ADDIS ABABA UNIVERSITY

FACULY OF INFORMATICS

DEPARTMENT OF INFORMATON SCIENCE

DEVELOPMENT OF MORPHOLOGICAL ANALYZER FOR AFAAN OROMOO TEXT

ASSEFA WOLDEMARIAM

ADDIS ABABA UNIV. LIBRARY
FACULTY OF INFORMATICS
BIBLIOGRAPHIC LAB

JULY, 2005

ADDIS ABABA UNIVERS
LIBRARIES
P.O.BOX 1176
ADDIS ABABA ETHIOPIA

ADDIS ABABA UNIVERSITY

FACULY OF INFORMATICS

DEPARTMENT OF INFORMATON SCIENCE

DEVELOPMENT OF MORPHOLOGICAL ANALYZER FOR AFAAN OROMOO TEXT

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

BY

ASSEFA WOLDEMARIAM

JULY, 2005

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
Faculty of Informatics
Department of Information Science

DEVELPING MORPHOLOGICAL ANALYZER FOR AFFAN OROMO TEXT

BY

ASEFA W/MARIAM

Name and Signature of Members of the Examining Board

Dr. Gashaw kebede, Chairman, Examining Board
Prof. B. R. Krishina Rao, Advisor
Dr. Nega Alemayehu , Examiner

Chairman, Faculty	Signature	Date
Chairman, Graduate Council	Signature	Date

This thesis is dedicated to

my children

OBSA ASSEFA

ILILLII ASSEFA *and*

BEEKAA ASSEFA

who lacked my follow-up and assistance during my study years

ACKNOWLEDGEMENT

I am most grateful to my thesis advisors Dr. B. Rama Krishna and Ato Kibur Lisanu who have been providing me technical, constructive suggestions and comments during the preparation of this thesis.

My indebtedness goes to my Linguistics advisor Ato Kebede Hordofa for his continuous comments, constructive suggestions and support throughout my work. He is also a provider of different materials on morphology of Afaan Oromoo that I could not find anywhere else. His encouragement and appreciation to my work helped me to develop a confidence generally on linguistic and specifically on morphology.

I would also like to thank my friend Ato Lemma Lessa who has been on my side for every technical assistance and comments as well as to respond my queries throughout the process of writing this thesis.

I would like to express my deepest thanks to my friends and family Ato Hussieen Dalecha, Ato Mesfin Gezahegn, Ato Hasan Waqayyoo and Ato Muhammad Hasanaa who have been my backbone and assisting me in so many ways during my study in terms of morale, financial and material support.

My deepest gratitude and respect goes to my beloved wife Worknesh Teferra who has taken all the responsibilities and burden of managing home, and my children who lacked my follow-up and assistance during the study years.

Finally, my thanks go to those who helped me in one or the other way and who stood at my side during the preparation of this thesis.

TABLE OF CONTENTS

ABBRIVATIONS AND SYMBOLS USED IN THE STUDY

→	=	becomes
1p	=	First person plural
2p	=	Second person plural
2s	=	Second person singular
3f	=	Third person feminine
3m	=	Third person masculine
3msc	=	Third person masculine
3sf	=	Third person singular feminine
AAU	=	Addis Ababa University
AFF	=	Affirmative
C	=	Consonant
cs	=	Causative
HMM	=	Hidden Markov Model
IMPER	=	Imperative
INFL	=	Inflectional
MDL	=	Minimum Description Level
N	=	Noun
NEG	=	Negative
NLP	=	Natural Language Processing
OCR	=	Optical Character Recognition
ORCB	=	Oromia Regional Cultural Bureau
OREB	=	Oromia Regional Education Bureau
Pl	=	Plural
POS	=	Parts of speech
PRES	=	Present tense
Sg	=	Singular
SUBJ	=	Subject
SUBORD	=	Subordinate
V	=	Vowel

LIST OF TABLES

page

LIST OF FIGURES

ABSTRACT

Afaan Oromoo, which belongs to a branch of Afro-Asiatic languages family, is spoken by more than 30 million people in Ethiopia and neighbor countries. It should have a good solid works on its computational aspect especially in storage, processing and retrieval. This study is an attempt on the development and implementation of morphological analyzer for Afaan Oromoo text.

Reviews of Afaan Oromoo morphology and its morphological analysis were made. Sample corpuses of different size ranging from 6,977-48,497 were gathered from three institutions. Documents were reviewed and discussions were made with experts in the field.

Emphasizing on the morphology of the language, a system that uses automatic morphological analysis is developed. The system uses neither stem dictionary nor morphological rules particular to the language. Rather it is based on corpus and learns morphology using heuristic rules to guess the result for from the corpus itself.

The developed analyzer uses Linguistica_beta2 as a main tool to decompose words with in the text in to *stem* + *affix* and analyzes them applying a series of heuristics. Different modifications and improvements were made on Linguistica_beta2 so as to analyze Afaan Oromoo words correctly.

Using Alchemist, a gold-standard of small size (1600 words) is developed to evaluate the performance of the system. On experimenting with different corpus sizes, the system has shown 92.8% of 48,497 words correctly, which is very encouraging and satisfactory.

CHAPTER ONE

INTRODUCTION

1.1. Natural language processing

Natural language is an integral part of our lives. Languages serve as the primary vehicle by which people communicate and record information. It has the potential for expressing an enormous range of ideas, and for conveying complex thoughts clearly and in a few words. Because it is so integral to our lives however, we usually take its powers and influence for granted (Grishman, 1986).

According to Grishman (1986), the potential for natural language processing was recognized quite early in the development of computers, and work in computational linguistics –primarily for machine translation -began in the 1950s at a number of research centers.

Most of human knowledge, according to Allen (1995), is recorded in linguistic form i.e. in the form of natural language texts and utterances. Natural languages are those languages that human beings learn from their environment and make use of them to communicate with each other in their day-to-day activities Harris (1985); Barr and Feigenbaum (1981). Accordingly **Afaan Oromoo, Amharic, English, French, Arabic, Japanese, Spanish** etc. are few instances of *natural languages*.

In every language, whether it is spoken or written, every meaningful pattern has its own structure and the elements of language relate to each other in understandable manner. Words are the basic elements of a language and are formed from morphemes which constitute the smallest meaningful unit of speech in a language. This is also true for Afaan Oromoo which has its own rules of words and/or sentence structure.

Words are central dimension of language. They have certain unique properties that they do not share with other elements of linguistic structure like sentences and speech sounds. The study of the internal structure of words in a language is very important in order to use the language for retrieval and further processing purposes. Moreover, morpheme, pattern of word formation, affixes and all other characteristics of words should be investigated and represented appropriately so that they could further serve for searching and retrieval purpose.

The Morphology of Afaan Oromoo, like most natural languages, is too complex to be understood by computers. Only computers that are made to have the capability to understand natural language can access all the information efficiently. Therefore, understanding Afaan Oromoo morphology, devising the necessary rules to analyze it and automating the process (procedure) make computers more efficient and understand Afaan Oromoo language for further processing and implementation.

1.2.Types of morphological processes

Morphological processes as discussed by Tesfaye (2002), can be categorized in to two– inflectional and derivational. The two processes differ in the type of words they produce. Derivational process create new words of different word class (for example generating a noun from a verb), while inflectional process does not change the word class of the word created but may extend the word. Examples as applied to Afaan Oromo are as shown below:

i) hattuu "thief" from ***hat*** "to steal" + *tuu* **Derivational process**

ii) fiigicha "running" from ***fiig*** "run" + *icha* **Inflectional process**

In example (i) above the *hattuu* 'thief' is formed from a stem *hat 'to* steal' and a suffix *tuu*. This change from noun to verb shows that the morphological process involved is

derivational, whereas in (ii) because both *fiigicha* and *fiig* are verbs, the process involved in forming *fiigicha* from *fiig* and *–icha* is inflectional.

Tesfaye (2002) described that inflectional morphology studies the inflectional changes in words that generally do not result in changing the classes of words. Rather, the inflectional change indicate tense (present, past, and future), number (singular, plural), gender/class (masculine, feminine, neuter), person (first, second, third), etc. It deals with the combination of stems with grammatical markers of suffixes such as –s, -ess, -ed, -ing in English for example. Generally inflectional morphology is very productive as all nouns have singular/plural distinctions; most verbs have tense distinctions, etc. More examples of inflectional morphology are as given below:

Male	**Female**	**Present**	**Past**	**Singular**	**Plural**	**1st/ 2nd person**	**3rd person**
actor	*actress	help	helped	boy	boys	clean	cleans

On the other hand, derivational morphology results in changing classes of words. As shown in the following example, a noun or an adjective may be derived from a verb.
Examples:

sing-er…….singer (verb to noun)

vaccine-ate…..vaccinate (noun to verb)

Similarly, the morphological analyzer for Afaan Oromoo can divide the word within a text into different components (prefix, stem, and suffix) and assigns different values for each component. For example, the word "***beeke***" (=he knows) is divided into "***beek***"(=to know) and "***-e***"(past tense indicator). It also provides the category of the word (as 3m.sg. = third person, masculine, singular).

Several criteria have put foreword in the literature in order to justify the division of morphology in to inflection and derivation. However, some studies deny the theoretical significance of the distinction and still other contributors argue that the separation of

* '⟶' sign means becomes

derivation from inflection, although useful from the description view- point, is not easily enforced. For instance, according to these studies there is no sharp contrast between the mechanism of both types, but rather a continuum ranging from phenomena which are manifestly inflectional to those ones which are undoubtedly derivational (Szymanek, 1989).

Kazakor and Manadhar (2002) broadly categorize approaches of computational morphology in to ***rule-based*** *and* ***corpus-based.*** A rule-based is based on theory of morphology laid down by an expert. This approach enables to incorporate sophisticated linguistic theories such as generative phonology in to computational morphology process. On the other hand, corpus-based approach does not strictly follow explicit theory of linguistics. It uses some algorithm to learn, say about the morphological segmentation of a language, from an input data (corpus). The knowledge acquired is then used to perform the morphological analysis task.

Belonging to the Afro Asiatic languages family, Afaan Oromoo (the Oromoo language) is one of the most widely used language in Africa. It is the third largest language in number of speakers, surpassed only by Arabic and Hausa. There are close to thirty million people who speak Afaan Oromoo mainly in Ethiopia, and the rest in Kenya, parts of Somalia and Tanzania (Muudee, 1995).

Besides having a large number of speakers the language is now serving as a working language in the local administration and courts in the Regional Government of Oromia (RGO). As a medium of instruction, the language is also used in primary and secondary schools, and training institutes and colleges in Oromia regional State. Furthermore, courses in Afaan Oromoo are offered at a BA level in two local colleges as a minor. Recently an MA program has also been launched at the Addis Ababa University.

More or less documents have also been produced and distributed in different formats. The documents are from all parts of knowledge (discipline), which include textbooks, reference books, journals and periodicals, fictions, government publications, etc. They

are mostly on language, literature, history, natural science, political, social and economic issues. Majority of them are available in hard copies while some of them are in electronic formats. The use of computers and the Internet are common in offices, private sectors and commercial enterprises in Oromia Regional State. Diversified and heterogeneous electronic information in Afaan Oromoo is also available through the Internet to be accessed by different category of user groups.

It is repeatedly said by different scholars that the main objective of information retrieval system is to find the most 'relevant' materials for the user query while reducing the least relevant ones. For instance, according to Salton (1983), the task of information retrieval system involves in addition to acquiring and representing, the retrieval of relevant documents and at the same time avoiding (excluding) the non-relevant ones from the relevant. The relevant documents might fully or partially matches with the users query depending on the request of each user need.

The capability of a retrieval system to retrieve relevant documents is partly dependent on a person's implicit knowledge of the rules of languages that make the production and understanding of an indefinitely large number of new utterances and the actual use of language in real situation linguistic knowledge (Katamba, 1993).

Thus, to generate or retrieve meaningful and accurate as well as relevant item of individual interest, both sentences and words should be kept appropriately following understandable and standard way that the speakers of the language understand easily. One job of the syntactic component of the grammar was thought as being to generate (i.e. to specify or enumerate explicitly) the constituents of words used in user's query and/or information items (documents). The output of the syntactic component (morphological analyzer) will later be used for searching and matching to broaden and narrow search terms to allow retrieval of more or less (specific) of the information items.

1.3. The Statement of the problem and its justification

The major development of Afaan Oromoo is in its linguistics aspects. The computational development of the language, especially in storage, processing and retrieval is at its infant stage. Some initiatives such as Oromoosoft (developed abroad) are at their rudimentary stage and yet not in standardized form.

Afaan Oromoo, which is widely used in Ethiopia, should normally have good and solid works on its lexicon and grammar ready to be used by anyone interested in learning the language, or those interested in researching about it. As to natural language processing of Afaan Oromoo, there are only few works by graduates of School of Information Studies for Africa (SISA). These include: Development of stemming algorithm for Afaan Oromoo text (Wakshum, 2000); Text to Speech for Afaan Oromoo, (Morka, 2001); and Sentence Parser of Afaan Oromoo texts (Diriba, 2002). Therefore, it is possible to say only a little has been done on developing a systematic representation of words in the language in facilitating further information storage and retrieval activities.

As pointed out by researchers listed above, their respective work is limited, being the starting material for those interested to extend the research to be used as a tool to the information retrieval environment for Afaan Oromoo. For instance, if we consider the work of Wakshum (2000), even if they are many in number in the language the stemming algorithm does not conflate compounds and some variants that have irregular patterns of word formation. The data utilized for the study is not and cannot be exhaustive; the word distribution needs more elaboration of the language's behavior, etc. All the three conductors of thesis research mentioned above did not hesitate to inform that their work is not exhaustive. They also recommended that the area of natural language processing for Afaan Oromoo is waiting for more researchers.

The grammar of Afaan Oromoo can take many forms. Since a language possesses an infinite number of words and sentences, one way of specifying such a language is by writing a program which reads a series of words, segment them into parts, analyze and

then generate an output. Since it is a highly inflected language, its morphological processing is difficult without having a robust morphological analyzer.

The absence of natural language understanding tools of Afaan Oromoo such as POS tagger and morphological analyzer as well as lack of online translating machine which can translate text from any other language to Afaan Oromoo and/or vise versa is becoming a bottleneck for those who are unable to read and write using other languages in which huge information is disseminated via the Internet. This problem needs localization and standardization. From the above discussion it is obvious that the absence of a morphological analyzer for the language will have an effect on researchers in the area of machine translation, spell-check, dictionary compilation, parts of speech tagging, automatic sentence construction, etc. Therefore, the need for NLP systems such as morphological analyzer is unquestionable for its automatic nature as well as a number of other relevant problems related in computational area.

The morphological analyzer that works for one language might not work for the other because of the diverse nature of the languages. For example the following are some of the attempts made in the area of Amharic language processing: Recognition of formatted Amharic text using optical character recognition (OCR) techniques by Ermias, 1998; Design and development of Amharic word parser by Abyot, 2000; Automatic part of speech tagging for Amharic Language: an experiment using Stochastic Hidden Markov (HMM) approach by Mesfin, 2001; Automatic Morphological Analyzer for Amharic :an Experiment Employing Unsupervised Learning and Autosegmental Analysis Approaches, by Tesfaye 2002; and Design and Development of Automatic Morphological Synthesizer for Amharic Perfective Verbs by Kibur, 2002. All these could not analyze Afaan Oromoo and solve the problem of users of the language regarding word formation.

Constructing a fluent and robust natural language understanding interface at once is a difficult and complex task. It needs step by step attempt and the contribution of researchers from Linguistics and Information Science fields. This helps to fill the gap by constructing simpler natural language understanding systems for Afaan Oromoo text.

As far as the researcher's knowledge is concerned, except few attempts discussed above, there is no work on morphological analyzer of Afaan Oromoo text. Therefore, this study is aiming at developing morphological analyzer for Afaan Oromoo text which can be used as an input for different purposes such as, spell-checking, dictionary compilation, parts of speech tagging, automatic sentence parsing and construction, etc.

1.4. Objectives of the study

1.4.1. General Objective

The general objective of this study is to develop a morphological analyzer for Afaan Oromoo text.

1.4.2. Specific Objectives

The following specific objectives are set to achieve the above general objective:

- review on the nature and word characteristics of Afaan Oromoo letters, words and word formation
- examine the lexicon of Afaan Oromoo text
- study the pattern of characters in words of Afaan Oromoo
- prepare a corpus that is used in developing morphological analyzer of Afaan Oromoo
- develop a gold-standard for Afaan Oromoo words
- develop an algorithm to design morphological analyzer for Afaan Oromoo text
- develop a prototype of morphological analyzer for the language
- test the developed prototype using variety of data for its effectiveness.

1.5. Methods

The following subsections briefly review the methodologies followed in this study. It includes literature review, discussions made with people, preprocessed data, tools and techniques and the experiment.

1.5.1. Literature review

Developing a morphological analyzer for Afaan Oromoo text needs good background knowledge of the language. Furthermore, the rules and principles that govern the language is studied in the area of phonology, lexical, morphology and parts of speech in language in general and Afaan Oromoo in particular. In order to perform these activities the output of other researches were reviewed. The sources include reference and text books, journals, research reports and materials on the Internet.

1.5.2. Data collection by discussion

In addition to review of literature, consultation and discussion with linguists have been made in order to share their experiences concerning words, characteristics of morphemes, and generally the nature of the language. Experts and professionals in the area of Information Science who have related experience and relevant background of the problem at hand were consulted throughout the research work.

1.5.3. Preprocessed data

The corpus used in this study is collected from three different institutions. Corpus of size 9,642 words was collected from Oromia Regional State Cultural Bureau to assist in speeding up the process. Teaching materials of morphology courses at Addis Ababa University, Institute of Language Studies were also used to get some manually processed data. The size of this corpus is about 6,977 words. 33,035 words are also used from Oromia Regional Education Bureau text books and different documents. Sample of texts are taken from different disciplines randomly to represent the language.

The corpus used in this study is collected from three different institutions. One of this is the Oromia Regional Cultural Bureau, which is currently working on standardization of the language to bring different dialects into common. Since Afaan Oromoo has short written history, consisting of many spoken words which are not in standardized form. The Education Bureau of Oromia Regional State, which is responsible for designing the curriculum, produce text books and teaching materials for various school levels has a number of texts and documents written in Afaan Oromoo, and it is the second source of corpus for this study. Teaching material of morphology course in the Institute of Language Studies of the Addis Ababa University is also used. Three sample texts from these institutions were selected from different areas of knowledge for this study.

1.5.4. Tools and techniques

Linguistica_beta2 and C^{++} are used for designing the prototype. *Linguistica* is a program which can be used to explore the *unsupervised learning* of natural language, with primary focus on morphology, which is to say, word-structure. It runs under Windows, and is written in C^{++}. It is used to decompose words into different components. C++ programming language is used for preprocessing purpose i.e., avoiding unnecessary characters (such as, ; " !"), and it is the program that I am comfortable with. Based on the investigation and analysis of the language using the above mentioned tools, morphological analyzer of Afaan Oromoo text was developed. The interface of the prototype is designed using VB.

1.5.5. The experiment

The data was categorized according to their relationship for further processing so as to develop the algorithm and gold-standard for the language based on the nature of the language's morphology.

The performance of the analyzer is measured by comparing the outputs of the analyzer with the result of the gold-standard which is designed for this language. This helps to see

how much the words are correctly segmented and each segment is given a correct value. As per the result of the experiment on the selected texts correlations were made on the values of parameters of Linguistica_beta2 repeatedly till no improvement achieved.

1.6. Application of the results

Morphological analysis, from the linguistics point of view, is used to know the meaning of the word and identify parts of speech that the word belongs to. According to Allen (1995), morphological analyzer is useful to compute the part of speech and inflectional categories of any word. This is to say, as some words are formed from another word forms by derivation, such a system is useful to get the common stem for derived words.

Morphological analysis is helpful as one of the component for developing electronic dictionary, spelling correction, grammar checking, thesaurus developing, e.t.c. Previous researches also show that morphological analysis can be used as a subcomponent in identification of content bearing terms and nouns in most of the systems, as well as for automatic abstract generation and sentence selections (Tesfaye, 2002).

Furthermore, the current tradition of using high speed electronic processing machine, the computer, everywhere (at home, offices, business centers, libraries, schools, health institutes, etc) may also demand for additional facilities to perform applications (for example word processing) effectively.

