# TOWARDS IMPROVING BRILL'S TAGGER LEXICAL AND TRANSFORMATION RULE FOR AFAAN OROMO LANGUAGE

Abraham Gizaw Ayana

Department of Geographic Information Science
Hawassa Universty
Hawassa, SNNPR, Ethiopia

**Abstract: -** The aim of this thesis is to improve Brill's tagger lexical and transformation rule for Afaan Oromo POS tagging with sufficiently large training corpus. Accordingly, Afaan Oromo literatures on grammar and morphology are reviewed to understand nature of the language and also to identify possible tagsets. As a result, 26 broad tagsets were identified and 17,473 words from around 1100 sentences containing 6750 distinct words were tagged for training and testing purpose. Transformation-based Error driven learning are adapted for Afaan Oromo part of speech tagging. Different experiments are conducted for the rule based approach taking 20% of the whole data for testing. A comparison with the previously adapted Brill's Tagger made. The previously adapted Brill's Tagger shows an accuracy of 80.08% whereas the improved Brill's Tagger result shows an accuracy of 95.6% which has an improvement of 15%.

**Keywords: -** Afaan Oromo, POS tagger, NLP, Brill's Tagger, AI 52%.

## i. INTRODUCTION

Natural Language processing is one of the current hot research areas for scientists and academic researchers. The goal is to parse and understand natural language, which is not fully achieved yet. For this reason, much research in NLP has focused on preprocess and intermediate tasks that make sense of some of the structures inherent in language without requiring complete understanding. One such task is part-of-speech tagging, or simply tagging.

In sentences, all words can be labeled with their Part-of-Speech tag. These tags denote the grammatical function of the word in the sentence. Some simple, but well-known part of speech tags are for instance nouns, verbs, adjectives, adverbs and determiners. Part-of-Speech tagging makes sentences easier to parse by a computer, and is therefore a preprocessing step frequently used in text-processing systems [2]. Over the years there has been a lot of research to automate Part-of-Speech tagging, where a computer program tries to label each word with the correct Part-of-Speech tag.

Different methods have been used so far for POS tagging, such as Transformation-based learning, statistical learning using Hidden Markov models, statistical learning using Maximum Entropy models, Neural Networks, Support Vector Machines. But in this study we use the Brill's tagger in order to improve the lexical tagging and transformational rule of Afaan Oromo Language.

In 1992, Eric Brill introduced a POS tagger that was based on rules or transformations, where the grammar is induced directly from the training corpus without human intervention or expert knowledge. According to Brill [3], there is a very small amount of general linguistic knowledge built into the system with no language-specific knowledge. The only additional component necessary is a sufficiently large and manually annotated training corpus which serves as input to the tagger. The system is then able to derive lexical/morphological and contextual information from the training corpus and 'learns' how to deduce the most likely part-of-speech tag for a word. Once the training is completed, the tagger can be used to annotate new unannotated corpus [4].

Even though several works have been done in POS tagging for Afaan Oromo, the performance of the tagger has not sufficiently improved yet. The work in [9] is the first attempt to use a transformation based Error-Driven Learning (TEL) for Afaan Oromo POS tagger. The researcher recommended future work on improving the lexical and transformational rule to improve the performance of the POS tagger. Besides, the researcher found out that adding more training dataset can improve the performance of the tagger since the experiment was carried out on small scale dataset. Hence, the aim of this thesis is to improve Brill's tagger lexical and transformation rule for Afaan Oromo POS tagging with sufficiently large training corpus.

## II. OBJECTIVES

- ➢ To review related works and collect training dataset prepared for the same purpose.
- ➢ To see the possibility of adapting the Brill tagger Lexical rule
- ➢ To see the possibility of adapting the Brill tagger transformation rule
- ➢ To prepare more training dataset from untagged Afaan Oromo corpus
- ➢ To model TEL based POS for Afaan Oromo
- ➢ To develop prototype TEL based POS for Afaan Oromo language
- ➢ To test and analyze the performance of the model built
- ➢ To recommend future directions

## III. METHODOLOGY

The Afaan Oromo balanced text corpus is collected randomly from different sources in a form of both hardcopy and softcopy. Those sources are considered to be under different domain or categories such as Afaan Oromo books, journals, publications, news, newspapers and previous research corpus. Accordingly, TV Oromia, Voice of America (Afaan Oromo service), Afaan Oromo FM radio,websites like www.oromiyaa.com (website of oromia regional state), www.gadaa.com,www.qalbesa.wordpress.com,online journals and publications, books like Seenaa Oromo Jarraa 16ffaa, Yaadanii, Hawii,newspaper like Bariisa, Kallachaand previous Afaan Oromo POS tagging research corpus from the work of [8] and [9] are some of the main data source.

