DESIGNING A STEMMER FOR GE'EZ TEXT USING RULE BASED APPROACH

ABEBE BELAY ADEGE, YIBELTAL CHANIE MANIE

Abstract— In this study, a stemmer of Ge'ez text was developed using rule based approaches. In designing processes, different concepts such as background for the thesis, literatures on conflation of the stemming algorithms, morphological nature of Ge'ez language, stemming techniques and other realted things were discussed in order to model and develop an automatic procedure for conflation or prototype. When inflectional and derivational morphologies of the language were discussed, affixations such as prefixing, infixing and suffixing are the main word formation processes in Ge'ez language. The language is morphologically complex. This is because different words can be formed due to the wide concatenations of affixes.For the experiment, two techniques were used: affix removal and morphological analysis techniques. To evaluate the stemmer, manually error counting technique was used. The stemmer was trained on 70% of the sample text and the test result was done on 30% of the sample data that were not included for training. From the experiment, three types of errors are observed: over stemmed (6%), under stemmed (4.27%) and structural problems (7.31%). When the stemmer runs on the unseen sample of 30% sample texts, it performed with an accuracy of 82.42%. The dictionary reductions of the stemmer were 29.9% to the stemmed words and 62.8% to root words on the test set. Lastly, the possible recommendations to future works and improvements of this work were reported.

Index Terms— affix removal, infix, information retrieval (IR), machine translation (MT), Morphology, n-gram stemmer, natural language processing (NLP), suffixes, and prefixes.

———————————— ◆ ————————————

1 INTRODUCTION

Morphology describes how various forms of words are created, and studies structures of words in the language. These are as a result of syntax, such as changes in person, number, tense and gender [7]. But, there are exceptionalism called derivational types may affect a word's meaning in part of speech. For example, affix changes from adjective to nouns, from verb to nouns, from noun to verb, and so on; like friend, friendly, friendliness and friendship.

From either type of morphologies, depending up on the complexity of the morphological natures of language types, very massive variants of words may be resulted from single word. Thus, there is a need of automated procedure that can reduce the size of the various words to manageable level, and also capture the strong correlations existing between different word forms [5].

There are four major types of automatic stemming strategies [18]: affix removal, table look up, successor variety, and n-gram. N-gram stemming is used based on the identification of di-grams and tri-grams and is more of a term used for clustering than stemming.

Unlike languages like English that has less morphological variants, there are languages like Amharic, Arabic and Ge'ez that are much rich in morphology. So, like Amharic [1] and Arabic [11], Ge'ez involves dealing with prefixes, infixes and derivatives in addition to suffixes, and it is the focus of this study.

Geez, the classical language of Ethiopia [13], is still used as a liturgical language by the Ethiopian Orthodox Tewahido Church, Ethiopian Catholic Church, the Beta Israel Jewish community of Ethiopia, and Eritrea. As Geez is morphologically complex language, there is a need for automated procedures that can reduce the size of lexicon to manageable level and improve the application of information retrieval and natural language processing.

Ther goal of this study is to review properties of the Ge'ez language in order to get familiar with the different aspects of the language and know how the separation of words is fashioned [14], affixing (adding suffixes, prefixes, and infixes), compounding, duplicating (reduplicating) and different vowel patterns are used to create various word forms in Ge'ez language.

- *Abebe Belay Adege is currently pursuing PhD degree program in International Graduate Program in Electrical Engineering and Computer Science in National Taipei University of Technology, Taiwan. E-mail: abbblybelay@gmail.com*
- *Yibeltal Chanie Manie is currently pursuing PhD degree program in International Graduate Program in Electrical Engineering and Computer Science in National Taipei University of Technology, Taiwan. E-mail: yibeshmamaru@gmail.com*

The pattern of vowels in a word can also create various word forms in Ge'ez text. For example, according to [12], verbs can be classified into three classes: Type A, Type B and Type C verb classes. They differ by vowel patterns for perfect and jussive tense descriptions. Class A is unmarked class. In the base past and jussive it has two sub classes: A1 has stem vowel /ä/ in the past and stem vowel /a/ in jussive, whereas A2 has stem vowel /a/ in the past and stem vowel /ä/ in the jussive.ieved. This is because verbs and/or nouns are very rich in morphological characters to agree with person, number, and gender of subjects in Ge'ez language, the discussions are mainly focused on noun and verb morphologies.