Furthermore, the corpus data gathered for this research, the gold-standard and the output of this research could also serve as an input for future research in the area. It can also be used for teaching learning process of Afaan Oromoo at higher institutions. Cultural Bureau of Oromia Regional State can also use the analyzer for standardization of the language. Thus, the primary beneficiaries of this study include researchers, Cultural Bureaus, teachers and students who have interest in the area of Afaan Oromoo language processing.

1.7. Scope and limitation of the study

The morphological analyzer developed for Afaan Oromoo analyzes the inflectional and derivational variants of Afaan Oromoo word structure like prefix, stem/root, and/or suffix. Because of the limitation of the tool used, it was not possible to identify categories of word classes in the language.

Most morphological analyzers need information about morphemes from morphologically preprocessed electronic data and a gold-standard of big size. Time constraint and the shortage of morphologically preprocessed data did not allow me to develop a gold-standard with more than 1600 words.

1.8.Organization of the study

The study is organized in six chapters. Chapter one introduces basic concepts of natural language processing, different types of morphological analyzes, the statement of the problem, objectives of the study, methods employed, application areas and scope & limitations of the study. The concept of morphological analysis is discussed in chapter two. This chapter consists of levels of language processing and the place of morphology in the grammar and relevant issues. Chapter three of this study deals with the structure and property of Afaan Oromoo morphology. It mainly focuses on words and word formation in Afaan Oromoo and the lexical categories of the language. The morphological process involved in the language is also given a subsection.

Chapter four deals with the development of the algorithms for the system. The experimentation, evaluation and testing procedures were presented in chapter five while chapter six concludes the study by presenting conclusions and recommendations.

CHAPTER TWO

MORPHOLOGICAL ANALYSIS

2.1. Introduction

The general concept of morphological analysis is discussed in this chapter. Its sub sections deal mainly with the levels of language processing and the place of morphology in the grammar, words and word formation, morphological analysis and other relevant issues.

2.2. Levels of language processing

A natural language system must use considerable knowledge about the structure of the language itself, including what the words are, how words combine to form sentences, what the words mean, how word meanings contribute to sentence meanings, and so on (Allen, 1995). To answer questions or to participate in a conversation, a person not only must know a lot about the structure of the language being used, but also must know the word in general and the conversational setting in particular.

There are different levels of language processing. As described by Salton (1983) the phonological, morphological and lexical levels as indicated below.

- The **phonological level** deals with the treatment of speech sounds as needed, for example, for the handling of speech understanding or speech generation systems.

- The **morphological level** of linguistic processing is concerned with the processing of individual word forms and of recognizable portions of words. The

recognition and removal of words suffixes and prefixes and the generation of word stems are based on morphological knowledge.

- The **lexical level** deals with the procedures operating on full words. In information retrieval this covers operations such as common words deletion, dictionary processing of individual words, and the replacement of words of thesaurus classes.

  There are two categories of lexical items: OPEN categories and CLOSED categories. ***Open*** categories represent the primary function of a word: nominal action, nominal modifier, and action modifier. The vast majority of words in the vocabulary fall into these categories which essentially correspond to the standard word classes: noun, verb, adjective and adverb. Words in open category are also called content words

  The ***closed*** categories, on the other hand, represent those words which include a finite, restricted number of instances; for example the class containing conjunctions or determiners in a language can not be extended. Words in this category are called function words (Abiot, 2000).

- The lexicon of natural language processing should provide the following three levels: the syntactic, semantic and pragmatic information.

  i. The **syntactic** representation of language are based on the notion of context free grammars, which represent sentence structure in terms of what phrases are subparts of other phrases. It indicates how words in the sentence are related to each other or how words are grouped together into phrases, what words modify what other words, and what words are of central importance in the sentence. In addition, this structure may identify the types of relationships that exist between phrases and can store other information about

the particular sentence structure that may be needed for later processing.

The syntactic level is designed to group the words of sentence into structural units such as prepositional phrases, and subject-verb-object groupings that correctly represent the grammatical structure of the sentence. A syntactic analysis is normally based on the surrounding structure in which the individual words are embedded in a sentence and on the use of syntactic features characterizing the individual words. Function words make up the skeleton of the sentence. They carry a great deal of information about the syntactic structure of the sentence. The part of speech category of a function word proved useful to distinguish more categories than the standard dictionary categories. The majority of content words not listed in the lexicon can be tagged using morphological information about suffixes.

ii. The **semantic level** adds contextual knowledge to the purely syntactic process in order to restructure the text in to units that represent the actual meaning of a text. Semantic interpretation of a language is a task of considering what combination of the individual word meanings can combine to create coherent sentence meanings. Exploiting such interconnections between word meanings can greatly reduce the number of possible word senses for each word in a given sentence (Allen, 1995).

iii. The **pragmatic level** according Salton (1983) uses additional information about the social environment in which a given document exists, about the relationships that normally prevail in the world between various entities, and about the world-at-large to help in the text interpretation. Pragmatic knowledge concerns how

sentences are used in different situations and how use affects the interpretation of the sentence.

2.3. Words

The definition of the word is as broad as possible. In different language there are sound complexes that are words from some angles but not from others. Loosely a word is defined by Huang (2001) as a lexical item, with an agreed-upon meaning in a given speech community that has the freedom of syntactic combination allowed by its type (noun, verb, etc.).

Words are the fundamental building block of language. Every human language, spoken, signed or written is composed of words. Every area of speech and language processing, from speech recognition to machine translation and information retrieval on the web, requires extensive knowledge about words (Jurafsky and Martion, 2000).

Every language enables its speakers to communicate each other. Human being can exchange ideas by using string of words arranged in specified manner. There are a number of points of view from which words can be defined one can attempt to characterize words as phonological units; as the irreducible terminal elements of the lexicon, and from a variety of other perspectives.

A "word" as a unit of language can be defined as "A single unit of language which means something and can be spoken or written". A word is an arbitrary symbol representing a physical object or a concept. Elaborating the meaning of words further, Abiyot (2000) says, the grammar of language dictates the way words are arranged and combined during communication. String of words arranged based on the grammar of the language will form the meaning of the individual words and the way they are arranged.

Kramsky (1969) also defines "word" broadly as "the smallest independent unit of language referring to certain extra linguistic reality or to a relation of such realities and characterized by certain formal features (acoustic, morphemic) either actually (as a unit of a lexical plan)".

According to Kramsky (1969), every word has two meanings: *actual* and *potential*. The *actual* meaning is the meaning of the word in a given context and the *potential* meaning is the meaning of the word as a lexical unit. In a lexical unit all topical meanings are potentially included; however none of these meanings are entirely complete without the contextual situation. It is the contextual situation by which they are fully determined. Paradigmatic properties of words includes part-of-speech, inflectional and derivational morphology, and compound structure.

2.3.1. Lexical part-of-speech

Lexical part-of-speech (POS) according Huang (2001) is a primitive form of linguistic theory that posits a restricted inventory of word type categories, which capture generalizations of word forms and distributions.

Assignment of a given POS specification to a word is a way of summarizing certain facts about its potential for syntagmatic combination. Additionally, paradigms of word-formation processes are often similar with in POS types and subtypes as well. The word properties upon which POS category assignments are based may include affixation behavior, historical development, productivity and generalizability, and others (Huang, 2001).

A typical set of POS categories would include noun, verb, adjective, adverb, interjection, determiner, preposition, and pronoun. POS tagging is the process of assigning a part-of-speech or other lexical class marker to each word in a corpus. There are many algorithms, for example, rule-based methods, hidden Markov models, and machine-learning methods, to automatically tag input sentences.

2.3.2. Morphology

Morphology is about the subpart of words, i.e., the patterns of word formation including inflection, derivation, which also subsumes compound word formation. Languages like English uses prefixes and suffixes to express inflection and derivational morphology.

2.3.2.1. Inflectional morphology

It deals with variations in word form that reflect the contextual situation of a word in phrase or sentence syntax, and that rarely have direct effect on interpretation of the fundamental meaning expressed by the word.
English inflectional morphology is relatively simple and includes person and number agreement and tense markings only. The variation in *cats* (vs. *cat*) is an example. The plural form is used to refer to an indefinite number of cats greater than one, depending on a particular situation. But the basic POS category (*noun*) and the basic meaning (***felis domesticus***) are not substantially affected.

Words related to a common lemma via inflectional morphology are said to belong to a common paradigm, with a single POS category assignment. In English, common paradigm types include the verbal set of affixes such as: *-s, -ed, -ing*; the noun set: *-s*; and the adjectival *–er, -est*.

Certain paradigms may consist of highly idiosyncratic irregular variation as well, e.g. *go, going, went, gone or child, children.* Furthermore, some words may belong to defective paradigms, where only the singular (noun: equipment) or the plural (noun: scissors) is provided for.

2.3.2.2. Derivational morphology

In derivational morphology, a given root word may serve as the source for wholly new words, often with POS changes. Table 2.1 illustrates some examples of stems and their related forms across POS categories.

Table 2.1: Some examples of stems and their related forms across POS categories

Stem	Noun	Verb	Adjective	Adverb
critic	*criticism*	*criticize*	*critical*	*critically*
fool	*fool*	*fool*	*foolish*	*foolishly*
industry	*industry, industrialization*	*industrialize*	*industrial, industrious*	*industriously*
employ	*employ employee employer*	*employ*	*employable*	*employably*
certify	*certification*	*certify*	*certifiable*	*certifiably*

Traditionally, linguists classify words into different categories based on their uses. According Allen (1995), two related areas of evidence are used to divide words into categories. The first area concerns the word's contribution to the meaning of the phrase that contains it, and the second area concerns the actual syntactic structures in which the word may play a role.

2.3.2.3. The place of morphology in the grammar

The fundamental question "where's morphology?" has met with a variety of divergent responses. The contention shared by most early generative theorists was that morphology was not a separate component of the grammar but rather formed in integral part of syntax. Accordingly, attempts were made to account for the derivation of complex words by

means of transformations. This view has come to be known as the transformation list approach (Szymanek, 1989).

The other approach according to (Szymanek, 1989) was the lexicalist approach. Special word formation rules assumed the role previously played by transformations in deriving complex words. A common feature of this approach is the emphasis on the formal links between morphology and phonology.

2.3.2.4. Stem and base

All affixes according to Szymanek (1989) may be divided into inflectional and derivational ones. A STEM "is the part of the word-form which remains when all inflectional affixes have been removed"; a polish participle like *placz(acego)* 'crying, gen. sg. masc.' where *placz-* is a verbal stem while *–ac* and *–ego* are inflectional morphemes (affixes).

A stem may be morphologically complex when, for instance, a given word-form represents a derived lexeme; cf. *placz-liw(y)* 'tearful'. A stem is complex also in the case of so-called "stem-forming" morphemes well- known form polish verb morphology; cf. the elements *–a-, -i-, -e-*, etc. in *plak-a-(c)* 'cry', *pal-i-(c)* ' burn', *krzycz-e-(c)* 'shout' (Szymanek, 1989).

The term BASE on the other hand is chiefly used in the derivational morphology to denote a lexeme (or, sometimes, a morpheme) from which another complex morpheme is formed. By extension, one can also use the term with reference to the class of the lexemes which undergo a particular word-forming process, e.g., the set of all nouns which may derive diminutive counterparts in polish (or just those nouns which derive diminutive in *–ik/-yk*).

2.3.2.5. Affix positions

Positional analysis (Grimis, 1983) is a way to determine the linear ordering of affix systems. In complex word structure, certain set of affixes conventionally come in a particular sequence before or after other affixes. He further elaborates the way an ordering scheme is discussed as follows:

In a language that has many prefixes, certain occur; they are the first order prefixes. If any first-order prefixes are present in a form, second-order prefixes come immediately before them; otherwise they come before the stem. There may be several relative orders of prefixes that can be determined on this basis.

Suffixes work the same way: first-order suffixes has always come immediately after the stem, and second-order suffixes after first order suffixes if there are any, otherwise after the stem, and so on.

2.3.2.6. Complications in affix positions

In both pre-affixation and suffixation, the order classes are relative, because in a given word there might be no reason for affixes of certain orders to be present. This makes it possible for, say, a fifth-order suffix to come right after the stem, and still be in the fifth relative order. It does not imply that a fifth-order suffix always has four other suffixes separating it from the stem. Nor does it imply that a fifth order suffix in a word that has only two other suffixes between it and the stem thereby slips into the third order. It implies only that if any lower-ordered suffixes are present, it will always come between them and the stem. The same is true for prefixes (Grimis, 1983).

Some affix systems have spanned orders. For example, a set of prefixes may behave like second order prefixes in that they can be separated from the stem and by first-order prefixes if at all, but at the same time the only things that can precede them may be, say,

prefixes that can be shown on other grounds to be sixth-order prefixes or higher. The order of these prefixes, then, is neither 2 nor 5, but the entire span 2-5, as with *sa-* and *za-* in Table 2.2 below. They occur before first-order *pa-* and *ta-*, but are preceded only by sixth order *ma-* and *na-*, and they can not occur with any of the ones in between.

Table2.2: Prefix systems of six order with spanning

<table>
<tr><td>ORDER</td><td>6</td><td>5</td><td>4</td><td>3</td><td>2</td><td>1</td><td>0</td></tr>
<tr><td rowspan="2">PREFIXES</td><td>ma-</td><td>ba-</td><td>da-</td><td>ga-</td><td>ka-</td><td>pa-</td><td rowspan="2">STEM</td></tr>
<tr><td>na-</td><td colspan="4">sa-
za-</td><td>ta-</td></tr>
</table>

The three complications that arise in performing positional analysis on affix systems according Grimis(1983), are:

- The complication that comes from alternate ordering, where the same affixes may occur in more than one order
- The complication that comes from layering, where there is a complex word structure in which one word is used as the core of another word form.
- The third complication is the complication that comes from cyclic data.
- Such a cycle inevitably tips up the positional analysis.

2.3.3. Relation of designation with meanings

There are languages which show great semantic differentiated semantically. For example in English every word, particularly verbs and nouns, has a number of meanings. It is possible to mention the abundance of meanings of such verbs as to *set, put, do, make*, etc.

One can distinguish between relation of one meaning with more designations and relations of one designation with more meanings. The first case according to Kramsky(1969), is that of synonymy and the second is that of polysemy and homonymy.

Words are made up on the one hand of sounds, on the other hand of meanings, and the relation they establish between sound and meaning essentially constitutes them. They have their own structural properties. In many cases according to Anderson (1992), they have a complex internal constituent structure. For example, for words like '*discontentedness*' there is more to be said than is made explicit. We might explicate its meaning as something like "the state of being discontented". This involves appeal to the meaning of discontented, which we could then explicate in its turn as something like "characterized by notable discontent (N)." Again, the meaning involves reference to that of another form, the Noun discontent, whose meaning is something like" the apposite of content. This noun, in turn, should probably be regarded as based on the adjective content "satisfied". He concluded by informing that, as there are several layers of reference to meaning, rather than a single homogenous association of single form with some semantic content.

The structuralist linguists in dealing with the analysis of morphology had already a model for their study: that of the 'phoneme' a minimal distinctive element of sound structure, as they understood it. In the process of interpreting the raw phonetic material of speech, the first step is to 'break the phonemic code': that is, to identify each of the phonemes, the distinctive element of sound structure, which are represented in succession by phonetic segments.

This phonemic model can be directly extended to accommodate morphological analysis, this morpheme of a language can be treated as distinctive elements of word structure, and regard them as related to their concrete phonemic instantiations in the some way as phonemes are related to phonetic segments.

Recent linguistic work has brought out the inherent and infectious ambiguity of the term word. Any language user according (Szymanek, 1989) is equally accustomed to thinking of words as they appear in any conventional dictionary. In order to get rid of confusing ambiguity of the term "word", modern linguists has introduced a sharp distinctive between word forms and lexemes.

2.3.5. The morpheme

A definition of the morpheme which has become particularly influential in modern linguistics is the one offered by Bloomfield (1933). He defined morpheme as "*a linguistic form which bears no partial semantic resemblance to any other form, is a simple form or morpheme*". According this and other authors the morpheme is the smallest unit of meaning that a word can be divided into: for example the word 'like' contains one morpheme, but '*unlikely*' contains three (*un-, like-, and -ly*).

It can be seen that a given word may be composed of several morphemes. The structure of such a complex form normally is reflection of the linear order in which individual morphemes are put together or added to another complex form, schematically: [[[*mean*]*ing*]*ful*]. In this example the innermost morpheme '*mean*' is termed a root, since it is not further analyzable into meaningful elements. The structurally more peripheral morphemes '*-ing*' and '*-ful*' are called affixes. Affixes are instances of bound morphemes i.e. morphemes which can never occur in isolation, so as to form an independent word. Here, the root '*mean*' may also serve as an instance of a free morpheme, since it can function as an independent word. This demonstrates as well that some words are morphologically simplex i.e. monomorphemic; e.g. mean, the, element (Szymanek: 1989).

According Anderson (1992), a consensus emerged on the substantive parallels between the role of morphemes in word structure and that of phonemes in sound structure. Just as utterances could be regarded as built by concatenating the atoms of sound structure, so words were regarded as forms be the concatenation of morphological atoms, or morphemes. The basic properties of this classical morpheme were the following:

- Morphemes are homogenous and indivisible atomic unit of linguistic form.
- Each morpheme in a given word is phonologically represented by exactly one morph, and each morph represents exactly one morpheme.

- The morphs themselves are consistently and uniquely (though not necessarily biuniquely) related to surface phoneme form.
- The morphemes are arranged into a structure of Immediate-Constituents which yields a sort of phrase maker as the analysis of words internal structure.
- Words are exhaustively composed of morphemes.

The morphology, on this account, is a set of statements about how these abstract elements are distributed with respect to one another and organized into Immediate-Constituents (the morphotactics); and about how each is realized, in terms of its morphological and /or phonological environment.

2.4. Morphological Analysis

Types of morphological processes found in human language vary. According to Anderson (1994) morphological processes are traditionally divided into two types: *inflection* and *derivation*. While this distinction remains controversial among linguists, this study considers the widely help view that derivational produces new words while inflectional produces forms of the same word. Thus the words *compute, computer, computerize, recompute, recomputerize, computerization* etc. are related by derivation since they are different words (lexemes), though based on the same root *compute*. Derivational affixes often change the part-of-speech of a word (compute is a verb, computer is a noun). They also may add semantic content (the prefix re- in recomputed means "again"). In contrast, the words *compute, computes, computing,* and *computed* are related by inflection since they are all forms of the same word (lexeme) compute. Inflection typically encodes categories such as number, tense, gender, and case which are relevant to syntax. Anderson (1994) concludes "'inflection' thus seems to be just the morphology that is accessible to and/or manipulated by rules of the syntax".

> "Parsing is standard technique used in the field of natural language processing. When we think of parsing, we likely think of syntactic parsing. But before syntactic parser can parse a sentence it must be supplied with

information about each word in the sentence. For instance to parse the sentence ***'the cat chased the rat'***, the parser must know that *cat* is a singular noun, *chased* is a past tense verb, and so on. In English such information may be supplied by a lexicon that simply lists all word forms with their part of speech and inflectional information such as tense and number. Of forms that must be listed in such as *cat* have only two inflected forms, singular and plural, and regular verbs such as *chase* have only four inflected forms: the base form, the *–s* form, the *–ed* form, and the *–ing* form. But an exhaustive lexical listing is simply not feasible for many other languages, which may have hundreds of inflected forms for each noun or verb. For these languages such as Finnish, Turkish, etc. one must build a word parser that will use the morphological system of the language to compute the part of speech and inflectional categories of any word" (Antworth, 1994).

The goal of the automatic morphological analysis is to perform automatically a morphological classification of an arbitrary word form. This includes identifying the base form of the word, its grammatical features and to which inflectional type (part of speech) it belongs (Krushkov). The decomposition process of analyzing a surface component of morphemes, which are the minimal meaningful element of words, such as prefixes, suffixes, and stem words themselves according to (Huang, 2001) is also referred as morphological analysis.

When a dictionary does not list a given orthographic form (all words and abbreviations arranged alphabetically in a lexicon) explicitly, it is sometimes possible to analyze the new word in terms of shorter forms already present. These shorter forms may combine as prefixes, one or more stem or roots, and suffixes to generate new forms.

The prefixes and suffixes are generally considered bound, in the sense that they can not stand alone but must combine with a stem. A stem, however, can stand alone. A word such as *establishment* may be decomposed into a "stem" *establish* and a suffix *–ment*. Thus, the morphological analyzer attempt to cover an input word in terms of the affixes and stems listed in the morphological lexicon. The covering(s) proposed must be legal sequence of forms, so that often a word grammar is supplied to express the allowable patterns of combinations.

