# American Journal of Computer Science and Engineering Survey

**Research** Article

# Investigating Afan Oromo Language Structure and Developing Effective File Editing Tool as Plug-in into Ms Word to Support Text Entry and Input Methods

## WORKINEH TESEMA<sup>1</sup>\* AND DURESA TAMIRAT<sup>2</sup>

<sup>1</sup>Department of Information Science Jimma University, Jimma, Ethiopia <sup>2</sup>Department of Information Technology, Medawolabu University, Robe, Ethiopia \*Corresponding author e-mail: workineh.tesema@ju.edu.et

### <u>ABSTRACT</u>

Afan Oromo is a member of the Cushitic branch of the Afro-Asiatic language family, which was the third most widely spoken language in Africa, after Hausa and Arabic. Its original homeland is an area that includes much of what is today Ethiopia and some parts of East African countries. Afan Oromo uses a Latin script which consists of thirty three basic letters, of which five are vowels, twenty-four are consonants, out of which seven are paired letters and fall together (a combination of two consonant characters such as 'CH', 'DH', 'NY', 'SH', 'TS' ). The idea behind this work is to open a chance to obtain computer software and file editing tool in Afan Oromo language. In order to develop this tool, unsupervised machine learning which was trained on unlabeled corpus. The training data were collected from government media, cultural, historical, sport news, political, and economical documents of Afan Oromo users were used. Once, the trained data was collected based on language structure N-gram algorithms (namely Unigram, Bigram, Trigram & Fourth Gram) were applied. Hence, Afan Oromo is one of the limited resources (small dataset) for training it restricted to use Unigram, Bigram and Trigram. Therefore, this work presents how we improve word entry information and input method as an assistive technology. Hence, this language uses double vowels (in this case waadaa) it needs integrated and independent file editor for native users of the language. As the developed system shows that it makes easy to text entry and improve the way to input files to computers. Finally, this work was brought the Oromo population, which is the first largest population of Ethiopian populations to get access the technology by their mother tongue language. The finding of the study was argued that the file editor indispensable to use technology by own language and especially for disable and typist to edit own file.

Keywords: File editor; Text entry; Word production; Afan promo; Input method; Assistive technology.

#### **INTRODUCTION**

Computers and technologies are becoming an integral and important part of life for most people, including those with physical disabilities as well as for normal also. Improving and enhancing text entry and interaction with computers for users has been investigated for many years, with many systems and interfaces proposed to facilitate and simplify text input process [1]. However, for Afan Oromo which is under resource language and not yet started the development of such technology was a great opportunity. Therefore, the aim of this work was to explore Afan Oromo language structure (spelling or *Qubee*, words, phrases and sentences) and then develop an effective file

American Journal of Computer Science and Engineering Survey

www.pubicon.in

editor which was plug-in to Ms Word 2007-2013. Actually, this file editor is helpful for those typists, disable and busy to write large text files. This work brought how to improve textual information entry, through file editor, as an assistive technology for Oromo people using the regular input device, the keyboard, to eliminate the overhead and the cognitive load needed for the learning and the familiarizing process associated with the new and specialized input method. Since the keyboard is the de facto universal interface for most general computing devices, it is more appealing to invent new techniques to allow faster and easier ways to interact with the keyboard for all users. Moreover, inputting text using regular keyboard, without the need of any special hardware or devices, enables such users to continue to use the everyday offthe-shelf computer applications and programs, like e-mail programs and word processors. While writing the files, especially using local languages, it takes time and resource, hence most of file editors are using the English and other language. For example, Afan Oromo which has more than half of the population of a country (Ethiopia) used as their mother tongue and which is long vowels and short vowels, needs assisting technology to edit files. In case of long vowels which is vowel repetition, it needs to file editor which predict and complete to save time and resources of users, particularly for Afan Oromo users with poor knowledge of spelling. Additionally, the speed of typing of many secretaries (busy peoples) when they write Afan Oromo text is very low and misspelled words that create miscommunication between users. Hence, the single letter may change the meaning of the word if it is misspelled. Furthermore the lack of Afan Oromo files editor impacts on Afan Oromo language speakers. In addition to lack of file editor, due to misspelling and low speed of typing, the new Afan Oromo speakers are shamed from practicing the language. Therefore, this study is undertaken to solve the problems by providing Afan Oromo file editor to the users. The proposed techniques are based on machine learning method and N-gram algorithms. In natural language processing research, most of the efficient and robust systems for processing natural language are designed and developed based on learning and training approaches [2]. Such learning approaches typically produce some knowledge bases or trained models that can subsequently be employed in the underlying natural language processing task. Unsupervised learning approach can be distinguished from supervised learning approach where supervised learning learns from the particular annotated corpus which is unlike this study.