2 Methodology

2.1 Programming Technique and Data sources

A prototype stemmer for Ge'ez language was written using the Python programming language (Python2.6 and Python3.0). A text corpus is one of the resources required in IR research. These sources are historical books such as Abune Habte Marrian History, cultural as well as religious books like Bible in Ge'ez and Wdase Marriam.

2.2 Expermental Methods

The selected data was divided into training set and testing set. Training set was 70% of the sample data and the rest 30% was used for testing set as relatively similar with different researchers used in this subject area. The stemmer works by conflating variants words in Ge'ez language (inflectional and derivational morphologies).

3 Related Works

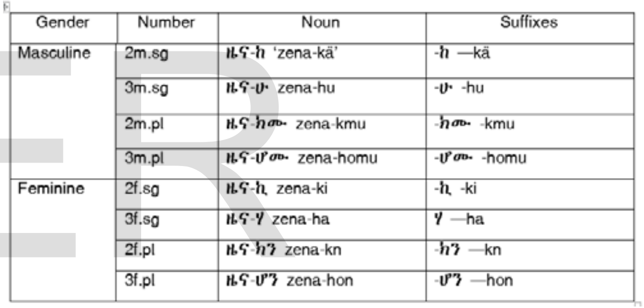
Morphological variants are the main problem by making a word to have various forms in the documents and queries, and these have problems in indexing and retrieving systems [3]. As discussed by [2], gender, number, tense, person, mood, or voice can characterize variants of words. For various languages, different stemmer researches are designed since years ago using different approaches like Amharic stemmer [16], Afaan-Oromo stemmer [17], Tigrigna language stemmer [9] and Wellayta language [14].

Table 3.5: Plural noun formations using affixes

Singular plural	Plural	Prefix	infix	suffix
ሰባኪ: säbaki	ሰባኪያን: säbakiyan			ያን -yan
ብሩህ: baruh	ብሩሃን: baruhana			ኣን-an
ፈረሳዊ: färäsawi	ፈሪሳዊያን:Färisawiyan		-እ- 'a-	ያን-yan
ጽጌ: śage	ጽገያት: śagäyat		-እ- -'ä-	ያት -yat
ብን : ban	□በን :bäna		-እ-, -'ä-	
መጸብሕ:mäśäbaḥa	መጸቢሕያን: äśäbiḥayan		-እ- 'ä-, እ-'a-	ያን-yan
ምሳሕ: masaḥa	ምሳሐት: masaḥät			ኣት-at
ቤት: bet	አብያት: 'bayata	አ 'ä-	ያ-ya-	

For instance, in Ge'ez verb: □□□ can be written in various morphological forms such as □□□□→□□□□, but □□□ does not follow similar pattern, rather it can be written as □□□□→□□□□, not change □ to □. This can be handled using exceptional rules. Hence, a rule based approach uses to design Ge'ez stemmer. When we come to Geez Morphology, Ge'ez language has a characteristic of conveying different messages with a single word alone [8]. A morpheme in Ge'ez can be free or bound, where a free morpheme can stand as a word on its own where as a bound morpheme cannot occur on its own as a word [4]. For example, □□□□□□□ 'killed each other'. This is because verbs and/or nouns are very rich in morphological characters to agree with person, number, and gender of subjects in Ge'ez language.