A morphological analysis system might be as simple as a set of suffix-stripping rules for English. If the word can not be found in the lexicon, a suffix stripping rule can be applied to first strip out the possible suffix, including *–s, -'s, -ing, -ed, -est, -ment,* etc. if the stripped form can be found in the lexicon, a morphological decomposition is attained. Similarly, prefix stripping rules can be applied to find prefix-stem decomposition for prefixes like *in-, un-, non-, pre-, sub-*, etc. although in general prefix stripping is less reliable (Huang, 2001).

Suffix and prefix stripping gives an analysis for many common inflected and some derived words such as *helped, cats, establishment, unsafe, predetermine, subword*, etc. It helps in saving system storage. However according Huang (2001), it does not account for compounding, issues of legality of sequence (word grammar), or spelling changes. There are different types of morphological analyzers reported in different literature. The reason for these diversity is that different languages have different morphological structure; the methods perfectly suitable for morphological poor language (like English) or agglutinative languages (like Finnish) are not the best ones for inflective languages (like Spanish) (Gelbukh and Sidorov, 2002).

The methods for automatic morphological analysis can be classified into dictionary-based and heuristic-based ones. The former one use a stem dictionary to guarantee the correct results for the words stored in the dictionary. The latter ones, which is the focus of this study, use heuristic rules to guess the result for previously unseen words, which is important since new words constantly appear in the language, not mentioning that no dictionary can be complete (Gelbukh and Sidorov, 2002). Both methods are as discussed below focusing on the second ones.

2.4.1. The Kimmo Parser

One of the most famous model for morphological analysis according Gelbukh and Sidorov (2002), is the two-level model (KIMMO) suggested by Koskenniemi (1983). PC-

KIMMO is an implementation for microcomputers of a program dubbed KIMMO after its inventor Kimmo Koskenniemi. It is of interest to computational linguists, descriptive linguists, and those developing natural language processing systems. The program was designed to generate/produce and/or recognize (parse) words using two-level model of word structure in which a word is represented as a correspondence between its lexical level form and its surface level form. A PC-KIMMO description of a language consists of two files;

1. *a rule file*: which specifies the alphabet and the phonological (spelling) rule, and
2. *a lexicon file*: which lists lexical items (words and morphemes) and their glosses and encodes morphotactic constraints.

The two functional components of PC-KIMMO are the *generator* and the *recognizer*. The *generator* accepts as input a lexical form, applies the phonological rules, and returns the corresponding surface form. It does not use the lexicon.

The recognizer accepts as input surface form, applies the phonological rules, consults the lexicon, and returns the corresponding lexical form with its gloss. Main components of PC-KIMMO are given in figure 2.1 below.

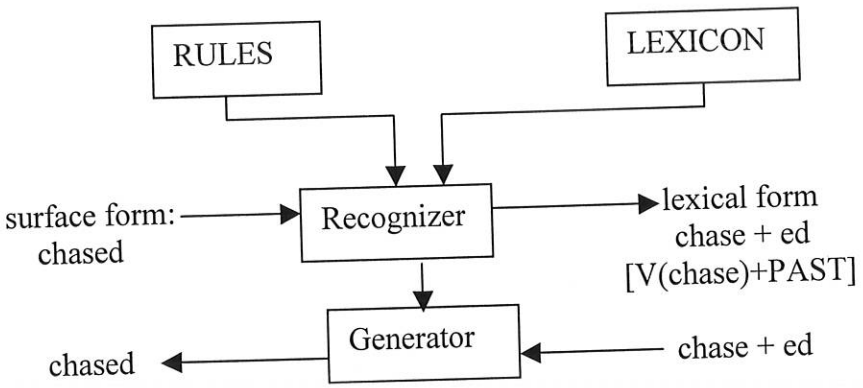


Fig 1.2: Main components of PC-KIMMO

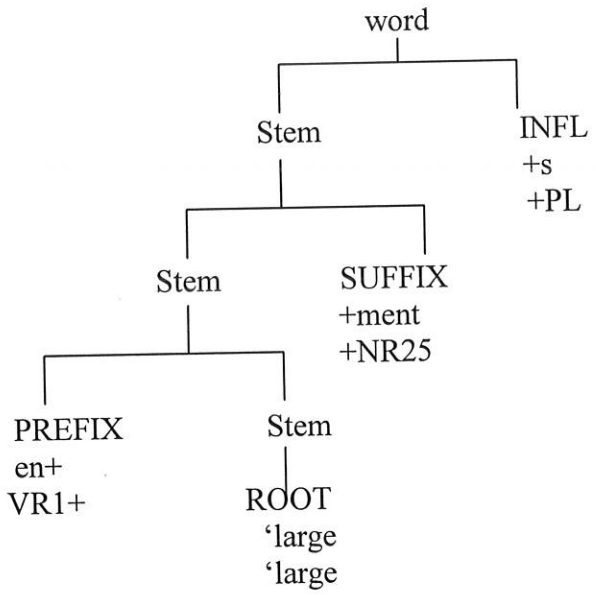
The rule and lexicon are implemented computationally using finite state machines. Till 1990 PC-KIMMO was quite good at what it was designed to do-tokenize a word into a sequence of tagged morphemes. But it had a serious deficiency: it could not directly determine the part of speech of a word or its inflectional categories. For example, given the word *enlargements*, the system could tokenize it into the sequence of morphemes *en+large+ment+s* and gloss each morpheme, but it could not determine that the entire word was a plural noun. This meant that it was nor adequate to act as a morphological front end to a syntactic parser- its most desirable application. In 1993 by adding a third analytical component, a *word grammar*, the deficiency was corrected.

2.4.1.1. Unification-based word grammar

The word grammar is a unification based chart parser that provides parses trees and feature structures. Just as a sentence parser produce a parse tree with words as its leaf nodes; a word parser produces a parse tree with morphemes as its leaf nodes. When we parse a sentence, it is normally already tokenized into words (since we put white space between words); but when we parse a word we must first tokenize it into morphemes. This tokenization is done by the rules and lexicon. When a surface word is submitted to recognized, the rules and lexicon analyze the word into a sequence of morpheme structures. A morpheme structure consists of a lexical form, its gloss, its category, and its features. For example, the word '*enlargements*' is tokenized into this sequence of morpheme structures as shown below:

Form:	en+	'large	+ment	+s
Gloss:	VR1+	'large	+NR25	+PL
Category:	PREFIX	AJ	SUFFIX	INFL
Feature:	[form-pos:AJ Head:[pos:v]]	[head:[pos:AJ]]	[form-pos:v head:[pos:N]]	[form-pos:N head: pos:N [number:PL]

This analysis is then parsed to the word grammar, which returns the parse tree and feature structure shown in figure 2.2 below:



word:
[head: [POS: N number: PL]]

Fig. 2.2. An example of parse tree for a word 'enlargements'

2.4.1.2. The two-level model of morphology

Koskenniemi's model of two-level morphology was based on traditional distinction that linguists make between morphotactics, which enumerates the inventory of morphemes and specifies in what order they can occur and morphophonemics, which accounts for alternate form of "spellings" of morphemes according to the phonological context in which they occur. Koskonniemi's model is "two-level" in the sense that a word is represented as a direct, letter-for-letter correspondence between its lexical or underlying form and its surface form. For example, the word chased is given this two-level representation (where + is a morpheme boundary symbol and 0 is a null character) (Antworth, 1994).

Lexical form:	c h a s e + e d
Surface form:	c h a s 0 0 e d

2.4.2. Unsupervised learning of natural language morphology

So far we have seen that morphological analysis that uses information about each word in the sentence before decomposing a word into different components. This section focuses on brief description on algorithm for automatic and unsupervised morphological analysis of a corpus of natural language that determines the prefixes, suffixes and stems of the language in question and generates morphological dictionaries from a given corpus using an unsupervised learning approach. There are various types of such approaches. An algorithm that consists of two levels (heuristics and MDL(Minimum Description Length) which were employed by Goldsmith (2001b) and implemented and tested in Linguistica are presented here for they are more relevant to this study.

The heuristics process of morphology divide naturally into an initial bootstrapping heuristic that is able to determine a finite pass analysis of the words of the corpus into stem and affix, and a set of incremental heuristics, which modify the analysis, leaving the decision to the MDL components as to whether the modifications are worth maintaining or should be dropped.

2.4.2.1.Minimum Description Length (MDL) analysis

Minimum Description Length (MDL) analysis according Goldsmith (2001b) is a form of analysis rigorously based on information theory. The central idea is that, given a corpus, an MDL model defines a description length of the corpus, given a probabilistic model of the corpus: the description length is the sum of the most compact statement of the model expressible in some universal language of algorithms, plus the length of the optimal compression of the corpus, when the probabilistic model to compress the data is used. The length of the optimal compression of the corpus is the base 2 logarithm of the reciprocal of the probability assigned to the corpus by the model.

Description length (Corpus C, Model M) = length (M) – $\log_2$ prob(C|M) 1

MDL analysis processes that the morphology M which minimizes the objective function is the best morphology of the corpus. Intuitively, the first term (length of the model, in bits) expresses the conciseness of the morphology, giving us strong motivation to find the simplest morphology possible, while the second term expresses how well the model describes the corpus in question. The morphology M spreads probability mass over a wide universe of possible words (by assigning a probability to all possible words in the language and by being subject to the requirement that the probabilities sum to 1). Instead of maximizing the probability, it wishes to maximize the log probability, and by multiplying the log probability by -1, it arrives at a quantity of information theoretic bits which can then be added to the first term in (1) above; and again, by multiplying the second term by -1, it is reasonably possible to add the two terms and attempt to find the analysis which ***minimizes*** the sums of the two terms. Hence the term: ***minimum*** description length Goldsmith (2001b).

2.4.2.2. Calculating the length of morphology

Most of the information that composes the morphology of a language like English is to be found in phonological (or logographic) content of the morphemes, but some of the information is contained in the information regarding the ordering of possible morphemes in the language. All of this information is condensed into essentially three components of morphology: a list of stems, a list of affixes, and a list of signatures which are structures including which stems may appear with which affixes. An example of Afaan Oromoo signature can be as visualized in fig. 2.3 below.

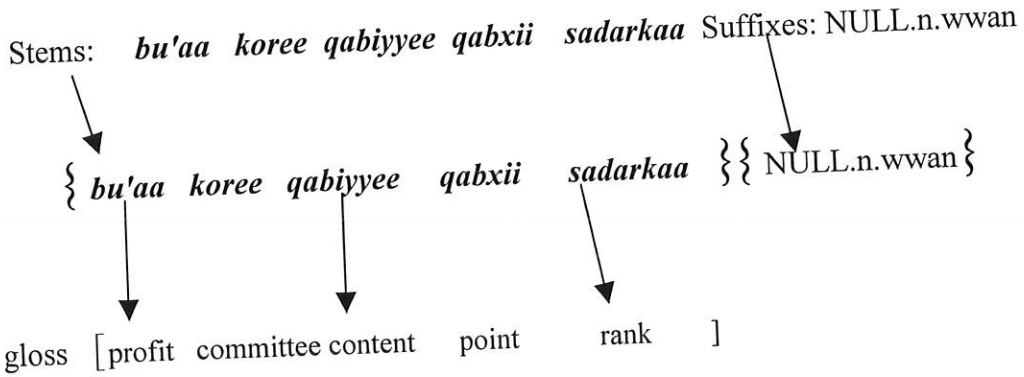


Fig. 2.3 An example of Afaan Oromoo signature

The morphology is constructed of pointers to more elementary units, and in the case of the above example the individual items in the stem list, the stem, are not themselves spelled out in the stem list, but are instead sequences of pointers to individual letters or phonemes. Then as indicated in section 2.4.2.1 above the probability of the corpus using the morphology is calculated. The general algorithm is discussed in chapter four of this study.

2.4.3. Word-based vs. morpheme-based morphology

According Arnoff (1976), morphology is based on (whole) words, rather than on morphemes, as the domain of the sign relation. Arnoff also noted that in the general case it is only at the level of whole words that form is associated with meaning. Another way of putting this is to observe that both the forms and the meanings of words are potentially internally divisible; but the relation between categories of meanings and aspects of form is often many-to-many rather than one-to-one.

3.2.1.1. Morphological changes

One of the main occurrences of morphophonemic changes is on verbs between the stem and derivational or inflectional suffixes. These produces the combinations of a variety of stem-final consonants followed by **t** (3sf, 2s and 2p), **n**(1p and nc) and **s**(causative-cs). In Table 3.2 below only stems ending in a vowel followed by a single consonant will be considered. The causative suffix has allomorphs conditioned by its phonological and syntactic environment. Intransitive verbs take *–s*; transitive verbs take *–sis* when the preceding syllable has a long vowel, and *–siis* otherwise (Stroomre, 1995).

Table 3.2: Main occurrences of morphophonemic changes on stems ending in a vowel followed by a single consonant (a-c)

a) Final consonant + t[+]

b + t = bd	arraabda	(arrab + ta)	"you lick"
d + t = dd	didda	(did + ta)	"you refuse"
dh + t = tt	haattii	(haadha + tii)	"mother + SUBJ"
f + t = ft	tufta	(tuf + ta)	"you spit"
g + t = dd	dhudda	(dhug + ta)	"you drink"
g + t = dd	midhaadduu	(midhaag + tuu)	"nice" (f)
k + t = tt	beetta	(beek + ta)	"you know"
l + t = lt	dhalte	(dhal +te)	"you gave birth"
m + t = mt	deemta	(deem + ta)	"you go"
n + t = nt	seenta	(seen + ta)	"you enter"
ph + t = pht'	supht'i	(suph + ti)	"she moulds"
r + t = rt	harta	(har + ta)	"you sweep"
s + t = ft	ijeefta	(ijees + ta)	"you kill"
t + t = tt	kutta	(kut + ta)	"you cut"

[+] In this data there are no examples of j + t and sh + t

b) Final consonant + n*

b + n = mn	qabna	(qab + na)	"we have"
d + n = nn	dinna	(din + na)	"we refuse"
dh + n = n	fuuna	(fuudh + na)	"we take"
f + n = fn	tufna	(tuf + na)	"we spit"
f + n = mn	hindhumne	(hin + dhuf +ne)	"I didn't come"
g + n = nn	dhunna	(dhug + na)	"we drink"
k + n = nn	beenna	(beek + na)	"we know"
q + n = nn	dhanne	(dhaq + ne)	"we went"
q + n = ndh	dhandhe	(dhaq + ne)	"we went"
l + n = ll	kofalla	(kofl + na)	"we smile"
m + n = mn	hindeemne	(hin + deem + ne)	"I did not go"
r + n = rr	harra	(har + na)	"we sweep"
s + n = fn	ijeefna	(ijees + na)	"we kill"
s + n = mn	ijeemna	(ijees + na)	"we kill"
s + n = nn	hinhanne	(hin + hat + ne)	"I did not steal"

c) Final consonant + s**

b + s = bs	c'absa	(c'ab+s+a)	"I break s.th., he etc."
d + s = c	duuca	(duud+s+a)	"I fill s.th., he etc."
dh + s = cl	nyaacisa	(nyaadh+sis+a)	"I feed s.o., he etc."
dh + s = s	fuusisa	(fuudh+sis+a)	"I make s.o. take, he etc."
f + s = fs	afsiisa	(af + siis +a)	"I make s.o. spread s.th., he etc."
g + s = ss	dhussisa	(dhug+sis+a)	"I make s.o. drink, he etc."
g + s = s	dhusiisa	(dhug+siis+a)	id.
g + s = s	midhaasa	(midhaag+sa)	"I make, he makes"
k + s = ss	tissa	(tik+s+a)	"I herd cattle, he etc."
k' + s = ss	dhissiisa	(dhiq+siis + a)	"I make s.o. wash, he etc."
l + s = lc	k'ajeelca	(k'ajeel+s+a)	"I guide s.o., he etc."
l + s = lc	awaalcisa	(awaal+sis+a)	"I make s.o. bury s.o, he etc."
m + s = ms	deemsisa	(deem+sis+a)	"I make s.o. go, he etc."
n + s = ns	c'iniinsa	(c'iniin+s+a)	"I make s.o. bite, he etc."
ph + s = phs	liphsadha	(liph+s+adha)	"I wink, twinkle, he etc."
r + s = rs	agarsiisa	(agr+siis+a)	"I show, he shows"
t + s = ch	hachisa	(hat+sis+a)	"I make s.o. steal, he etc."

---

* In this data there are no examples of j + n and sh + n

** In this data there are no examples of c + s, ch + s, j + s,sh + s.

d) (h)in- or (h)in- + initial consonant

There is total assimilation of *n* to a following: *l, m, n, or s*, e.g.:

Hilloola	(hin+lool+a)	"I, he will fight"
Himmara	(hin+mar+a)	"I, he will turn"
Hirriik'a	(hin+riik'+a)	"I, he will grind"
Hissirba	(hin+sirb+a)	"I, he will dance"

3.2.1.2. Geminated and ungeminated consonants

Consonants may occur as geminated (*'jabaa'*) or as ungeminated (*'laafaa'*) to form different words. Examples of geminated and ungeminated consonants are given in Table 3.3 below.

Table 3.3: examples of ungeminated and geminated consonants

ungeminated consonants	**gloss**	**geminated consonants**	**gloss**
baruu	'to learn'	*barruu*	'palm of hand'
*dhi**b**a*	'to be difficult'	*dhi**bb**a*	'hundred'
*ba**d**aa*	'hot ash'	*ba**dd**aa*	'highland'

3.2.1.3. Morphophonemic Process in Consonant Cluster

Afaan Oromo has a rich set of morphophonemic rules. Almost all operate on consonant-consonant sequences or vowel-vowel sequences across morpheme boundaries. Consonant clusters across morpheme boundaries originate in the following environments.

In Verb forms:

i) The last consonant of the verb stem can be followed by the initial consonant of person makers such as *–ta, -ti, -na, -tani, -tu, -nu, and –tana* and the negative suffix *–ne* (Stroomer, 1995). Examples:

Table 3.4(a). Inflectional suffixes

Present main clause				**Present subordinate clause**	
		AFF	**NEG**	**AFF**	**NEG**
Sg	1	*-a*	*-u*	*-u*	*-ne*
	2	*-ta*	*-tu*	*-tu*	*-ne*
	3m	*-a*	*-u*	*-u*	*-ne*
	3f	*-ti*	*-tu*	*-tu*	*-ne*
Pl	1	*-na*	*-nu*	*-nu*	*-ne*
	2	*-tani*	*-tani*	*-tani*	*-ne*
	3	*-ani*	*-ani*	*-ani*	*-ne*

Table 3.4(b). Inflectional suffixes

Past main clause				**Past subordinate clause**	
		AFF	**NEG**	**AFF**	**NEG** *
Sg	1	*-e*	*-ne*	*-e*	*-ne, -in(n)i, -in(n)u*
	2	*-tea*	*-ne*	*-te*	*-ne, -in(n)i, -in(n)u*
	3m	*-e*	*-ne*	*-e*	*-ne, -in(n)i, -in(n)u*
	3f	*-te*	*-ne*	*-te*	*-ne, -in(n)i, -in(n)u*
Pl	1	*-ne*	*-ne*	*-ne*	*-ne, -in(n)i, -in(n)u*
	2	*-tani*	*-ne*	*-tani*	*-ne, -in(n)i, -in(n)u*
	3	*-ani*	*-ne*	*-ani*	*-ne, -in(n)i, -in(n)u*

* the negative indicator *"**-ne**"* is the standard negative maker of past subordinate clause whereas, *"**-in(n)i, -in(n)u**"* are different dialects used at different regions.

Table 3.4(c): Inflectional suffixes

	Imperative	
	AFF	**NEG**
sg	*-i*	*-in(n)i*
pl	*-aa*	*-in(n)aa*

Imperative clauses can only occur as independent clauses, other than when they are reported speech. The verb ending remains imperative and unlengthened on the coordinate clause. They are used to give commands and may take objects and obliques; the subject is always omitted (Payton, 1989).