#### **Overview of Afan Oromo Script**

Afan Oromo, also called Oromiffaa or Afaan Oromoo, is a member of the Cushitic branch of the Afro-Asiatic language family [3]. It is the third most widely spoken language in Africa, after Hausa and Arabic. Its original homeland is an area that includes much of what is today Ethiopia, Somalia, Sudan and northern Kenya and some parts of other East African countries [4]. Currently, it is an official language of Oromia Regional State (which is the biggest region among the current Federal States in Ethiopia). It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 50 % of the total population in 2007 (2015 Census statistic of Ethiopia). With regard to the writing system, Qubee (a Latin-based alphabet) has been adopted and become the official script of Afan Oromo from 1991 [5].

Among the major languages that are widely spoken and used in Ethiopia, Afan Oromo has the largest speakers. It is considered to be one of the five most widely spoken languages among the roughly one thousand languages of Africa [6]. Afan Oromo, although relatively widely distributed within Ethiopia and some neighboring countries like Kenya, Tanzania and Somalia, is one of the most resource scarce languages [7]. It is part of the Lowland East Cushitic group within the Cushitic family of the Afro-Asiatic phylum, unlike Amharic (an official language of Ethiopia) which belongs to Semitic language family. Although it is difficult to identify the actual number of Afan Oromo speaking societies (as a mother tongue), due to lack of appropriate and current information sources, according to the census taken in 2007 it was estimated that 50% of Ethiopians are ethnic Oromo [8].

#### Afan Oromo Qubee

Afan Oromo uses *Qubee* (Latin based alphabet) that consists of thirty three basic letters, of which five are vowels, twenty-four are consonants, out of which seven are paired letters and fall together (a combination of two consonant characters such as '*ch*'). The Afan Oromo alphabet characterized

by capital and small letters as in the case of the English alphabet. In Afan Oromo language, as in English language, the vowels are sound makers and are sound by themselves. Vowels in Afan Oromo are characterized as short and long vowels. The complete list of the Afan Oromo alphabets is found on the manuscript by [3]. The basic alphabet in Afan Oromo does not contain 'p', 'v' and 'z', because there are no native words in Afan Oromo that formed from these characters. However, in writing Afan Oromo language, they are used to refer to foreign words such as "*polisii*" ("police").

#### WORD AND SENTENCE BOUNDARIES

In Afan Oromo, like in other languages, the blank character (space) shows the end of one word. Moreover, parenthesis, brackets, quotes are being used to show a word boundary. Furthermore, sentence boundaries punctuations are almost similar to English language i.e. a sentence may end with a period (.), a question mark (?), or an exclamation point (!) [9].

#### WORD SEGMENTATION

The word, in Afan Oromo "jecha" is the smallest unit of a language. There are different methods for separating words from each other. This method might vary from one language to another. In some languages, the written or textual script does not have whitespace characters between the words. However, in most Latin languages a word is separated from other words, by white space characters [10]. Afan Oromo is one of Cushitic family that uses Latin script for textual purpose and it uses white space character to separate words from each other's. For example, "Bilisummaan Finfinnee deeme". In this sentence the word "Bilisummaan", "Finfinnee" and "deeme" are separated from each other by white space character. Therefore, the task of taking an input sentence and inserting legitimate word boundaries, called word segmentation, is performed using the white space characters.

#### LANGUAGE STRUCTURE

Afan Oromo has a very rich morphology like other African and Ethiopian languages [3]. With regard to the writing system, *Qubee* (Latin-based alphabet) has been adopted and became the official script of Afan Oromo since 1842 [6]. The writing system of the language is straightforward, which is designed based on the Latin script. Thus letters in the English language are also in Afan Oromo except the way it is spelled. A detailed description of Afan Oromo Writing System can be found in any text related to the language, but [11] discussed writing system of the language.

