. However, generally such techniques are rewarded by high accuracy and speed, and simplicity in implementation [4].

## 3. OVER VIEW OF AFAN OROMO

Afaan Oromo is one of the major African languages that is widely spoken and used in most parts of Ethiopia and some parts of other neighbor countries like Kenya and Somalia. Currently, it is an official language of Oromia state (which is the largest Regional State among the current Federal States in Ethiopia). It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 34.5% of the total population [5]. With regard to the writing system, Qubee (a Latin-based alphabet) has been adopted and become the official script of Afaan Oromo since 1991.

Like a number of other African and Ethiopian languages, Afaan Oromo has a very rich morphology (Oromoo, 1995). It has the basic features of agglutinative languages where all bound forms (morphemes) are affixes. In agglutinative languages like Afaan Oromo most of the grammatical information is conveyed through affixes (prefixes, infixes and suffixes) attached to the roots or stems. Both Afaan Oromo nouns and adjectives are highly inflected for number and gender. In contrast to the English plural marker s (-es), there are more than 12 major and very common plural markers in Afaan Oromo nouns (example: -oota, -ooli, -wwan, -lee, -an, -een, -oo, etc.) [2]. Afaan Oromo verbs are also highly inflected for gender, person, number and tenses. Moreover, possessions, cases and article markers are often indicated through affixes in Afaan Oromo. Since Afaan Oromo is morphologically very productive, derivations and word formations in the language involve a number of different linguistic features including affixation, reduplication and compounding [2].

In this research, the morphological analysis of the language is organized in to 6 categories. The categories are: nouns, pronouns and determinants, case and relational concepts, functional words, verb and adverbs. Almost all Oromo nouns in a given text have person, number, gender and possession markers which are concatenated and affixed to a stem or singular noun form. Like wise, determinants have number, gender, adjectives, and quantifier markers similar to Afan Oromo nouns. Afaan Oromo verbs are also highly inflected for gender, person, number, tenses, voice and transitivity. Furthermore, prepositions, postpositions and article markers are often indicated through affixes in Afaan Oromo.

## 4. RELATED WORKS

Very limited works have been done in the past in the areas of stemming in relation to Afaan Oromo. One of the stemmer is developed by Kekeba, Varma and Pingali[7] and it used to develop an Oromo-English CLIR system that enable user to access and retrieve online information sources that are available in English by using Afan Oromo queries.

The stemmer used a rule based suffix-stripping algorithms focusing on very common inflectional suffixes of Oromo language. This light stemmer is designed to automatically remove frequent inflectional suffixes attached to headwords (base-word forms) of Afaan Oromo. Some of the common suffixes that have been considered in their light stemmer include gender (masculine, feminine), number (singular or plural), case (nominative, dative), possession morphemes and other related morphological features in Afaan Oromo.

This kind of stemming techniques can have several shortcomings when applied to heavily inflection language such as Afaan Oromo. These shortcomings can include:

- Improper removal of some affixes (part of a word might appear to be a prefix or suffix). Their stemmer simply strips of any end of a word that matches one of the affixes in a list without any condition to be tested. Afaan Oromoo suffixes are not quite different from non suffix endings. Suffixes like -aa as in the case of maqaa,mucaa;-ee,-lee as in the case of killee, eelee, itilee, mee, kee, ree; -an,-n in the case of aannan, ilkaan, Afaan, shan, kan; -tuu, -tu as in the case of utuu, hatuu, nyaatu, baatu, kaatu are part of a root word as indicated in the above words that should not be conflated. There fore most word endings can be part of a word and suffix as well. So simple removal of endings can create invalid stems. But the above stemmer conflates the words since the endings of the words matches one of the suffixes.
- The stemmer didn`t considered words formed by duplication of some characters at all. But Afaan Oromo is rich in this kind of word formation. Most of the adjectives form the plural by reduplication

of the first syllable. For example words like, jajjabaa (stron-plural), gaggabaabaa (short-plural) are formed from -jabaa and -gabaabaa by duplicating the first syllabus, respectively.