The preverbal elements are:

(h)in- : This optional element precedes main clause present and past affirmative verbs. It emphasizes the predicate when followed by main clause present affirmative verb it can be translated as a present or as a future. Initial consonant of a verb stem can be followed by the *n* of preverbal *(h)in* (Stroomer,1995).

hin + bad + e ……*hinbade* *('he/it has lost')* 3msg PAST
hin + hoj +at + e ……*hinhojate* ('he/it worked) 3msg PAST
hin + dhuf + ti……*hindhufti* ('she will come') 3fsg PRES

(h)in- : this obligatory element is unstressed clitic that immediately precedes the negative verb forms in main and subordinate clauses, in the present, in the past, and in the imperative. The first vowel of the verb root is stressed suffix, such as *–in(n)uu, -in(n)i, -in(n)aa* (morpheme bound stress). The hin- prefix occurs before all negative verbs. It undergoes a wide range of assimilation to the initial consonant of the verb root. Some examples are:

hin- beek - u → * *himbeeku* (NEG - know -1sg NEG-PRES) "I don't know"

hin-raf –ini → *hirrafini* (NEG – *sleep*- 2sg NEG-IMPER) "Don't sleep(you)"

ii) The last consonant of the verb root can be followed by the initial *s* of the causative verb extensions *–s-, -si(i)s-*.

Examples:

The element *–s-* is attached to the verb root to derive transitive verbs from *s* adjectival stems (Stroomer, 1995).

gabaabsa..... "I make (sth.) short ,he, etc." (cf. *gabaabaa* (m) "short")

owwisa..... "I make (sth.) warm, he, etc." (cf. *owwaa* (m) "warm, hot")

The suffix *–s-* is also used to derive transitive verbs from intransitive verbs.

Examples:

cabsa ..... "I break, he breaks (trans.)" (cf. *cab-* "to break" (intrans.))

kaasa "I make (s.o, sth.) stand, he etc. (trans.)" (cf. *kaa-* "to stand" (intrans.))

Also *–si(i)s* is used to derive transitive verbs from intransitive and transitve verbs. Some examples are:

argsiisa ..... "I show, he shows" (cf. *arg-* "to see")

kofalchiisa .... "I make (s.o) laugh, he etc. (cf. *kofl* "to laugh")

iii) The last consonant of a verb stem can be followed by nomen agentis suffix *–tuu.*

Examples:

tissituu	"cowherd"	(from *tiss+ tuu*	"to herd")
tumtuu	"blacksmith"	(from *tum+ tuu*	"to pound")

* → = becomes

hattuu	"thief "	(from *hat+ tuu*	"to steal")
sirbituu	"singer"	(from *sirb+ tuu*	"to sing")

3.2.1.4. Verb paradigms

The verb paradigms include verb roots ending in C, ending in (V) V, and ending in –w and –y. Verb roots are indicated by a final hyphen (Stroomer, 1995).

1) Main clause affirmative present:

		beek- "to know" "I etc.know"	oli kaa- "to stand on" "I etc. stand on"
sg	1	beeka	oli ka'a
	2	beetta	oli kaata
	3m	beeka	oli ka'a
	3f	beetti	oli kaati
pl	1	beenna	oli kaana
	2	beettani	oli kaatani
	3	beekani	oli kaani

		machaw-"to be drunk" "I etc. am drunk"	**dhagay- "to hear" "I etc. Hear"**
sg	1	machaa'a	dhagga'a
	2	machoofta	dhageeta
	3m	macha'a	dhaga'a
	3f	machoofti	dhageeti
pl	1	machoomna	dhageena
	2	machooftani	dhageetani
	3	machaa'ani	dhaga'ani

2) Main clause negative present:

		beek-"to know" "I etc. Don't know"	c'ee-"to jump" "I etc. don't jump"
sg	1	himbeeku	hinc'eu
	2	himbeettu	hinc'eetu
	3m	himbeeku	hinc'eu
	3f	himbeettu	hinc'eetu
pl	1	himbeennu	hinc'eenu
	2	himbeettani	hinc'eetani
	3	himbeekani	hinc'eani

		machaw-"to be drunk" "I etc. am not drunk"	dhagay-"to hear" "I etc. don't hear"
sg	1	himmachau	hindhagau
	2	himmachooftu	hindhageetu
	3m	himmachau	hindh agau
	3f	himmachooftu	hindhageetu
p1	1	himmachoomnu	hindhageenu
	2	himmachooftani	hindhageetani
	3	himmachaani	hindhagaani

3) Main clause affirmative past:

		beek- "to know" "I etc. knew"	jajjii- "to cut meat in strips" "I etc. have cut"
sg	1	beeke	jajjie
	2	beette	jajjiite
	3m	beeke	jajjie
	3f	beette	jajjiite
p1	1	beenne	jajjiine
	2	beettani	jajjiitani
	3	beekani	jajjiani

		machaw- "to be drunk" "I etc. am drunk"	dhagay- "to hear" "I etc. Heard"
sg	1	machaa'e	dhaga'e
	2	machoofte	dhageete
	3m	machaa'e	dhaga'e
	3f	machoofte	dhageete
p1	1	machoomne	dhageene
	2	machooftani	dhageetani
	3	machaa'ani	dhaga'ani

4) Main clause negative past, subordinate clause negative past and subordinate clause negative present have forms that are invariant for person:

beek-	"to know"
himbeenne	"I (etc) didn't know; that I (etc.) don't didn't know"
jajjii-	"to cut meat (into strips)"
hinjajjiine	"I (etc.) didn't cut meat (into strips)"
machaw-	"to be drunk"
himmachoofne	"I was not drunk; that I am not, was not drunk" (etc.)
dhagay-	"to hear"
hindhageene	"I (etc.) didn't hear; that I (etc.) don't didn't hear"

5) Subordinate clause negative past have forms that are invariant for person:

dhaq-	"to go"
hindhaqu	"that I (etc.) didn't go"

6) Imperatives

sg	**p1**	
dhaabi	dhaabaa	"cook (it)!"
hojadhu	hajadhaa	"work!"
hindhaabini	hindhaabinaa	"don't cook (it)!"
hinhojatinni	hinhojatinnaa	"don't work!"

7) Adhortatives

eegi aanini haa fooni c'iru
(eeg + i aanini haa fooni c'ir+u)
wait+ sgIMPER I (SUBJ) adhortative meat cut + 1sgPRES.SUBORD
"Wait, may I cut the meat"

aanini haa mina haru
(aanini haa mina har+u)
I (SUBJ) adhortative house sweet+ 1sgPRES.SUBORD
"May I sweep the house?"

bisaani kadhannoo
(bisaani kadh+at+noo)
water ask + MIVO+1p1ADHORT
"Let us ask for (some) water"

soodaa dhunnoo
(soodaa dhug+noo)
Soda drink+1p1ADHORT
"Let us drink a soda"

Beesee sii kenninoo*
(beesee sii ii kenn+noo)
money you LIN=SCOPE give+1p1ADHORT
"Let us give you money"

In this data the following imperatives are regularly used to express an adhortative; they may be followed by a main clause finite verb:

sg

nuu kaasi
(nuu kaa + s + i)
us stand-up+CAUS+sgIMPER
"make us stand up"

nuu yaasi
(nuu yaa+s+i)
us go + CAUS + sg IMPER
"Make us go"

p1

nuu kaasaa
(nuu kaa + s + aa)
us stand-ulp+CAUS+p1 IMPER
"make us sand up "

nuu yaasaa
(nuu yaa + s + aa)
us go +CAUS+pl IMPER
"make us go"

wolini nuu yaasi shay dhunna
(woli ii + ni nuu yaa + s + i shay dhug +na)
Together LIN + ni=INSTR us go + CAUS + sgIMPER tea drink+ 1plPRES
"Let us go and drink tea together"

In Nouns:

The last consonant of a noun stem can be followed by consonant initial subject markers`-*ni or* focused subject marker-*ti*.

Examples:

("girl + focused SUBJ") *intala + ti* ⟶ *intalti*

("navel + focused SUBJ") *handhuura + ti* ⟶ *handhuurti*

Examples of -*ni* after nouns or adjectives ending in a long vowel:

oboleettii + ni ⟶ *obboleetti(i)ni* ..... "sister + SUBJ"

guddaa + ni ⟶ *gudda(a)ni* ..... "big + SUBJ"

oduu + ni ⟶ *odu(u)ni* ..... "news + SUBJ"

In Adjectives:

The final consonant of the adjectival stem can be followed by the feminine gender suffix -*tuu*. Examples:

Masculine	**Feminine**	
diimaa	*diimtuu*	"red"
dhalaa	*dhaltuu*	"female"
dheeraa	*dheertuu*	"long, tall"

The environments as given in verb forms, noun forms and adjectivals discussed in this section can be summarized as follows:

- Final consonant + t
- Final consonant + n
- Final consonant + s
- (h)in- + initial consonant

3.2.1.5. Vowels

Afaan Oromoo has basically 10 phonemic vowels, five short and five long. A long vowel is interpreted as a single unit and occurs everywhere a short vowel can occur. Long vowels reduce to short voiced vowels and short vowels reduce to short voiceless vowels or zero word finally. Vowel phonemes can appear in initial, medial and final positions.

Short vowels

Front		Back
i		u
e		o
	a	

Long vowels

Front	Back	
ii	uu	HIGH
ee	oo	
	aa	LOW

The following examples show some samples of vowels at word initial, medial and final positions.

Initial position	Medial position	Final position
eelee "pan"	*fii*xaa "family, relatives"	gar***aa*** "belly"
uume "he/it creates"	k***ee***nna "our"	daar***aa*** "ash"

3.3. Word formation in Afaan Oromo

Surprisingly, as stated in chapter two of this study, there is no exact and common definition of a term 'word'. Linguists look at the term '*word*' differently. To syntax analyzers '*words*" is the unit that makes up sentence. Words are grouped according to their function in their sentential structure. Each group gets a tag usually called part-of-speech or word category- and grammar deals with this tags only, omitting the details of specific words. Morphological analyzers are concerned with the inner structure of

'words'. They attempt to uncover the rules that govern the formation of words from smaller units as discussed earlier in chapter two of this study.

Word formation in Afaan Oromo refers to the question of whether the bases of nominalization processes, are *roots, stems* or *words*. In Afaan Oromo roots are bound as they can not occur on their own like in *dhug-* "drink", and *beek-* "know", which are pronounceable only when other completing affixes are added to them. In other words, these roots serve as base stems in Afaan Oromoo since they possess non-verbalized glosses (meanings).

There are two ways of word formation in Afaan Oromo. These processes according Temesgen (1995) are *affixation* and *compounding*; both are discussed separately in the following sections.

3.3.1. Affixation

Different lexical categories are derived by adding various affixes to bases belonging to different lexical categories, and that the addition of the affixes entails different properties of phonology, morphology, syntax and semantics.

Affixation is therefore, the process by which new words are formed from different bases. It is the most productive way of word formation in the language. Although in most cases words are derived by additions of overt affixes, the process may also involve zero affixation also known as zero derivation or compilation. Like the root an affix is also a morpheme that can not occur independently. It is attached in some manner to the root, which serves as a base.

Affixations are of three types. They are generally described as the addition of affixes at the beginning, in between and/or at the end of a root or stem depending on whether the affix is prefix, infix or suffix.

For Oromo prefix and suffix occur at the beginning and at the end of a root or a stem respectively in forming a word. A suffix is an affix that is attached after a stem. For example in ***gaarummaa***, 'fineness' *–umma* is a suffix and *gaar-* 'fine' is a stem. By far the most common affixes are suffixes in the language.

Prefix: A prefix is an affix that is attached in front of the root/stem. Prefixes in Afaan Oromo includes *(h)in*. With high or low tone following positive or negative sentence structure these prefixes are used to change root into negative, positive or interrogative based on falling or rising of stresses on the vowel. Unlike other affixes *(h)in* is some times not attached to the word they express.

Examples: *hín bèknè* "we know", *hìn bééknè* "didn't know"
hín dèèmnè "we went", *hìn déémnè* "didn't go". However some evidences indicate that the assimilation of *(h)in* with stems are common in different texts.

Examples: *nigala* "he will be returned"
hingalle "he,she,etc. didn't returned". Further study is thus needed to give detail explanation for the variations in the usage of the prefix. But in this study *hin* is treated as the only prefix in the language.

3.3.1.1. Process of Suffixation

Besides inflectional processes Afaan Oromo is also rich derivationally. Baye (1986) quoted in Temesgen says that "to the exclusion of adpositional all the other categories have derived forms in addition to their simple forms." Some derivational processes are discussed in the following subsections.

According Katamba (1993), the English word-form ***'see'*** has three letters and the word ***'seeing'*** has six. There are also other word-forms of the same family. And, if we were counting the number of words in a passage, we would gladly count ***see, sees, seeing, saw,***

and ***seen*** as five different word forms (belonging to the same lexeme). In languages like Arabic, according Katamba (1993), whose morphology is largely concatenating, very often words are formed, not simply by concatenating affixes and roots, but rather by infixing, geminating and other changes taking place internally within the root. But in Afaan Oromo the case is some how different. For example a number of words can be derived from the stem '***beek***' which means 'to know' as indicated below.

beekaa	*beektota*	*beektaniirtu*
beeke	*beekta*	*beektii*
beekuu	*beektuu*	*beeki*
beekaniiru	*beekte*	*beektanii*, etc.
beekumsa	*beekeera*	

3.3.1.2. Nominalization

This is the process of forming nominal from different categories. In Afaan Oromo there is a large stock of nonminals derived from adjectival, adverbial and nominal bases. In Afaan Oromo the verb root can take the following nominalizing suffixes: *-uu, -aa, -iitii, -achuu, (-ach+uu), -achaa (-ach + aa), -achiitii (-ach + iitii).* Verb roots plus *–uu* or *–achuu* function as infinitives (Temesgen 1995, Stroomer,1995). This is as can be seen in the examples below:

ajjeesuu	"to kill, killing"	(cf. *ajjeesa*)
barbaaduu	"to look for, looking for"	(cf. *barbaada*)
gammachuu	"to rejoice, rejoicing"	(cf. *gammada*)
hojjachuu	"working"	(cf. *hojjadha*)

Some more nominals are as discussed in the following section (i-vii).

i. Abstract nominals

Abstract nominals are derived from adjectival and nominal bases by the addition of different suffixes as shown in Table 3.6 below.

Table 3.6: Abstract nominals

Base		Affix	Derived nominals	
gaarii	'fine'	*-ummaa*	*gaarummaa*	'fineness'
gamna	'wise'	*-uma*	*gamnuma*	'wisedom'
diimaa	'red'	*-ina*	*diimina*	'redness'
adii	'white'	*-eenna*	*add-eenna*	'whiteness'
durba	'girl'	*-ummaa*	*durbummaa*	'girlhood'

Abstract nominals may derive from non-abstract nominals or adjectivals with an affix.

ii. Process/action nominals

Process nominals refer to 'the fact, the act, the quality or occurrence of' the base from which they are derived. In Afaan Oromo, such nominals are derived from verbal roots by adding different suffixes of which */-cha(-choo) /, /-sa/, /-umsa/, /(-a) –a/, and /-taa/* are just a few as the example in Table 3.7 illustrate.

Table 3.7 Some affixes of process/action nominals

Verbal base		Affix	Process/Action nominal (words)	
fiig-	'run'	*-cha*	*fiigicha*	'running'
ilaal-	'see'	*-cha*	*ilaalcha*	'seeing'
ijaar-	'build'	*-sa*	ijaar*sa*	'building'
qot-	'till'	*-sa*	*qotiisa*	'tilling'
gurgur-	'sell'	*-taa*	*gurgurtaa*	'selling'
burraq-	'jump'	*-a*	*burraqa*	'jumping'
kadh-	'beg'	*-aa*	*kadhaa*	'begging'

To predict the distribution of such affixes is not simple, for instance, the alteration between */-aa/* and */-a/* may be accounted for in terms of syllable structure in that */-aa/* is found when the vowel of the base is short and */-a/* otherwise. This phenomenon is quite common in the language.

iii. Result nominals

Some of the affixes used in process/action nominals such as: */-umsa/, /-sa/, /(-a) -a/,* and */aatii/* are also used in the formation of result nominals from verbal roots. These may be homophones. In addition to these are others like */-tee/, /-ii/, /-chuu/, /-oo/,* and */-suu/* which are also used to derive other result nominals.

Some examples of result nominals are as indicated in Table 3.8.

Table3.8. Some affixes of result nominals

Base		Affix	Result nominals	
beek-	'know'	*-umsa*	*beekumsa*	'knowledge'
abaar-	'curse'	*-sa*	*abaassa*	'cursing'
kenn-	'give'	*-aa*	*kennaa*	'gift'
dhug-	'drink'	*-aatii*	*dhugaatii*	'drink*(n)*'
mur-	'cut'	*-tee*	*murtee*	'decision'
dadhab-	'exhaust'	*-ii*	*dadhabii*	'exhaustion'
gammad-	'be happy'	*-chuu*	*gammachuu*	'happiness'
arrabs-	'insult'	*-oo*	*arrabsoo*	'insult*(n)*'
dallaan-	'be sad'	*-suu*	*dallansuu*	'sadness'

iv. Gerundive nominals

These are derived from verbal roots by the addition of */-uu/* as in the following examples:

Table 3.9: Examples of gerundive nominals

Base		Affix	Gerundive nominals	
bit-	'buy'	*-uu*	*bituu*	'buying/to buy'
deem-	'go'	*-uu*	*deemuu*	"going/to go"
nyaat-	'eat'	*-uu*	*nyaatuu*	'eating/to eat'

v. Manner nomoinals

These are nominals which refer to the means or ways of doing something. They are derived from verbal roots with the suffixes */-ii/, /-umsa/* and */-aatii/* as shown in the following examples.

Table 3.10: Some examples of manner nominals

Base		Affix	Manner nominals
ijaajj-	'stand'	*-ii*	*ijaajj*ii 'manner of standing'
taa-	'sit'	*-umsa*	*taaumsa* 'manner of sitting'
dhug-	'drink'	*-aatii*	*dhugaatii* 'manner of drinking'

vi. Instrumental nominals

Instrumental nominals are nominals which are formed with */-ata/,* and */-tuu/*. Examples of such nominals include:

Table 3.11: Examples of instrumental nominal affixes

Base		Affix	Instrumental nominals	
har-	'sweep'	*-ata*	*harata*	'broom'
hidh-	'tie'	*-ata*	hidh*ata*	'armament'
hodh-	'suck'	*-tuu*	*hodhtuu*	'milk feeder'
ham-	'harvest'	*-tuu*	*hamtuu*	'harvesting machine'

vii. Agent nominals

Agent nominnals are derived from verbs of action and have a meaning like one who does the action of a verb. In Afaan Oromo agentive nominals are derived with */-aa/* and */-tuu/*. The following are examples:

Table 3.12: Examples of agent nominal affixes

Base		Affix	Agentive nominals	
barsiis-	'teach'	*-aa*	*barsiisaa*	'teacher'
eeg-	'keep'	*-tuu*	*eegduu*	'keeper'
nyaat-	'eat'	*-tuu*	*nyaattuu*	'eater'

In conclusion, the bases for abstract nominals are words, whereas in the case of others the suffixes are attached to either roots or stems. Thus, one can say that nominalization takes *roots, stems* or *words* as its domain. Distribution of suffixes is unpredictable. Some nominals are formed with different affixes. In such cases the distribution is difficult to account for.

3.3.1.3. Verbalization

Derived verbs in Afaan Oromo include causatives, statives, reflexives, and passives.

i) Causatives

Causatives are said to be derived from verbal roots or stems by the addition of the variants of the causative morpheme */-(si) is-/, /-s-/, /-ess/,* as illustrated below:

Table 3.13: Examples of causatives

Base		Affix	Causative verb	
mur-	'cut'	*-siis-*	*mursiis-*	'make cut'
raf-	'sleep'	*-ls-*	*raffis-*	'make sleep'
gub-	'burn'	*-siis-*	*gubsiis-*	'make burn'
malaa	'pus'	*-s-*	*malaas-*	'discharge pus'
dheer-	'tall'	*-ess-*	*dheeress-*	'make tall'
gudd-	'big'	*-is-*	*guddis-*	'make big'

In Afaan Oromo when a form that ends in the glottal stop /-ʻ/ is followed by a suffix, the glottal stop is dropped and a compensatory lengthening of vowel in the stem takes place.