Table 3.2: Gender marker for possessive pronouns



Gender	Number	Noun	Suffixes
Masculine	2m.sg	ዜና-ከ 'zena-kä'	-ከ —kä
	3m.sg	ዜና-ሁ zena-hu	-ሁ -hu
	2m.pl	ዜና-ክሙ zena-kmu	-ክሙ -kmu
	3m.pl	ዜና-ሆሙ zena-homu	-ሆሙ -homu
Feminine	2f.sg	ዜና-ኪ zena-ki	-ኪ -ki
	3f.sg	ዜና-ሃ zena-ha	ሃ —ha
	2f.pl	ዜና-ክን zena-kn	-ክን —kn
	3f.pl	ዜና-ሆን zena-hon	-ሆን —hon

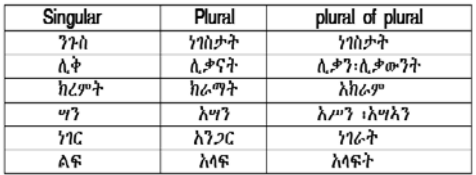
The suffix –□, -□, -□□, -□□, -□, -□, -□□ and -□□ are possibly employed Gender markers. Suffix, prefix, infix or their combinations form the plural nouns as shown above.

Nouns that are created by external plural markers are carried out by adding suffixes and/or prefixes at a stem. The following affixes are some of plural number indicates like prefix □- 'ä-, suffixes like -□□ –yan, -□□ –'an, -□□-yat, -□□–'at, -□ -w, -□□ -wat , □ -mu and □ -t.

Different patterns of letters for nouns have theis patters to dublicate or prulalized for example äCCuC: nouns with this pattern precede a pattern as vowel ä + consonant + consonant + u. äCC(a)C-t: this pattern follow a sequence of Vowel □ + consonant + consonant + vowel □ + consonant followed by □. äCaCC/t/: it has a pattern of vowel □ + consonant + vowel □ + consonant + vowel □ +consonant /□/.

Verbal nouns can be derived from verbs which have not more than three radicals [10]. Even Ge'ez verbs have an ability to create new words that are not similar with the original verb form. There are also compound words which are created

Table 3.4: plural nouns formation



Singular	Plural	plural of plural
ንጉሥ	ነገሥታት	ነገሥታት
ሊቅ	ሊቃናት	ሊቃን፡ሊቃውንት
ክረምት	ክራማት	አክራም
ሣን	አሣን	አሥን ፡አሣእን
ነገር	አንጋር	ነገራት
ልፍ	አላፍ	አላፍት

by combining two different words.

The common negation prefix in Ge’ez verb is □/‘i-/. This is when it comes with perfective form of verbs.

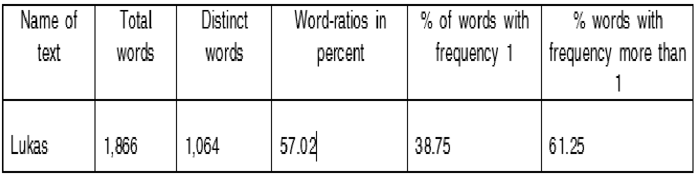
The writing system of different Ge’ez alphabets of similar sound are written differently in the early times.The different writing of alphabets with similar sounds would raise the question. Since verbs and nouns are very rich in morphological characters to agree with person, number, and gender of subjects in Ge’ez language, the discussions are mainly focused on noun and verb morphologies.

4. Results

For this work, sample texts were prepared from different sources: □□□ □□□□ ‘wdase maryam’ (prayer book), history of Abune Habtä Marriam and Bible in Ge’ez (□□□□ □□□□). To come up with the whole test set words, the first word was randomly selected and then every 3rd word is selected after the last selected word in the sorted word lists. This helps to avoid multiple selections of a word and protect multiple passing around sample texts i.e. it helps to get the whole test data in one pass through the sample data without coming back for another round.

As discussed by [9] and [14], word distribution in text documents of a language helps to study language’s behavior, and this distribution can be shown using word-ratio (numbers of distinct words to total numbers of words), and percent frequency ratios (e.g. total words which have frequency equals to 1 to total numbers of words). These help to show how much words are morphologically distributed within a document.