Another stemmer is developed by Wakshum[10] which use suffix table in combination with rules that strips off suffix from a given word by looking up the longest match suffix in the suffix list[10]. List of suffixes are compiled automatically by counting the most frequent endings. Other linguistically valid suffixes are also included manually. The stemmer fined the longest suffixes that matchs the end of a given word and remove. The problems he faced are similar with that of Kekeba, Varma and Pingali[7]. Some of these are: irregular formation of variants from root word, the challenge to increase the number and complexity of the rules and words formed by duplication of some characters. The lists of the suffixes are not linguistically valid. Out of 342 suffixes compiled by the stemmer only 70 are linguistically valid. This is because the lists of suffixes are compiled statistically and requires no analysis of the language. The suffixes are compiled by counting and sorting the most frequent endings. One great problem occurred with this kind of compilation of suffixes is that during conflation frequently occurring endings which are part of root word is considered as suffixes and removed.

Another problem is that, the compilation of stop words is also done statistically and frequently occurring content bearing words are also included. For example,barannoo,barattoo,barnoota are varieties of the root barat(to learn) ,duree(rich),fayyadam(to use),dhiyeess(to approach),barsiisu(to teach), barsiisa (teacher), agarsiis (to show) and the like are include as stop word. And this indicates that frequently occurring content bearing words of Afan Oromo are not considered by the stemmer. More than 96 content bearing words that occur frequently are included as stop word.

This necessitates the need for the detailed knowledge of the language to use such frameworks to generate a stemmer that handles words formed by duplication of some characters and other under/over stemming problems.

## 5. AFAN OROMO STEMMER

### 5.1. Introduction

It is not possible to apply the stemming algorithms developed for English or other languages like porter's [3] to Afan Oromo due to differences in the patterns of word formations and differences in their morphologies. Some of the concepts from Porter stemmers are however adopted to develop a stemmer for Afan Oromo. Specifically, Concepts about measure, arranging the rules in clusters ,analyzing word formation based on the nature of their endings(for example words that attaches the suffixes *–de* must end with b/g/d in Afan Oromo) are taken from porter algorithm.

Afan Oromo stemmer is based on a series of steps that each removes a certain type of affix byway of substitution rules. These rules only apply when certain conditions hold, e.g. the resulting stem must have a certain minimal length. Most rules have a condition based on the so-called measure. The measure is the number of vowel-consonant sequences (where consecutive vowels or consonants are counted as one) which are present in the resulting stem. This condition must prevent that letters which look like a suffix but are just part of the stem will be removed. Other simple conditions on the stem are:

- Does the stem end with a vowel?
- Does the stem end with a consonant?
- Does the stem end with specific character?
- Does the 1st syllabus of the stem duplicated?

### 5.2. Extensions to Porter Stemmer`s Implementation

Most Afan Oromo words form repetition by duplicating some of the starting characters. Because this affix exhibits certain pattern that can be recognized, the algorithm has been extended to handle them. The original Porter stemmer [3] only treats suffixes.

### 5.3. Definitions of Afan Oromo Stemmer

Define Afan Oromo vowel as one of
a e i o u
Define Afan Oromo consonants as one of
b c d f g h j k l m n p
q r s t v w x y z `
Define a valid de, du, di, do, dan-ending as one of
b g d
R1 is the region after the first non-vowel following a vowel. If the word starts in vowel it is the region before the next consonant.
R2 is the region after the first non-vowel following a vowel in R1
C1 is the firs character of a word.

### 5.4. Compilation of Stop Word List

As can be seen from the table, the stop word list consists of prepositions, conjunctions, articles, and particles. The stop word list is collected and compiled based on information in A Grammatical sketch of Written Oromo [6] and *Caasluga Afaan Oromoo, Jildi I* [2]. The list of linguistically valid Afan Oromo prepositions, conjunctions, articles, particles are available on the above mentioned books.

| Number | word | English |
|---|---|---|
| 1 | kana | This |
| 2 | sun | That |
| 3 | ani | Me? |
| 4 | inni | He |
| 5 | isaan | They |
| 6 | ise | She |
| 7 | akka | Like |
| 8 | Ana | Me |
| 9 | fi | and |

**TABLE 1:** Afan Oromo stop word list examples

There are no any content-bearing words in the stop word list.

