St. Mary's University

Faculty of Informatics

Department of Computer Science

Design of Anaphora Resolution for Afaan Oromo Personal Pronoun

Moti Teshome Tolera

December 2017

St. Mary's University

Faculty of Informatics

Department of Computer Science

Design of Anaphora Resolution for Afaan Oromo Personal Pronoun

By

Moti Teshome Tolera

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

Addis Ababa, Ethiopia

December 2017

ACCEPTANCE

Design of Anaphora Resolution for Afaan Oromo Personal Pronoun

By

Moti Teshome Tolera

Accepted by the Faculty of Informatics, St. Mary's University. In partial fulfillment of the requirements for the Degree of Master of Science in Computer Science

Thesis Examination Committee

____

Mr. Michael Melese

Internal Examiner

____

Dr. Million Meshesha (Ph.D.)

External Examiner

____

Mr. Asrat Mulatu

Dean, Faculty of Informatics

December 2017

Declaration

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for this work have been duly acknowledged.

Moti Teshome Tolera

____

Addis Ababa, Ethiopia

This thesis has been submitted for examination with my approval as an advisor

Mr. Hafte Abera

____

Addis Ababa, Ethiopia

Acknowledgement

First and foremost, I would like to thank my almighty God for giving me the strength, knowledge, ability and opportunity to undertake this research study and to persevere and complete it satisfactorily. Without his blessings, this achievement would not have been possible.

Next, my deepest gratitude goes to my advisor Ato Hafte Abera for his continuous follow-up, guidance, suggestions and his constructive comments were very essential. Moreover, many thanks to my colleague Henok Assefa who sacrifices more at work to allow me to study.

My acknowledgment would be incomplete without thanking the biggest source of my strength, my family. I appreciate and heartily thank my wife Ayuu and my beautiful daughter Sifan for great love, support and prayers you have given me during my entire study. My parents Ato Teshome Tolera and W/ro Bayush Abate I thank you for support, love and great sacrifices you made to educate and raise me, especially my brother Mamush who showed me the way and tell me to work determinedly to make my study perfect. I also want to thank my brother Bonsa and my sisters Hawi and Ayantu for their love, appreciation and support from the beginning to the end. Thank you so much for everything.

My sincere thanks must also go to my friend Tekalegn who helped me in manual POS tagging of collected data sets and supported me to understand a grammatical behavior of Afan Oromo language.

Finally, I want to thank all who helped me in completing this work but not mentioned here in name. Those who were always there to keep me on the line even when there were hard situations. God bless you all abundantly.

Table of Contents

List of Acronyms

AI – Artificial Intelligence

AOARM – Afan Oromo Anaphora Resolution Model

BART – Beautiful Anaphora Resolution Tool

CO – Coreference Relation

GUI- Graphical User Interface

ID – Identification

IE – Information Extraction

MARS – Mitikov Anaphora Resolution System

MUC – Message Understanding Conference

NE – Named Entity

NER – Named Enity Recognition

NLP - Natural Language Processing

POS – Parts Of Speech

RAP – Resolution of Anaphora procedure

SDD – Solid State Drive

ST – Scenario Template

TE – Template Element

TR –Template Relation

List of Figures

List of Tables

Abstract

Anaphora is defined as the linguistic phenomenon of pointing back to a previously stated item in the text. The pointing back word or phrase is called an anaphor and the entity to which it refers or for which it stands is its antecedent. Anaphora resolution is the process of determining the antecedent of anaphor. The scope of this resolution can be Intersentential or Intrasentential. The implementation of Anaphora resolution improves most of NLP applications such as machine translation, question answering, and text summarization and information extraction.

Most of Anaphora resolutions are studied for the English language. Nowadays research on anaphora resolution has been studied for other languages, such as Norwegian, Estonian, Spanish, Arabic, Turkish and Amharic. As anaphora resolution, system for one language is not directly adapted to another language, because it requires specific design for Afan Oromo based on the grammatical behavior of the language.

This study presented a model for resolving anaphora occurrences in Afan Oromo text using knowledge poor approach. The approach is implemented without any sophistication of linguistic knowledge and its core method is a list of multilingual antecedent indicators like a subject place, recency, frequency and constraints rules like gender, person and number agreement.

The proposed model focuses mainly on pronominal anaphora types and specifically on third personal pronouns. The models deal with Intrasentential and Intersentential types of anaphor. These personal pronouns can be hidden anaphor that resides in verbs and independent anaphor that occurs as personal pronouns. The proposed model follow different sub tasks, These are: preprocessing text which includes POS tagging, locating independent anaphor in a sentence in the text, extracting hidden personal pronouns, identifying possible antecedent candidates in defined range of preceding sentences, application of eliminative rule – constraint rules and optional rule - preferential rules, and selection of the candidate with the highest aggregate score.

Data used as datasets for our experiment were collected from Afan Oromo Holly Bible and Fiction. The evaluation of the prototype is performed on 330 sentences and conducted for two different scenarios. First, the hidden intrasentential anaphor algorithm scored a success rate of 57.84% and for independent intrasentential, anaphor the algorithm scored 47.51% success rate. For both Intrasentential anaphor, the algorithm scored a success rate of 55.20%.

On the other scenario, the algorithm scored success rate of 98.28% for hidden intersentential anaphor algorithm and 98.85% for independent intersentential algorithm. For both Intersentential anaphor, the algorithm scored success rate of 98.43%.

The challenging tasks in the study are extracting hidden anaphor from the verb word class because there are ambiguity of words in the language that are extracted based on the meaning of the sentence. Therefore, further works focusing on the knowledge of pragmatic is the major direction with regard to Afan Oromo anaphora resolution.

Keywords: - Afan Oromo anaphora resolution, knowledge poor anaphora resolution approach, hidden anaphor, antecedent indicators, independent anaphor

CHAPTER ONE

INTRODUCTION

1.1. Background

Language is the way of communication between humans, with basic aspects of human behavior both in written and spoken form. In written form, it helps us tracking of knowledge from one generation to the other, whereas in spoken form it serves as a primary means of communication in day-to-day activities [32].

There are various academic disciplines that studies language linguistics, computational linguistics, Psycholinguistics and philosopher. Natural Language Processing (NLP) a branch of computational studies, mainly concerned about the interaction between human and computers [32]. Though digitalization of information has also grown rapidly in this era, this trend continues with the globalization of information sharing. Theoretical understanding of the automated treatments of computer processes based on natural language has great benefit.

To solve the main challenge of the computer in understanding of human language, NLP is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech. It helps to do valuable things such as information retrieval, Information extraction, cross lingual, Coreference resolution and other applications [3].

In NLP system, knowledge about the structure of the language is important to what the words are, how words combine to make sentences, what the words mean, how word meaning affects sentence meaning and so on. However, having general world knowledge and reasoning ability is also crucial. According to Allen James [32], there are different forms of relevant knowledge sources for natural language understanding [32] [34].

i. **Phonetic and phonological knowledge:** phonetic and phonology knowledge is concerned about how words are related to the sounds that realize or creates them.
ii. **Morphological knowledge**: is concerned about how words are constructed from basic meaning units called morphemes.
iii. **Syntactic knowledge**: deals with how words can be put together to form correct sentences and what is the role of the words in the sentence.
iv. **Semantic knowledge**: is concerned about what words mean and how these meanings combine in

sentence to form sentence meaning.

v. **Pragmatic knowledge:** is concerned about how sentence used in various concepts and how its use affects the interpretation of the sentence.

vi. **World knowledge**: deals about the general knowledge about the structure of the world that language speakers do have.

vii. **Discourse knowledge:** defines how immediately preceding sentences affects the meaning of the next sentence.

From the above types of relevant knowledge sources, anaphora resolution is the most commonly appears as pronoun resolution in the problem of reasoning reference to earlier or later items in the discourse [33], which solves the relationship between two words in the given texts that need to use Coreference NLP tasks. One of the commonly studied research areas of Coreference NLP task is anaphora resolution system [2].

According to Advanced Oxford dictionary, the term “anaphora” means “the use of a word that refers to or replaces another word used earlier in a sentence”. Though, anaphora resolution is the process of determining the antecedent of an anaphor-where anaphor is a word or phrase points back and antecedent is the entity where anaphor it refers back [3].

Research in anaphora resolution broadly falls into two different approaches; knowledge poor approach and knowledge rich approach. Knowledge rich approach is commonly employed a rule based and algorithmic approach that requires deep knowledge of the language [9]. On another hand, knowledge poor approach that is implemented in various local and foreign languages [7, 12, 13, 15] is more corpus based and depend on NLP tasks like POS taggers and sentence splitter to make the system fully automated. As this approach tested and implemented widely, is adaptable to any language with minimum modification.

The application of anaphora resolution exists in different NLP areas like Machine translation, question answering, text summarization, text classification and others [3].

Afan Oromo language is spoken by more than 30% of Ethiopian population [1] and declared as the official language of Oromia Regional State. Currently, it is well recognized in the newspaper, different publication, radio and television media. However, the development of Afan Oromo language in Computer technology is not much satisfactory, which is also factual in the development of natural language processing applications. As it is well described by various studies [3, 7, 8] the identification of the antecedents of anaphors is crucial for different NLP tasks in any language. For instance, in machine translation, it is vital to resolve the anaphoric relation in translating into languages, which mark the gender of pronouns [7, 24], in question answering, information extraction, and other many applications. The challenge faced during anaphora resolution development is, as the POS tagging of our dataset is conducted manually it is error to prone and requires checking repeatedly.

According to Ruslan Mitkov [36], several problems can be encountered during the implementation of anaphora resolution system.

- Less efficiency of preprocessed data – during the anaphora resolution system inaccurate pre-processing problem can lead to a tangible drop of overall performance of the system. This requires stable POS taggers and morphological analyzer.
- Lack of Annotated corpora – there are not widely available corpora annotated with anaphoric or coreferential links. Mainly, this issue exists due to unavailability of specific annotation scheme and user based annotating tools.
- Factors in Anaphora resolution – it is not possible to propose generic set of factors used in anaphora resolution and any other factors not still used by researchers. In addition, four questions remain unresolved related to factors. These are how dependent are factors to each other, are preference rules better than constraints, do factors hold good for all genres and which is the best order to apply factors [37].
- Computational Strategy – even if different approaches use a similar set of factors, the computation power or strategy they use for application of these factors may differ. It is all about the way factors are employed and weights are formulated.
- Multilingual anaphora resolution required – as other NLP work as a whole is implemented for multilingual, anaphora resolution also required consideration for multilingual. It requires multilingual tools and resources, as the process is language dependent.

The anaphora resolution is typically recognized as a very difficult problem in NLP because it highly depends on both linguistic knowledge and world knowledge [33]. In addition, most anaphora resolution identifies antecedents as a noun or noun phrases for anaphors because identifying anaphors for verb phrases, clauses, sentences or paragraphs as antecedents is an another complicated task of anaphora resolution. The main objective of this work is to design a knowledge poor anaphora resolution model for Afan Oromo language.

1.2. Statement of the problem

On this information technology era, development of language is highly dependent with development of technology. So that, this study enables development of Afan Oromo language to grow with current information technology support. Anaphora occurs very frequently in written texts and spoken conversations. The majority of the algorithms that solve the anaphora were initially designed for English and other European languages. However, due to grammatical differences, morphological richness of Afan Oromo language and dependency of anaphora resolution on specific language it requires specific study of anaphora resolution for Afan Oromo language.

Moreover, research on NLP for Afan Oromo language is being one of the major study areas in Ethiopia; these efforts include Question Answering for Afaan Oromo [25], Parts of Speech tagger for Afan Oromo Language [41] and Afaan Oromo Text Retrieval System [46]. However, these tasks are ignored the anaphoric relation in the text. In Machine Translation, the establishment of the antecedents of anaphors have a crucial importance for correct translation. When translating into languages that mark the gender of pronouns for example, it is essential to resolve the anaphoric relation. So that, the proper implementation of anaphoric relation in machine translation will enable the translation of discourse instead of isolated sentences [7].

Similarly, in Question Answering systems, which are capable of extracting the answer to user's queries directly from documents, implemented using different approach. The approach for Question Answering such as noun-phrase extraction systems, which try to search the precise information requested by question whose answer is defined typically by a noun phrase. As the approach uses information from different sentences, paragraphs and even different documents to determine the answer (relevant noun-phrase) to the question. However, this system does not consider the information referenced pronominally, which means simply ignored anaphoric relation in documents [47]. As Chaltu [25] also points out, integrating an Afan Oromo co-reference resolution would be helpful in Question Answering system. In addition, Information extraction system that used for extracting of relevant information from vast text resources have critical steps like identification of entities followed by identification of relations. Though most of the pronominal expressions present contributing to relations refers to some other entity, the exact antecedent of each co-reference resolution mention has to be resolved for extraction of exact relation.

This study is therefore, as there is no effort in dealing with Afan Oromo anaphora resolution as far as the knowledge of the researcher, initiated to studying and addressing the anaphoric relation of Afan Oromo language.

1.3. Research Questions

- What type of antecedent indicators are required for Afan Oromo anaphora resolution?
- How does the Afan Oromo language resolution algorithm work for both independent and hidden anaphors?