Table 4.1: Number of words and their distributions in Ge’ez documents

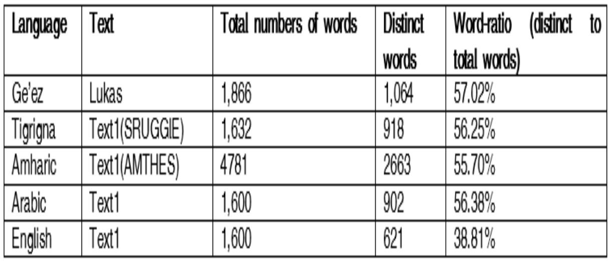


Name of text	Total words	Distinct words	Word-ratios in percent	% of words with frequency 1	% words with frequency more than 1
Lukas	1,866	1,064	57.02	38.75	61.25

As tried to figure out at table 4.1, one-third (1/3) of the total words in the sample texts composed of frequency equals to one. More than half of the sample texts were also distributed uniquely. This implies that there are existences of more variants of words in Ge’ez language

To compare Ge’ez word distributions with other the whole data sets were taken and shown in table 4.2.

Table 4.2: Comparison of word distribution ratios



Language	Text	Total numbers of words	Distinct words	Word-ratio (distinct to total words)
Ge’ez	Lukas	1,866	1,064	57.02%
Tigrigna	Text1(SRUGGIE)	1,632	918	56.25%
Amharic	Text1(AMTHES)	4781	2663	55.70%
Arabic	Text1	1,600	902	56.38%
English	Text1	1,600	621	38.81%

The ratios obtained from table 4.2 are almost similar but slightly greatest to Ge’ez from of all Amharic, Tigrigna and Arabic texts. However, it is absolutely different from English text.

Larger numbers of unique words are found in Ge’ez document when it was compared with other languages, especially with English. Hence, Ge’ez language has more distinct words and is morphologically very complex language. The stemmer removes the affixes by applying various rules to each affix and this was done using an application of context sensitive rules. These rules are designed based on morphological natures of a language to each sequence of activities. There are also affixes that are stripped using iterative approach with rules of the language. For example; lists of characters such as □ ‘wä’, □ ‘lä’, □ ‘zä’ and □ ‘kä’ can be appeared by concatenating each others as prefix of words.

The root of Ge’ez text can be obtained by stripping out all the vowels from the stemmed words [17] and [18]. But, there are some words whose vowels can be considered as consonants. For example, when □ ‘ä comes at the beginning of a word, if it is not removed as prefix, it is not considered as vowel and is not removed from a word. For instance from a word □□□□ ‘änst’ , ‘ä’ is considered as consonant. Basically three actions are taken in the stemming processes using AFFIX-REMOVAL TECHNIQUES only the two actions are applied to affix removal. These are:

Action1: do not remove any affix

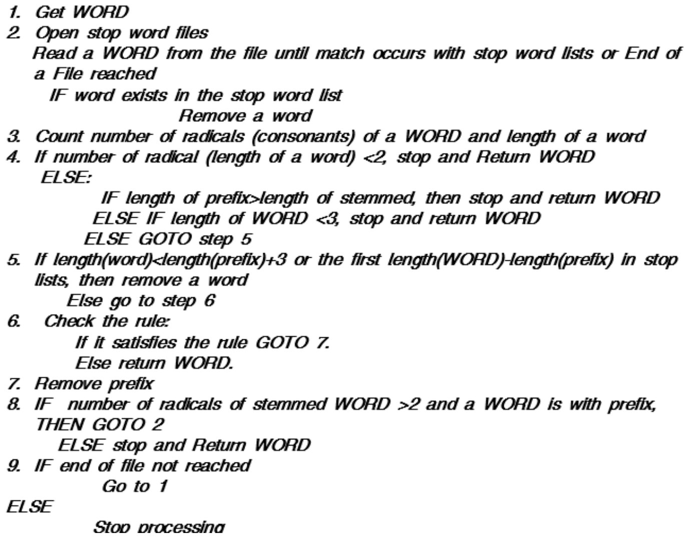
Action2: remove the concerned affix

To take any one of the above actions, there are conditions that are used to check the rules and apply an action 1 or action 2. These are:

Condition 1. After getting the assumed prefix or suffix, if number of radicals is not more than two or length of words less than three for a word without affix, take action 1.

Condition 2. If part of the assumed affix is obtained and number of radical without assumed affix is greater than two or length of a word greater than three and is not in stop wordlists, take action 2.