#### SENTENCE STRUCTURE

Afan Oromo and English are different in sentence structuring. Afan Oromo uses subject-object-verb (SOV) language. SOV is the type of language in which the subject, object and verb appear in that order. Subject-verb-object (SVO) is a sentence structure where the subject comes first, the verb second and the third object. For instance, in the Afan Oromo sentence "Mooneeraan bilisa bahe". "Mooneeraa "is a subject, "bilisa" is an object and "bahe" is a verb. Therefore, it has SOV structure. The translation of the sentence in English is "Mooneeraa has got freedom" which has SVO structure. There is also a difference in the formation of adjectives in Afan Oromo and English. In Afan Oromo adjectives follow a noun or pronoun; their normal position is close to the noun they modify while in English adjectives usually precede the noun. For instance, namicha gaarii (good man), gaarii (adj.) follows namicha (noun).

#### **PUNCTUATION MARKS**

Punctuation marks used in both Afan Oromo and English languages are the same and used for the same purpose with the exception of the apostrophe. Apostrophe mark (') in English shows possession, but in Afan Oromo it is used in writing to represent a glitch (called hudhaa) sound. It plays an important role in Afan Oromo reading and writing system. For example, it is used to write the word in which most of the time two vowels appeared together like "ba'e" to mean ("get out") with the exception of some words like "ja'a" to mean "six" which is identified from the sound created. Sometimes apostrophe mark (') in Afan Oromo interchangeable with the spelling "h". For instance, "ba'e", "ja'a" can be interchanged by the spelling "h" like "bahe", "jaha" respectively still the senses of the words is not changed.

The motivation behind this work was to bring technology to the users and allow the users to use file editing tool by presenting into their local language.

#### LITERATURE REVIEW

The task of helping people to interact with computers has been investigated for a long time and many tools, techniques, and devices have been proposed to help users with special needs (physical deficiencies) to successfully use computers. Several research projects [12] have been working on developing tools and solutions to assist computer utilization for disabled users and to provide them with an effective and efficient means to interact with computers, which eventually will lead to broaden the participation of disabled and elderly people in the fast growing information technology world [13].

In general, the problem of facilitating computer interaction has been addressed as follows:

Improving the design of the input interface, and usually, introducing new devices (head-stick, head-pointer, eye-tracker, stylus touch pad) [14-16]. New devices result in extra hardware cost and usually require special training, which often hurt its adoption by disabled users.

a. Reducing the input keystrokes needed to produce a given text using the existing regular input devices [17].

The research conducted by [18] on word prediction text entry for Afan Oromo on a mobile phone which is limited to only hand devices. His work presents a word prediction approach based on context features and machine learning. As the result, it shows that the accuracy performance of his system 56.8%. This work was limited to only hand held devices which has a problem when users have no smart phone and since cannot plug-in to Ms Word. Additionally, his work was not functional when users want to edit files on their computer.

According to Workineh (2017) on Afan Oromo sense clustering at word level states that Afan Oromo is the richest morphology hence double vowels (long vowels) are used. The main aim of this work was to cluster the word that has multisense (meaning) for the purpose of Information Retrieval. As the finding of the study was argued that from the root of words, many words can be produced. Based on this, our study was strongly supported that hence this language was morphological rich it needs a file editor to edit the users file by their own on their computer.

Unsupervised methods are based on unlabeled corpora, and do not exploit any manually tagged corpus to provide a predicted choice for a word in context unlike supervised method [19]. Unsupervised methods have the potential to overcome the knowledge acquisition bottleneck [20], that is, the lacks of large-scale resources manually annotated with word predict. They do not rely on labeled training text and, in their purest version, do not make use of any machinereadable resources like dictionaries, thesauri, ontology. However, the main disadvantage of fully unsupervised systems is that, as they do not exploit any dictionary, they cannot rely on a shared reference inventory of senses [21]. Most of the time, supervised approaches are superior to unsupervised in terms of accuracy of prediction when used on the same type of words that the systems were trained on [22]. This is especially evident in the creation of corpora that is manually annotated (tagged) with the senses, which are used for training machine learning classifiers in a supervised setting. There are two important problems during manual tagging of a corpus: low inter annotator agreement (IA) and high cost of the annotation process. IA is a way of measuring how much an annotation assigned by one annotator differs from annotations assigned by another annotator. IA is used for the estimation of an upper bound on performance in automatic prediction, but there is also another measure [23].