1.4. Objective of the study

1.4.1. General Objective

The general objective of this study is to design and develop anaphora resolution for Afan Oromo language.

1.4.2. Specific Objectives

To come up with the stated general objective, the following specific objectives are performed:

- Review previous works on the area of anaphora resolution which is designed for different local and foreign languages
- Study and analyze the language specific behavior of Afan Oromo language
- Collect data from online web content and offline document sources used for the experiment
- Adopt an algorithm for Afan Oromo anaphora resolution model
- Evaluate performance of Afan Oromo anaphora resolution system using test data source..

1.5. Scope and Limitation

Taking time and other constraints into consideration, this study is limited to anaphora resolution of Afan Oromo language for the following points.

- The algorithm works only for text documents
- The search scope to find the antecedent is in current sentence and back four sentence
- Datasets only from Afan Oromo bible and fiction genre are collected
- The Cataphora resolution not considered
- The study focuses on word level to find the anaphor and antecedent
- In this study, we use knowledge poor approach to develop anaphora resolution for Afan Oromo
- The model assumes full sentences that are grammatically, syntactically and semantically correct.

Taking time and other constraints into consideration, this study mainly focuses only on Afan Oromo pronominal anaphora type's specifically third person pronoun. In addition, for hidden anaphor only suffix of verbs are considered which means this study does not consider prefix of verbs in extracting pronouns.

1.6. Methodology

Methodology delivers an understanding of the way a research is conducted and studied. This study was conducted in order to finding challenges in implementing Afan Oromo anaphora resolution model. To achieve the main objective, the following systematic procedure are followed.

1.6.1. Literature review

In this study relevant published text documents, books, journal articles, and previous related works have been reviewed. Mainly, it is researched for the English language, but recently there are researches for different languages like Estonian, Turkish, Swedish, Norwegian, Amharic, Spanish, Indian and Portuguese. To conduct this study, foreign and local language previous works are intensely studied from different papers and adopted into Afan Oromo language as required. Additionally, different Afan Oromo books and references were reviewed for a better understanding of language specific grammaticalbehavior.

1.6.2. Data collection

To get a better result of anaphora resolution system and due to its dependency on the availability of documents it is required to collect data that have adequate noun phrases and personal pronouns with the assumption of finding tagged data source. In this study, the proposed data collection used for the primary source is Afan Oromo Bible and Fiction text documents. These two genres are selected because the anaphor and antecedent relation occurred frequently and they are rich in anaphora. To get the understanding of data collection method the experience of previous related works are seriously considered. Preparing POS dataset for this work was done manually and it is a very challenging task. So that, to accomplish this study we have collected 200 third personal independent pronouns from different data sources. In addition, we have collected 300 verbs, which helps us to extract hidden personal pronoun. The collected tagged document is used for measuring the performance of the prototype.

1.6.3. Tools

In this study, we use Java programming language along with other on demand tools to develop the prototype for Afan Oromo anaphora resolution model. This programming language is selected because Java is an Object oriented language and the researcher is familiar with it.

1.6.4. Prototype development

After the preliminary tasks conducted such as reviewing previous work, a collection of data, selecting suitable tool and design approach, the researcher starts to develop a prototype using suitable and efficient java programming language. This prototype is used to test the effectiveness of the model.

1.6.5. Evaluation of Prototype

In the evaluation of anaphora resolution as it is discussed by different researchers [7, 13] there are no standard guidelines. Nevertheless, to evaluate the performance of the prototype, we tested using the success rate metric that was used in different previous works. Where success rate is the ratio of a number of correctly resolved anaphor over a total number of anaphors, the correctly resolved anaphor is anaphors’ who have antecedent according to the algorithm. It is defined mathematically as follows.

$$\text{Success Rate} = \frac{\text{Number of Correctly Resolved anaphors}}{\text{Number of All Anaphors}} \text{ X } 100\ \%$$

1.7. Significance of the study

The finding of this study contributes greatly to the benefit of the Afan Oromo NLP researches that plays an important role in the development of the language in computer technology today. The greater demand for anaphora resolution is to improve the performance of many NLP applications if integrated properly. The proposed anaphora resolution model can be used by many NLP applications [3, 9, 10, 11, 12, 13]. Some of these application areas are listed below whose performance can be improved if anaphora resolution is applied. In machine translation, the establishment of the antecedents of anaphors is very often of crucial importance for correct translation. When translating into languages that mark the gender of pronouns for example, it is essential to resolve the anaphoric relation.

In addition, in Systems for information extraction and text summarization contain techniques for obtaining information from relevant parts of the text. However, it is often the case that certain portion of the chosen information is realized by pronouns, the antecedents of which are in otherwise irrelevant parts of the text. The pronouns need to be expanded with referent auto sematic phrases, so that the acquired information is complete. Moreover, this thesis work will motivate researchers to work in this area for further study for another and same language.

1.8. Organization of Thesis

The rest of the thesis is organized in the following way: Chapter 2 presents an overview of Afan Oromo language, words classes including verbs, adjectives, prepositions, postpositions, pronouns, and Afaan Oromo morphology types including derivational and inflectional. Chapter 3 discusses related works done so far for different languages, such as Estonian, English, Amharic and Turkish, in the area of anaphora resolution task. The great focus in this chapter is given to the approaches followed, the algorithm and features used, and performance of the systems.

Chapter 4 discusses the architecture and design of the Afaan Oromo anaphora resolution system. The architectural components of our system are briefly discussed in this chapter. In addition, the algorithm for both independent anaphora and hidden anaphora resolution system are also discussed. Chapter 5 presents about dataset preparation, the testing technique used and performance result of the model. Chapter 6 sum up our work by presenting a conclusion and future works recommendation for improvement of the system.

CHAPTER TWO

LITERATURE REVIEW

In this chapter, we presented the overall concepts of Anaphora Resolution System based on the availability of journals, articles, books and previous research works. In the first two sections, we were explained about the importance of Anaphora Resolution in NLP applications and knowledge sources for natural language understanding. Next, we discussed basic concept of anaphora, process, different types and presented phases of anaphora resolution system and the approach used in the process and finally we discussed about related works related to anaphora resolution.

2.1. Anaphora Resolution in NLP applications

These days, anaphora resolution is implemented in many NLP applications. Proper usage of anaphoric relations shapes the performance of applications such as machine translation, information extraction, text summarization, or question answering [7].

2.1.1. Machine Translation

In machine translation when handling of anaphora is ignored, the translated text may be not only unnatural and incoherent, but also possibly factually incorrect. As pronouns in a language are required to match their antecedent in number and gender, which is language specific the antecedent of an anaphor in the source language can be translated by a phrase with a different gender. Therefore, it is not appropriate to base the translation of an anaphor on its form in the source language, but rather on the translation of its antecedent.

2.1.2. Information Extraction

Information extraction (IE) is the process of identifying text instances to make it more accessible for further processing [22]. Information Extraction is also defined as the process that takes text or speech as input and produce unambiguous data as output. Overall, the goal is to create a machine-readable text from given sentences [34].

The process of IE development is knowledge intensive and computationally intensive task compared to other NLP tasks like Information Retrieval. At the time of last Message Understanding Conference event in 1998, Information Extraction definitions are categorized in to five independent tasks [23].

2.1.2.1. Named Entity recognition (NE)

It is a task of identifying and classifying words of a sentence, paragraphs or a document into predefined categories such as names of persons, organization, locations, expression of times, quantities, monetary values and percentages. In addition, the process is weakly domain dependent [23]. The following example shows name entity type LOCATION that is Ambo.

Example, Abdiin gara *Ambo* deeme. / Mr. Abdi went to *Ambo*/

2.1.2.2. Template Element production (TE)

It is a task builds on named entity recognition and Coreference resolution, associating descriptive information with the entities. This task also weakly domain dependent same with Named Entity [23]. For example, the 'Bush administration' is also referred to as 'government officials', and adds this as an alias.

2.1.2.3. Template Relation production (TR)

This task requires the identification of a small number of possible relation between the template elements identified in the template element task. Extraction of relations between entities is a major feature of any information extraction task. Same with other tasks it is also weakly domain dependent task [23]. For example, an employee relationship between a person and a company, a family relationship between two persons, or a subsidiary relationship between two companies shows a template relation.

2.1.2.4. Scenario Template production (ST)

Scenario templates are the classical outputs of Information systems. They have ties with template element production and template relation production into vent description. Therefore, the result of Named entity, Template Relation and Template Element feed into scenario template. Unlike other types of information extraction, it is both domain dependent and tied to the scenarios of interest to users [23]. For example, Template Element may have identified Isabelle, Dominique and Francoise as people entities present in the Robert edition of Napoleon's love letters.

2.1.2.5. Coreference resolution (CO)

The importance of Coreference resolution in information extraction has led to inclusion of Coreference resolution task to in MUC program; which brought at the time the considerable growth of Coreference resolution algorithm and new emerging system. The task of Coreference resolution involves the identification of the expressions in a text that refer to the same real world entity in the same document. Coreference resolution is considered a hard and remaining problem and a challenge in artificial intelligence (AI) [24].

Based on the entity that refers the expression Coreference resolution is sub-categorized into anaphoric resolution and proper noun Coreference. The proper noun Coreference finds the same entity represented by different word order. On the other hand, anaphoric resolution is the process of pointing back a previously mentioned entity in the text [7].

2.2. Knowledge source for Natural Language Understanding

A natural language system should have considerable knowledge about the structure of the language [32]. The following are some of the different forms of knowledge relevant for natural language understanding.

2.2.1. Phonetic and phonological knowledge

It is the knowledge about how words or phrases are related to the sounds that realize them. Computational phonology is the application of formal and computational techniques to the representation and processing of phonological information [34]. Mainly, phonetic and phonology knowledge are important for speech based system.

2.2.2. Morphological knowledge

It is a source of knowledge concerned about the formation words from basic meaning units [32]. This smallest unit of meaning in language is called morpheme. There are different forms of morpheme [28].

Free morphemes – morpheme that stands alone as a word

Bound Morpheme – attached to free morpheme, which means creating new word

Inflectional morpheme – it is the composition of free morpheme and bound morpheme

Derivational morpheme – it is kind of morpheme which creates new word, this word is different in meaning to the original. Moreover, composed of free morpheme and bound morpheme.

2.2.3. Syntactic knowledge

It is a source of knowledge concerned about how word come together to form correct sentences and determines what type of role each word plays in sentences and what type of phrases are part of other phrases [32].

2.2.4. Semantic knowledge

Semantic knowledge is another type of source knowledge that studies meaning of linguistic. Moreover, it is concerned about what a word mean and how these meanings are correlated in sentence to form sentences meanings. The meanings of the sentences are studied regardless of the context that it is used [32].

2.2.5. Pragmatic knowledge

Pragmatics knowledge is the extension of the meanings or semantics knowledge. It deals with the contextual aspects of meaning in particular situations. It concerned how sentences are utilized as a part of various circumstances and how utilize influences the understanding of the sentence [32].

2.2.6. Discourse knowledge

It is a source of knowledge concerned about how immediately preceding sentences affect the meaning or interpretation of the next sentence. This type of knowledge is important in interpreting pronouns. Mainly, discourse knowledge is a study of chunks of languages that are bigger than a single sentence [32, 34].

2.2.7. World knowledge

It is a general everyday knowledge that all users of the language share about the world and the structure of world users of the languages must have to maintain conversation [32].

So that, Anaphora is a discourse level phenomenon in which the interpretation of one expression is dependent on another previously mentioned expression, also known as the antecedent.

2.3. The Concept of Anaphora Resolution

In the rapid growth of studies in computational linguistic field anaphora resolution has significant places in the history of NLP developments. Anaphora is the ancient Greek word that is translated as "*anajora*" with composition of two words *ana* and *jora* – back upstream and the act of carrying respectively; anaphora means the act of carrying back upstream. It can also paraphrased as expression which points back to some previous item [7]. The pointing back word or phrase is called anaphor and the expression to which it refers or for which it stands is its antecedent. In case when the pointing back expression refers to forward item in the text, which means the anaphor is occurred before the antecedent it is called Cataphora; which is also not common as anaphora. The process of determining the referent of an anaphor is called anaphora resolution [3].

Example: Anaphora: *John* studied hard for *his* test.

Yohannis qormaata *isaa* cimse dubbise jira.

The anaphor *his/isaa* refers back the antecedent *John/Yohannis*

Cataphora: Because she studied hard, Ebise passed her exam.

Sirritti waan *isheen* dubbisteef, *Ebiseen* qormaaata dabarte.

The anaphor *she/ishee* refers forward the antecedent *Ebise*

Hidden Anaphora: When Herod the king had heard these things, he was troubled, and all Jerusalem with him.

Yerusaalemis guutummaatti isa wajjiin ni *rifatte*.

In the above hidden anaphora example, the verb '*rifatte'* refers back to the word on subject place '*Yerusaalemis'*. Finding the verb '*rifatte'* refers back to '*Yerusaalemis'* solved with getting of end stem from the word and matching it with predefined pronouns. So that, the word '*rifatte'* is composed of '*rifa*-' and '*-tti'*. Moreover, in defined mapping list '*-tti'* is mapped to '*ishee'* third person pronoun.