Eg. d*u'*- 'die' + *-tuu* →*duutuu* 'deceased'

Truncation does not take place in the derivation of causatives from verbal roots ending in similar segments, for example

n*yaat-* 'eat' + *-siis-* → *nyaat- chis-* 'cause to eat'

xabat- 'play' + *-siis-*→ *xabat- chis-* 'cause to play'

The causative suffix is attached without any truncation with some verbs derived from nominals by addition of /*-at-*/. Example:

sodaa 'fear']N + *-at-*]V + *-siis*→ *sodaat-chis-* 'frighten'

dubbii 'talk']N + *-at-*] V + *-siis* → *dubbat-chis-* 'cause to talk'

ii) Stative verbs

Stative verbs (according Temesgen, 1995), are verbs which "denote qualities or attributes possessed by the subject of the clause in which they appear". In Afaan Oromo these verbs are derived from adjectival and nominal bases with the suffix /*-at-*/. Example:

Table 3.14: Examples of derived statives

Base		Affix	Derived statives	
diimaa	'red'	*-at-*	*diimat-*	'become red'
furdaa	'fat'	*-at-*	*furdat-*	'become fat'
hojji	'work'	*-at-*	*hojjet-*	'to work'
dheebuu	'thirst'	*-at-*	*dheebot-*	'become thirsty'

iii) Middles

Middle verbs are identified by their subject that performs the action or participates in the event denoted by the verb expressly for his own benefit. Hence, they are different from reflexive verbs only in function, i.e. where as in the case of statives the subject of the clause undergoes some changes of state, with middles the subject participates in the action of the verb and as a beneficiary from the action. Here the morpheme is the same /*-at-*/ as can be shown below:

Table 3.15: Examples of benefactives

Base		Affix	Benefactives	
bit-	'buy'	*-at-*	*bitat-*	'buy for oneself'
qab-	'catch'	*-at-*	*gabat-*	'catch for oneself'
haad-	'shave'	*-at-*	*haddat-*	'shave oneself'
dhiq-	'wash'	*-at-*	*dhiqat-*	'wash oneself'

In some cases, the base to which the affix is attached is a stative verb as in the following:

ulfaataa 'heavy' cf. * ulfaa-

gabbataa 'fat' cf. *gabbaa-

Other adjectives are marked with */-ssa/, /-ttii/* endings as in the following:

sooressa/ttii 'rich (masc/fem)'

qabeessa/ttii 'wealthy (masc/fem)'

Still others end in */-ee/, /-uu/ , /-ii/* and */-oo/* as shown below.

gadhee 'bad'

gobbuu 'dense'

fagoo 'far'

dheedhii 'raw'

Even though such endings show consistency of form, it is difficult to assume that all vowels in the language are used to derive adjectivals. Furthermore, the bases with which these vowels appear cannot be categorized into any of the lexical categories except adjectivals, therefore the only plausible thing to do is to regard them as simple (non-derived) adjectives. But we have to recognize that some adjectivals are derived on the analogy of such simple ones.

Generally, the productive adjectivization process in Afaan Oromo is the addition of */-aa/ -uu/* which, at the same time also marks the gender of the derived form.

3.3.2. Compounding Process

Compounding is another word formation process by which significant number of compound words formed. It is the process of forming new words by combining different lexical categories which functions independently. However, it is not the case that every two words (stems) combine to form a compound form. Every language follows certain rules by which it forms its compounds.

The process of compounding different lexical categories are formed by combining two words or stems, and that such a process also entails different phonological, morphological, syntactic and semantic characteristics.

3.3.2.1. Afaan Oromo compound words

There are different types of Afaan Oromo compounds. One of the criterions for the classification of compounds in this language is by considering the role they play in sentences as nouns, verbs, adjectives, etc. Using this criterion some of them are briefly discussed in this section.

Noun-Noun Compounds- Two nouns can combine to form various kinds of compound nouns. For example the noun *abbaa* 'father' or *haadha* 'mother' can occur as a first member in a compound like the following.

abbaa buddeenaa 'step father'
father food

haadha manaa 'wife'
mother house

Instrumental compounds are formed by combining two nouns of which the first member is instrumental for the realization of the thing designated by the second member.

dhagaa daakuu 'millstone'
stone flour

xuwwee marqaa 'porrige-pot'
pot porridge

Names of certain parts of the body may also be combined with other nouns to form compounds designating names of disease.

mata cabsaa 'headache'
head breaker

dugda kutaa 'backache/pain'
back cutting

Other compound nouns may be formed by combining nouns referring to locations where activities take place.

m*ana barumsaa* 'school'
house learning

bakkee waraanaa 'battlefield'
field fighting

Noun-Adjective compounds: Some examples of such compounds include:

sanbata guddaa 'sunday'
sabath big

muka guraacha 'a kind of a tree'
tree black

Adposition-Noun compounds: A combination of an adposition and a noun may result in a compound noun.

gadi qabaa 'opression'
down having

dura taa'aa 'chairman'
front he who sits

keessa deebii 'revision'
in returning

As can be seen from the above examples a large number of compounds occur in the language. The underlying formation process of compounds is irregular. Because of this determining the stem from which the words are derived become very difficult. As a result it requires more efforts and time to develop an analyzer that identifies such words with in the language.

3.4. Afaan Oromo word classes

Like any other natural languages Afaan Oromo has its own word classes. There are five categories of word classes in Afaan Oromo. These major lexical categories serve as heads in phrasal constructions. According to Baye (1986) the five major lexical categories are nouns, adpositions, adverbs, adjectives and verbs.

3.4.1. Nouns ("*Maqaa*")

Afaan Oromo nouns are words used to name or identify any of categories of things, people, places or ideas or a particular one of these entities. In simple words nouns are "words used to name a person, place or thing". Two numbers (singular and plural) are recognized in Afaan Oromo. A singular noun is marked by zero morphemes where as a plural in marked by various forms. Here are some examples:

Singular		**plural**	**plural marker**
saree	'dog'	*saroota*	*/-oota/*
muka	'tree'	*mukkeen*	*/-een/*
obboleessa	'brother'	*obboleeyyii*	*/-yyii/*
ilma	'son'	*ilmaan*	*/-an/*

As can be observed form the above example nouns are pluralized by adding various forms of suffixes such as *-oota, -yyii, -an, -lee, -wan* etc. It is also possible to use more than one type of plural markers in some nouns. For instance *jaarsa* 'old man' can be

pluralized both as *jaarsota* or *jaarsolii*. However, most nouns of the language prefer one plural maker to the other. When nouns stem and plural making suffix combine, various morphological processes take place among which the last vowel of the stem drops and are compensated by gemination of the preceding constant.

Noun roots end in a short or long vowel. They have grammatical gender (masculine or feminine), and are either countable or uncountable (rare, e.g. ***bishaan*** 'water'). Some nouns have the ***–esa*** 'masc' or ***–ettii*** 'fem' suffixes: a given noun may exist in any combination of the unaffixed, masculine or feminine format.

The noun will take the nominative suffix unless the noun phrase includes:

i) a demonstrative and any other constituent, even if the demonstrative is in an embedded phrase; and

ii) two adjectives.

In Afaan Oromoo there are two genders (feminine and masculine). In addition to feminine and masculine there is also another special form of number indicator called singulative gender that refers to 'exactly one'. Such variants according Bender et al (1976) are usually formed by addition of the suffix-***icha*** to a noun for masculine and ***–ittii*** for feminine as in *nam-icha* 'the man' and *nam-ittii* 'the woman'.

3.4.2. Verbs ("*xumura*")

Verbs in Afaan Oromo are forms which occur in clause final position and belong to a distinct category from that of nouns (nominals). Different forms of verbs are: intransitive, transitive, detransitive, modals and auxiliary verbs. The intransitive verbs are those verbs which do not take any phrase as their complement. The detransitive verbs took two complements in Afaan Oromo.

Verb roots end in a consonant, or in a vowel sequence or long vowel. There are three derivational sequences which change the voice of the verbs: ***-siis***, and ***–s*** for causative, ***-et***

and ***–adh***(1s) for subject reflexive, the so called reflexive middle voice, and –am for passive stative.

An inflected verb may also take one of two case suffixes. ***–fi*** 'been', where a benefactive object in a sentence is omitted or unmarked, and ***–ni*** where an accompanying object (including instrument) in the sentence is omitted.

A verb takes the prefix ***–hin*** 'neg' (with following floating low tone) when it is negative. Indicative finite verbs inflect for person, gender, number, clause type, polarity and aspect. Imperative mood verbs inflect for polarity and number.

There is no consistent way to split the verb agreement marker down into two or more separate affixes. Instead, the verb ending will be considered as a single fused suffix, containing all the information. According Payton (1989), ten percent of the paradigm consists of affixes which are single phonemes and over sixty percent are identical with another form, splitting the affixes would add complexity without increasing clarity. Some rough splits have however, been included in Table 3.4 a-c above.

3.4.3. Adposition ("*Dur duubee*")

Adpositions are words which will have meaning only when they are attached or used together with other words such as nouns, verbs, pronouns, and adjectives. Adpositions are characterized by having no inflectional or derivational morphology and belong to the closed system.

Examples of adposition in Afaan Oromo are of two types depending on whether they are bound or free.
Abbabaa waliin, gara manaa, harka-an, ummata-af are some of the examples of adpositions.

3.4.4. Adverbs ("*Ibsa xumuraa*")

Words which are used to modify verbs in Afaan Oromo are considered as adverbs. There are different types of adverbs. These are adverbs of time, place, manner, frequency, degree, reason, etc. In general adverbs are treated as the subclass of verbs. There are only a few adverbs of manner, most ideas of manner being expressed in another way, e.g., a phrase of manner, or adding the quality to the subject by an adjective or genitive construction. There are however, a variety of adverbs of time.

Example: ***gurba –n*** ***kaleessa*** ***dhuf –e***

boy –nom yesterday come -3msc

'The boy came yesterday'

3.4.5. Adjectives ("*Ibsa maqaa*")

Adjectives are specifiers of noun or a pronoun: they usually come after the noun they specify. *Gurracha, adii, dheeraa, xinnaa,* etc are some of adjectives in Afaan Oromo. Adjectives end in a short or long vowel, and can also end in a vowel sequence v1, v2v2. Adjectives can inflect for gender although they are less likely to do so as the second half of an equative sentence than when embedded in a noun phrase. They are inflected for gender and number. The suffixes *–sa, -aa, -tuu, -aawaa*, and *–oo* are gender sensitive adjectival endings.

CHAPTER FOUR

DEVELOPMENT OF ALGORITHM

4.1. Introduction

In this chapter the approach adopted to develop the algorithm for morphological analyzer, justification for the adoption of the morphological analyzer and the gold-standard prepared for the language are discussed.

The chapter is organized into two sections. The first section discuses the algorithm of morphological analyzer; and the second part discusses the preparation of gold-standard for the analyzer.

4.1.1 Cleaning the corpus

A corpus which is found in any natural language and printed in different information sources such as text books and newspaper are in raw forms and huge in size. It is available both in text and electronic forms. But the text needs to be cleaned to remove some unnecessary characters, such as exclamation marks (!), quotation marks (“) and the like.

Before decomposing a word into different components such as prefix, stem and suffix, a program should clean all the mentioned characters which are not needed in the design of the gold-standard as well as during analysis using Linguistica_beta2*. These characters, if not cleaned by appropriate tools, may result in poor performance of the analyzer. Because of the big size of the corpus, a program is written to perform the job automatically.

* Linguistica_beeta2 is the improved version of inguistica2001.

Supported by different tools the analyzer takes a text file as its input and produces a partial morphological analysis of the words in the corpus giving an output (stem + affix with different statistical information). It also separates the linear ordering of affix system if they come in a particular sequence before or after other affixes. The way an ordering scheme designed is defined in Alchemist**.

4.2. The algorithm for preprocessing text data

In this section the algorithm used in preprocessing the text file is presented.

1. Open source file for reading a character
2. Open destination file for writing a character
3. Read a character from a source file while not EOF reached
4. Write the character on the destination file while it is different from , , !, " , etc
5. Repeat step 3 and 4.

Fig. 4.1 The algorithm used for cleaning the text file

The morphological analyzer works with text file. Thus, a file recorded in plain text format is prepared. The file is cleaned by automatic means using a C++ program that read one text file and eliminate unnecessary characters in order to get a pure text that in turn can improve the performance of the analyzer. The program works on character basis in order to easily identify and remove the unnecessary characters through out the sample corpus. Whatever size of text file one can have it can easily be managed by the program automatically.

4.3. Afaan Oromoo affixes

Affixation as discussed in chapter three of this study is the process by which new words are formed from different bases. In most cases words are derived by additions of overt

** a GUI based tool for creating morphological gold-standards (detail discussion of Alchemist is presented in section 4.4 of this study)

affixes, the process may also involve zero affixation also known as zero derivation or compilation. Like the root an affix is also a morpheme that can not occur independently. Affixations are generally described as the addition of affixes at the beginning, in between and/or at the end of a root or stem depending on whether the affix is prefix, infix or suffix.

The formation of prefix and suffix is discussed in the previous chapter. In this section the formation of prefix-suffix pairs and list of common suffixes in Afaan Oromoo is presented. As indicated in section 3.3.1 ***hin*** is the only prefix in the language and therefore considered in this study; the list of suffixes is presented in Table 4.1.

4.3.1. Prefix-suffix pairs

The occurrence of prefix-suffix is frequent in Afaan Oromoo text. For example *hindhufne* "he,she etc did not come", *hindhufnee* "doesn't he, she etc come yet?" contain both prefix and suffixes. Accordingly, *hin-* is a prefix in both examples, *-ne* and *–nee* are suffixes added to the stem *–dhuf-*.

Table 4.1: List of Afaan Oromoo suffixes

a	aa	aadh	aaf	aatiin	aawwaa
aawwan	adh	am	an	at	ata
aw	cha	dha	dhaa	e	een
eeny	eess	eettii	eyyii	f	fa
fi	i	icha	ichi	ii	iin
iitti	ittee	itti	ittii	lee	llee
ma	maa	me	mee	mma	n
na	ne	ni	nne	nni	nnis
nu	NULL	o	olee	olii	om
oma	oo	ooftu	oolee	oolii	ooma
oon	oonni	oota	ota	s	sa
si	siis	sis	ssa	t	ta
taa	taaf	te	teen	ti	ticha
tii	toota	tte	tti	ttu	tu
tuu	u	ullee	umaa	umma	umsa
umsi	uu	uudh	wwa	wwan	yyu
yyuu					

4.4. Algorithm for morphological analysis

There are a number of natural language morphological analyzer algorithms based on the method of the analysis. This section describes briefly the process that takes place in the development of algorithm for the unsupervised learning of Afaan Oromoo morphology. It does not need dictionary and the morphological rule of the language. Since Linguistica uses automatic and unsupervised morphological analysis of a corpus of a natural language, it determines the prefixes, suffixes and stems of the language and generates morphological dictionary of its own (called signature) using the corpus. From that dictionary (***Mini lexicon*** according to Linguistica_beta2), it generates a partial analysis by producing stem and affix with its different components and statistical information about the corpus in general and the analyzed words in particular.

The algorithm used for this study is not far from that of an algorithm for the unsupervised learning of morphology. It divides the words of the corpus into stem and affix, and a set of incremental heuristics, which modify the analysis; leaving the decision to other part of the program i.e. Minimum Description Length (MDL) discussed in chapter 2, section 2.4.2 of this study. As indicated earlier among a number of algorithms of morphological analysis, the algorithm of unsupervised learning of natural language morphology is relevant to this study. It identifies the morpheme components of stems and maintains the association among morphemes in different signatures.

An algorithm for the unsupervised learning of Afaan Oromoo morphology (adopted from Goldsmith, 2001b) is as shown in fig 4.2 below. The detail and complete description of the algorithm is given in appendix I.

1. **Get a corpus from a text file**
2. **Find successor frequency** (to cut the words into two pieces – stem and affix)

   If a word ending is one of the following: a, aatiin, am, an, at, aw, cha, dhaa, e, eyyii, f, fi, i, ichi, lee, ittii, maa, mma, n, na, nne, nu, o, oolee, oota, sa, sis, ta, t, tte, ttu, u, umsa, uu, etc

   Organize the pieces into signature, put in respective list (stem, suffix)

Else go to 8

3. Check signatures

- incorporate the insights of MDL

- compute the entropy of the set of stem-final letters

4. Extend known stem to known suffix

5. Extend known signature

6. Extend known stems in finding new suffixes

7. Extend known suffixes ("loose fit")

8. Find singleton signatures

9. Detect rules of allomorphy

Repeat 2 to 9 until EOF

10. End

Fig. 4.2: An algorithm for the unsupervised learning of Afaan Oromoo morphology (adopted from Goldsmith, 2001b)

4.5. Linguistica_beta2

Linguistica is a program designed to explore the unsupervised learning of natural languages, with primary focus on morphology (word-structure). It runs under Windows and Linux, and it is written in C++ within the Qt* development framework. Linguistica2001 needs a large corpus size for analysis. The default setting for Linguistica2001's corpus input is 5,000 words: this is the number of words from a corpus that the program will read. The maximum words that can be read by it are about 1,000,000. If one wishes to change the default setting, select "***Words requested***" in the Lower Tree on the left. A pop-up window will appear in which one can specify a different number of words to be read from the corpus. This number refers to the total number of word (tokens) read, not word types or unique words. The program used in this

* Qt is a C++ program toolkit for multiplication of GUI and application development. It provides single-source portability across MS Windows, Mac OS X, Linux, and all major commercial Unix variants.

study (Linguistica_beta2) has the ability to analyze more words than the previous version (Linguistica2001). The default is 10,000 and one can set based on his/her text size.

As indicated in the previous sections, unsupervised learning refers to the computational task of making inferences (and therefore acquiring knowledge) about the structure that lies behind some set of data without any direct access to the structure. In the case of unsupervised learning of morphology, Linguistica_beta2 explores the possibilities of morpheme-combinations for a set of words based on the internal knowledge of the language from which the words are drawn.

Based on the MDL discussed in section 2.4.2, an unsupervised learning algorithm of natural language morphology has used heuristic to develop the analysis system. Major steps used in Linguistica_beta2 are discussed below.

4.5.1. Segmentation

Segmentation is a process of identifying a set of candidate word breaks (successor frequency). It is an optimal division of each word within a corpus into stems and affixes based on Harris algorithm. Harris algorithm helps to detect morpheme boundaries (given a phonemic representation) by asking, in between each letter of a word, how many different ways there were to finish a word, given all of the letters (or phonemes) up to that point in the word.

Goldsmith made a modification on Harris's algorithm by giving more emphasis and weights on word breaks that are around the middle or end of words than those at the beginning. This is done by including length information into the calculation of word break points. Two tasks involved at this point are: *computing the word break point values AND distributing the corpus frequency of each word to their respective word breaks.* Examples of Afaan Oromoo segmented words is given in figure 4:3.

4.5.2. Identification of signature

After finding appropriate cuts of some of the words into two pieces, it then treat the first piece as a stem and the second as a suffix, and for each stem it organize the entire set of suffixes with which it appears in the corpus as an alphabetized list (i.e. ***a suffix list***). It then creates a list of such affix-lists, and associate with each such list the set of stems that appears with precisely the set of suffixes. It is this association that we call a signature. In this process Linguistica decides what the stems are, what the suffixes are, and so forth. Most of Linguistica's functionality, at this point, goes in to making these decisions. In the process of identifying signatures word splits are put in separate lists and a signature is created to link stems with the associated suffixes or prefixes.

Signatures are used to organize the stems and suffixes in such a way that the words in the corpus can be generated from the broken components. For example Afaan Oromoo words ***alagaa, biiroo, kaluu,*** and ***nagayaa*** appear to have no suffixes (i.e., with NULL suffix). Thus they have a common signature NULL.n.tti.

However, such signature is created only for stems and suffixes that are optimal at least to one word in the language. This is to mean that ***n*** or ***tti*** is kept in signature as suffix only if it appears at least with one word. This is done to eliminate spurious parses. Such word breaks are identified automatically by computing ***log (stem count)*log (affix count)*** and taking only those word breaks with non-zero values. Such signatures are referred as *regular signatures* and the associated suffixes as regular affixes assuming them as possible morphemes. But they are not quite the affixes that one would like to establish for a natural language. To make them linguistically acceptable there are two options:

i) using heuristics associated with MDL test designed to ascertain the improvement, and

ii) designing the gold-standard with other programs such as Alchemist. Both options are used in this study, and therefore are as presented below for comparison purpose.

4.5.2.1. Optimizing the morphology with heuristics

Lists of stems, suffixes and prefixes are established in the steps segmentation and identification of signatures. According to Goldsmith 2000a, these lists can be considered as the morphology of a given language-the language of the corpus. This identifies a word having a stem and suffix components with a set of pointers that point to the stem list and the suffix list.