#### **Corpus - based Approaches**

A major challenging face WP research is the ability to acquire a large number of words with their frequency. Corpus-based approaches came up with an alternate solution to the challenge by obtaining information necessary for WP directly from a corpus. A corpus provides a bank of samples which enable the development of numerical language models, and thus the use of corpora goes hand-in-hand with empirical methods.

#### MATERIALS AND METHOD

Machine learning approaches, as indicated in the above section, have shown significant robustness and efficiency in natural language processing [24]. In this research, one class of learning method investigated, designed, and implemented is unsupervised learning method. Then, this class of the method was integrated into a comprehensive learning architecture and N-gram algorithms that were capable of acquiring reliable and relevant knowledge automatically and efficiently. We have applied the learned knowledge to the problem of file edition for text entry (text entry method). The key objective is to allow the system to learn from prior (training) texts (unsupervised learning by n-grams algorithms), and from the user (take the first spelling of the word), so it can reliably predict the words the user intends to input.

The following steps summarize the proposed method:

• Developing machine learning algorithms that focus on the syntactic/words/phrases information in the corpus.

• Developing N-gram algorithms (Unigram, Bigram & Trigram) focusing on the particular user/domain specific information.

• Integrating the two learning mechanisms of (1) and (2) into a robust learning model that can easily be applied to other tasks similar to the one described in (4).

• Applying the model of (3) to the task of word prediction and completion, with the goal of generating a very small number of suggestions. This task will offer a faster and more efficient way of interaction and communication with the computer with not much extra cognitive load.

• The devised system will be available for physically disabled users to test and evaluate the efficiency and effectiveness of the approach.

#### **Training Data**

A corpus (plural corpora) is a large collection of natural language text. It can be collected from different sources such as novels, newspapers, discussion forums, government media, cultural, historical, sport news, political, and economical documents etc. Some corpora are only collected from one source while others come from multiple sources. Different corpora apply to different needs, e.g. an application used in the financial industry would not benefit from a corpus based on sports articles since the different lingo match poorly. A suitable choice of the corpus is therefore important and has an impact on the application.

#### **Simulation Tools**

As it discussed in the above sections, to develop this system the algorithms were implemented in Java, NetBeans IDE 8.02 version software which was open and run on the prepared corpus. The reason why the researcher used this tool is hence Java is free and helps us to develop user interfaces. And also, it is a general purpose and open source programming language. Moreover, it is optimized for software quality, developer productivity, program portability, and component integration. Lastly, the reason why Java selected for this work was it is a platform independent language which is after application developed, we can run on different operating system of the users.

#### IMPLEMENTATION AND EXPERIMENT

The main contribution of this work would be integrated unsupervised learning with n-gram algorithms to produce robust learning techniques that are capable of performing file editing with high accuracy to achieve maximum keystroke savings. The value of this work can be viewed in four aspects; the first one is its universality as it uses the keyboard that is the universal text entry device to computers, thus no training needed and no new devices introduced. The second aspect is it enhance the functionality of Ms Word of the plugin file editor to support local language(Afan Oromo). The third aspect is its focus on reducing cognitive load by offering, in most cases, only one or more suggestion for word prediction. The fourth aspect is the positive impact of machine learning integration on other fields like Human Computer Interaction (HCI).

As the finding of the study was argued that Afan Oromo is the plentiful morphological unit of a language. Since, this language has long vowels which are needed to type repeatedly the same or different vowels and it consume the time of the typist to edit one simple file. For instance; the word "dagaagaa" has too long vowels "a" which is the same vowel in the word but need to type again and again (Figure 1). Therefore, such kind of words in the langauge needs a file editor tool which makes easy to type and create files in a short time. The development of this tool has brought the golden chance for Oromo peoples and speakers, hence any volunteer human or government officials, TV studio, students, secretary, typist of the language can create their files/document within limited minute. As it is shown in the Figure 2 below, the tool also has additional functionality like word prediction by pressing the starting letter and complete with a word completion button.