Unlike Coreference resolution that operates across documents, anaphora resolution operates within a document. The successful interpretation of anaphora is crucial for different NLP tasks like machine translation, Information extraction, text summarization and others. The task of anaphora resolution is successful if any of previous entity in coreferential chain is identified as antecedent [23].

2.4. Types of Anaphora

Before resolving the identified anaphor, we need to know the type of anaphor mentioned. This section is committed to discuss the different varieties of anaphor. The varieties of anaphora resolution is based on lexical form of anaphor as listed in pronominal anaphora, Definite noun phrase anaphora, verb anaphora, adverb anaphora, zero anaphora and one anaphora [3].

2.4.1. Pronominal anaphora

It occurs when the anaphoric word is a personal pronoun, possessive pronoun, reflexive pronoun, demonstrative pronoun or a relative pronoun [3].

Example, Har'a *Siifaan* argeen ture anis kitaaba *ishee* kenneef.

[Today I meet *Sifan* and I gave *her* book.]

In the above example, the personal pronoun anaphor '*ishee'* ('*her'*) refers to the antecedent '*Sifan'*.

Pronominal anaphora can be occur in the form of different sub classes.

- Personal pronoun: This includes pronouns like he, him, she, her, it, they, them)

  John walked into the house; *he* is in a hurry to drink water.

- Possessive pronoun: This includes pronouns like his, her, hers, its, their, theirs)

  Smith is using my mobile charger; I think he will buy *his* own soon.

- Reflexive pronoun: This includes pronouns like himself, herself, itself, themselves)

  Lensa scored good grade this semester, she really improved *herself*.

- Demonstrative pronoun: This include pronouns like this, that, these, those)

  Bob owns another *house* in this city. *This* is his second house.

- Relative pronoun: This include pronouns like who, whom, which, whose)

London who attacked by terror act last week decided to leave European Union.

2.4.2. Definite noun phrase anaphora

It happens when the antecedent is referred by a definite noun phrase or a proper name anaphor [3]. Example, *Guddinaan* sa’aa booda dhufa kanaaf *Qaallichicha* eeguu qabna.

[*Gudina* will come afternoon so we will wait *the priest*.]

In the above example, the definite noun anaphor “*Qaallichicha*” (“the priest’”) refers to the antecedent ‘*Gudina’*.

2.4.3.Verb anaphora

Verb anaphora occurs when the anaphor is a verb and antecedent is verb or verb phrase [3]. Example, Eebbaan kaleessa bari *kaate*. Obbolesssi ishees akkasuma *godha*.

[Eba *woke up* early yesterday. So *did* her brother]

In the above example, the verb anaphor “*godha*” (“did”) refers to the antecedent “*Kaate*” ‘woke up’.

2.4.4. Adverb anaphora

Adverb anaphora occurs when the anaphor is an adverb [3].
Example, Isheen iftaan *Jimmaa* deemti. Guyya muraasa *achi* turti.

[She will go to *Jimma* the day after tomorrow and will stay *there* some time]

In the above example, the adverb anaphor “*achi*” (“there”) refers to the antecedent “*Jimma*”.

2.4.5. Zero anaphora

Zero anaphora is the case when any types of anaphor are not defined (omitted) in the sentence [3]. *Example,* Amy had made a special effort that night but (*she*) disappointed with the results. In the above example, anaphor ‘*she’* is omitted from the sentence.

Furthermore, this type of anaphora sub classified to zero pronominal anaphora, zero noun anaphora, zero verb anaphora and zero verb phrase anaphora.

2.4.6. One-anaphora

In one anaphora, the anaphor is the word one [3].
Example, He *fell* from the bicycle two times yesterday, and another *one* today. In the above example, anaphor ‘*one’* refers the word ‘*fell’*

2.4.7.Intrasentential and Intersentential anaphora

The anaphora are also categorized into two according to their location in the same sentence or not with the antecedent. Anaphor occurred in the same sentence with the antecedent is called Intrasentential anaphor and the anaphor occurred in different sentence with antecedent is named as Intersentential anaphor [7].

Example, *Liban* brings the document today. *He* is very committed person.

As shown in the above example, the anaphor '*He'* is Intersentential anaphor type that refers to '*Liban'.* For more examples, see Appendix A

Example, *Helena* presented *her* thesis yesterday.

As shown in the above example, the anaphora '*her'* is intrasentential anaphor type that refers to '*Helena*'. For more examples, see Appendix C.

2.5. The process of anaphora resolution

Different level of knowledge is required to solve the problem of anaphor resolution that helps the challenging tasks of the disambiguation of anaphors. The required knowledge includes low-level lexical information to high-level pragmatic level of the language [3]. One of the most essential source of knowledge in identifying anaphor is morphological and lexical information that includes gender and number agreement. In some cases, it is possible to identify antecedents for anaphor using only morphology and lexical information. The other knowledge required is Syntactic that helps to provide information about the boundaries of the sentences, clauses and noun phrases [15]. This type of knowledge is used extensively in anaphora resolution in addition to morphological and lexical information.

Semantic knowledge is also a valuable source of knowledge in case of both Syntactic; morphological and lexical knowledge is not enough to identify the relation between anaphor and antecedent. It is devoted to find the interpretation of the expression between anaphor and antecedent is meaningful or not. On other hand, discourse knowledge that is devoted to investigate the relationship between form and function in communication is important in resolving anaphora. The high level source information called real-world (common-sense) knowledge mainly uses additional information about the social environment is another used knowledge for anaphora resolution which is uncommon due to its difficulty in implementation.

In general, according to Ruslan Mitkov [7] the process of anaphora resolution consists of the following generic main stages.

Stage 1: Identification of Anaphors

This stage involves identifying anaphors that occurred in the given dataset. The identification involves all anaphors types such as pronoun anaphoric and noun phrase anaphoric using the part of speech tagger or morphological analyzer.

Stage 2: Identifying the antecedent

After the anaphor is detected the next steps is to identify potential candidates from the dataset. This includes listing of possible candidates mainly includes head of nouns and noun phrases antecedents as other types of candidates like verb phrases, clauses, sentences or sequences of sentences are more complicated task. Most of the time the search scope for the finding antecedent candidates of pronoun anaphoric type ranges from current sentences to back three sentences [3]. To do this task part of speech tagger and noun extractor are helpful.

Stage 3: Selection of the antecedent

Finally, to resolve the relationship between identified anaphor and listed potential candidate antecedents different resolution rules are used. Gender and number agreement, proximity and parallelism are some of known key rules that acts as eliminative and preferential rules used for this selection process, which frequently called anaphora resolution factors.

Eliminative Constraints rules

An anaphor and antecedent must agree in certain attributes to generate a relationship. These includes gender (male or female or neuter), number (singular or plural) and person (first, second, or third). Pronouns are mostly marked for number and gender which is useful in the resolution, but sometimes there is some exceptions where anaphor and antecedent are not agree on gender and number but semantically correct [7].

Preferential rules

Unlike eliminative rules that are obligatory, these rules are preferential or optional rules. It will help to select correct antecedent when more than one-listed potential candidates are remain after eliminative rules are applied. A set of preference indicators applied to candidates are based on salience, structural matches, referential distance and preference of terms [14]. The candidates will get scores for the above indicators and antecedent scored high in total will be selected as referent for the anaphor. Below usually used indicators are described briefly [7].

Definiteness: Definite noun phrases are more likely antecedents of the anaphors than the indefinite ones. Specifically in English language, the noun phrase is regarded as definite if the head noun is modified by a definite article ‘the’. But in Afan Oromo language definitiveness is marked by the suffix
-icha (-ticha) for masculine and -ittii (-tittii) for feminine [5],

Example:

Root word	**Masculine**	**Feminine**	**English**
Qaalluu	qaall-**icha**	qaall-**ittii**	The Priest
Jaarsa	Jaars-**icha**	Jaar-**titti**	The old Man/ the old woman
Mootii	Moot-**icha**	Moot-**itti**	The King/The Queen

Obliqueness: a noun phrase that represents the 'theme' or 'given information' in specified discourse is considered more likely as antecedent of the anaphor [7].

Indicating Verbs: there is number of verbs that give high salience for the noun phrase following them. In English language, noun phrases following verbs like {discuss, present, illustrate, identify, summarize} are considered a reasonable antecedent [7].

Lexical Reiteration: lexically repeated items within the same paragraph are believed to be likely antecedents. Apart from exact match of the lexical items synonymous noun phrases are also counted in the process [7].

Referential Distance: a candidate that is near to the anaphor are the best candidate for the antecedent of an anaphor than candidates that located far from the anaphor [7].

Boost Pronoun: as pronoun represents additional information of an entity, they can be considered as possible antecedent of the anaphor. It is valuable in case when the noun phrase corresponding to an antecedent may be beyond the range of the algorithm, appearing only before the two sentences preceding the one in which the pronoun appears [11].

2.6. Anaphora Resolution Approaches

Based on the computational strategy used, approaches in the development of the anaphora resolution system is categorized into two- namely knowledge rich and knowledge poor approach. They are also named as traditional and alternative approaches respectively [7, 15]. Earlier systems lean to knowledge-rich. However, with need to develop fully automated systems, and the advent of cheaper and more corpus based NLP tools like POS taggers and parser, and growth of machine learning and statistical technique, modern anaphora resolution systems tend to use a knowledge-poor approach [9].

2.6.1. Knowledge-rich approach

Previous research in anaphora resolution regularly engaged a rule based, algorithmic approach and was generally knowledge-rich. By definition, knowledge-rich approach is a rule-based approach that requires a knowledge of syntax, semantics and discourse of the language and domain knowledge with fully corrected and parsed input data. This traditional approach integrates knowledge sources to the indicators that discount unlikely candidates until minimal set of plausible candidates is obtained and then makes use of center or focus, or other preference [7]. According to Tejaswini Deoskar, [9] the approach categorized into syntax-based approach and discourse based approach. Mostly, this approach is labor intensive and time-consuming tasks. Evaluation was typically carried out by hand on a small set of evaluation examples. Following discussions are presented based on [9].

i. Syntax-based approach

These approaches assume the presence of a fully parsed syntactic tree and traverse the tree that considers the antecedents and applying appropriate syntactic and morphological constraints on them. One of the classical result of this approach is Hobbs 1997 [9].

Hobbs Algorithm

Hobbs 1977 used a naïve algorithm to obtain an impressive accuracy in pronoun resolution that works by surface parse trees of the sentences in the text. Normally, the algorithm performs a left to right search that means every node of depth n is visited before any node of depth n+1, gives preference to closer antecedent. With this steps algorithm collects possible antecedents and checks for them if they agree with pronoun in gender and number. The algorithm evaluated on 300 personal pronouns covered he, she, it and they and the algorithm resolved 88.3% of the cases correctly [8].

ii. Discourse-based Approaches

Alternative traditional approach method of obtaining the reference of pronoun is discourse based, called centering theory. This theory models discourse coherence based on each utterance features topically most prominent entity-center. The main idea is certain entities mentioned in an utterance are more central than others are – become center of attention. One of the most well known result using this approach is Brenan, Friedman and Pollard (BFP) 1987 [39]. They uses syntactic and morphological knowledge like number and gender agreement to eliminate unfitting candidates, but uses centering principles to rank potential candidates.

iii. Hybrid Approaches

These type of approaches takes into considerations a number of knowledge sources, including syntactic, discourse, morphological and semantic to rank the potential antecedents. Lappin and Leass [19] is one of the most well-known systems using this approach. Unlike costly semantic and real world knowledge in evaluating antecedent, they use syntactic and morphological constraint filter to eliminate candidates that do not satisfy gender and number. Salience is obtained based on different factors that are integrated to the algorithm such as sentence recency, subject place, Existential emphasis, Accusative emphasis, and Indirect Object and Head noun emphasis. Each factors have their own weight associated to them. Therefore, the Lappin and Leass approach gives weights to these factors, unlike centering approaches. The factors that the algorithm uses to calculate salience are given different weights according to how relevant the factor is which ranges from 50 to 100.

iv. Corpus based Approaches

This approach is a statistical method type of resolving pronoun antecedent relation. Chariak et al [40] uses this method. They implemented on training corpus marked with co-reference resolution and based on Hobb's algorithm with a probabilistic model. The information that this probabilistic model bases are distance between pronoun and antecedent, syntactic constraint, actual antecedent-to give information on number and gender, interaction between the head constitute of pronoun and antecedent, and antecedent mention count. In their work, they assume these above all factors are independent. The experiment they conducted first calculates the probabilities from training corpus and apply on test corpus to resolve pronouns in test corpus. Evaluation method is based on 10-fold cross validation and obtained results of 82.2% correct [9].

Thus, knowledge-rich approach summarized, as there are two types of approach one of them works by first eliminating some antecedent based on constraints-syntactic constraint and then choosing the best of the remaining based on some factors like centering. The other one is, considers all candidates as equal but makes decision on how possible a candidate is based on different factors [9].