Goldsmith also include a triage to determine how many stem must occur in order for the data to be strong enough to support the existence of a linguistically real signature. Triage is done by removing certain signatures and observing the improvement gained.

Accordingly, any signature with a total number of stem letter is less than 5, and any signature consisting of a single, one letter suffix are deleted; then only the signature for which the savings in letter counts is greater than 15 (where savings in letter counts is simply the difference between the sum of the length of words spelled out as a monomorphemic word and the sum of the lengths of the stems and the suffixes); 15 is chosen empirically.

Words	Mini-Lexicon 5	Stem	Mini-Lexicon 4	Mini-Lexicon 3	Mini-Lexicon 2	Mini-Lexicon
hinqabaanne	hin	qab			aa	nne
hinqabaanneetti	hin	qab		aa	nne	e
hinqabaatan	hin	qab			aa	tan
hinqabaatiin	hin	qab			aa	tiin
hinqabaatu	hin	qab			aa	tu
hinqabaman	hin	qab			am	an
hinqaban	hin	qab			a	n
hinqabani	hin	qab			a	ni
hinqabatin	hin	qab			a	tin
hinqabdu	hin	qabd				u
hinqabduu	hin	qabd				u
hinqabin	hin	qab				in
hinqabne	hin	qab				ne
hinqabnee	hin	qab			ne	e
hinqabneef	hin	qab			ne	ef

Fig 4: 3 Segmented words of Afaan Oromoo inclding prefix, steem ans suffixes

4.5.2.2. Designing the gold-stand with Alchemist

Alchemist is a GUI based tool for creating morphological gold-standards in XML* format. The main purpose of Alchemist is to allow reading in raw and text files and creating morphological gold-standards in XML format.

XML format is a highly structured way of organizing information so that it can be easily read by software. It has become the standard way to keep track of information in files. Using Alchemist one can identify morphemes, along with a number of characteristics of the morphemes, such as whether they are roots or affixes, the degree of analyst certainty and allomorphs of the morpheme.

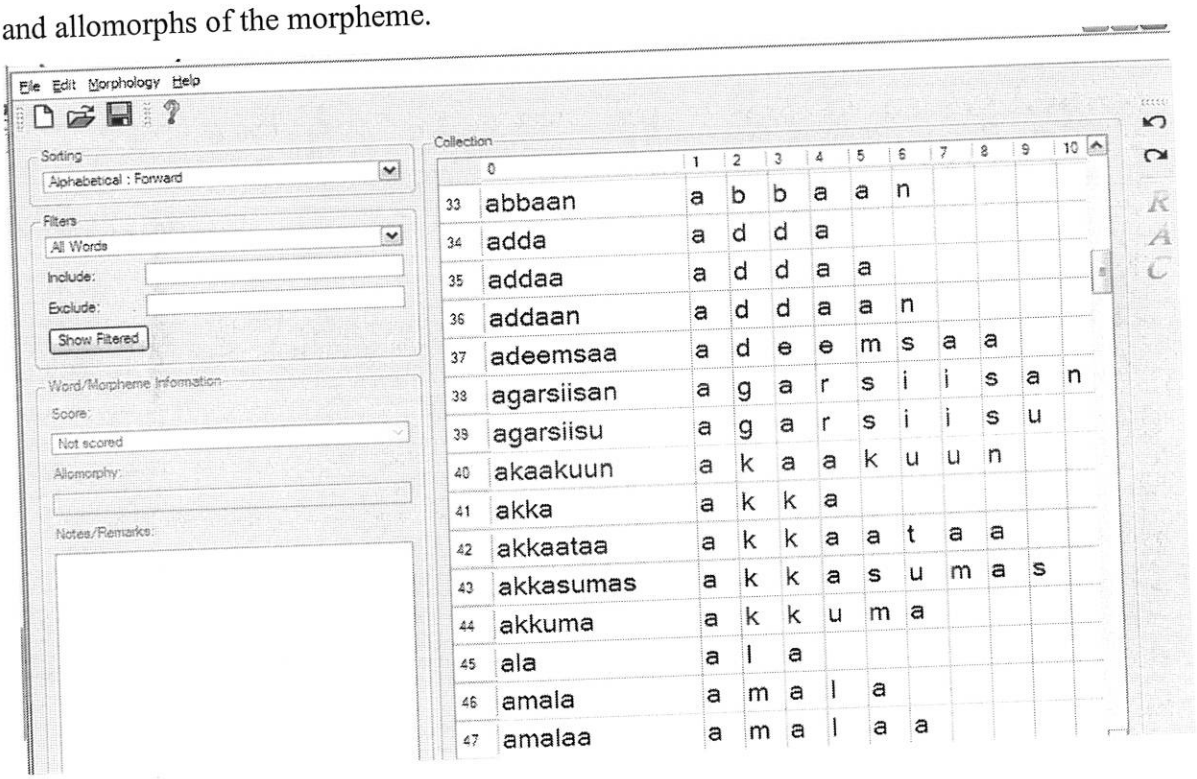


Fig: 4.4: A shot of the Alphabetical forward sorting order of the Alchemist

* XML(extensible markup language) is a meta mark up language that provide a format for describing a structured data. This facilitates more precise declaration of content and more meaningful search results across multiple platforms. In addition, XML will enable a new generation of web based data viewing and manipulation applications.

Alchemist facilitates rapid creation of morphological gold-standards. These standards can be used as performance rating tools for morphology learning applications. It can sort data either from left-to-right, right-to-left in ascending or descending order. Four sorting options in Alchemist are: ***Alphabetical forward, Alphabetical backward, Reverse alphabetical forward*** and ***Reverse alphabetical backward***. For example figure 4.4 the screenshot of the alphabetical forward sorting order of the Alchemist used in Afaan Oromoo words.

Alchemist can also filter data by limiting the sequence of letters searched for in the collection. For example if we want all words that contains the letters a particular affix such as – *sis, -ummaa,* then we enter that string in the filter box. It can then present with only the words containing that string, and we can more easily identify the words among them that contain the morpheme one is searching for. The screenshot indicated in figure 4.5 shows the search result of the words consisting the string *–sis*.

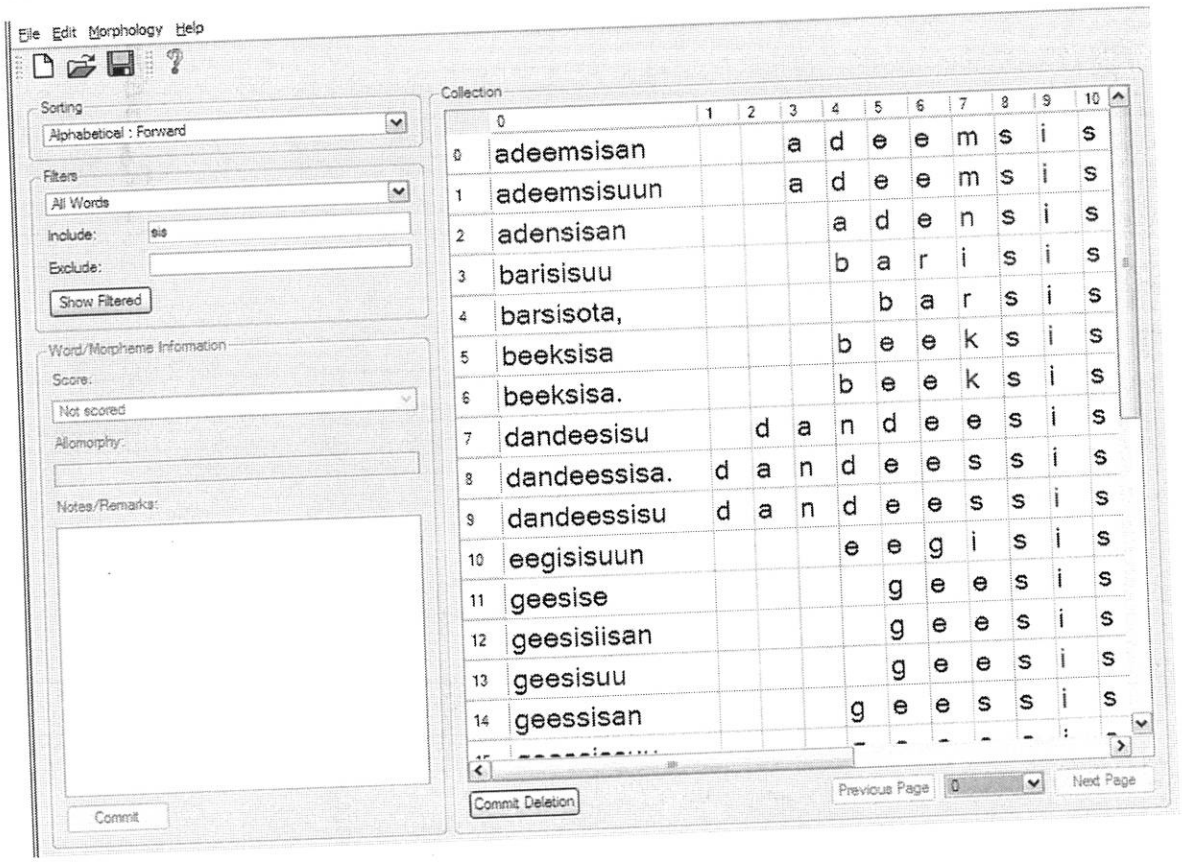


Fig: 4.5: A shot of the search result of the words consisting of the string *–sis*

CHAPTER FIVE

THE EXPERIMENT

5.1 Introduction

This chapter deals with the experimentation done using Lingustica_beta2 on the results of the preprocessed text and the gold-standard. The experiment is conducted using Linguistica_beta2. The process of experimentation and the prototype designed is presented in the following sections.

5.2 preprocessing

The text is given to the program for clearing natural text that consists of a number of different unwanted characters which are not considered during analysis. Every text consists of uppercase letters, lower case letters, punctuation marks, different signs and symbols, numbers etc. Before running such a text using Linguistica_beta2 all these characters are removed using the program written in C++ as discussed in section 4.2. The cleaned text is then given to Alchemist (a program used for creating morphological gold-standards -see section 4.5.2.2).

5.3 Test data

The texts colleted are from different institutions and various areas of disciplines so as to represent the language sufficiently. They are as indicated under sample A,B, and C.

Sample A

The first text is from Oromia Regional State Education Bureau. This is a report generated reporting about the structural organization of the institution. The words used in this text are terms of general purpose. They are not limited to a certain field of study or specialization. Since they are not subject dependent they can be used for generic purpose. It consists of 33,035 words.

Sample B

The second text was from Oromia Cultural Bureau. It consists of about 6,977 jargons of cultural and historical significancethat are used in different field of study.

Sample C

The third text is form the Institute of Language Studies. It consists of about 9,642 words.

The main reasons for the selection of various types of texts rely on:

- their availability in electronic form (all categories- Sample A-C)
- their multidisciplinary nature (the first and second categories-Sample A and B)
- their emphasis on morphology of the language (the third category -Sample C)
- they are form different disciplines, and it is believed that they (all categories) represent the language, and
- their literal quality get approval from linguists and experts from Education Bureau of the Oromia Regional State (all categories- Sample A-C).

Since the study is mainly on the morphology of the language, the text from the Institute of Language Studies of the Addis Ababa University could be more helpful. It consists of different examples in almost all parts of speech of the language, sample of stems and affixes in the language and issues related to the morphology of Afaan Oromoo. For comparison the actual size of the texts in terms of words is presented in Table 5.1.

Table 5.1: Sample texts used in the study

Sample category	**Total word count**	**Unique words**	**Source**	**Type-token ratio***
Sample A	33,035	6,832	OREB	4.8353
Sample B	6,977	1,777	ORCB	3.9263
Sample C	9,642	2,309	AAU	4.1758

* Word ratio is the ratio of total word count to unique words within the text

5.4 The experimentation process

Linguistica_beta2 as discussed in chapter four, section 4.5 of this study is used as the main tool to develop the morphological analyzer for Afaan Oromoo text in this study. It is used to identify different components of words within a text of a given corpus. These morphemic components produces a morphological dictionary called mini-lexicon which consists of another mini-lexicon with list of analyzed and unanalyzed words, list of stems, list of suffixes, signature, foreward trie, list of prefixes, total number of words read and distinct numbers of words read. It passes through various stages as discussed in the following sections.

The raw texts used as an input is not given to Linguistica_beta2 as it is. After clearing/removing the impurities(unwanted characters) using a program written in C++, the raw text data in different categories as listed in Table 5.1 are given to Linguistica_beta2. It analyzes them and generates a report. Accordingly the result is as summarized in Table 5.2.

Table 5.2 Summary of sample texts using only Linguistica_beta2

Sample category	Total word count	Unique words	Analyzed words in ML1*	Analyzed words in ML2	No. of suffixes	No. of prefixes	Signature	Stem suffixed
Sample A	33,035	6,832	4,378	145	164	54	361	2879
Sample B	6,977	1,777	379	27	40	23	58	279
Sample C	9,642	2,461	1,083	25	68	16	113	747

From the data in Table 5.2 one can understand that, given a text data to Linguistica_beta2, as it has the ability of analyze and generate some sort of information whether that information is correct or not for Afaan Oromo. As evidence, the analysis indicated by the screenshot of Sample A from Table 5.2 is as indicated in figure 5.1

* ML stands for Minilexicon which is generated by linguistica while running a sample text. ML1 consists of a list of suffixes, ML2 a list of prefixes.

below. In addition to the statistical information, it displays the prefixes identified with their descriptive information as well.

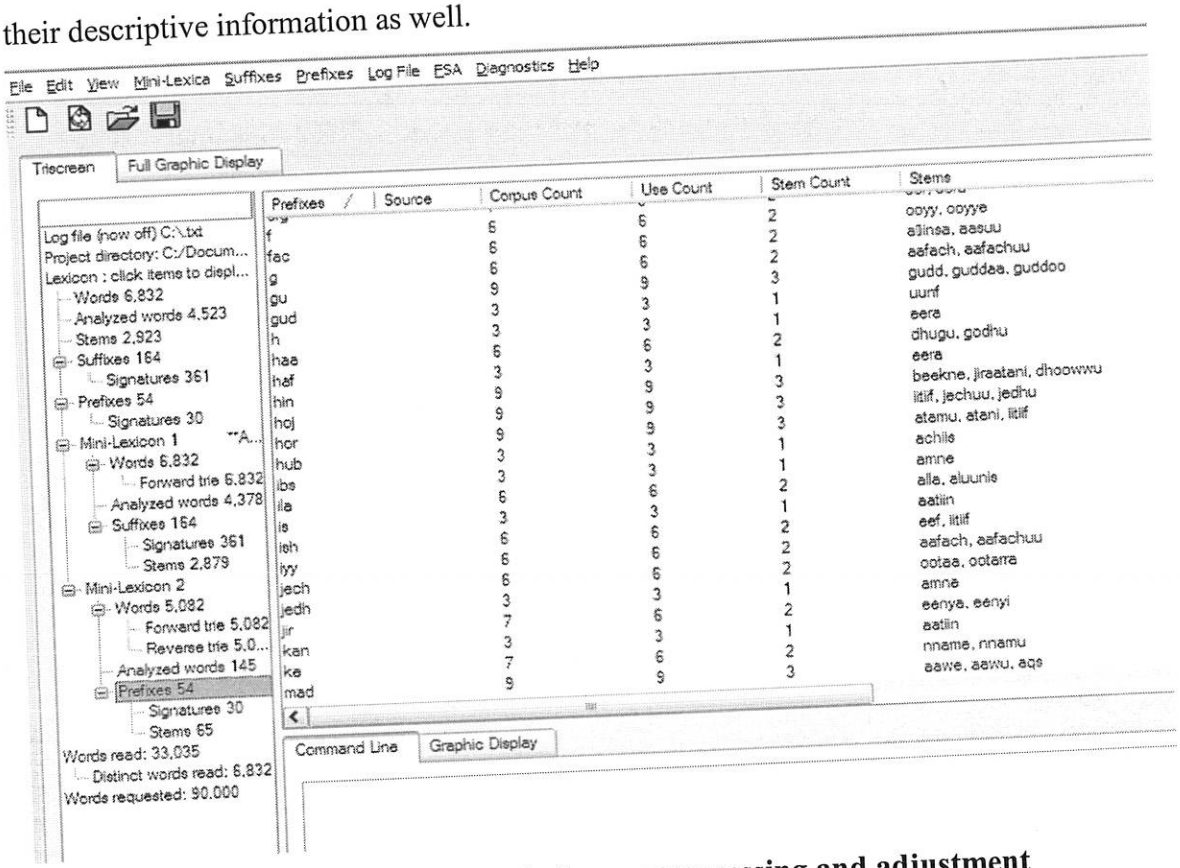


Fig 5.1 Screenshot of Sample A text before preprocessing and adjustment

From this one can understand that Linguistica_beta2 by itself is not sufficient to analyze words of Afaan Oromoo correctly. As discussed earlier throughout the study there is only one prefix ***h(in-)*** in Afaan Oromoo. But while running Sample A text as indicated in Table 5.2, 54 prefixes are wrongly identified. As can be seen from the screenshot more than 50 stems are reported wrongly as a prefix. In each sample above the prefixes identified by Linguistica_beta2 are more than one, ranging from 16 to 54 (see Table 5.2). Linguistica considers some stems as prefixes, others as suffixes. Therefore the number of stems, suffixes and prefixes indicated in Table 5.2 can not represent the components of words in the language. It needs some modifications and adjustments so as to analyze words of Afaan Oromoo correctly.

5.4.1 Improvements made on Linguistica_beta2

To solve the problem identified on the first run of Linguistica_beta2 the following modifications on the value of parameters are made:

- There are five diagraphs or consonant clusters (**ch, dh, ny, ph, sh**) in Afaan Oromoo that cannot be separated from each other when they occur within a word in this order. Each of these characters should be considered and read as a single consonant producing a single sound. Linguistica_beta2 cannot handle them as a single sound before modification is made to it. So these letters are passed to linguistica so as to be interpreted as one letter consisting single sound. This process is called character filtering.
- Since there is one prefix *h(in)* that most of the linguists agreed on, and reported in this study, its minimum and maximum length is set to three. This information is passed to Linguistica_beta2 so as to consider it. This is done by setting the value of the parameters of MaximumPrefixLength to 3 and value of MinimumPrefixLength also to 3.
- Concerning suffixes, after a repeated fine-tuning the result of the analysis, it becomes possible to come across the best value for relevant parameters for them.

These improvements, which are made in the Linguistica preferences, allow altering preferences so as to improve the quality of the analysis or to improve the readability of the program's output.

After making the necessary modifications and adjustments, (setting filtering, changing the value of the parameter of Linguistica_beta2, fine tuning different results of the analysis) each of the three sample texts are run using Linguistica_beta2 to see the effect of the improvements done on the performance of the analyzer. Accordingly the result is summarized in Table 5.3.

Table 5.3: Summary of the analysis after making some modification in Linguistica_beta2

	Total No. of words read	Distinct words read	Analyzed words ML1	Analyzed words ML2	Suffixes	Prefix	Signature	Stems suffixed
Sample A	32,772	6,734	4,313	219	158	2	341	3,027
Sample B	6,977	1,763	339	6	40	2	61	229
Sample C	9,642	2,309	636	11	59	2	88	439
Total	***49,341***	***10,806***	***5,288***	***236***	***257***	***6***	***490***	***3,695***
Merged (A+B+C)	48,497	8,834	6,176	259	197	3	468	3,921

In Table 5.3 the number of total words read is the sum of all samples in Table 5.2. As it can be seen from the table, the total number of distinct words of merged text (8,834) is less than the actual sum of the three sample texts (10,806) because some words may appear repeatedly in more than one sample. There are also few additional 'prefixes' such as (***ded-***) which are wrongly generated by the analyzer. *ded-* for example is generated from ***deddeebi'ee*** 'come again and again'. This might be because the word ***deddeebi'ee*** is formed by reduplicating the initial morph.