An incremental approach is used to prepare the tagged corpus. First we took the 258 previously tagged Afaan Oromo corpus for training the Brill tagger. Then this trained tagger takes untagged text as an input and tags the words based on the knowledge that it has acquired during the training and gives tagged text as an output. The output of the tagger is taken and given to the language professionals for correction and approval. After the corrected and approved tagged text is obtained the corpus is updated which is used in turn for training of the final POS tagger model. This process is repeated until adding the corpus can have insignificant effect on the performance of the tagger.

A Leaning curve is used to analyse the effect of the size of the training corpus on the performance of the tagger. First the tagger is trained on the 10% of the training corpus, which result in a small performance. Then we added another 10% of the training corpus and saw a little increment on the tagger performance. The process continues until the increasing the size of the training corpus does not show significant improvement on the tagger performance.

## IV. RELATED WORKS

The first work on Afaan Oromo language part of speech tagging, which uses statistical approach with Hidden Markov Model [8], was done in Addis Ababa University in 2009. A transformational error driven learning approach was used in the work of [9] in Addis Ababa University in 2010 for Afaan Oromo language. In this work, the researcher has adapted the Brill Transformational error driven learning with some modifications on the tagger template. The researcher has used 233 sentences (1708 distinct words) of Afaan Oromo language which he divided into training set and testing set. He used 18 tagsets to tag the 233 sentences. Accordingly, he has got 80.08% accuracy for the modified Brill tagger.

## V. AFAAN OROMO TAGS AND TAG SETS

In this section, the actual tags used in this thesis work are discussed. The identification of the tags is made by taking 11 word classes namely: noun, pronoun, verb, adjective, adverb, preposition, conjunction, numerals, punctuation, introjections, and negation as basic tags and others are derived from combination of or these basic classes. List of all Afaan Oromo tags is shown in table 3.4 below

***Table 3.8: Afaan Oromo Tags set***

| ***S/N*** | ***Basic*** | ***Derived*** | ***Description*** |
|---|---|---|---|
| 1. | Noun | NN | Noun |
| 2. | | NPROP | Proper noun |
| 3. | | NC | Noun + conjunction |
| 4. | | NP | Noun + Preposition |
| 5. | Pronoun | PP | pronoun |
| 6. | | PS | Preposition + pronoun |
| 7. | | PC | Pronoun + conjunction |
| 8. | | PREF | Reflexive pronoun |
| 9. | | PD | Demonstrative pronoun |
| 10. | | PDPR | Preposition + demonstrative pronoun |
| 11. | Verb | VV | verb |
| 12. | | AX | Auxiliary |
| 13. | | VC | Verb + conjunction |
| 14. | Adjective | JJ | adjective |
| 15. | | JC | Adjective _ conjunction |
| 16. | | JP | Preposition + adjective |
| 17. | Adverb | ADV | adverb |
| 18. | | ADV PREP | Preposition + adverb |
| 19. | | ADVC | Adverb + conjunction |
| 20. | Preposition | PR | Preposition |
| 21. | Conjunction | CC | conjunction |
| 22. | Numerals | ON | Ordinal number |
| 23. | | JN | Cardinal Number |
| 24. | Punctuation | PUNC | Punctuation |
| 25. | Interjection | II | Interjection |
| 26. | Negation | NG | Negation |

## VI. APPROACHES AND TECHNIQUES

As it is mentioned in section one, this work is an extension of the work done in [9], which uses TEL for Afaan Oromo Tagger. The researcher has tried to customize the original Brill Tagger for Afaan Oromo with a bit modification. Even though the performance of the modified Brill Tagger is better than the default, it has also got varies drawback. Most of the words are incorrectly assigned to a single tag (noun) the initial state tagger is assign for untagged Afaan Oromo texts. Moreover, the transformational rules were trained on very small training corpus that lacks knowledge to generalize and perform proper change of tags based on the learnt rule.

Thus this research is designed to mitigate the limitation of the work done in [9] by doing the following amendments that the researchers believe will enhance the performance of the TEL POS for Afaan Oromo. The first one is to use sufficiently large corpus to train the transformation rules so as to capture detail knowledge of tag transformation of the language. The second is to replace the initial state annotator in the Brill Tagger with HMM based POS tagger. This would have the following impacts that improve the Brill Tagger lexical and transformational rules for Afaan Oromo.