### 5.5. The Test Set

A corpus is a collection of texts or speech stored in an electronic machine-readable format [11]. Balanced corpus is needed to process natural language processing tasks like stemming. Balanced corpus is a corpus that represents the words that are used in a language. As indicated in [11] texts collected from a unique source, say from scientific magazines, will probably be biased toward some specific words that do not appear in everyday life. Such types of corpuses are not balanced corpus so that they are not appropriate for many natural languages processing in general and stemming in particular except for special purposes. However, developing a balanced corpus is one of the difficult tasks in NLP research because it requires collecting data from a wide range of sources: fiction, newspapers, technical, and popular literatures. As a result it requires much time and human effort.

For this particular study, corpus was collected from different popular Afaan Oromo newspapers (Bariisaa, Bakkalcha Oromiyaa and Oromiyaa) and bulletins (Qabee and Oromiyaa) to balance the corpus. Newspapers, bulletins and public magazines are considered as consisting different issues of the community: social, economical, technological and political issues. So that they are a potential source for collecting balanced corpus for natural language processing tasks.This corpus is used for evaluating the performance of the stemmer. The corpus consists of 159 sentences (the total of 1621 tokens).

### 5.6. Afan Oromo Stemmer Rule Clusters

Based on the information written on A Grammatical sketch of Written Oromo [6] and Caasluga Afaan Oromoo, Jildi I [2] six rule clusters were created for Afan Oromo stemmer.

The rule-clusters are defined by similarity of their pattern in word formation, the level at which the affixes occur in the word formation process (the most common order/sequence of Afaan Oromo suffixes (within a given word) is: <stem> <derivational suffixes> <inflectional suffixes> <attached suffixes> [7]) and the length of the affixes.

Thus, this stemmer removes (from the right end of a given word) first all the possible attached suffixes, then inflectional suffixes and finally derivational suffixes step by step. This is done to reduce computational time. Morphemes that are formed out of this sequence can also be removed though it takes additional computational time. Complex affixes are thus removed in consecutive steps. For example, *baratootarratti* (on the students) has four suffixes*: -itti, -rra, -oota* and *-at*. Therefore firs *-tti*, then *-rra*, then *-oota* and finally *-at* is removed to get the root *“bar-“.*

In addition to the affix-rules, context sensitive conditions are also designed to cover some specific phenomena. Examples of these conditions are, *Endswith V/C*, i.e. when remaining stem ends in a vowel or consonant; *Ends with B/G/D*, i.e. when remaining stem ends in the characters *B* or *G* or *D*; *Endswith CC*, i.e. when remaining stem ends in double consonant.

The affix-rules have the following general form:

***Affix*** → ***substitution measure-condition <additional conditions>***

***Where:***

***Affix*** is a valid Afan Oromo prefix or suffix

*In Afan Oromo repetition (plural) is formed by duplicating the first syllabus and it is also considered as prefix.*

***Substitution*** is a string which is substituted with a given affix to produce valid stem.

***Measure-condition*** is the number of vowel-consonant sequences (where consecutive vowels or consonants are counted as one) which are present in the resulting stem.

***Additional conditions-*** additional conditions are also designed to cover some specific phenomena. Examples of these conditions are, *Endswith V/C,*

The sixth cluster contains derivational suffixes whose measure is greater or equal to 1 and ends with consonant.
e.g.

| | | | | |
|---|---|---|---|---|
| “eenya” | | | | |
| “offaa” | → | 0 | measure > =1 | EndsWith c |
| “annoo”, | | | | |
| “ummaa” | | | | |

Explanation for the sixth rule cluster:
If a word ends with one of the suffixes: “eenya”, “offaa” , “annoo”, “ ummaa” and measure> =1 and the remaining stem ends with consonant, delete the suffix.
E.g. qabeenya(asset) ends with the suffix “eenya”,measure>=1 and the remaining stem *qab-* ends with the consonant *b.* Therefore it is correctly stemmed to *qab-*.

The fifth cluster contains suffixes that are removed if measure is greater or equal to 1 or substituted with the suffix -` if measure equal zero.
e.g.

| | | | |
|---|---|---|---|
| “`aa” | | | |
| “’e”, | | | |
| “’u” | → | 0 | measure > =1 |
| “’ee”, | | | |
| “suu” | | | |
| “sa”, | | | |
| “sse” | | | |
| “nya” | → | ` | measure > =0 |

Explanation for the fifth rule cluster:
If a word ends with one of the suffixes: “`aa”, “’e”, “’u”, “’ee”, “suu”, “sa”, “sse” ,“nya” and measure> =1, delete the suffix. If measure =0, substitute the suffix with `( *glottal*).