In the stemming process, based on the above conditions the appropriate action is taken. The basic rules that check whether the assumed prefix is true prefix or not are checking word length (represents the length of a word without a prefix), prefix structure (represents a prefix and its follower, end of alphabet that could be attached to a word and so on) and whether the word is part of stop word list (represent whether a stemmed word is part of stop lists or not). These rules are used to minimize the over stemming and under stemming problems[6]. To strip prefixes, the following algorithm is used.

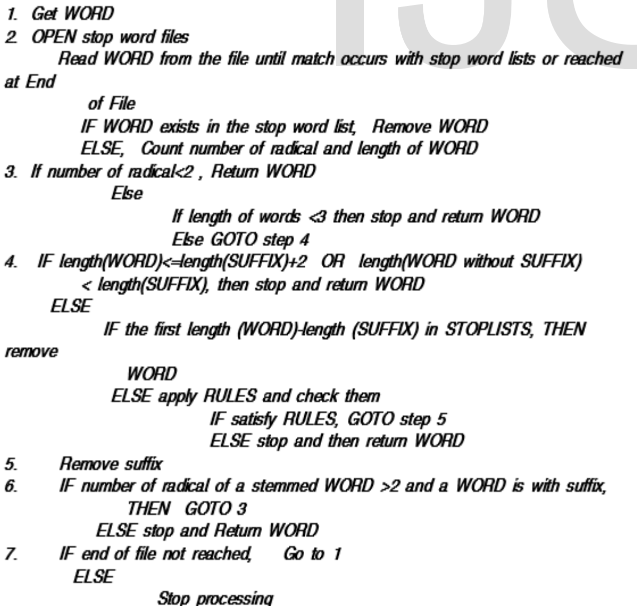


```
1. Get WORD
2. Open stop word files
   Read a WORD from the file until match occurs with stop word lists or End of
   a File reached
     IF word exists in the stop word list
                   Remove a word
3. Count number of radicals (consonants) of a WORD and length of a word
4. If number of radical (length of a word) <2, stop and Return WORD
    ELSE:
            IF length of prefix>length of stemmed, then stop and return WORD
           ELSE IF length of WORD <3, stop and return WORD
          ELSE GOTO step 5
5. If length(word)<length(prefix)+3 or the first length(WORD)-length(prefix) in stop
   lists, then remove a word
        Else go to step 6
6.  Check the rule:
         If it satisfies the rule GOTO 7.
         Else return WORD.
7. Remove prefix
8. IF  number of radicals of stemmed WORD >2 and a WORD is with prefix,
   THEN GOTO 2
      ELSE stop and Return WORD
9. IF end of file not reached
            Go to 1
ELSE
           Stop processing
```

Figure 4.1: Prefix striping algorithm

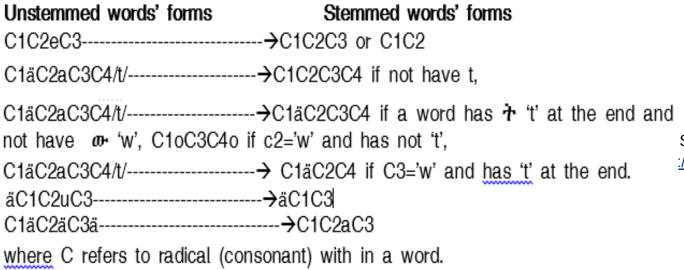
In the process of suffix striping, word length and number of radicals which represent the length of a word without suffix. Suffix checker (represents a condition that checks whether the assumed suffix is true suffix or part of a word). If the above conditions are fulfilled and a word without part suffix is not found in stop word lists, suffix striping is done based on the assigned rules. The following algorithm is used to strip suffixes.