1. The initial state annotator almost will have the appropriate tag of each lexicon in the given corpus and hence will

improve performance as it minimizes the wrongly assigned initial tags to words.

2. The transformation rule requires less knowledge to make corrective actions and hence the required knowledge would easily be captured from the corpus.

## VII. PREFORMANCE ANALYSIS

The Brill tagger with modifications is used for conducting experiments in the rule based tagger. Ten different experiments are conducted on the Brill's tagger using different size of the training set and different initial state annotators. The experiment starts from the first 10% of the training corpus, repeatedly adding 10% of the corpus until the entire corpus is used. Table 6.1 and figure 6.1 shows the different experiments conducted using different portions of the training set with the corresponding performance of the rule based tagger for the different initial state annotators.

### i. Brill's Tagger Versus Intial State Tagger

***Table 6.1 Brill's Tagger performance using different initial state taggers***

| **Initial State Tagger** | | **Size of the Training set** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
| *Performance per different initial state taggers (%)* | Default Tagger | 54.2 | 59.4 | 60.3 | 64.2 | 67 | 72.1 | 76.5 | 83.2 | 84 | 89.6 |
| | HMM Tagger | 71.4 | 72.9 | 74.6 | 75.8 | 79.1 | 79.9 | 81.7 | 87.0 | 92.4 | 93.35 |

The default tagger assigns a specific part of speech tag for each word. In this work, when it takes the default tagger as an initial state annotator, the Noun (NN) and Proper noun (NNP) if capitalized part of speech tagger is selected to be default tag. The HMM tagger assigns the most optimum tag sequence given the word sequence. A significantly higher performance is achieved when the HMM tagger is used as initial stated tagger, which implies that HMM tagger simplifies the learning work of the Brills training as well as the accuracy of the rule generated. The Following diagram shows the learning curve during the training of the Brill's tagger using the HMM tagger as initial state tagger.

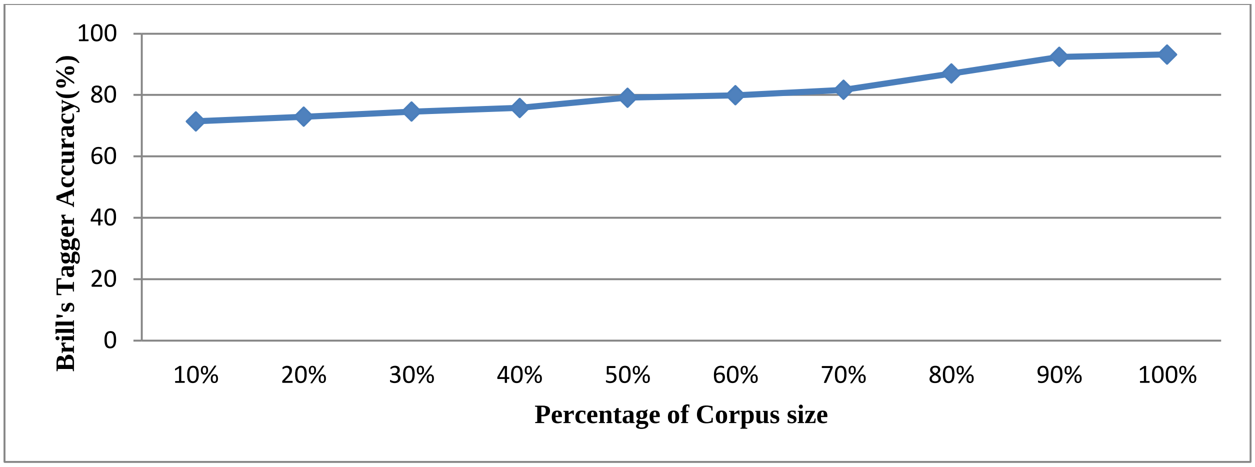

*Figure 6.1 Learning curve of the Tagger*

ii. **Brill's Tagger Versus Corpus Size**

The tagger is also checked on the size of training corpus used. The size depends on numbers of words in the corpus. Accordingly, it is shown that the size of the corpus used for Brill's Tagger has a significant effect on the accuracy of the Tagger. Figure 6.2 shows the increasing Brill's Tagger accuracy with the increase in the size of Afaan Oromo corpus.