2.6.2. Knowledge-poor approach

On the other hand, Knowledge-poor approach requires less linguistics knowledge of the language and less domain knowledge that uses machine-learning techniques. This type of approaches called alternative approach can computes the most likely candidate based on statistical or artificial intelligence techniques/models. It is achieved by automating the pre-process of input data such as part of speech tagging, pronoun identification, noun phrase identification. Mainly it finds the candidates by eliminative or preferential techniques, and commonly uses a combination of those two [9].

Knowledge poor approach can be implemented without any sophisticated linguistic knowledge by avoiding complex syntactic, semantic and discourse analysis or even without parsing the corpus. Instead, it is benefiting from corpus based NLP techniques such as sentence splitting and parts of speech tagging and rely on the efficiency of sentence segmentation, noun or noun phrase identification and the high performance of the antecedent indicators. Therefore, the core of the approach is a list of multilingual antecedent indicators after identifying candidates from the current and preceding sentence under specified scope, based on the gender and number agreement. Prior to this, it requires text pre- processing by sentence splitter to determine the sentence boundaries, parts of speech tagger to identify the parts of speech and simple phrasal grammar to detect the noun phrase. In addition, the approach is easily adopted to other languages with minimum modifications [38].

2.7. Related Works

A great effort has been made to design anaphora resolution in many languages even before Coreference resolution is introduced at Message Understanding Conference (MUC) in 1998
[21] as a new independent task. As anaphora resolution is a complicated problem in natural language, processing considerable research has been done by computational linguists for local and foreign languages.

2.7.1. Design of Amharic Anaphora Resolution Model

The work of Temesgen [15] presents Anaphora resolution system for Amharic. The researcher developed Amharic Anaphora resolution system for third person personal pronoun. The model is designed for both independent and hidden anaphors that focused on resolution of pronominal anaphoric entity. The proposed Amharic system has five major components: low-level knowledge extraction, identification of independent anaphors, identification of hidden anaphors, identification of candidate antecedents and anaphora resolution [15].

The first component is low-level knowledge extraction that is identifying independent anaphors and hidden anaphors by analyzing the morphology of verbs and identifies nouns for antecedents with the aim of POS tagging-classification of word classes in a text. The second component is identification of Independent anaphors to search words tagged as personal pronoun in POS tagged Amharic text and it is achieved by having personal pronoun database. The third component is identification of hidden anaphors as the Amharic language has unique complexity of anaphora resolution identifying hidden anaphor is crucial. The hidden anaphors are identified by morphological analyzer to identify verbs in chunked text for which it is marked. Even though the researcher is not satisfied with result, he uses HornMorpho morphological analyzer as the identification of hidden anaphors in Amharic language. The fourth component is identification of candidate antecedent is a way of listing words in which the correct antecedents of anaphors are selected. The final component in Amharic prototype model is anaphora resolution in which the relation between anaphor and antecedent is established by applying constraint and preference stored rules. Where constraint rules are about gender, number and person agreement and preference rules are applied when list of antecedents compute after implementation of constraint rule. The criteria to establish preference rules used are subject place, definiteness, recency, mention frequency, and boost pronoun [15].

Generally, the researcher collected 311 sentences having 315 verbs that used to extract hidden verbs and process the hidden anaphora resolution. And they collected 163 sentences to get 110 personal pronouns for independent pronoun resolution process. The average success rate and critical success rate of hidden anaphora resolution part are 81.79% and 76.07% respectively whereas average success rate and critical success rate of independent personal pronouns resolution part are 70.91% and 57.58% respectively.

2.7.2. Knowledge-poor Anaphora Resolution System for Estonian

Mutso [14] presents knowledge poor Anaphora resolution system for Estonian. The researcher developed Estonian anaphora resolution system for pronominal anaphora type using knowledge poor approach and primarily focused on third person pronouns.

The proposed system passed through different stages. At first the input text file is read in line by line and parsed according to the file type and then the text file has been read into the memory, the program starts to search for the anaphora to resolve. It looks only for the third person personal pronouns. After the anaphor is found, the program looks for the appropriate candidate nouns for the anaphora. The range for finding the possible antecedents is three sentences. After the list of candidate words has been processed by all the indicators, they are multiplied with the recency constant defined by the researcher. In the last step, the candidates are sorted by their total score.

To evaluate performance of the system the researcher used training and testing datasets. Collected data having 646 third person personal pronouns is used for training purpose whereas 856 third person personal pronouns is used for testing purpose. In training phase, out of 646 third person personal pronouns, the system resolved 440 third person personal pronouns correctly which is around 73.6%. In testing phase, out of 856 third person personal pronouns, the system resolved 549 third person personal pronouns correctly which is around 73.7%.

2.7.3. Automatic Anaphora Resolution for Norwegian (ARN)

Holen [13] developed a system called Automatic Anaphora Resolution system with rule-based approach based on RAP (Lappin and Leass) and MARS (Mitikov) existing English systems for third person pronouns Norwegian language. The anaphora resolution module works in parallel with reading of input files and making of sentence and word objects. It contains four major steps. The first steps includes finding the anaphor, in this stage the system checks if the sentence object contains an anaphor. If one or more anaphor is found, the list of anaphora to be resolved and pass to next step. Else, if the sentence does not have anaphor the system proceeds with next sentence. Secondly, a list of candidate antecedents is made, consisting of all the nouns and pronouns in the current sentence object. When all candidates are listed, the one that have higher ID number than anaphor is removed to avoid Cataphora resolution. Then, on identified and listed candidates, the factors are applied one by one. Each factor gives positive or negative points to each candidates in stack. Finally, the system rearrange candidates based on the score and the candidate with the highest score is proposed as the antecedent.

To evaluate the performance of the system, the researcher uses two different corpus Oslo and BREDIT files and prepared 15 pairs of equivalent files each set of files having 46972 words. In addition, divided into training data of 21800 words and test data with 25172 amount of words.

2.7.4. Robust pronoun resolution with limited knowledge

The work by Mitkov [26] implemented knowledge poor anaphora resolution approach for pronominal anaphora resolution in English language. The researcher discovered knowledge poor anaphora resolution approach for the first time to avoid labor intensive, expensive and time-consuming tasks to resolve anaphors. To avoid complex syntactic, semantic and discourse analysis, the robust and poor knowledge approach is vital and it does not parse and analyze the input in order to identify antecedent of anaphors. Mainly, the approach works as follows. It takes as an input the text preprocessed by POS tagger, and then identifies the noun phrases that located before anaphor as candidates with in two sentences distances. Thirdly, it checks the candidate and anaphor gender and number agreement and applies the genre-specific antecedent indicators for those passed the constraint rules. Finally, the noun phrase with the highest aggregate score is proposed as antecedent of anaphor.

The researcher also described the approach can be successfully adapted for other languages with minimum modifications. The Evaluation reports shows this new approach scores a success rate of 89.7%. It is also applied to other languages like Polish and Arabic, and resulted in 93.3% and 95.2% success rates respectively.

2.7.5. Automatic Pronoun Resolution for Swedish

Gustav [12] presents a system called SwePron anaphora resolution for Swedish third person singular pronoun which developed by Java using knowledge-poor concept of Mitikov [17]. The researcher uses separate existing tools text analyzer and parser for preprocessing the input text.

This work composed of five phases. In the first phase, the text is parsed syntactically and information about parts of speech, lemmas, syntactic functions and dependency relations is extracted. Next, the algorithm identifies nominal anaphoric pronoun based on machine learning method. Thirdly, for each pronoun identified potential candidates are extracted from the text with in three sentences of boundaries search scope. After the gender and number agreement of candidates comply with identified pronoun the next step is application of preference factors. In this phase, algorithm applies preference factors for each selected candidates by giving numerical score that helps to determine the candidate probability as antecedent, the score is added to the composite score. Finally, for each identified anaphoric pronoun the candidate with the highest score is selected as the antecedent, if more than one candidates gets equal composite score the candidate closest to the anaphor is selected.

To evaluate the performance of the algorithm the researcher tested on text from Stockholm Umea corpus with 257 pronouns for both masculine and famine pronoun resolution, whereas 254 pronouns are selected for running neuter pronoun resolution. The success rate for the system is 61.87% for masculine and famine pronoun resolution, and success rate of 66.14% for neuter pronoun resolution.

2.7.6. Knowledge-Poor Pronoun Resolution System for Turkish

The main motive of the researcher Kucuk [2] is to propose and implement knowledge-poor pronoun resolution system for Turkish text. The resolution system attempt only the third person pronoun and possessive pronouns which refers to person names.

Design and implementation of the system includes preprocessing of text and architecture of the system. The preprocessing module helps to freed resolution system to resolve non-anaphoric. After the input text is preprocessed, the actual pronoun resolution system will continue. This includes splitting input text into sentences, extraction of third person pronoun and reflexive pronoun, creating the candidate list and determining the antecedent of each extracted pronoun by applying constraint and preferences. The first step is sentence splitting which is a process to split the sentences. The researcher uses dot (.), exclamation mark (!) and question mark (?) as sentence separators. Next, the system extracts pronouns that are marked as third person pronoun and reflexive pronoun during preprocessing stage. The pronouns are marked with overt (0) and zero (z) pronoun signs. On third step, the system attempt to form the list of candidate antecedents. Since proper names in Turkish text are the only words capitalized, the researcher identifies those proper nouns with in three previous sentence boundary including current sentences as candidates. Those identified proper names are checked against Turkish names dictionary to avoid any initial word in the sentence which is also capitalized. On final step, the pronoun resolution system determine the antecedent from list of candidates. The system will apply the constraints and preferences to select the potential antecedent. The constraint for Turkish includes number agreement, reflexive pronoun constraint, personal pronoun constraint and selection restrictions. If more candidates remain after constraints are applied, then preferences are applied to remaining candidates. These preferences includes quoted/unquoted, text, recency, nominative case, first noun phrase, nominal predicate, repetition, punctuation, and antecedent of zero pronoun preferences. Each preferences rule has an associated score that is used to determine the correct antecedent. After implementation, the system is tested on two different test samples. The first test scored 85.2% and the second test scored 73.6%.

Table 2. 1 Summary of related works

	MARS [11]	**RAP [19]**	**MOA [7]**	**ARN [13]**	**Mutso [14]**	**Temesgen [15]**	**SwePron [12]**
Data Set	Technical manual	Computer Manual	Technical Manual	Oslo and BREDT Corpus	Newspaper texts, fiction, scientific texts and legal texts	WIC and Amharic Holly Bible	Stockholm Umeå Corpus
Prerequisite	Preprocessing done automatically	data manually checked	The data manually checked and corrected	Manually tagged corpora	Data annotation is done automatically	Manually tagged and automatically chunked data	Different parser tools are used
Purpose	Third personal Pronoun	Third personal Pronoun	Third personal Pronoun	Third personal Pronoun	Third personal Pronoun	Third personal Pronoun Both hidden and independent	Third personal Pronoun
Language	English	English	English	Norwegian	Estonian	Amharic	Swedish
Algorithm	Rule-based (syntax, morphology)	Rule-based (syntax, morphology	Rule-based (morphology)	Based on MARS and RAP	Based on MARS	Based on MARS	Based on MARS
Succe ss Rate	61.55%	84.10%	89.70%	70.50%	73.60%	81.79%	61.87%

2.8. Summary

In this chapter, we have discussed basic concept of Anaphora resolution and defined as the process of determining the noun or noun phrase that refers to anaphor in the text. Similarly, discusses different types of anaphora based on lexical form of anaphor; these are pronominal anaphora, definite noun phrase anaphora, verb anaphora, adverb anaphora, zero anaphora and one anaphora. In addition, the anaphora resolution module consists different major steps; checking sentence for anaphor, making candidate list, application of constraint and preference factors and choosing the most appropriate candidate. On top of that, we also discussed approaches in anaphora resolution falls into two broad categories; knowledge rich and knowledge poor approaches. From a review of works, we learned that knowledge poor anaphora resolution approach is effectively implemented to different local and foreign language. In addition, the researchers uses different types of constraints and preferences rules based on language specific grammatical behavior. It is understood that there are constraint and preference rules that can be applied to all languages and there are some factors cannot be applied to all languages.

As Afan Oromo language has hidden anaphor unlike other languages like English, it requires special consideration of resolving anaphora. In addition, as shown in table 2.1 to the best knowledge of the researcher, there is no effort made for Afan Oromo language. This research therefore has a great contribution in resolving the relationship between anaphor and antecedent in Afan Oromo texts.

CHAPTEER THREE

AFAAN OROMO LANGUAGE

In this chapter, we focused on highlighting the Afan Oromo language basic grammar concepts. In the first section, we presented a general overview of the language. The next section discusses the classification of words in Afan Oromo language. Finally, we presented the morphology of Afan Oromo language, Specifically types of morphology in Afan Oromo.

3.1. Overview

Afan Oromo is a Cushitic Language and mother tongue of Oromo people who lives in Ethiopia and some parts of neighboring countries like Kenya and Somalia [5]. Oromo people who are the largest ethnic group in Ethiopia amounts to 34% of the total population [27]. The Oromo people in Ethiopia are dominantly occupied in Oromia region. Currently Afan Oromo is the official language of the regional state of Oromia being used as a working language in offices, educational language for all non-language subjects in junior-secondary schools (1-8 grades). The writing system, "Qubee" (Latin-based alphabet) has been adopted and became the official script of Afan Oromo since 1991 [5].