As a technology on the way of growing, many users have tried to use different technology for different purposes. File editor is one of the technologies that plays a great role in our day to day activity especially in relation to processing natural language and that most of us used in our office to edit our files. The writer is intending to produce when creating text documents using varies text editors. Similarly, typing of Afan Oromo language with huge alphabets requires suitable and efficient methods to access all the letters with a QWERTY keyboard. In all application areas, various text entry and input methods have been suggested to provide a more efficient input of texts with poor spelling and disability [1].

As the finding, argued that this file editor was ease of use and high speed when compared with the normal Ms Word. It is one that manages the string (sequence of characters) or list of the words that represents the current state of the file being edited. As it is shown that the desire for this file editor is that could more quickly insert text and delete text by this language. Therefore, this file editor is a computer program that lets a user enter and usually change (characters and numbers, special symbols and input using the existing keyboard devices, arranged to have meaning to users).

The other aim of this study was to enhance the functionality of Ms Word used to edit our files by adding additional plug-in file editor tool in Afan Oromo. This increase the usability of Ms Word which is plug-in by their own local language (Afan Oromo). Since, Ms Word was not supported Afan Oromo language the role of this work was great and brings a chance to Afan Oromo speakers to edit their own files by their own language. As the result shows that people with physical disabilities and motion impairments, especially who can't write and low speed typing where the first beneficiary of this work. One of the main difficulties faced by such people in interacting with computers is that their word entry is very slow, and the typing process can be tiring [16].

As the experiment shows that the heart of this work is file editing task. The potential user would be direct touch with, this GUI interface. Its goal is quite simple as accept input from the user, show him predicted continuations of the text and allow him unobtrusive acceptance of the predictions. N-gram language model is an important technique for such task. We use the large data corpus for training in N-gram language model for editing the correct word to complete Afan Oromo word with more accuracy [25-37].

Based on the figure 2 above, the users can type and edit any file which was plug-in into on Ms Word. As the user open Ms Word this tool displayed in the provided graphical user interface. Within limited time we can create/type large files since the assistive tool was displayed the necessary options as we press the first letter of the word (as it shown on the above figure). Additionally, this tool was used as word prediction and then completion. If the needed words, does not listed in the predicted list, the user must know as the word is not present in the prepared corpus and the user can still write the next character of the word.

Based on the result of the study the words were displayed on GUI based on their frequency occurrences which are ranked (highest to lowest occurrence) in the corpus. Sometimes, hence the corpus collected from multi resources, there are spelling errors. Unfortunately, our system cannot recognize the problem of misspelling due to its needs Afan Oromo spelling checker system. However, if the word is spelled correctly and available in the corpus the performance is high.

#### CONCLUSION

The overall focus of this research is to investigate Afan Oromo language structure and to develop a file editor tool which addresses the problem of poor spelling, misspelling and completion. Ideally, this can speed up and ease the user's typing of word production. This work is improving and enhancing textual information entry for disabled and typist users has been investigated, with unsupervised approach and user interface proposed and implemented to facilitate and simplify text input for such people. In this work the unsupervised machine learning achieved an accuracy of 85%, 80.34% of precision and recall respectively.

#### REFERENCES

- 1. Quiroga L, Crosby ME, Iding MK, Reducing cognitive load, International Conference on System Sciences. 2004, 319.
- 2. Fazly A (2002) The Use of Syntax in Word Completion Utilities.
- Tullu Guya, CaasLuga, Afaan Oromoo Jildii-1, (2003) Gumii Qormaata Afaan Oromootiin Komishinii "Aadaa fi Turizimii Oromiyaa", Finfinnee.
- 4. Abera N (1988) Long vowels in Afan Oromo: A generic approach.
- 5. Debela Tesfaye (2010) Designing a Stemmer for Afan Oromo Text: A hybrid approach, Int J of Comp Ling, 1:2.
- 6. Gragg, Gene B (1996) Oromo of Wollega non-

semetic languages of Ethiopia.