The fourth cluster covers a special case
e.g.

| | | | | |
|---|---|---|---|---|
| “du” | | | | |
| “di” | → | 0 | measure > =1 | EndsWith B/G/D |
| “dan” | → | d | measure =0 | |
| | | | | |
| “wwan” | → | 0 | measure > =1 | EndsWith VV |

Explanation for the fourth rule cluster:

If a word ends with one of the suffixes: “du”, “di”, “dan” and measure> =1 and the remaining stem ends with B/G/D, delete the suffix. If measure =0, substitute the suffix with d.
If a word ends with the suffix “wwan” and measure> =1 and the remaining stem ends with double vowel delete the suffix.

The seventh cluster contain rules that conflate words formed by duplication of the first syllabus

| | | | | | |
|---|---|---|---|---|---|
| R1 | → | 0 | measure > =0 | Startswith C | R1=R2 |
| | → | 0-1 | measure > =0 | Startswith C | C1+R1 =R2 |
| C1+` | → | 0 | measure > =0 | Startswith V | `+R1=R2 |

Explanation for the seventh rule cluster:
If R1 and R2 are the same delete R1.
If C1+R1 and R2 are the same delete R2.
If the word starts with vowel and if glota (`)+R1 equals R2 delete R2.
E.g. gaggabaaba(plural form of short),R1 is ga , R2 is gga and C1 is g. Therefore gaggabaaba is correctly stemmed to gabaaba by removing R2, since C1+R1 equals R2.

To stem a word the, the stemmer first checks if a given word is in the stop list or not. If found in the list, the word is excluded from further processing and nothing returned to the calling routine; stop and process the next word if any. If the word is not in the stop list, the word is checked for any match in the rule clusters. If a match is found, the respective action for that rule will be taken. As described earlier the rule has conditions like measure (the number of vowel consonant sequence), ending of the remaining stem with specific character, ending of the remaining stem with consonant, ending of the remaining stem with short or long vowel, matching of the ending of the word with one of the suffixes and the detail for each rule cluster is described in the privies section. The actions taken includes removing the suffixes, substituting the suffix with another one, removing of the reduplicated characters in the case of words formed by reduplication of some of the characters as described in the 7th rule cluster.

### 5.7. Evaluation of the Stemmer

In this report, error counting approach is adapted to evaluate the algorithm in terms of the number of accurately conflated results. The number of correctly conflated words and incorrectly conflated ones are counted for analysis. The output from the stemmer was then checked against the respective expected valid stem. These errors were then described in terms of under stemming and over stemming:

- Over stemming
  Over stemming occurs when too much of the term is removed.
- Under stemming
  Under stemming occurs when too little of the term removed.

Evaluation of the effectiveness of stemming algorithms is necessary to reveal specific error patterns. This information can subsequently be used to improve the algorithm where possible. Some error types, however, are inherent to the suffix-stripping method and without the additional information provided by, for instance, a dictionary, these errors cannot be avoided [8].

This stemmer is run on the test set of 5000 words which is assumed to be balanced. The literature from which the rule of the stemmer developed is totally different from the test set. This was done deliberately in order to predict the performance of the stemmer in the real world data.

The output from the stemmer indicates, Out of 5000 words 38 words (0.77%) were under stemmed and 220 words (4.39 %) were over stemmed. Totally this stemmer generates 258 words (5.16 %) stemming error. As a result, the accuracy of the stemmer becomes 94.84%.

In terms of compression, i.e., reduction of dictionary size, percentage of compression is calculated using the formula [9]:

C = 100 * (W - S)/W

Where,

C is the compression value (in percentage)<br>
W is the number of the total words<br>
S is a distinct stem after conflation

Accordingly,

Size of the data = 5000<br>
Number of distinct stem after conflation = 1654

Hence, the percentage of compression for Afan Oromo text based on the evaluation text for this stemmer becomes 100 * (5000- 3100) / 5000 = 38%.

Reasons for the under stemming and over stemming problems are:

1. It was difficult to come up with the complete rule because of the complexity of the language. More conditions/rules are required based on more study of the morphology of the language.

2. Homographs are words which are spelled identically but nevertheless have a different meaning. The algorithm does not have access to information about, for instance, word categories; the different senses of these types of words are not distinguished.