Afan Oromo has its own phonetic language, which means that it is spoken in the way it is written. It uses simple Latin script, which makes it straightforward in its written system. Afan Oromo language has vowels and Consonants. The vowels are signified by the five basic letters such as a, e, i, o, u. There is also doubled form of the mentioned vowels in Afan Oromo language like "aa", "ee", "ii", "oo", "uu". The consonants letter are not differed mainly from the English language, but there are few special combinations such as "ch","dh", "sh","ny" and "ph". In general, Afan Oromo has 37 letters (32 consonants and 5 vowels) [25].

The Sentence is a group of words come together to express completes thoughts. The typical word order in Afan Oromo sentence structure is subject–object–verb, strictly verb is at the end [5] but in English sentence word order is Subject-Verb–Object. Modifiers, articles, and pronouns follow the nouns they modify.

3.2. Afan Oromo Word Classes

The Afan Oromo language words can be classified into nouns, verb, adverb, adjective, pronoun and prepositions [5].

3.2.1. Nouns

A noun is a class of word that denotes person, animal, place, thing or idea. Nouns are expressed by gender, number and definiteness.

3.2.1.1. Gender

In Afan Oromo nouns gender is marked by suffix like, -**eessa** and **–eettii** denotes masculine and feminine respectively [5].

Dur**eessa** (m)	Dur**eetti**(F)	[Rich]
Og**eessa** (m)	Og**eetti** (F)	[Expert]

There are notable exceptions where nouns derived from verbs, the masculine noun adds an **-aa** suffix and the feminine noun adds a **-tuu** suffix to the verb root.

Teacher	Barsiisaa (m)	Barsiistuu (F)	Barsiisuu (Verb) – to teach

3.2.1.2. Number

In Afan Oromo plural forms of nouns are rarely used, plurality is shown by suffix like **-oota**, **-aan**, **-wan**, and **–en** [5].

Farda (sg.)	[Horse]	Fardoota (Pl.)	[Horses]

3.2.1.2. Definiteness

Like other languages, Afan Oromo language does not have special word to denote definiteness, it is marked by the suffix **-icha** (**-ticha**) for masculine and **-ittii** (**-tittii**) for feminine [5].

Qaalluu	Qaalli**ticha** (m)	Qaalli**tittii** (F)	[The Priest]

And Indefinitiveness of words in Afan Oromo is denoted by numeral ‘tokko’ which gives meaning of ‘a certain’ [5].

Gurbaa **Tokko**	[**one/a** boy]

3.2.2. Verbs

Verbs are words or compound of words that expresses action, a state of being and/or relationship between two things. In their normal position, they are found at the end of the sentence as shown below.

Example, Caalaan mana ***bite***. [Chala bought a house.]

Ayyaantuun ***dhufte***. [Ayantu has come.]

There are four derived stems, the formation of which is still productive, Autobenefactive, Passive, Causative and Intensive.

Autobenefactive

The Afan Oromo autobenefactive is formed by adding -(a)adh, -(a)ach or -(a)at or sometimes -edh, -ech or –et to the verb root. This stem has the function to express an action done for the benefit of the agent himself.

Example: bitachuu - to buy bit- root verb

Passive

The Oromo passive corresponds closely to the English passive in function. It is formed by adding -am to the verb root. The resulting stem have different form regularly.

Example: Jaar- built Jaaram- be built

Causative

The Afan Oromo causative of a verb corresponds to English expressions such as 'cause ', 'make ', 'let '.With intransitive verbs, it has a transitiving function. It is formed by adding -s, -sis, or -siis to the verb root.

Example: erguu - to send ergisiisuu - to cause to send

Intensive

It is formed by duplication of the initial consonant and the following vowel, geminating the consonant.
Example: Waamuu - to call wawwaamuu - to call intensively

3.2.3. Adverbs

Afan Oromo adverbs are words that are used to modify verbs. Adverbs usually precede the verbs they modify or describe. They have the function to express different adverbial relations such as relations of time, place, and manner or measure.

Adverbs of time:

Amma/now, booda/later, boru/tomorrow,dura/at first, har’a/today, etc

Adverbs of place:

Achi(tti)/there, as(tti)/here, dhiyoo/near, fagoo/far, gama/on other side, etc

Adverbs of Manner:

Dansa/fine, sirritti/correctly,suuta/slowly, wayya/better, etc

Example, Obboleessi Koo ***kaleessa*** dhufe. [My Brother came yesterday.]

Amma hojiin qaba ***booda*** kotu. [Now I am busy, come back later.]

3.2.4. Adjectives

An adjective is a word that describes or modifies a noun or pronoun. It specifies to what extent a thing is as distinct from something else. The masculine form terminates in one of the following suffixes –*aa*, -*eessa*, or -(a)acha, and the feminine form terminates in one of the following suffixes *–oo, -tuu, -eettii,* or *–aattii.*

Example, Dursaan ***gabaabaa (m)*** dha. [Dursa is short]

Hawwiin ***furdoo (f)*** dha. [Hawi is fat.]

The number form of adjectives in the plural occurred by reduplication of first syllable.

Example:	Xinnaa (m) (s)	xinno (f) (s)
	Xixinnaa (m) (p)	Xixinnoo (f) (p)

3.2.5. Preposition

In Afan Oromo prepositions are much less numerous than postpositions [5]. The most common are:

akka - according to, like, as . ***eega*** - since, from, after. ***eegasu*** - in that case, therefore

gara - in the direction, towards, side. ***haga –*** until. ***hamma*** - upto, until such that, as much as.

Example, ***Akka*** isaa jabaan namu hinjiru. [There is no one as strong as he is.]

Inni ***gara*** manaa deema. [He goes (towards) home.]

3.2.6. Postposition

Unlike European languages, Afan Oromo language uses frequently postpositions [5]. The most common postpositions in Oromo are:

ala - out, outside. ***bira*** - beside, ***booda*** - after ***jala –*** under

Example, lsheen obboleessa ishee ***bira*** dhaabatti. [She stands beside her brother.]

3.2.7. Pronoun

In Afan Oromo, like in other languages, a pronoun is a word that is used instead of a noun or noun phrase. They are characterized based on number and gender. The Pronoun in Afan Oromo can be independent or hidden with the verb based on their existence in a sentence. Independent pronouns are pronouns exist in a sentence as a separate word in the sentence. In the following example, "*Inni*" is an independent personal pronoun.

Example: Yohaannis yeroo dhufe homaa hin nyaatu homaas hin dhugu ture; Isaanis ***Inni*** dhukuba qaba jedhu.

However, hidden pronouns are pronouns attached to the word in the sentence. In the following example, the word "*deemte*" indicates the pronoun "*Ishee*" hidden in it.

Example: Toltuun kaleessa magaala Finfinnee ***deemte*** ture.

i. Personal Pronouns

Personal pronouns are one type of pronouns that we may use to express particular person. In Afan Oromo language, personal pronouns are formal with everyone except close friends and other relative persons [5]. Concerning the form and usage of a personal pronoun, in Afan Oromo as a subject, it comes at the beginning of the sentence and as an object, it comes before the verb, unlike English, it comes after the verb.

Table 3. 1 Lists of Afan Oromo personal pronouns

English object	**Afan Oromo Object**	**English Subject**	**Afan Oromo Subject**	**Gender**	**Number**	**Person**
Me	Ana	I	Ani	Male/Female	Singular	1st Person
Us	Nu/Nuu	We	Nutii/Nuyi	Male/Femal	Plural	1st Person
You	Si	You	Ati	Male/Female	Singular	2nd Person
You	Isin	You	Isin	Male/Female	Plural	2nd Person
Him	**Isa**	**he**	**inni**	**Male**	**Singular**	**3rd Person**
Her	**Ishee/Isii**	**she**	**Isheen/Isiin**	**Female**	**Singular**	**3rd Person**
		It	**Isa/Ishee**	**Neuter**	**Singular**	**3rd Person**
Them	**Isaani**	**They**	**Isaan**	**Male/Female**	**Plural**	**3rd Person**

There is no special word for ‘**it**’ in Afan Oromo, instead ‘*inni*’ /him/ used formally or informally. In addition, gender markers (verb) for other things that are not human being [5].

Example: That Book is New. It was bought last week

Kitaabni sun haaraadha. Torban darbe bitame.

The verb ‘*bitame*’ shows that the subject is the third person ‘*he’*.

ii. Demonstrative Pronouns

A demonstrative pronoun is a pronoun that is used to point something specific with in a sentence. They can be singular or plural and points out items in time or space. Some these pronouns are: *kun(i)/this, tun(i)/this, sun/that ,kana/this, tana/this, sana/that, kunniin/these,kanneen/those.*

iii. Possessive pronouns

Possessive pronouns are pronouns that designate possession. These pronouns include *koo, kiyya, isaa, ishee, keenya, isaani, keessan.*

iv. Reflexive pronouns

A reflexive pronoun indicates the person who realize the verb action is same with receipt of the action. These includes *of/self, uf/self* pronouns.

Example: Isheen **of** laalti [she looks at **herself**]

v. Interrogative pronoun

Interrogative pronoun is used to make asking question. Some of these pronouns include *eenyu/who, kan eenyuu/whose, maal(i)/what, maaliif/why, eessa/where, meeqa/how much, akkam/how, yoom/when.*

3.3. Afan Oromo Morphology

Morphology is a branch of linguistics that deals about the knowledge of the meaningful component of words [28]. Jurafsky and Martin [28] defined morphology as the study of the way words are built up from smaller meaningful units called morphemes. Word is the most basic unit of linguistic structure. Like other Ethiopian languages, Afan Oromo has complex and rich morphology.

3.3.1 Types of morphology in Afan Oromo

In Afan Oromo language, we have two broad types of morphology namely derivational morphology and inflectional morphology [30].

i. Inflectional Morphology

Inflectional morphology is the processes when words are adapted to their proper functions within a given sentence without changing the meaning of base words. Alternatively, can be defined as the change the form of a word for grammatical usage. It occurs in the form of different word classes. Inflection of verbs, Inflection of Nouns and Inflection of Adjectives [30].

Inflection of Nouns – most of Afan Oromo nouns end with a vowel except few which ends with consonants like *n,l,t* . [Inflection Morphology Oromo]. Inflectional categories under nouns exists mainly in the forms of marking number, definite and gender.

In Number, marking inflections distinguish plural and singular. Several types of suffixes are attached to nouns to make plural forms.

Example: The plural –(o)ota attached on Nouns

Waggaa [Base form] Wagg-oota [Inflected form] Years

The plural –lee attached on Nouns

Baatii [Base form] Baatiilee [Inflected form] Months

In definite marker or singulative marker shows noun is marked for being used as single form.

Example: Nama [Base form] Namicha [Inflected form] The man

In gender marker, we use inflection to identify masculine and feminine through gender suffixes.

Example, boonaa [base form] boont-aa [inflected form -m] / proud boy boont-uu [inflected form –f]/proud girl

Inflection of Verbs – verbs are the most classes in which inflection are occurred. Mainly verb inflection happens in the form of inherent and agreement properties. Inherent properties is a verb inflection that triggers inflection on that word class includes aspect, mood, and voice. However, agreement properties indicate inflection of word class for properties out of its members include person, number, gender and case. In the Afan Oromo language, the roots or stems of verbs, usually end in consonant, take inflectional morphemes showing distinction between aspects or mood or gender or number.

Example, Mur – [root form] Mur-te [Inflected form]

Inflection of Adjectives – are the same with that of nouns. Adjectives are inflected for number, gender and singulatives like nouns. If adjectives occur within sentences, number is marked on both of them. Inflection for number of adjectives occurred in the form lexical, reduplication and–(o)ota.

Example: Inflection for Numbers

Lexical:	Sooressa (s)	Sooreeyyi (p)
Reduplication:	guddaa (s)	gud-guddoo (p)
-(o)ota:	hamaa(s)	ham-oota (p)

In Afan Oromo, the base forms of adjectives are normal to be used as masculine, but inflection occurs when we make them for feminine.

Example: Hamaa (m) Hamtuu (f)

In Afan Oromo, singulative markers are not used on both noun and adjective at the same time. Which means, when nouns is marked, adjective is not and vice versa.

Example: Muk-ni(n) dheer-icha(adj.). The long stick

ii. Derivational Morphology

Derivation morphology is the creation of new words from already existing words in the language. And is the alternate word form which changes the meaning and word class. Different derivational suffixes are attached to the root or stem of the word [30].

It occurs in the form of different word classes. Derivation of verbs, Derivation of Nouns and Derivation of Adjectives.

Derivation of verbs – is the creation of new verb word class from other given word class stem.

Example: arguu [base form] / to see argachuu [derivated form] / to get, find

Derivation of Nouns – is the creation of new noun word class from other given word class stem.

Example: bulchuu [base form] / to administer bulchiinsa [derivated form] / administration

Derivation of Adjectives – is the creation of new adjectives from nouns or from other adjectives or from verbs.