The result of the analysis can also be saved for further reference. Accordingly one can have a minimum of thirteen (if only one iteration exists) readable text files that can be viewed using WordPad. As the number of iteration increases the number of readable text files that one can save also increases. Figure 5.2 shows one of the files (general summary of the analysis) saved while running ***Sample A*** after the necessary enhancements are made.

```
LEXICON ('sample')
      Number of mini-lexica: 5
      Number of word tokens: 32772
      Number of word types:  6734
-----------------------
MINI-LEXICON 1
      Number of words:                                  6734
      Number of stems:                                  2835
      Number of regular suffixes:                       160
      Number of signatures with regular suffixes:       336
-----------------------
MINI-LEXICON 2
      Number of words:                                  5017
      Number of stems:                                  682
      Number of regular suffixes:                       47
      Number of signatures with regular suffixes:       95
-----------------------
MINI-LEXICON 3
      Number of words:                                  4401
      Number of stems:                                  79
      Number of regular suffixes:                       15
      Number of signatures with regular suffixes:       15
-----------------------
MINI-LEXICON 4
      Number of words:                                  4321
      Number of stems:                                  9
      Number of regular suffixes:                       3
      Number of signatures with regular suffixes:       2
-----------------------
MINI-LEXICON 5
      Number of words:                                  4312
      Number of stems:                                  183
      Number of regular prefixes:                       2
      Number of signatures with regular suffixes:       2
```

Fig. 5.2: Summary of Sample A text after improvement

As can be seen from figure 5.2 there are five mini-lexicons. In the scheme a set of highrarchically ordered mini-lexicon comprise a lexicon. Most information in the given figure is self explanatory. The first four lexicons are for analysis of suffixes generated from the text, while the last one consists of information about suffixes. One layer of morphology can be found using a mini-lexicon by populating the 'words' collections with stems and/or unanalyzed words. It is used to derive a new signature, stems, and affixes. The stems and unanalyzed words from this (the first mini-lexicons) can then iteratively be 'inherited' by a new mini-lexicon which can use the same methods to find another layer of affixes. In this way, layers of affixes are removed from words and stems and the resulting stems are re-analyzed using the same strategies to find the next layer this

process is continued until no new affixes and/or stems are found in the latest mini-lexicon. So, in the figure the fourth mini is the latest for suffixes (four iterations for suffixes), while the fifth one is the only mini-lexicon that handles information on prefixes (one iteration to generate prefixes after analyzing all the suffixes).

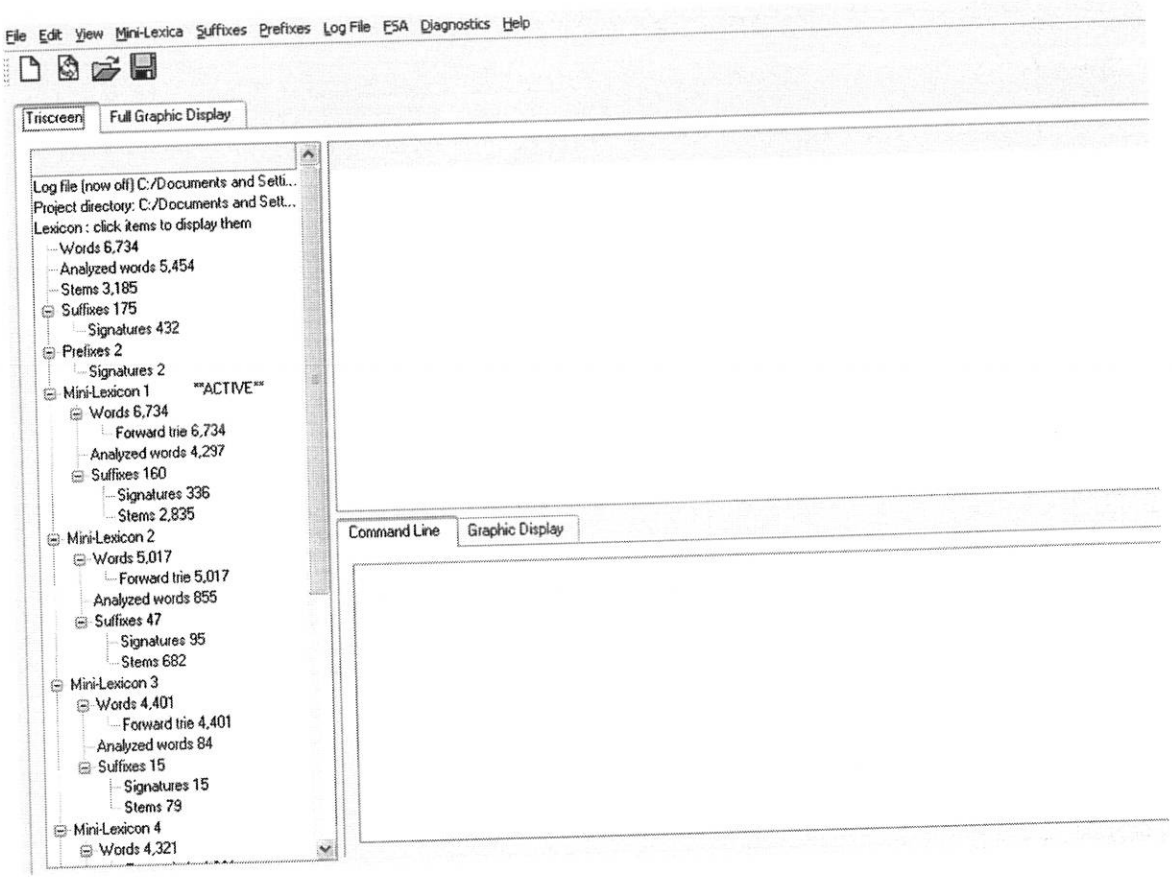


Figure 5.3: A screenshot of the tree area of sample A text

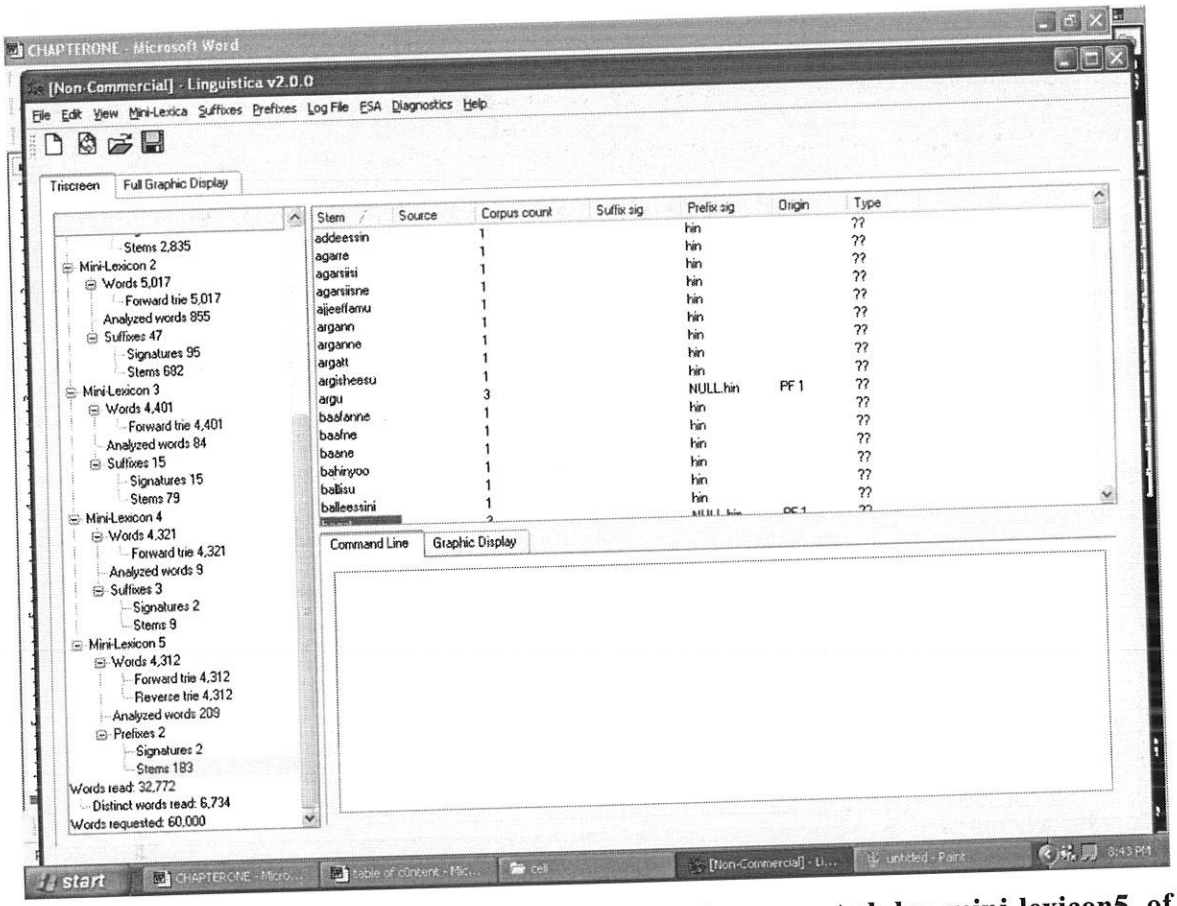


Figure 5.4: A screenshot of the analyzed words generated by mini-lexicon5 of sample A text

5.4.1.1 Another corpus size

All the three sample texts were merged together to get a corpus of big size. The total size of the corpus now becomes 48,497 words. In this text there are 8,834 distinct words. After analyzing the merged text the result of the analysis is recorded as follows.

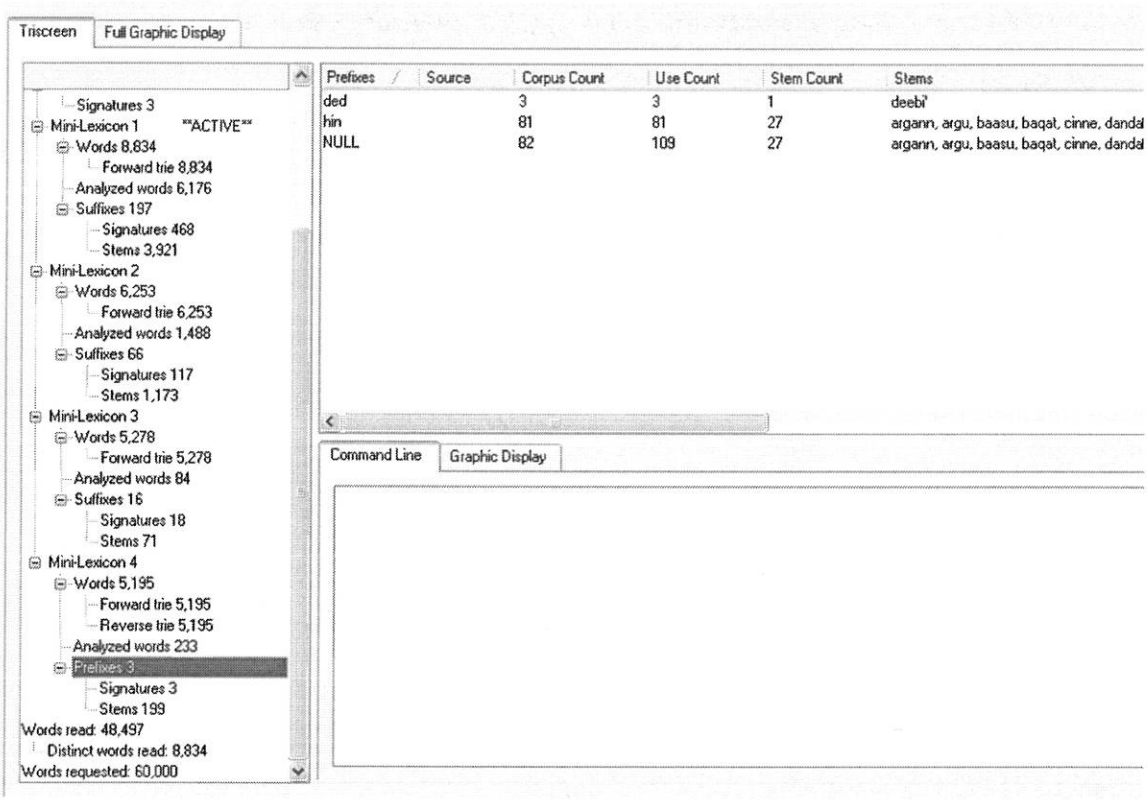


Fig 5.5: Screenshot of mini-lexicon of the merged text of all three sample texts

As can be seen from the screenshot of merged file the number of words raised to 48,497. But the number of iterations which were 5 while running a text of 32,772 words in the case of Sample A is decreased to only three to analyze suffixes. This can save both running time and storage space for saving readable text files. From this one can observe and come in to a conclusion that Linguistica_beta2 needs a large corpus size so as to give quality (improved) analysis. Summary of the merged text that consists of four mini-lexicons is as presented in figure 5.6. The number of regular prefix here is three because of the '*ded-*' string which the system considered as prefix wrongly.

```
LEXICON ('merg')
      Number of mini-lexica: 4
      Number of word tokens: 48497
      Number of word types:  8834
------------------------
MINI-LEXICON 1
      Number of words:                              8834
      Number of stems:                              3921
      Number of regular suffixes:                   197
      Number of signatures with regular suffixes: 468
------------------------
MINI-LEXICON 2
      Number of words:                              6253
      Number of stems:                              1173
      Number of regular suffixes:                   66
      Number of signatures with regular suffixes: 117
------------------------
MINI-LEXICON 3
      Number of words:                              5278
      Number of stems:                              71
      Number of regular suffixes:                   16
      Number of signatures with regular suffixes: 18
------------------------
MINI-LEXICON 4
      Number of words:                              5195
      Number of stems:                              199
      Number of regular prefixes:                   3
      Number of signatures with regular suffixes: 3
```

Fig. 5. 6: Summary of the merged text

5.4.2. The development of gold-standard

The gold-standard which can be used either with Alchemist itself or with Linguistica is developed. It allows Linguistica to automatically test itself to see how well it is doing with its analysis. In addition to this, it gives a room for sorting words within a text in various ways (ascending, descending, foreword and backward order). Furthermore the gold-standard helps to segment words into morphemes. This includes decomposing a word into a stem and affixes (if it has any). If a word does not have any affix it gives zero affixation (NULL).

The gold-standard developed for this study consists of about 1,600 words. It is saved in XML format for the purpose mentioned earlier in section 4.5.2.2. Its accuracy gets approval by linguists. It can be used with Linguistica_beta2 either to analyze the text it was designed from or to analyze other text documents in the language. It can be run at any time using Linguistica to compare results between the Linguistica_beta2 analyses with the designed gold-standard. The result of the comparison can be opened with WordPad. The following section presents the result of the experiment while using the developed gold-standard.

In the experimentation, the researcher has taken corpora as small as 6,083 running words to as large as 48,497 running words (token count rather than type count). Accordingly, from the experimentation it is observed that good result typically comes from corpora of larger size. This result is checked automatically by comparing the output of the analyzer with the gold-standard which is generated automatically and can be read using WordPad.

5.5. Using the designed gold-standard to test the system

In this phase the analyzer was tested using the designed gold-standard. The program performs unsupervised learning in the sense that the program's sole input is the corpus. The program requires preprocessed text. It does not need dictionary and no morphological rules particular to the language. Since its goal is restricted to provide the correct analysis of words into component pieces (morphemes) with a rudimentary categorical labeling.

The performance of the system is evaluated using the gold-standard that consists of 1600 words. By running different texts the number of correctly and incorrectly analyzed words in each sample is observed. This was done by comparing the result or the output of Linguistica_beta2 against the developed gold-standard. The result is summarized in Table 5.4 as follows:

Table 5.4 Performance rating of the system using the gold-standard

SAMPLE TYPE	Size of corpus	Types of prefixes reported	Maximum No. of mini-lexicon	Match with the gold-standard in %	Amount of error recorded in %
A	32,772	2	4	84.5	5.5
B	6,977	2	3	68.1	21.9
C	9,642	2	3	72.4	17.6
MERGED	48,497	3	4	92.8	7.2

The achieved accuracy level seems satisfactory. As can be observed from Table 5.4, the performance of the system is more satisfactory as the size of the corpus increases and the amount of error generated decreases. This is because of Linguistica_beta2, which is based on unsupervised learning, requires large data to learn better. In addition to the smallness of the size of the corpus the following are also among possible means's for the occurrence of errors and the difference observed in Table 5.4:

- The availability of miss spelt words in the text, i.e. spelling errors of words
- Confusion of glottal stop " ʿ " (hudha) that is considered as a consonant with single quotation mark
- The inconsistency of some words in the text
- The error in the gold-standard itself
- Improper modification of some of the values of parameters during modification

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusion

Among different strategies to develop morphological analyzer for Afaan Oromoo text, the unsupervised learning of natural language processing system was employed in this study to perform automatically a morphological classification of arbitrary word form.

The central idea of unsupervised learning of natural language is that, given a corpus as an input to the system, it analyzes each word in the text and generates a report. The system uses neither dictionary nor morphological rules particular to the language.

The complexity of morphology of Afaan Oromoo at different levels was discussed in detail. Accordingly, the pattern of word formation including inflection, derivation, and the formation of compounds in the language was presented. Some complications are observed in the process of affixation. Because of such complexity of the language, conflating words manually become cumbersome and needs to develop an automatic morphological analyzer using unsupervised learning of natural language processing. Accordingly, the algorithm for the unsupervised learning of morphology was adopted from Goldsmith (2001a) for this study.

Few errors are reported during the experimentation process. Some of the possible means's for the occurrence of these errors includes miss spelt words in the text, confusion of glottal stops(') with single quotation mark by the system, inaccuracy in the construction of the gold-standard, inconsistency of some words in the text and the like.

A corpus of different size ranging from 6,977 to 48,497 is used in the experimentation process to test the system. A gold-standard of small size (1600 words) is also developed to evaluate the performance of the system.

The result obtained raised from 68.1% to 92.8% for a corpus of 6,977 and 48,497 respectively. This indicates that as corpus size increases the accuracy of the system will also increase, which is quite satisfactory result.

6.2 Recommendations

The developed morphological analyzer for Afaan Oromo text that uses Linguistica_beta2 as its basic tool is not a full-fledged one. It needs further improvements and research so as to get a robust analyzer that performs better. The performance of the current analyzer is satisfactory. The following recommendations were forwarded so as to extend the efficiency of the analyzer.

- The corpus which is used to design the gold-standard for this study is of a limited size. The accuracy reported is based on this small size corpus for the test data. Since unsupervised learning of natural language requires large corpus size, more corpus size in the gold-standard is needed for developing an improved morphological analyzer.

- An attempt is taken to include most of the affixes used in the language. But one can identify and come across the affixes that are not considered in the study or disprove the existence of some of the affixes currently used in the study.

- A number of irregularities (such as the prefix ***hin*** is assimilated to a stem or written as a separate word or found even mixed in some texts) have been observed in sample texts. Furthermore, *hin* is used interchangeably with *him* and *ni* within

the text. Therefore, more research has to be done to overcome such irregularities by standardizing the words in the language.

- Morphological analyzer is not an end by itself. It is one of the components that serve for natural language processing. Therefore, further study should also be done focusing on developing systems such as parts of speech tagger for Afaan Oromoo text. In addition, the development of morphological analyzer using other approaches (for example rule-based) for comparison of the results obtained with that of the unsupervised learning approach used in this study is also another research area.

- Because of the limitation of Linguistica_beta2 that is used in this study as main tool for decomposing words, it was not possible to identify categories of word classes in the language. Therefore, more research has to be done to identify categories of word classes in the language.

- Some other mechanism should be investigated to solve the problem of few 'prefixes' (such as ***ded-***) which are wrongly generated because of the reduplication of the initial morph.

- The procedure followed in this study to develop Afaan Oromoo text can be adopted in developing morphological analyzer for other local languages that use the Latin alphabet.

- Furthermore the output of this study can be used in developing application tools such as spell-checkers, dictionary compilation, etc for Afaan Oromoo.

- Further research should be undertaken to develop a full-fledged morphological analyzer for Afaan Oromoo text.

References

Abera Nefa. Long Vowels in Oromoo: a Generative Approach. Addis Ababa University (MA Thesis), 1988.

Abera Nefa. Oromoo Verb Inflection. Addis Ababa University (Senior essay), 1982.

Abiyot Bayou. Developing Automatic Word Parser for Amharic Verbs and their Derivation. Addis Ababa: AAU (SISA MSIS Thesis), 2000.

Akkadaamii Afaan Saboota Itoophiyaatiin Ministeera Beeksisaa fi Aadaā. Galmee jechoota Afaan Oromooo. Finfinnee: Akkadaamii Afaan Saboota Itoophiyaatiin Ministeera Beeksisaa fi Aadaa, 1996.

Allen, James. Natural Language Understanding. 2nd ed. California: The Benjamin/Cummings publishing company, 1995.