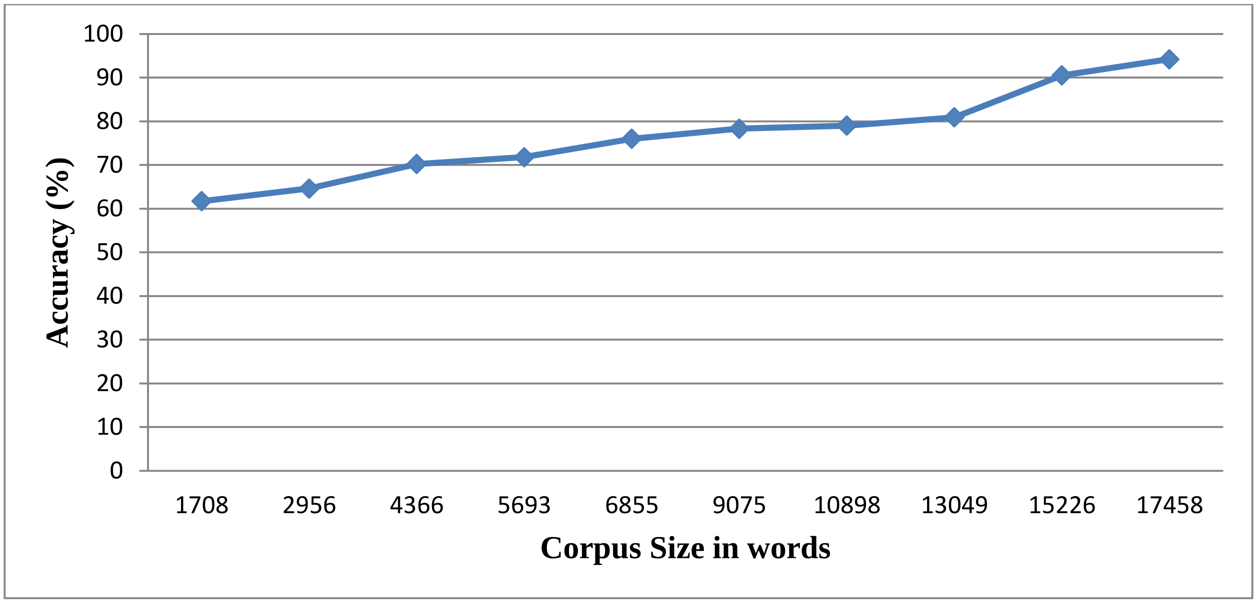

*Figure 6.2 Brill's corpus size tagger versus*

VIII. DISCUSSION

Different experiments are conducted for the Afaan Oromo Brill's tagger. Comparison with the Brill's tagger developed for Afaan Oromo in the work of [9] is done. Accordingly, different performance is obtained: the improved Brill's Tagger performed better than previously adapted Brill's Tagger. The performance of the Original

Brill's Tagger and Improved Brill's tagger is 89.8% and 95.6% respectively, which results with the difference of 5.8%.

This performance improvement is made because of the improvement on the size of the training and testing corpus, the choice of HMM tagger as initial state tagger and the rule generating system in the lexical rule learner.

In general, a 10 fold validation system is used to evaluate the accuracy of the tagger. This is done by dividing the entire corpus randomly into ten parts. The nine fold is used for training and the remaining tenth fold is used for testing the tagger that was trained on the previous nine folds. The Process was repeated ten times by taking other nine as training and the tenth one as testing corpus.

A performance comparison for each part of speech tagger for the previously adapted Brill's Tagger and Improved models is given in table 6.3 to see the performance improvement through making improving the Brill's Tagger for Afaan Oromo Language. The Comparison is made with the 10 fold validation system.

***Table 6.4 Comparison of Original Brill's Tagger [9] and Improved Brill's tagger***

| **S/N** | **No of words** | **Original Brill's Tagger** | **Improved Brill's Tagger Accuracy (%)** |
|---|---|---|---|
| 1 | 2450 | 92.4 | 97.4 |
| 2 | 2406 | 91.6 | 97.6 |
| 3 | 2381 | 94.9 | 98.9 |
| 4 | 2112 | 91.2 | 96.7 |
| 5 | 1850 | 92.8 | 93.4 |
| 6 | 1297 | 86 | 92.6 |
| 7 | 1587 | 88.5 | 94.9 |
| 8 | 1310 | 84.5 | 93.6 |
| 9 | 1091 | 91.5 | 94.8 |
| 10 | 997 | 84.8 | 96.1 |
| **Average Accuracy** | | **89.8** | **95.6** |

Previously, the accuracy of HMM Afaan Oromo Tagger is 87.58% for Unigram and 91.97% for Bigram for the work of [8]. The Afaan Oromo Brill's tagger has got 80.08% accuracy from the work of [9]. In this work, the Brill tagger is with Average of 89.8% accuracy while the Improved Brill's Tagger is with 95.6% accuracy with the same corpus size.