Example: jabaa [base form] / strong jabaacuu [derivated form] / to be strong

3.4. Summary

Afan Oromo language is an Afro-asiatic language in family of Cushitic language spoken by people live in Ethiopia and neighboring country. Afan Oromo word classifications are important in the writing system of the language, these includes noun, verb, adjectives etc. The morphology of Afan Oromo language is same as other language in broad categorization, like derivational and inflectional where inflectional is adapting new forms of word but the same in meaning and derivational is the creation of new words from already existing words. Unlike other language like Philippines, Afan Oromo does not have infix morpheme. The challenge of anaphora resolution in Afan Oromo language is identifying hidden anaphor from verb word classes. The other challenge encountered in this work is the absence of POS tagged corpus, so that we are forced to tag manually.

CHAPTER FOUR

AFAAN OROMO ANAPHORA RESOLUTION

In the previous chapters, we have discussed different concepts of Afan Oromo language behaviors and related works that is implemented for other foreign and local languages on anaphora resolution. This chapter gives a detail description of the architectural design and detailed implementation of Afan Oromo anaphora resolution Model (AOARS). In the first part, we discussed about architecture of the resolution system. Finally, we described about development of the prototype and the algorithm we were implemented.

4.1. Architecture

The design and implementation process of AOARS is developed for the resolution of independent personal pronouns and hidden personal pronouns inside words attached to a verb. The architecture depicted in Figure 4.1 has two main modules namely, resolution of independent anaphors and resolution of hidden anaphors.

The model consists of the following major components that are data preprocessing, identifying independent anaphors, identifying hidden anaphors, identifying candidates, applying resolution factors and choosing most appropriate candidate. Figure 4.1 below shows the general architecture of Afan Oromo anaphora resolution model. The architecture depicts the flow of process from input to output of the prototype.

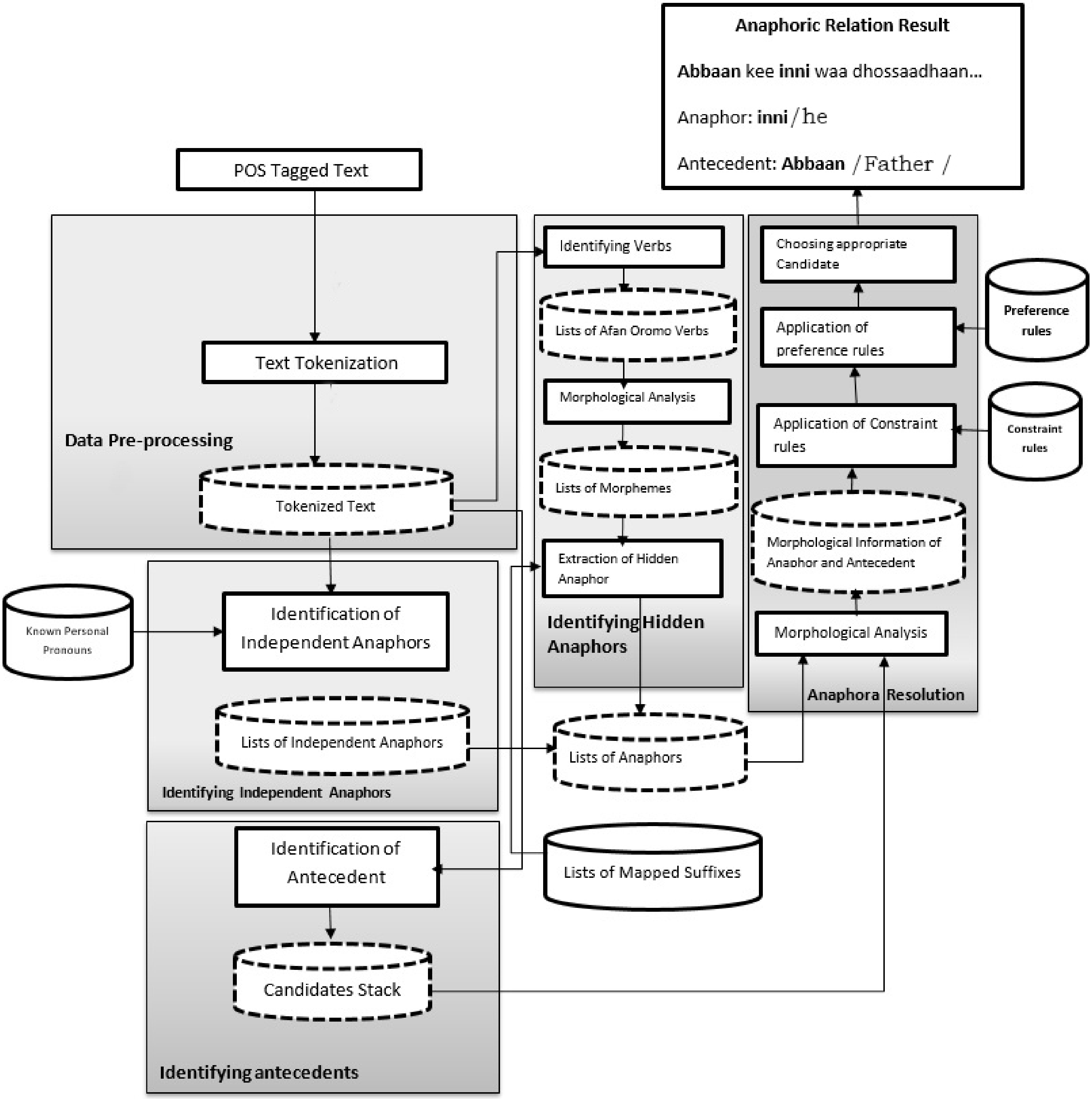

Anaphoric Relation Result
Abbaan kee inni waa dhossaadhaan...
Anaphor: inni /he
Antecedent: Abbaan /Father /
POS Tagged Text
Text Tokenization
Data Pre-processing
Tokenized Text
Identifying Verbs
Lists of Afan Oromo Verbs
Morphological Analysis
Lists of Morphemes
Extraction of Hidden Anaphor
Identifying Hidden Anaphors
Choosing appropriate Candidate
Application of preference rules
Preference rules
Application of Constraint rules
Constraint rules
Morphological Information of Anaphor and Antecedent
Morphological Analysis
Anaphora Resolution
Known Personal Pronouns
Identification of Independent Anaphors
Lists of Independent Anaphors
Identifying Independent Anaphors
Lists of Anaphors
Identification of Antecedent
Lists of Mapped Suffixes
Candidates Stack
Identifying antecedents

Figure 4. 1 General architecture of Afan Oromo anaphora resolution model

4.1.1 Data Pre-processing

The data for this thesis work is collected from Afan Oromo Bible and Fiction text documents. Before resolution starts the anaphor and antecedent relation different pre-processing tasks are involved. The main pre-processing techniques we have used are tokenization and POS tagging.

a. POS Tagging

The first steps in data pre-processing is POS tagging, which is the process of classifying word with the correct grammatical tag based on the context [34]. We have used POS tagging because it helps us to identify candidates and identify anaphor in addition it assist us to get number and gender of antecedent and anaphor from given input text. In addition, the output of POS tagging in our work is words accompanied by their POS tags and their number and gender expression. In anaphora resolution, POS tagging is helpful in a way that it facilitates and simplifies resolution by focusing only on words 'pronoun', 'noun' and 'noun phrase'. Even though, there are research works done on POS tagging for Afan Oromo; the works of (Getachew [41] and Mohammed [42]), we do not get fully implemented and freely available Afan Oromo POS tagger so far. So that, we tagged Afan Oromo sentences manually which we have used in resolution process.

Example,

Dhalachuuni_NP Yesuus_NN_S_M Kiristoosi_NN_S_M akkana_AD ture_VV.
[This is the genealogy of Jesus Christ the Messiah]
Yerusaalemis_NP_S_MF guutummaatti_JJ isa_PP_S_M wajjiin_PR rifatte_VV.
[When Herod the king had heard these things, he was troubled, and all Jerusalem with him.]
Hanga_AD dua_NN Herodisiittis_NP achi_AD jiraate_VV.
[Until the death of Harrods he stayed there]

In the above example, words in the sentence, *Dhalchuuni Yesuus Kiristoosi akkana ture*, are tagged with the correct word categories of Noun phrase, noun, adverb and verb respectively. The suffix NP, NN, AD and VV are tags of Noun phrase, noun, adverb and verb respectively. In addition, the _S and _M codes depicts the number and gender of the word, in this case singular and male respectively. For more examples, see Appendix G

b. Tokenization

The next steps in data pre-processing stages that is used, as input for our resolution process is tokenization of texts. Tokenization is the process of breaking up texts into different number of levels; paragraphs, sentences or words. In our cases, POS tagged text have broken up into sequence of sentences based on the three punctuations that are period (.), question marks (?) and exclamation marks (!). In addition, to assign each word unique word ID and Sentence ID we have broken up even the sentences to word level.

Similarly, other basic tasks in data pre-processing is removing non-alphanumeric characters from input texts. This task mainly helps us to clean our data from different unused special characters like comma, single and double quotation. Single quotation (‘) is common Afan Oromo words like ‘*taa’i / sitdown*’, this characters is called ‘*hudhaa*’. The algorithm used for the splitting of POS tagged senetnces presented on Figure 4.2.

```
Given POS Tagged Corpus
Read All POS Tagged Documents
Define Sentence delimiters: period, question mark and exclamation mark
For each sentence in the text
    Check sentence boundary
When there is delimiter
    Put each sentence on separate line and add to list
```

Figure 4. 2 Sentence Splitting Algorithm

4.1.2. Identification of independent Anaphors

After the data pre-processing task performed on input text provided by user completed, the identification of independent anaphora module started. In this module, the model checks if the current sentence contains an independent anaphor or not, in our case it looks only for the third person personal pronoun. These pronouns are identified by the morphological information that was previously tagged and saved by POS tagging for each word, i.e. if the word tag is “_PP_” then it is a pronoun. After those words are identified then we can match them against lists of known third person personal pronoun of Afan Oromo. Simultaneously, the identified anaphor gender and person are stored along with it. Once the model found one or more pronoun, the list of anaphora to be resolved is made and the resolver proceed to the next module that is finding appropriate candidate nouns for the anaphora. In case, the selected sentence does not contain any independent anaphor, the resolver will find on the next sentence.

4.1.3. Identification of Hidden Anaphors

Unlike English and other foreign languages only resolve independent anaphors, there are hidden anaphors in Afan Oromo language. The following sub components help us to find the hiddenanaphors from the given verb words.

Identification of verbs

As Afan Oromo anaphors hides under verbs, we need to first identify words that are marked as verb, _VV, by POS tagger that we have seen in previous section.

Morphological Analysis

After we have identified verbs, we proceed to analyzing their morphology like what suffixes they have and how their ending suffixes is interpreted in Afan Oromo language grammar. We have used our own identifier which do first take suffixes of verbs and compare it against the lists of personal pronouns they match. To check the performance of our analyzer we have used HornMorpho [43] that is freely available Afan Oromo, Amharic and Tigrigna Morphology analyzer to verify the result and see 85% of extracted hidden verbs are denoted with correct personal pronouns.

```
1. Given POS Tagged Corpus
2. Given suffixes
3. String listofVerbs
4. String suffixforhidden
4. If (words ends with _VV)
      Add to listofVerbs
5. Else
      break
6. If (listverbs.endswith= suffixes)
      Add to suffixforhidden
7. Else
      break
```

Figure 4. 3 Suffix Identification Algorithm

The following examples illustrates how our own identifier process the morphological information of verbs. Example,

1. Verb: **rifate**

Mapped Pronoun: inni

Number: Singular
Gender: Male

2. Verb: rifatte

Mapped Pronoun: ishee

Number: Singular
Gender: Female

3. Verb: **fudhatu** Mapped

Pronoun: isaan Number:

Plular Gender:

Male/Female

In the above example 1, the word ‘rifate is a verb and have suffixes of ‘-ate’ and the character before the last suffix ‘-te’ is ‘-a-’, it is Afan Oromo vowel, so the hidden anaphor will denote male third personal pronoun. If the character before the suffix ‘-te’ is Afan Oromo consonant letters it will show female personal pronoun; for instance, the word in example 2, ‘rifatte have suffix ‘-te’ and character before this suffix is vowel, ‘-t-’, so it can be considered as feminine personal pronoun ‘ishee/she’. Therefore, the verb ‘rifate’ in example 1; have ‘inni/he male third personal pronoun. On other hand, the verb ‘rifatte’ in example 2, have ‘ishee/she female third personal pronoun.

In the above example 3, the word ‘fudhatu is a verb and have suffix ‘-tu’ and In Afan Oromo grammar if a verb ends with ‘-tu’ (‘-tuu’ to show question) or ‘-tu’ it denotes the plural and neutral gender. So the verb ‘fudhatu in example 3, have ‘isaan/they neuter third personal pronoun. In addition, identified hidden anaphor have been assigned with the same word ID with the original verb.