Figure 4.2: Suffix striping algorithm



```
1.  Get WORD
2   OPEN stop word files
        Read WORD from the file until match occurs with stop word lists or reached
at End
          of File
         IF WORD exists in the stop word list,   Remove WORD
         ELSE,   Count number of radical and length of WORD
3.  If number of radical<2 , Return WORD
             Else
                     If length of words <3 then stop and return WORD
                     Else GOTO step 4
4.   IF length(WORD)<=length(SUFFIX)+2   OR   length(WORD without SUFFIX)
         < length(SUFFIX), then stop and return WORD
      ELSE
           IF the first length (WORD)-length (SUFFIX) in STOPLISTS, THEN
remove
             WORD
            ELSE apply RULES and check them
                        IF satisfy RULES, GOTO step 5
                        ELSE stop and then return WORD
5.      Remove suffix
6.      IF number of radical of a stemmed WORD >2 and a WORD is with suffix,
             THEN   GOTO 3
           ELSE stop and Return WORD
7.      IF end of file not reached,     Go to 1
          ELSE
                    Stop processing
```

Words which have prefix but not striped from the word are also held and stemmed with it. For example; □□□ →□□□□ 'däbr→ 'ädbar', □□□→□□□□ 'färäs→ 'äfras', and so on. In these types of nouns, the prefix □ 'ä' not removed using prefix striping technique because the condition is not fulfilled, but this technique can solve this problem. See the following examples how one structure of words is changed to other form:



Unstemmed words' forms **Stemmed words' forms**

C1C2eC3-------------------------------→C1C2C3 or C1C2

C1äC2aC3C4/t/-----------------------→C1C2C3C4 if not have t,

C1äC2aC3C4/t/-----------------------→C1äC2C3C4 if a word has ት 't' at the end and not have ው 'w', C1oC3C4o if c2='w' and has not 't',

C1äC2aC3C4/t/-----------------------→ C1äC2C4 if C3='w' and has 't' at the end.

äC1C2uC3-----------------------------→äC1C3

C1äC2äC3ä-------------------------------→C1C2aC3

where C refers to radical (consonant) with in a word.

Table 4.3 Sample structural analysis of Ge'ez words (verbs and nouns)

Using manual assessments of the stemmer, under stemmed holds 4.27 %, over stemmed covers 6 % and due to structural problems 7.31% from sample data sets are observed. Totally this version of the stemmer generates 17.58% stemmer errors. Consequently, the accuracy of the stemmer becomes 82.42 %.

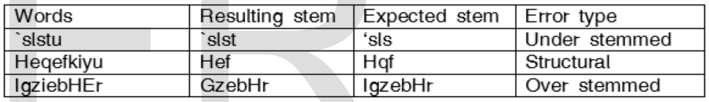
For calculating the compression rate(C), the expression used by [15] was used.

$$\boxed{C = \frac{100*(W-S)}{W}}$$

The dictionary size and the compression obtained for stem and root as Number of stems 29.90 % reduction and Number of roots 64% reduction.

Some stemming errors like over stemming and under stemming problems are observed from affix removal technique, and structural problems are mainly observed from morphological analysis technique. For these data sets of words, 29.90 % compression of stemmed words and 62.8% compression of root words were found.

Table 4.4: Some stemming errors



Words	Resulting stem	Expected stem	Error type
`slstu	`slst	'sls	Under stemmed
Heqefkiyu	Hef	Hqf	Structural
IgziebHEr	GzebHr	IgzebHr	Over stemmed

5. Conclusions and Recommendations

The analysis of word ratio of distinct words to total words calculated from sample text showed that Ge'ez is highly morphologically complex language than English. In the experiment, the accomplished result of this work is comparatively balanced when compared with other stemmers that are developed for other languages.

For future works, the study in this work is done on the limited size of sample texts and not tested in IR environment due to time constraints and limitation of freely available Ge'ez texts in soft copy. One can add more rules in order to increase the accuracy of this stemmer by designing taggers which able to differentiate part of speeches and stemmed each based on their assigned rules. Different approaches such as N-gram approach, Iterative approach, and/or their combinations can be applied to see whether better performance can be achieved or not. After improving the algorithm to its appropriate level, the stemmer can be an important tool for those researchers who are interested to study the Ge'ez language morphology. It is possible to use the stemmer by incorporating other components for developing other computational tools like morphological analyzer, parser, spell checker, thesaurus, word frequency counting, and summarizers.