## IX. CONCLUSION

With the increase on the size of training corpus, the accuracy of the tagger increases. This is shown with the choice of the initial state tagger, which has a significant effect on the accuracy of the tagger. Accordingly, HMM tagger is chosen to be the one with best performance. This implies that using HMM tagger as initial state tagger increases the accuracy of the rule generated during the learning phase of the Brill's tagger.

The comparison of the improved Bill's tagger is made with the Original Brill's tagger with 10 fold validation system. Accordingly, the overall accuracy for Original Brill's Tagger is 89.8% while the improved Brill's tagger is 95.6%.

## X. RECOMMENDATION

There are lots of research areas in natural language processing that can be done for different languages in Ethiopia. The same thing holds true for Afaan Oromo language. Therefore, to assist researchers, it will be of great paramount if a standard corpus for Afaan Oromo language is developed that will be available for NLP researchers in Afaan Oromo language.

Finally this research work suggests the following items as a future work:

- ➢ Using morphologically analyzed corpus for training of Brill's tagger's to consider the inflectional properties of the language.
- ➢ Comparison of two hybrid approaches: the hybrid of rule based and HMM tagger and the hybrid of rule based and ANN for Afaan Oromo language
- ➢ Extending this work by training in using tagsets that can identify gender, number, tense etc with different feature set
- ➢ Conducting similar researches for other local languages by adapting this work.

## ACKNOWLEDGMENT

First and foremost, I would like to express my heartfelt gratitude to the almighty God. All of my efforts would have gone for naught if it had not been for his importunate help. Then I offer my sincerest thanks to my supervisor, Dr Sebsibe H/Mariam, who has supported me throughout my work with his patience and knowledge whilst allowing me the room to work in my own way.

## REFERENCES

[1] Christopher D. Manning HinrichSchutze. Foundations of Statistical Natural Language Processing, 2nd Ed. The MIT Press Cambridge, Massachusetts London, England, 2000.

[2] Tarveer S. Natural Language Processing and Information Retrieval.Published by Oxford University press in Indian Institute of Technology, Allahabad, India, 2008.

[3]Brill, E. A simple rule-based part of speech tagger.Department of Computer Science, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A, 1995.

[4] Megyesi B. Brill’s POS Tagger with Extended Lexical Templates for Hungarian. Master’s thesis, Department of Linguistics, Computational Linguistics, Stockholm University, Stockholm, Sweden, 1999.

[5] Abdulsamad M. ‘SeerlugaaAfaanOromoo’. Bole Printing Enterprise, Addis Ababa, Ethiopia.1997.

[6] Mohammed S. & Pedersen T. Guaranteed Pre Tagging for the Brill Tagger.University of Minnesota, Duluth, USA.

[7] GamtaTilahun. Forms of Subject and Object in AfaanOromo.Journal of Oromo Volume 8 Number 1&2, July 2001.

[8] GetachewMamo. Part-of-Speech Tagging for Afaan Oromo Language.Master’s thesis, Addis Ababa University, 2009.

[9] Mohammed-Hussen. Part Of Speech Tagger for Afaan Oromo Language using Transformational error driven learning (TEL) approach.Master’s thesis, Addis Ababa University, 2010.

[10] Robin. Natural Language Processing.Article on Natural Language Processing.Published on December 16th, 2009.

[11] Wolfgang Teubert. Corpus Linguistics and Lexicography,JohnBenjamins Publishing Co. International Journal of Corpus Linguistics Volume 6, 2001, 125-153

[12] Fahim Muhammad Hasan, NaushadUzZaman, Mumit Khan, (2006). Comparison of Different POS Tagging Techniques (n-grams, HMM and Brill’s Tagger) for Bangla, International Conference on Systems, Computing Sciences and Software Engineering (SCS2 06) of International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CIS2E 06), pp: 4-14.

[13] Hall, Johan. A Probabilistic Part-of-Speech Tagger with Suffix Probabilities.Master’s thesis, School of Mathematics and Systems Engineering, Växjö University, 2003.