Extraction of Hidden anaphors

In this sub component, we extracted hidden anaphors from identified verbs using the suffix information we obtained during morphological analysis. In the following example, we explain how hidden anaphors are extracted.

```
1. Given identified verb with suffix
2. Extract the suffix
3. Check suffix mapped to personal pronoun
4. Get personal pronoun
5. Store as hidden anaphor
```

Figure 4. 4 Extraction of Hidden Anaphors

Example 1,

Dhuf**te**:

Biliseen kaleessa **dhufte**. [Bilise came yesterday.]

POS: verb, root: <dhuf> Subject: 3, sing, fem

In the above example, the word ‘**dhufte**’ is a verb and have suffixes of ‘**-te**’ and the character before this suffix is ‘**-f-**’, is not Afan Oromo vowel, so the hidden anaphor will denote female third personal pronoun. If the character before the suffix ‘**-te’** is Afan Oromo vowel letters it will show male personal pronoun; for instance the word ‘**nyaate**’ have suffix ‘**-te’** and character before this suffix is vowel, ‘- **a’**, so it can be considered as masculine personal pronoun **‘inni/he’.**

Therefore, the verb ‘**dhufte**’ have ‘**ishee/she**’ female third personal pronoun. In addition, this hidden anaphor have same word ID with the original verb.

Example 2, Dhuf**te**

Baga nagan **dhufte**. [Wecome.]

POS: verb, root: <dhuf> Subject: 2, sing, neut.

Unlike previous example, in the Example 2 the word ‘**dhufte’** is a verb and have suffixes of ‘**-te’** and even if the character before this suffix is ‘-f-’, is not Afan Oromo vowel, the hidden anaphor will denote neutral second personal pronoun. Therefore, this type of verb class words can challenge the algorithm to extract the hidden anaphor because it requires the meaning of the sentence.

4.1.4. Identification of Antecedent

When a sentence contains anaphor, a resolver find antecedents to make a stack of candidates. Range of the identification for antecedent is four sentences: the sentence where the anaphora was found, the previous sentence, the sentence before the previous sentence and the sentence before previous of the former sentences. In the sentence where anaphor itself is found only the words preceding or words having word ID less than anaphor are considered (i.e. to avoid the attempt to resolve Cataphora). The word ID are generated globally for each words exists in the text. The identification of candidates are on the basis of their morphological data stored by POS tagging which looks for nouns (words of tag “_NN_”), noun phrases (words of tag “_NP_”) and pronouns (words of tag “_PP_”). For each identified antecedent their gender and number also saved along with it. After the candidate stack is, made resolution process continues by applying resolution factors.

4.1.5 Applying resolution factors

To get correct antecedent for the identified anaphor from candidates stack we need to apply some constraint and preference rules. The resolution process apply these two different resolution rules for each candidates in the list.

a. Constraint rules

Constraint rules are considered as obligatory conditions that are enforced on the relationship between antecedent and anaphor. In this work, we have used two basic constraints rules called gender and number agreement. Since the resolution requires both anaphor and antecedent must agree in number and gender it used as eliminative factor that drop out those candidate from candidate stack that does not agree. With the output of POS tagging all the nouns, noun phrases or pronouns annotated with number (singular or plural) and gender (feminine or masculine or neuter). After gender and number agreement, constraint is applied to the list of candidates and if only one candidate is match with gender and number of anaphor, this candidate can be declared as antecedent of the anaphor. If more than one candidate pass the matching, the model can apply the preference rules. However, in case with in search scope if no noun or noun phrase with a matching number and gender is found, then the process can be complete without antecedent.

b. Preference rules

In knowledge poor resolution systems, after application of the constraint rules, preference rules are used to evaluate the remaining candidates. Each preference rule has a score value associated with it, which is given to the scores of the candidates satisfying the preference to define the overall score of a candidate. At the end of the process, the candidate with the highest aggregate score is proposed as the antecedent.

The score assigned to candidates for preference rules differ language-to-language and researcher to researcher. Some researcher argued as what matter is the comparative relation between factors and supports the award can be arbitrary value. For English language, score ranges from -1 to +2 [7] and for Norwegian language points in the range of -100 to 100 was used [13]. In the same way, Algorithm for Pronominal Anaphora Resolution of English language uses range of 0 to 100 [19]. The positive points are awarded to factors to promote antecedent candidates, while the negative points are awarded to factors to discourage candidates, like indefinite nouns in English.

There has been a broad agreement in the literature that range of the points that a system awards is arbitrary [45] and that what matters is the comparative relation among factors [19]. In the following section, we discussed the preference rules we have used in our work. In general, the factors that the algorithm uses to calculate salience are given different weights according to how relevant the factor is. Afan Oromo resolution uses points in the range of +0.125 to +1.0.

Recency: is a preference factor which checks if the identified candidate is more recent to the anaphor or not. The most recent one is one closest to the anaphor after calculating using their word ID we assigned +1.0 point to the candidate. This point is assigned to the indicator because from our collected datasets the highest number, around 40% of anaphor-antecedent relation occur due to recency of the candidate.

Subject place: the first noun or noun phrase in a sentence is more considered as antecedent than the one not found on the first position. In our work, the fist noun or noun phrase is considered as the subject of the sentence so the resolution system assign +0.75 point to those subject placed candidates. The point is assigned to this indicator because from our collected datasets the second highest number of anaphor-antecedent relation is occur due to subject placed of the candidate.

Boost Pronoun: is the preference factor when the pronoun candidates in the current and previous sentences are regarded as antecedent. The reason Pronouns are considered as candidates for antecedent is they are more salient and the antecedent noun phrases can be out of rang. So that, we awarded +0.50 point to the pronoun candidates. This point is assigned to the indicator because from our collected datasets the third highest number of anaphor-antecedent relation is occur due to pronouns occurred as candidate.

Definiteness: in Afan Oromo language there is no special marker to denote definiteness, rather we use different suffixes attached to nouns [5]. As definite noun phrases are seen as more likely antecedents than indefinite ones. We assign +0.25 point to the definite candidates. This point is assigned to the indicator because from our collected datasets the fourth number of anaphor-antecedent relation is occur due to the definiteness of the candidate.

Frequency Indicator: the nouns and pronouns that appeared frequently in the text are preferred over nouns and pronouns appeared only once. Since this factor is number of repetition of words, the candidate mentioned repetitively (more than once) is awarded +0.125 point. The point is assigned to this indicator because from our collected datasets the least number of anaphor-antecedent relation is happens due to frequent occurrence of the candidate.

4.1.6. Selecting appropriate Candidate

After all factors, which are constraints and/or preference are considered, the candidate stack is reordered based on score. The candidate with the highest aggregate score is selected as the correct antecedent. If two or more candidates score the same points, the one recent or closer to the anaphor is chosen.

4.2. Designing Afan Oromo Anaphora Resolution Model (AOARM)

Development of AOARM prototype consists the design and algorithm development for Independent and hidden anaphora resolution based on Mitikov knowledge poor approach [7]. The AOARM development are based on the Java programming language. As such, a Java Runtime Environment is required. The AOARM consists of five basic classes each of them have different type of methods that are called in process accordingly. One of these classes is Anaphora controller that helps us to provide lists of anaphora in the given sentences. The second class is Antecedent Controller, which are important in identifying nouns and head nouns including pronouns as candidates. The other important class is Gender Agreement, which assists us in checking gender for both identified anaphora and antecedent. In addition, the fourth is Number Agreement, which helps us to get the number of anaphor and antecedent. Finally, we have the main class called Resolve, which includes the main logic to resolve the relationship between anaphor and antecedent. In the following example, we demonstrate the resolution process of independent anaphor.

Example, **Herodis Mootichi** yommuu waan kana dhagayetti ni rifate.Yerusaalemis guutummaatti ***isa*** wajjiin rifatte. [When **Herod the king** had heard these things, he was troubled, and all Jerusalem with **him**.]

In the above sentence, '***isa***'/'**him**' is identified anaphor and '***Herodis'***, '***Mootichi'*** and '***Yerusaalemis'*** is identified candidate antecedent. First, the prototype gets the gender and number value of the anaphor and antecedent. In this case, the gender and number value of '***isa'*** anaphor is ***masculine*** and ***singular*** respectively. Next, it will find the value of gender and number for candidates, for '***Herodis'*** it is ***masculine*** and ***singular***, When we check Candidate '***Mootichi'*** it is ***masculine*** and ***singular.*** In addition, candidate '***Yerusaalemis'*** have value of '***neuter'*** and ***Singular.*** From the constraint module,it will get only two candidate antecedent as output namely '***Herodis'*** and '***Mootichi'*** , because identified candidate antecedent '***Yerusaalemis'*** does not fulfill the gender constraint rule against anaphor. Having both candidate antecedent as input, it will apply preference rule for each of them. So that, '***Herodis'*** is at subject place awarded total score of 0.75 = (SUBJECT=0.75, RECENCT=0.0, BOOST=0.0, DEFINITENESS=0.0, FREQUENCY=0.0). Similarly, '***Mootichi***' is more recent to the anaphor and definite type of noun so it awarded total score of 1.25= (SUBJECT=0.0, RECENCT=1.0, BOOST=0.0, DEFINITENESS=0.25, FREQUENCY=0.0). With this value, candidate '***Mootichi***' have higher aggregate score than candidate '***Herodis'.*** The noun phrase '***Mootichi***'/***The King*** selected as antecedent for anaphor ***'isa' / him*** due to its highest score identified by the prototype. The independent pronoun resolution procedure algorithm described in the following Fig4.5.

```
Given number of sentences (i)
For i to size of sentence array
      If sentence[i] contains _PP tagged word
            Add to list of pronouns (j)
      Else
            Go to sentence [i+1]
      For j to size of list of pronouns
            If word [j] = known pronoun
                  Add to lists of independent anaphor (k)
            Else
            Go To Word [j+1]
      End for
      For k to size of list of independent pronoun
            If sentence [i] contains words tagged by _NN, _NP or _PP
                  Add to candidate list (l)
            Else
                  Go to sentence [i-1] with in defined scope
            End if
            For l to size of candidate list
                  If gender and number agreement of anaphor and antecedent equal
                        Add to preferred candidate list (m)
                  Else
                        Drop the candidate
                  End If
            End for
            For m to size of preferred candidate list
                  Execute preference rule on preferred candidates
            If one member score highest score
                  Declare the member as antecedent
            Else if two member score equal highest aggregate score
                  Declare member recent to the anaphor as antecedent
            Else
                  Declare as no antecedent found
            End if
            End For
        End For
End For
```

Figure 4. 5 Independent Anaphor Resolution algorithm

In the process of hidden anaphor resolution, like independent anaphor resolution and starts by reading the first sentence of the dataset. After loading current sentence the algorithm searches for verbs to identify any hidden anaphors with in sentence. This is accomplished by extracting words that is previously tagged as '_VV' to denote words as verb category. For identified verbs, the algorithm finds ending character and compare it with predefined known ending, if this ending character mentioned in the list it categorized as respective hidden personal pronouns. The hidden pronoun resolution procedure algorithm described in the following Fig 4.4. The following example demonstrates how hidden anaphor is extracted from verbs.

Example, 1. **Yooseefis** kahee mucichaa fi haadha mucichaa fudhatee halkaniin gara Gibxitti **sokke**.

Yospeh take the son and his mother to live in Egypt.

From above sentence, the identified verb is the word '**sokke'**. The ending character of this word is **'e'** and the character before the ending character is not '**t'** ; then it is mapped to masculine and singular personal pronoun '**inni/he**' in the list of ending character to personal pronoun record. So the identified hidden anaphor '**inni/he**' which is hides in verb '**sokke**' is refers to the antecedent '**Yooseefis**'.

2. Haati isaa **Maariyaam** kaadhimaa Yooseefi **turte. [**His mother Merry is a wife of Yoseph.]

In above sentence identified verb is '**turte**'. The ending two character of this word is '**te**' and the character before this endings, which is on third position is not vowel **– 'r'.** So from mapping of ending characters with pronouns, the verbs contains the feminine and singular personal pronoun '**ishee/she**'. Finally, the antecedent '**Maariyaam**' is selected as referent to the hidden pronoun '**ishee/she**'. The hidden pronoun resolution procedure algorithm described in the following Fig 4.6.

```
Given number of sentences (i)
For i to size of sentence array
    If sentence[i] contains _VV tagged word
        Add to list of Verbs (j)
    Else
        Go to sentence [i+1]
    For j to size of list of verbs
        If word [j] ending= pre-defined suffix
            Add to lists of suffixes (k)
        Else
            Go To Word [j+1]
    End for
    For k to size of list of suffixes
        If suffixes [j] mapped to known pronoun
            Add to lists of hidden anaphor (l)
        Else
        Go To suffixes [k+1]
    End for
    For l to size of list of hidden pronoun
        If sentence [i] contains words tagged by _NN, _NP or _PP
            Add to candidate list (m)
        Else
            Go to sentence [i-1] with in defined scope
        End if
        For m to size of candidate list
            If gender and number agreement of anaphor and antecedent equal
                Add to preferred candidate list (n)
            Else
                Drop the candidate
            End If
        End for
        For n to size of preferred candidate list
            Execute preference rule on preferred candidates
        If one member score highest score
            Declare the member as antecedent
        Else if two member score equal highest aggregate score
            Declare member recent to the anaphor as antecedent
        Else
            Declare as no antecedent found
        End if
        End For
     End For
  .
```

Figure 4. 6 Hidden Anaphor Resolution algorithm

The extraction of hidden anaphor from verbs are implemented on different types of Afan Oromo language verb categories. The most widely known and occurred in Afan Oromo texts are Regular Verbs, Double-consonant Ending Stems, -chuu Verbs and Vowel-Ending Stems (Irregular Verbs). Table 4.1 below summarizes types of Afan Oromo verbs we have implemented in our prototype.