Anderson, Stephen R. A-Morphus morphology. Cambridge: Cambridge University press, 1992.

Antworth, Evan L. "PC-KIMMO a two-level processor for morphological analysis" available at http://www.sil.org/pckimmo/about_pc-kimmo.html

Arnoff, Mark. Word formation in generative grammar. Cambridge, MA: MIT press, 1976.

Baye Yimam. Oromoo Sustentative: some Aspects of their Morphology and Syntax. Addis Ababa University (MA thesis), 1981.

Baye Yimam. The Phrase Structure of Ethiopian Oromoo. School of Oriental and African University of Landon, 1986.

Curriculum Design for Undergraduate Programs Department of Afaan Oromoo at Alemaya and Jimma Universities. (Unpublished), Adama, 2002.

Diriba Megersa. Development of an Automatic Sentence Passer for Afaan Oromoo Texts. Addis Ababa: AAU (SISA MSIS thesis) 2002.

Ermias Ababe Kassa . Recognition of Formatted Amharic Text using Optical Character Recognition (OCR) Techniques, 1998.

Gelbukh, Alexander and Sidorov, Grigori "Morphological analysis of inflective languages through generation", ***Procesamento del lenguaje natural*** No.2: 105-111, 2002.

Girmay Berhane "Word Formation in Amharic". ***Journal of Ethiopian Language and Literature*** No. 2, 1992, 50-74.

Goldsmith, John. "An algorithm for unsupervised learning of the morphology"2000b also available at: http:/www.humanities.uchicago.edu/faculity/goldsmith/

Goldsmith, John. "Unsupervised learning of the morphology of natural language", ***Computational linguistics***. 27(2): 153-198, 2000a also available at http:/www.humanities.uchicago.edu/faculity/goldsmith.linguistica2000/paper.htm

Grimes, Joseph E. Affix positions and occurrences: the PARADIGM program. Dallas: Summer Institute of Linguistics, Inc., 1983.

Grishman, R. Natural Language Processing. ***Journal of American Society for Information Science*** 35 (5): 1984; 291-296.

Huang, Xuedong; Acero, Alex and Hon, Hsiao-Wuen. Spoken language processing: a guide to theory, algorithm and system development. New Jersey: Printice Hall PTR, 2001.

Jurafsky, Daniel and Martin James H. Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition, New Jersey: Prentice Hall, 2000.

Katamba, Francis. Morphology. London: Francis Katamba, 1993.

Kebede Hordofa. Nominilization patterns in Oromoo. Addis Ababa: Addis ababa University, 1981(Unpublished BA Thesis)

Kibur Lisanu Wudineh. Design and Development of Automatic Morphological Synthesizer for Amharic Perfective Verbs,2002.

Kinfe Tadesse Mengistu. Sub-Word Based Amharic Word Recognition: An Experiment Using Hidden Markov Model (HMM), 2002.

Kramsky, Jiri. The word as a linguistic unit. Paris: Mouton, 1969.

Krushkov, Hristo. "Automatic morphological synthesis and analysis for Bulgarian language" available at http://www.uni-plovdiv.bg/hdk/morfeng.htm

Mesfin Getachew. Automatic part of speech tagging for Amharic Language an experiment using Stochastic Hidden Markov (HMM) approach, 2001.

Mudee, Mahdi Haamid. Hamid Muudee's Oromoo Dictionary. Vol. I, Atlanta, Georgia: Sagalee Oromoo Publishing Co., Inc., 1995.

Oromoo Orthography. ***Bariisa*** Vol.1 No.1, September 1995.

Pao, Miranda Lee. Concepts of Information Retrieval Englewood. Colorado: Miranda Lee Pao, 1989.

Payton, George. Orma grammar report, 1989 (unpublished)

Salton, Gerard. Introduction to modern information retrieval. New York: McGraw Hill, 1983.

Silzer, P. Morphology. Available at: http://www.people.biola.edu/faculty/peter/linguistics/morphology.htm

Stoomer, Harry. A Grammar of borana Oromoo (Kenya): phnology, morphology, vocabularies. Köln:Köppa, 1995.

Stoomer,Harry. A comparative study of three southern Oromoo dialects in Kenya: phonology, morphology and vocabulary. Köln:Köppa, 1987.

Syzmanec, Bogdan. Introduction to morphological analysis. Warszawa:Panastwowe Wydawnictwo Naukaowe,1989.

Takkele Taddese. "Are s' (o) and t' (m) variants of Amharic variables? : A sociolinguistic analysis" Journal of Ethiopian Language and Literature No.2, 1992,103 - 121.

Temesgen Negassa. Word Formation in Oromoo. Addis Ababa University: Addis Ababa, 1993.

Tesfaye Bayou. Automatic Morphological Analyzer for Amharic Text: an Experiment Employing Unsupervised Learning and Auto-Segmental Analysis Approaches (Master Thesis at SISA), 2002.

Tilahun Gemta "Oromoo-English Dictionary: Problems of Indicating Pronunciation using the Amharic Syllabary" ***Proceedings of the Ninth International Congress of Ethiopian studie***s. Vol. 5, Moscow: NAUKA Publisher, 1988 131-137.

Tilahun Gemta. "Afaan Oromoo", ***Journal of Oromoo Studies, 1994*** also available at: ***http://www.sas.upenn.edu/african.studies/Hornet/Afaan_Oromooo_1977 7html***

Tolamariyaam Fufaa. Seera Afaan Oromoo: sagalee, dhamsaga, latoo fi caasima. Finfinnee, 1992 (unpublished).

Van Rijsbergen, C.J. Information Retrieval 2nd ed. Landon: Butterworth's 1979.

Wakshum Mekonnen. Developing Stemming Algorithm for Oromoo Texts (MA Thesis) 2000.

APPENDIX I

AN ALGORITHM FOR THE UNSUPERVISED LEARNING OF MORPHOLOGY

(Adopted from Goldsmith, 2001b)

1. **Get a corpus from a text file**
2. **Find successor frequency** (to cut the words into two pieces – stem and suffix)
   - **Discover morpheme peaks in successor frequency**
   - **Cut words into pieces**[***] (only those words which can be decomposed)

Organize the pieces into signature put in respective list (stem, suffix)

Else go to 8

- **create a stem list**
- **associate each suffix list with corresponding set of stems to form signatures**
- **alphabetize the lists**
- **eliminate certain implausible signatures**
- **prefer accuracy over recall**
- **set a threshold for minimum number of words an affix may appear in**
- **eliminate signature based on the apriori likelihood of the suffix**

[***] A word consists of an obligatory root (or indivisible stem) preceded by zero or more prefixes and followed by zero or more suffixes. In the case of Afaan Oromoo there is only one prefix (*hin*) a negative maker and many other suffixes. These are not perfectly identified by Linguistica, because it is designed for European Languages. By changing parameters and designing the gold-standard that can be read by Linguistica, further analysis is simplified.

- set maximum number of stem permitted, minimum number of affixes
- find the robustness of signature

3. Check signatures
   - incorporate the insights of MDL
   - compute the entropy of the set of stem-final letters
   - if entropy > threshold value do nothing
   - if entropy < threshold value
     - set entropy = threshold
   - set entropy of stem-final strings
   - calculate the change in the morphology's description length of an individual signature
   - consider alternative signatures
   - consider the length of the pointers to stems
   - construct a list of all places in the morphology
- If no new signature occurs independently, calculate the relevant parts to its DL, (i.e. length of its pointers to its individual suffixes)

4. Extend known stem to known suffix
   -If suffixes began with the same letter, scan through list of discovered stems
   -If unanalyzed words appear
   -Unanalyzed and divide into stem and suffix
   -If similar unanalyzed words occur choose the one with more common stem

5. Extend known signature
   - sort the signature
   - look for the most robust first

-if the signature matches with set of words analyze the words into stem and suffix

6. Extend known stems

-If stems analyzed help in finding new suffixes, consider each stem and the signature

7. Extend known suffixes ("loose fit")

- use previous knowledge to get more new stems and signatures
- look at unanalyzed words
  - If found divide it into stem and suffix,
  - If it is already recognized signature,
    - accept the new stem else calculate the effect

      If effect decreases DL

      - accept new signature and new word divisions

    If cost of new analysis < cost of current analysis,

    - select new analysis
    - Concomitant word-analysis

8. Find singleton signatures

If a word appears only once in a corpus

If it ends with known suffix (including NULL)

- compute the prob. of each stem length
- take the unanalyzed words to set of stems with NULL suffixes

9. Detect rules of allomorphy

Repeat 2 to 9 until EOF

10. End

APPENDIX II

STEMS OF SOME COMMON AFAAN OROMO VERBS AND THEIR RESPECTIVE BASE SUFFIXES

{a, e, ta, ti, tani, ani, te, tu, na, nu, ne, i, ini, uu, sise, sisa, sisan, siste, sisna}

aan	falam	ilaam	quuf	urguf
af	faxuul	kitim	quphan	uum
aggaam	fuf	kuf	raf	waam
akeek	funaan	kurruuf	roorr	waraan
alaak	gaafat	kut	rukut	warroom
baas	gan	laaqam	saam	xalaf
aman	godaan	lubbam	saf	yaam
atoom	goom	luf	seen	dhaam
baan	guduunf	lugaam	soddoom	dhaan
baat	hamat	magan	somsom	dhab
ban	hammaar	mak	soof	dharoom
beek	haraam	mammaak	soom	dheekkam
biif	haram	masak	sokkuum	dhiit
birat	harq	maseen	suuf	dhidhim
birmat	har	miidham	tiif	dhuf
booji	heerum	moo'	tilmaam	dhum
cuf	hiik	moof	tim	dhuuf
daak	him	muddam	tuf	shaf
deem	hinaaf	namoom	tum	shallam
doom	hooggan	of	udaan	sharaf
dorgom	hordof	oof	udum	shuqun
dulloom	hubam	ordof	ukam	shuum
duroom	huum	qaam	ukkaamam	
ergif	id	qeeqam	un	
eyyam	if	qot	unkut	

{ata, ate, atti, atani, ani, atte, atu, attu, anna, annu, anne, adha, adhe, adhu, atini, adhuu, achise, achisa, achisan, achiste, achispa}

bayyan ka'	naafmar'	rakk	seer

{a, e, da, di, dani, ani, de, du, na,nu, ne, i, ini,uu, sise,sisa, sisan, siste, sisna}

araad
arrab
bad
barbaad
bareed
barood
calqab
dalag
deeg
dib
did
dig
duud
duug

eeg
fag
fid
findig
gammad
gob
gog
goob
gub
hagab
hag
hareed
herreg
jig

kaab
kabaj
kab
katab
kijib
koob
luug
midhaag
mug
muud
nagad
ood
qab
qadaad

qimmiid
qood
qoqqob
raag
ramad
reeb
rig
roob
sabab
sassaab
sosob
taajjab
utub
wag

waraab
wareeg
yaad
dhaab
dhaqqab
dheed
dhib
dhiib
dhimbiib
dhug
dhukkub
dhuub
shallag
shigid

{a, e, ta, ti, ani, tani, te, tu, nna, nnu, ne, dha, dhe, dhu, ini, uu, chisise, chiisisa, chisisa, chisisan, chisiste, chisisna}

bilchaat
bit
boqat
bulgaafat
but
dubbat
gabbat
galaafat
gangalat
gat
guddat
gufat

guut hat
hedat
hinqirfat
hisat
hojjet
injifat
jaal
jabaat
jallat
jifat
kakat

kadhat
komat
kottonfat
laat
lagat
liqeeffat
lulluqqat
marat
marxifat
mul'at
obbaaffat
odeeff

onnat
owwaat
qalbifat
qusat
raajefat
rifat
riiqat
salfat
salphat
sassat
simat

sodaat
tarkaanf
taphat
tirat
toohat
tuffat
uffat
ulfaat
warwaat
wayyat
yaabbat
yarat

dhaamot
dheebot
dheerat
dhihaat
dhiphat
dhungat
dhuunfat
nyaat

{a, e, da, di, dani, ani, de, du, na, nu, ne, i, ini, uu, se, sise, sisa, sisan, siste, sisna}

cab coolag xoolag dhagoog dhiig

{a'e, eta, eti, etani, a'ani, ete, etu, enna, ennu, ene, a'i, a'ini, a'uu, esise, esisa, esisan, esiste, esisna}

dand

{a, e, ita, iti, itani, ani, ite, itu, ina, inu, ine, i, ini, uu, isise, isisa, isisan, isiste, isisna}

akkeess	borc	dirm	elm	gadd	karkars
alalch	burkuteess	dirq	erg	gonf	marq
arg	callees	durs	faal	hundeess	mirg
balf	cich	duumess	falm	irb	sokk
bobees	danf	duuch	gaabb	iyy	

{a, e, ifta, ifti, iftani, ani, ifte, iftu, ifna, ifnu, ifne, i, ini, uu, ise,isa, isan, ifte, ifna}

foosis

{a, e, ita, iti, itani, ani, ite, itu, ina, inu, ine, i, ini, uu, ,ise, isa, isan, iste, isna}

balaalees	himims	mimirs	sarm	xill
cururs	kireess	mork	sifees	xinneess
fafeess	kolf	nuff	siileess	xiqqeess
fagees	labs	obs	silliks	yakk
fakkeess	liqims	odeess	sirb	dheeff
fayy	luuccess	qirc	sirribs	dheess
foks	mars	qirqirs	soneess	dheedhess
gaggabs	madeess	qixxeess	summeess	dhiyeess
geess	mallatteess	qoors	suuleess	dhooww
gingilch	maqs	qulqull	suph	dhork
gondolch	mars	qurc	tiks	shakk
gors	maxxans	qurx	tinf	shokoks
gowwoms	mi'eess	raabs	tirs	
gumaach	milk	sams	wadda	
heddummeess	mindeess	sang	waaqeess	

{a, e, ta, ti, tani, ani, te, tu, na, nu, ne, i, ini, uu, ise, isa, isan, iste, isna}

albaas baas	balis	badhaas	booress	buus
bakkis	barreess	bobbas	borxoxxeess	
			bukeess	

calaas
calqqis
cee'
daars
dabars
dabs
dagsis
dayyas
dimimmis
doorsis
eebbis
facaas
fannies
fokkis
fullaas
furors
futtaas
gabaas
gombis
hafars
hambis
haqqis
harkis
harrabs
hadhees
haphis
hebbis
huqqis
hurris
ibs
kaas
kiis
kolaas
korris
kuus
laaf
leellis
loos
luqqis
malaas
marmars
misoom
mitt
moggaas
obaas
qoraas
quncis
saffis
taasis
teessis
tus
us
uwwis
wallis
weeddis
weellis
wuxxis
yaas
yuus
dhiis
dhimmis
dhis
dhiis
dhoos
dhukaas
shullis

{a, e, ta, ti, tani, ani, te, tu, la, lu, le, i, ini, uu, ,sise/chise, sisa/chisa, sisan/chisan , siste/chiste, sisna/chisan}

afeel
aqanqaal
awwaal
bilbil
bul
dabal
daldal
duul
eegal
falfal
fincil
hokkol
huddeel
kal
kuul
lol
madaal
ofkal
qal
qol
sakaal
saqal
tul
tajaajil
uddeel
ungulaal
utaal ushaal
wallal
wanjal
xalal
xiinxal yaal
dhaaldhal
shaakal

{a, e, ta, ti, tani, ani, te, tu, la, lu, le, i, ini, uu, , che, chisise, chisisa, chisisan, chisiste, chisisna}

caal
calal
eel
gal
gargal
haleel
haqanqaal
hoffal
ilaal
okkol
ool
qajeel
qubeel
tol
waxal
shallal

{a'e, ofta, ofti, oftani, a'ani, ofte, oftu, ofna, ofnu, ofne, a'i, a'ini, a'uu, ese, esise, esisa}

beel
bush
cinc
finc
lakkaa'

mad	mi'a	orbobb	qabbana'
mach	misoom	qaana'	

{a, e, ta, ti, tani, ani, te, tu, ra, ru, re, i, ini, uu, se, sise, sisa, sisan, siste, sisna}

aar	goror	kur	qaxxaamur	tortor
agur	gor	maagar	qor	tur
bar	gurgur	magar	quucar	unkur
bohaar	guur	makkaraar	rar	ur
cir	hafuur	mar	safar	waqar
darar	handhuur	marar	sakar	wuxir
deegger	hir	miciir	sarar	xumur
diriir	hiriir	mur	seekkar	xuruur
dongor	ijaar	qaar	soor	
gabbar	jir	qancar	sor	
gargar	kasaar	qar	suntuur	
gatantar	koor	qarqaar	taraar	
geerar	kor	qaxar	tar	

{a, e, ta, ti, tani, ani, te, tu, na, nu, ne, i, ini, uu, se, sise, sisa, sisan, siste, sisna}

bokok	caam	ergis	sham
boon	ciis	gadoom	shoom
buu	cim	waldhaan	

{a, e, ta, ti, tani, ani, te, tu,na, nu, ne, i, ini, uu}

cit	duu'	dut

{a, e, xa, xi, xani, ani, xe, xu, na, nu, ne, i, ini, uu, sise, sisa, sisan, siste}

cuuph	hooq	qarax	soq	xuux
fix	hulluyq	saaq	suuq	dhaq
habbuuq	lix	sassaaq	tuq	dhiq
haq	muux	solloq	xax	

{a, e, xa, xi, xani, ani, xe, xu, na, nu, ne, i, ini, uu, se, sise, sisa, sisan, siste, sisna }

baqaq	camad	miliq	quuq	warraaq
bulluq	carraaq	munyuuq	siiq	dhommoq
burq	macalaq	naq	sunquuq	shuruq

APPENDIX III

SUFFIXES (WORD ENDINGS) OF AFAAN OROMO

a
a'ani
a'e
a'i
a'ini
a'u
a'ut
aa
aa'u
aachi
aachu
aadh
aadhe
aaf
aaf
aafi
aaiif
aam
aan
aani
aann
aannu
aata
aatan
aate
aati
aati
aatte
aatto
aattu
aatu
aaw
aawaa
aawwan
acha
achaaa
achi
achii
achiis
achis
achisa
achisan
achise
achisna
achiste
achu
achuf
adh
adha
adha
adhe
adhee
adhu
adhuu
am
ama
amaa
amaa
amaan
aman
amanii
amarr
ame
amee
amee
ameh
amen
amett
amin
amiss
amn
amne
amni
amo
amoo
amt
amta
amtan
amte
amti
amto
amtu
amu
amus
amut
amutt
amuu
amuuf
an
ani
anii
aniif
aniin
aniir
anin
anir
anis
anitt
anna
annaa
anne
anni
annoo
annu
annuu
anuu
as
asin
at
ata
ataa
ataa
ataan
atam
atan
atani
ate
atee
atin
atinn
att
atte
attee
atti
atto
atto
attu
atu
atus
atuu
atuuf
aw
awa
awwaa
cha
chisa
chisan
chise
chisisa
chisisan
chisise
chisisna
chisistan
chisiste
chisna
chiste
chu
da
dani
de
dh
dha
dhaa
di
du
e
echi
echu
ee
eef
eeff
eefi
een
eeni
eenn
eeny
eenya
eenyi
eenyu
ees
eesi
eess
eessa
eessi
eet
eeti
eetii
eett
eeyyu
en
ene
eni
enna
enne
ennu
es
esisa
esisan
esise
ess
etam
etan
ete
etee
ette

DECLARATION

This thesis is my original work and has not been submitted for a degree in any other University.

Assefa Woldemariam Jegeno
July, 2005

This thesis has been submitted for examination with our approval as University advisors.

Dr. B. RAMA KRISHNA RAO
JULY, 2005

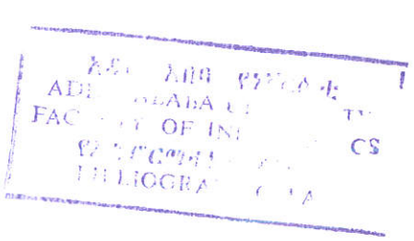
DECLARATION

This thesis is my original work and has not been submitted for a degree in any other University.

Assefa Woldemariam Jegeno
July, 2005

This thesis has been submitted for examination with our approval as University advisors.

Dr. B. RAMA KRISHNA RAO
JULY, 2005



DECLARATION

This thesis is my original work and has not been submitted for a degree in any other University.

Assefa Woldemariam Jegeno
July, 2005

This thesis has been submitted for examination with our approval as University advisors.

Dr. B. RAMA KRISHNA RAO
JULY, 2005

ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
BIBLIOGRAPHY