[14] Blunsom Ph. Hidden Markov Models: pcbl@cs.mu.oz.au, August 19, 2004

[15] KhineZin, (2009). Hidden Markov model with rule based approach for part of speech tagging of Myanmar language, World Scientific and Engineering Academy and Society (WSEAS) Stevens Point, Wisconsin, USA.

[16]Brants, T. TnT - a statistical part-of-speech tagger. In Proceedings of the 6th Conference on Applied Natural Language Processing, Seattle, Wash., 29 April – 4 May 2000, pp. 224–231.

[17]Gerold S and Martin Volk. Adding Manual Constraints and Lexical Look-up to a Brill Tagger for German, Computational Linguistics Group, Department of Computer Science, University of Zurich, 2000.

[18] Schmid, H. Part-of-speech tagging with neural networks. In Proceedings of COLING-94, Kyoto, 1994.

[19]Nuno C. & Gabriel Pereira, Neural Networks, Part of Speech Tagging and Lexicon.Technical Report DI-FCT/UN, University Nova de Lisboa– Faculty of Technology, Department of Informatics, Portugal, 1998.

[20] TeklayGebregzabihe. Part of Speech Tagger for Tigrigna Language.Master's thesis, Addis Ababa University, 2010.

[21] Solomon Asres, (2008). Automatic Amharic Part-of-Speech Tagging Using Hybrid Approach (Neural Network and Rule-Based). Master's thesis, Addis Ababa University.

[22] Megyesi B. 1998. Improving Brill's POS Tagger for an Agglutinative Language.Thesis in Computational Linguistics, Department of Linguistics, Stockholm University, Sweden.

[23] Petasis G, Paliouras G, Vangelis, Karkaletsis D and Androutsopoulos, Resolving Part-of-Speech Ambiguity in the Greek Language using Learning Techniques, Institute of Informatics and Telecommunications, N.C.S.R, "Demokritos", 2002.

[24] FDRE Population census Commission, Summary and Statistical report of the 2007 population and housing census.Printed by United Nations Population Fund (UNFPA) Addis Ababa, December 2008.

[25] TilahunGamta. QubeAfaan Oromo: Reasons for Choosing the Latin Script for Developing an Oromo Alphabet. Published on the Journal of Oromo studies Volume I Number I Summer 1993.

[26] Http://www.Wepeadia.com/ Wiki: Oromo language (1/3), visited on Aug 3 2010.

[27] Brill E and Marcus M. 1992. Tagging an Unfamiliar Text with Minimal Human Supervision.In Proceedings of the Fall Symposium on Probabilistic Approaches to Natural Language, 1992.

[28] Andrew Roberts. Machine Learning in Natural Language Processing, October 16, 2003

[29]Hassan S. Statistical Part of Speech Tagger for Urdu. Thesis, National University of Computer and Emerging Science, Department of Computer Science. Lahore, Pakistan, 2007.

[30]Qing Ma, KiyotakaUchimoto, Masaki Murata,and Hitoshi Isahara. ElasticNeuralNetworksfor Part of SpeechTagging, Communications Research Laboratory, MPT, Japan

[31] DiribaMerga, Automatic Sentence Parser for Oromo Language, Thesis, School of Graduate studies, Addis Ababa University, 2001.

[32] Daniel Bekele. AfaanOromo-English Cross-Language Information Retrieval.Master's thesis Addis Ababa University, 2011.

[33] Clark, S., J. R. Curran & M. Osborne. Bootstrapping POS taggers using unlabelled data. In Proceedings of the Seventh CoNLL conference held at HLT-NAACL, Edmonton, Alberta, Canada, 27 May –1 June, 2003, pp. 49–55.

[34] Jurafsky, D and Martin H. James. Speech and Language Processing, Prentice Hall, 2000.

[35] Church, K (1988) A stochastic parts program and noun phrase parser for unrestricted text. In: Proceedings of the second conference on Applied Natural Language Processig, ACL.

[36] Cutting, D, Kupiec, J, Pederson, J, and Sibun, P (1992) A practical part-of-speech tagger. In: Proceedings of the third conference on Applied Natural Language Processing, ACL.

[37]http://www.scribd.com/doc/78614218/MODERN-AFAAN-OROMO-GRAMMAR. Visited on January, 2012.

[38] Bryan Jurish. A Hybrid Approach to Part-of-Speech Tagging, Berlin, German, 2003