6 Acknowledgments

First, we need to thank our mother St.Mary (the Blessed ever vergin Mary) with her son 'Jesus christ' who help us by making things easier and give persistency to all faces up in our life.

We have deepest gratitude to Dr.Million Meshesha who sows the seeds of IR into our mind and initiated the area of this paper. M/r Dessie Qeleb also deserves special thanks for his contribution and encouragement. Finally, our thank goes to our families, colleagues and work partners.

Addis Ababa, Ethiopia

7. REFERENCES

[1] Alemayehu,N and Willett,P. (2002), Stemming of Amharic words for information retrieval , Literary and Linguistic computing ,vol.17,No.1, pp.1-17

[2] Wakshum Mekonnen. (2000). Development of Stemming Algorithm for Afaan Oromo Text. M.Sc. Thesis . Addis Ababa Uniersty

[3] AnanthakrishnanRamanathan & DurgeshD.Rao, (no year available), ALightweightStemmer for Hindi, National Centre for Software Technology, NaviMumbai400614, India,pp.1-6

[4] Atelach Alemu Aregaw and Lars Asker, June 2007, Amharic Stemmer: reducing words to their citation forms, Association for computational linguistic proceeding of the 5th work shop on important unresolved matters, Stockholm university, Swiden, Czech Republic, pp.104-110

[5] Basem O. Alijla , April 6, 2009, Stemming in Natural Language, Faculty seminar, Department of Information Technology System, Faculty of Information Technology, The Islamic university of Gaza

[6] Birungi, P. (1995). Improved Strategies for Employment and Human Resources Utilization in the Information and Documentation Sector. Strategies for Human Resources Development for information Management in Africa, Ed. 49-57. Addis Ababa: UNECA, PADIS.

[7] Ethiopian Orthodox church teaching, The Ge'ez Language and Kine (Poetry, http://www.eotc-patriarch.org/index.htm

[8] Gabriella F. Scelta, (Dec 2001), The Comparative Origin and Usage of the Ge’ez writing system of Ethiopia, pp.1-9

[9] Girma Berhe, 2001, Stemming Algorithm Development for Tigrigna Language Text Document, Thesis, Addis Ababa University.

[10] Grishman,R. 1984, Natural language processing JASIS 35(5), pp.291-296

[11] Hayder K. Al Ameed, Shaikha O. Al Ketbi, Amna A. Al Kaabi, Khadija S. Al Shebli,Naila F. Al Shamsi, Noura H. Al Nuaimi, Shaikha S. Al Muhairi, ARABIC LIGHT STEMMER: ANEW ENHANCED APPROACH ,(no year available), The Second International Conference on Innovations in Information Technology (IIT’05), College of Information Technology, UAE University

[12] James Mayfield and Paul McNamee, July 28-August 1, 2003, Information Storage and Retrieval content analysis and indexing, Single N-gram Stemming.

[13] Jelita A.Hugh , E. Williams & S.M.M. Tahaghoghi, (2005), Stemming Indonesian language, 28th Australasian Computer Science Conference (ACSC2005),Conferences in Research and Practice in Information Technology, Vol. 38, pp.1-8

[14] LEMMA LESSA FEREDE, (2003), development of stemming algorithm to wolytta text, Addis Ababa University, department of information science.

[15] M.Taufik Abdullah, F.Ahmad, R.Mahmod, T.Mohd & T.Sembok. , (February 2009), Rules Frequency Order Stemmer for Malay Language , IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.2, pp.433-438

[16] Nega Alemayehu and Peter Willett, (2003), the effectiveness of stemming for information retrieval in Amharic words for information retrieval. Journal: electronic library and information systems, Vol.37, num.4, pp.254-259

[17] Haidar Harmaneni, Walid Keirouz, Saeed Raheel (2006), A rule based extensible stemmer for information retrieval with application to Arabic, The International Arabic Journal of Information Technology, Vol.3, No.3,pp. 265-272

[18] Ricardo Bauza-Yates & Bertnier Ribeiro-Neto, (1999), Modern information Retrieval, India

[19] Salton, Gerald (1983), ‘‘An Introduction to modern Information Retrieval’’, New York.