3. The rule designed for the suffix *-s* is not general and conflates some terms incorrectly. Designing general rule for Words ending in *-s* is challenging.

4. Some Compound words are not conflated correctly. This stemmer didn`t include any rule that handles compound words. Even though there is no rule included to conflate compound words, rules that are designed for non compound words can be applied and produce correct result for most compound words. Examples of compound words that are conflated correctly are: *karadeemaa(*passenger) is correctly conflated to *“karadeem-*”, *biyyalafaa*(world) is correctly conflated to “*biyyalaf-*“. But *manabaate*(married(f)) is incorrectly conflated to “*manab-*“.

5. *“ni”* and *“hin”* are considered as prefixes in some literatures and independent terms in other literatures. There fore, this stemmer didn`t include the rule that remove them when they appear as prefix. Both cases are possible in the language.

## 6. CONSLUSION & FUTURE WORK

Stemming is important for highly inflected languages such as Afan Oromo for many applications that require the stem of a word. In this work, a context sensitive rule based stemmer was developed that attempts to determine the stem of a word according to linguistic rules. According to the evaluation of the experiments, it can be concluded that an overall accuracy of about 94.84% is an encouraging figure. The proposed method generates some errors. Indeed, it is possible to anticipate such considerable contributions and positive effects of the stemmer since Afaan Oromo is one of the morphologically rich and complex languages. These errors were analyzed and classified into two different categories (under stemmed words and over stems). The error rate is about 4.27%. This shows that rule based stemmer can be performed with low error rates in high inflected languages such as Afan Oromo.

A big step in the future improvement of the Afan Oromo stemmer can be a study on how the word compounding and suffixes affect Afan Oromo words and their stems, and how one can include new rules that do not affect the effectiveness of the stemming process.

Besides, further study is required to increase the effectiveness of the rule based stemmer with no or little decrease in efficiency. Extracting other additional context sensitive conditions which is based on the study of the morphology of the language can increase the accuracy of the stemmer.

Moreover, the stemmer has to be tested with large amount of texts to prove its real performance. To succeed this we need to apply Afan Oromo stemmer in a web search engine, which retrieves information from Afan Oromo texts. Then we can have a complete view of the stemming system and the returned results after every search request. In this case we can do extended evaluation tests, we can measure the precision and recall in various texts and we can estimate the errors distribution in the stemming results.

All the rules described in this work can be a base for this further research and it can support extended stemming rules covering most of the terms in the Afan Oromo.

## 7. REFERENCES

[1] L. JB. "*Development of a stemming algorithm"*. Mechanical Translation and Computational Linguistics, 11: 22-31,(1968)

[2] G. Q. A. Oromoo. "*Caasluga Afaan Oromoo* Jildi I", Komishinii Aadaaf Turizmii Oromiyaa, Finfinnee, Ethiopia, pp. 105-220 (1995)

[3] M. F Porter. "*An algorithm for suffix stripping*". Program, 14(3):130–137,(1980)

[4] S. Jacques. "*Stemming of French Words Based on Grammatical Categories*." Journal of American Society for Information Science 44(1): 1-9, (1993)

[5] Census Report. "*Ethiopia's population now 76 million*". (2008) available at: http://ethiopolitics.com/news

[6] C. G. Mewis." *A Grammatical sketch of Written Oromo"*, Germany: Koln,pp. 25-99 (2001)

[7] K. K. Tune, V. Varma and P. Pingali. "*Evaluation of Oromo-English Cross-Language Information Retrieval*", Language Technologies Research Centre IIIT, Hyderabad India, (2007)

[8] W. Kraaij, R. Pohlmann."*Porter's stemming algorithm for Dutch*". Bioinformatics, 25: 1412–1418, (1997)

[9] L. Lessa. "*development of stemming algorithm for wolaytta text*", Masters Thesis Addis Ababa University, Fuculty of Informatics, Department of Information Science, (2003)

[10] W. Mekonen. "*Development of stemming algorithm for Affan Oromo anguage text*", MSc thesis faculty of informatics, Addis Ababa University, Addis Ababa,(2000)

[11] S. Dandapat, S. Sarkar and A. Basu. "*A Hybrid Model for Part-f-Speech Tagging and its Application to Bengali"*, Journal of world information society, 43(6):384–390, (2004)