- 7. Finfinnee (1995) Boolee Tilahun Gamta Seera Afaan Oromoo, Press.
- 8. Kula Kekeba Tune, Vasudeva Varma, Prasad Pingali (2007)" Evaluation of Oromo English Cross-Language Information Retrieval".
- 9. Getachew Rabirra Furtuu (2014) Seerluga Afaan Oromoo, Finfinnee Oromiyaa press.
- 10. Atelach Alemu Argaw, Lars Asker (2010) An Amharic Stemmer: Reducing Words to their Citation Forms.
- 11. Wakshum Mekonnen (2000) Development of Stemming Algorithm for Oromo Texts. MA
- 12. Thesis.
- 13.Glinert EP, York BW (1992) 'Computers and people with disabilities'. 35: 32-35.
- 14. Steriadis C, Constantinou P (2003) 'Designing human-computer interfaces for quadriplegic. 10: 87-118.
- 15.Keates S, Clarkson J, Robinson P (2000) 'Investigating the applicability of user models for motion-impaired users', The Fourth International ACM SIGCAPH Conference on Assistive Technologies. 129-136.
- 16.Manaris B, Mc Cauley R, Mac Gyvers V. (2001) 'An intelligent interface for keyboard and mouse control', Proc. 14th Int 1 Florida AI Research Symposium (FLAIRS-01), FL pp 182-188.
- 17.Cockburn A, Siresena A (2003) Evaluating Mobile Text Entry with the Fasta Keypad.
- 18. Wobbrock JO, Myers BA Kembel, JA Edge (2003) Write: a stylus-based text entry method designed for high accuracy and stability of motion, 61-70.
- 19.Gudisa T, (2013) Design and Implementation of Predictive Text Entry Method for Afan Oromo on Mobile Phone.
- 20.Roberto Navigli, Word Sense Disambiguation: A Survey, ACM Computing Surveys. 2009, 41.
- 21.Gale W, Church K, Yarowsky D (1992) A method for disambiguating word senses in a large corpus. Computers and the Humanities, 26: 415-439.
- 22.Doina Tatar, Gabriela Serba (2001) A New

Algorithm for Word Sense Disambiguation, Studia Universitatis "Babes-Bolyai", Seria Informatica, 16.

- 23.Ide N, Veronis (1998) J. Introduction to the special issue on word sense disambiguation: the state of the art. Comput. Linguist, 24: 2- 40.
- 24. Artstein R, Poesio (2008) M Inter-coder agreement for computational linguistics. Computational Linguistics, Vol 34.
- 25.Felici G, Sun F, Truemper K (2006) 'Learning logic formulas and related error distributions', in Triantaphyllou, E. and Felici, G. (Eds): Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques, Massive Computing Series, Springer, Heidelberg, Germany, pp.597-628.
- 26. Masudul Haque M T(2015) Automated Word Prediction In Bangla Language Using Stochastic Language Models. International Journal, 4-9.
- 27.ACM Transactions on Computer-Human Interaction 1994, 10: 87-118.
- 28.Fazly A, Hirst G. (2003) 'Testing the efficacy of part-of-speech information in word completion', Workshop on Language Modeling for Text Entry Methods, 11th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, April, pp 9-16.
- 29. Felici G, Sun F, Truemper K (1999) 'A method for controlling errors in two-class classification', Proc. 23rd Annual Int'l Computer Software and Applications Conference, 186-191.
- 30. Girma Debele (2014) Afan Oromo News Text Summarizer, Master's thesis, Pohang University of Science and Technology.
- 31.N, Veronis J (1998) Word Sense Disambiguation, The State of the Art.
- 32.B Mc Caul, A Sutherland (2004) Predictive Text Entry in Immersive Environments. Proceedings of the IEEE Virtual Reality.\_
- 33. Tesema W Afan (2016) Oromo Sense Clustering in Hierarchical and Partitional Techniques. J Inform Tech Softw Eng, 6: 191.
- 34. Tesema W, Tesfaye D, Kibebew T (2015) Towards the Sense Disambiguation of Afan Oromo Words using Hybrid Approach.

Ethiopian Journal of Education and Sciences, 12.

- 35.Tilahun Gamta (1989) Oromo-English Dictionary: Addis Ababa. University Press.
- 36. Yarowsky D. (2007) Unsupervised word sense disambiguation rivaling supervised methods.
- 37. Workineh T, Duresa T (2017) Enhancing the Text Production and Assisting Disable Users in Developing Word Prediction and Completion in Afan Oromo J Inform Tech Softw Eng 7.