Table 4. 1 Afan Oromo verb categorization

Types of Verbs	**Root Word**	**Example of Verbs**
Regular Verbs	Deemuu – 'to go'	Deem**ti**, Deem**a**, Deem**u**
Double-consonant Ending Stems	Arguu – 'to see'	Arg**iti**, Arg**a**, Arg**uu**
-chuu Verbs	Nyaachuu – 'to eat'	Nyaa**tti**, Nyaat**a**, Nyaat**u**
Irregular verbs	Du'uu – 'to die'	Duu**ti**, Du'**a**, Du'**u**

Based on the above Afan Oromo language verb category grammar concept; the mapping of verbs suffix to personal pronouns are presented in Table 4.2 [5].

Table 4. 2 The mapping of verb suffix to personal pronouns

Suffix	**Personal Pronoun**		**Example**	**Remarks**
	Afan Oromo	**English**		
-a	Inni	He	Deem**a**	
-ta	Inni	He	Jaalla**ta**	
-te	Inni/Ishee	He/She	Nyaa**te,** Nyaat**te**	Based on '**t**' we will identify as male or female
-tee	Inni/Ishee	He/She	Nyaa**tee,** Nyaa**ttee**	To show Continuative and based on '**t**' we will identify as male or female
-e	Inni	He	Arg**e**	
-ee	Inni	He	Arg**ee**	To show Continuative
-ti	Inni	He	Beekee**ti**	
-e	Ishee	She	Booss**e**	
-ee	Ishee	She	Booss**ee**	To show Continuative
-te	Ishee	She	Deem**te**	
-tee	Ishee	She	Deem**tee**	To show Continuative
-ofti	Ishee	She	Haas**ofti**	
-eessi	Ishee	She	Dand**eessi**	
-ssi	Ishee	She	Boo**ssi**	
-uti	Ishee	She	Duu**ti**	
-tti	Ishee	She	Nyaa**tti**	
-iti	Ishee	She	arg**iti**	
-ti	Ishee	She	Deem**ti**	
-u	Isaan	They	Dandah**u**	
-tu	Isaan	They	Nyaa**tu**	

CHAPTER FIVE

EXPERIMENT

The focus of this chapter is on testing and evaluating of the proposed model. In this chapter, we discussed about dataset gathering and preparation, implementation of the prototype, evaluation of prototype and finally experimental results.

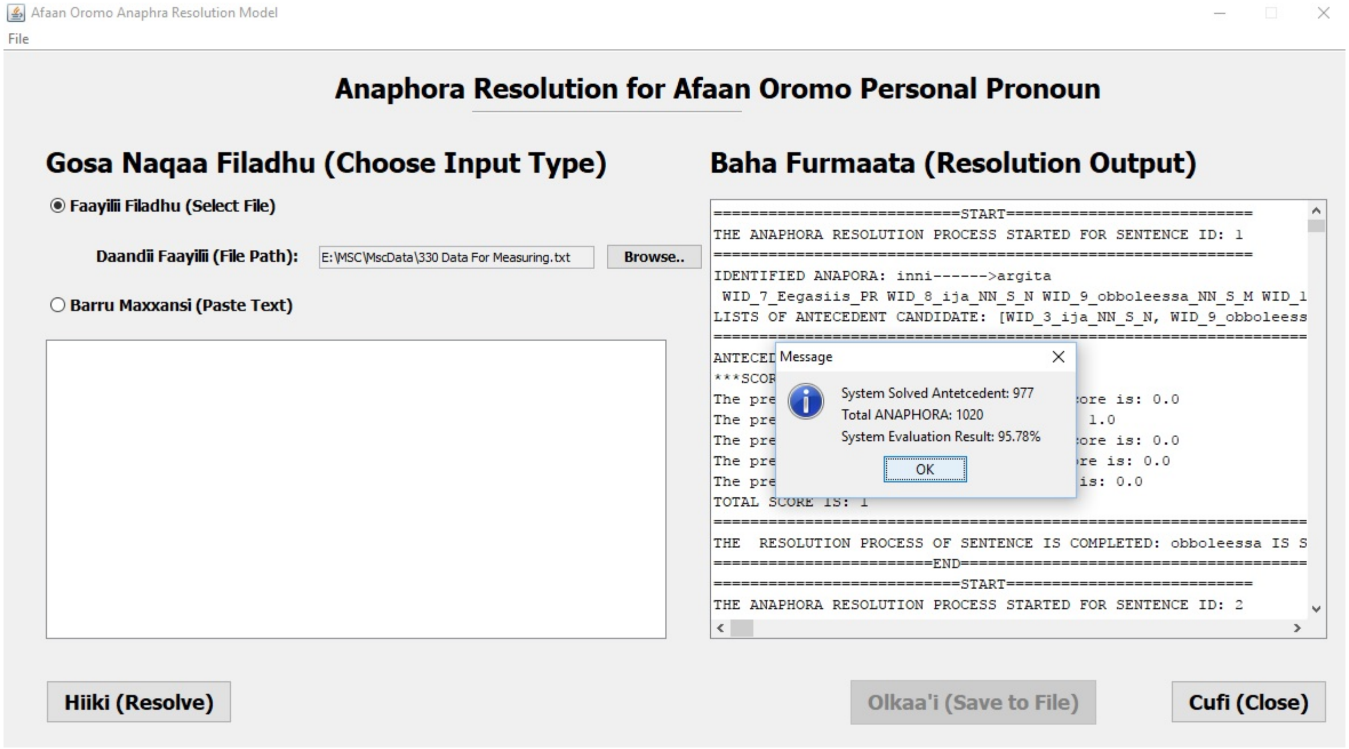
5.1. Dataset Gathering and Preparation

The anaphora resolution requires dataset that contains grammatically correct sentences. However, there is no authorized and publicly available tagged Afan Oromo text for any work of NLP. This is also true for anaphora resolution also. As a result, we developed Afan Oromo dataset. Our first task before developing the dataset was deciding the domain of our study. We decided to work on the Holy Bible domain and fiction. Accordingly, Afan Oromo sentences gathered from Afan Oromo Holy Bible and small amount of Afan Oromo Fiction gathered from online resource for testing and training. We have collected 330 sentences (307 from bible and 23 from fiction) with 4383 words and having 261 independent pronouns plus 1174 verbs, which is used to extract hidden anaphors. The data was first tokenized in accordance with the consistent tokenization procedure; this includes removing special characters and numeric characters. Since the researcher does not get the POS tagged Afan Oromo corpus, we have tagged our dataset manually with the help of linguists.

For instance, the POS tagging morphological category representation described in Appendix H; the word attached with suffix _VV denotes verb and _NN indicate noun, and second representation is suffix with _S to indicate word is singular and with suffix _M to indicate the gender of the anaphor and candidate is male. Before the actual resolution process starts, there is an algorithm, which generates word ID and Sentence ID for each of words in the dataset, which helps us later on preference rule application. Both word ID and Sentence ID of each word is stored in hash map for later use throughout the process.

5.2. Implementation of Prototype

We have proposed a prototype, which can take user's natural language tagged sentences and after going through all mentioned steps, it will delivers resolution result to the user. The development environment is the NetBeans 8.2 editor, which is an open-source platform. A Graphical User Interface (GUI) is developed for the model using the Swing package of Java. This GUI enables the user to input a text file (.txt file extension) in the local machine containing the preprocessed text to be resolved, the output of the model is written to a text file in the path that is predefined. In addition, GUI also enables the user to enter the text to be resolved to an input text area and the output text is written to an output text area. The model is developed and tested on a model with Intel® Core™ i5-6200U CPU @ 2.30GHZ, 8GB RAM and 500 SDD on Window 10 operating system. `


Afaan Oromo Anaphra Resolution Model
File
Anaphora Resolution for Afaan Oromo Personal Pronoun
Gosa Naqaa Filadhu (Choose Input Type)
Baha Furmaata (Resolution Output)
Browse..
○ Barru Maxxansi (Paste Text)
Hiiki (Resolve)
Olkaa'i (Save to File)
Cufi (Close)
Message
OK

● Faayilii Filadhu (Select File)

Daandii Faayilii (File Path): E:\MSC\MscData\330 Data For Measuring.txt

```
==========================START===========================
THE ANAPHORA RESOLUTION PROCESS STARTED FOR SENTENCE ID: 1
==========================================================
IDENTIFIED ANAPORA: inni------>argita
 WID_7_Eegasiis_PR WID_8_ija_NN_S_N WID_9_obboleessa_NN_S_M WID_1
LISTS OF ANTECEDENT CANDIDATE: [WID_3_ija_NN_S_N, WID_9_obboleess
=================================================================
ANTECED
***SCOR
The pre                                  ore is: 0.0
The pre                                   1.0
The pre                                  ore is: 0.0
The pre                                  re is: 0.0
The pre                                  is: 0.0
TOTAL SCORE IS: 1
=================================================================
THE  RESOLUTION PROCESS OF SENTENCE IS COMPLETED: obboleessa IS S
========================END=======================================
==========================START===========================
THE ANAPHORA RESOLUTION PROCESS STARTED FOR SENTENCE ID: 2
```

System Solved Antetcedent: 977
Total ANAPHORA: 1020
System Evaluation Result: 95.78%

Figure 5.1 Screenshot of Afan Oromo Resolution GUI

5.3. Evaluation Result

To measure the performance prototype algorithm we have used success rate metric, which is the ratio of number of successfully resolved anaphors with number of all anaphors expressed as a percentage [7]. Since this measure focuses on the performance of the algorithm and not on any pre- processing component, the precise success rate will be obtained if the input to the anaphora resolution algorithm is either post-edited by humans or extracted from an already tagged corpus.

$$\text{Success Rate} = \frac{\text{Number of Correctly Resolved anaphors}}{\text{Number of All Anaphors}} \times 100\ \%$$

The overall performance of the prototype is measured for two different scenario based on anaphor types, identified based on location of anaphor and antecedent.

5.3.1. Performance Of Intrasentential Anaphor

The evaluation of the prototype is performed on dataset and the algorithm considers only candidates that exists in the same sentence with anaphor, which means all candidates resides in the previous sentence of anaphor sentence is dropped.

The overall performance of Afan Oromo Anaphora Resolution prototype for independent personal pronouns are conducted for Intrasentential. The evaluation result for the intrasentential sentence is scored a success rate of 47.51%. Similarly, the overall performance of Afan Oromo Anaphora Resolution prototype for hidden personal pronouns are conducted for Intrasentential. Evaluation for Intrasentential is conducted on the proposed algorithm and success rate obtained is 57.84%. Finally, the performance of Afan Oromo Anaphora Resolution prototype for both independent and hidden personal pronouns are conducted for Intrasentential. The evaluation result for the intrasentential sentence is scored a success rate of 55.2%. The detail description for the result is presented in Table 5.1.

Table 5. 1 Performance result of algorithm for intrasentential anaphora resolution

Anaphor type	**Total Anaphor**	**Non-resolved anaphor**	**Resolved Anaphor**	**Success Rate**
Independent	261	137	124	47.51%
Hidden	759	320	439	57.84%
Both	1020	457	563	55.2%

5.3.2. Performance Of Intersentential Anaphor

The evaluation of the algorithm is performed for test data and it favored candidates exists in current sentence and previous sentences. Therefore, the search scope for the algorithm we conducted or evaluation is twenty-three sentence. For better understanding, the experiment is conducted for each of the following scenarios. Moreover, the result is explained in the following Table 5.2.

Table 5. 2 Performance result of algorithm for intrasentential anaphora resolution

Number of Previous Sentence	Anaphor Type	Total Anaphor	Non-resolved anaphor	Resolved Anaphor	Success Rate
One	Independent	261	29	232	88.89%
	Hidden	759	185	574	75.63%
	Both	1020	214	806	79.02%
Two	Independent	261	7	254	97.32%
	Hidden	759	135	624	82.21%
	Both	1020	142	878	86.08%
Three	Independent	261	4	257	98.47%
	Hidden	759	102	657	86.56%
	Both	1020	106	914	89.61%
Four	Independent	261	5	256	98.08%
	Hidden	759	75	684	90.12%
	Both	1020	80	940	92.16%
Five	Independent	261	3	258	98.85%
	Hidden	759	64	695	91.57%
	Both	1020	67	953	93.43%
Six	Independent	261	3	258	98.85%
	Hidden	759	56	703	92.62%
	Both	1020	59	961	94.22%
Seven	Independent	261	3	258	98.85%
	Hidden	759	53	706	93.02%
	Both	1020	56	964	94.51%
Eight	Independent	261	3	258	98.85%
	Hidden	759	45	714	94.07%
	Both	1020	48	972	95.29%
Nine	Independent	261	3	258	98.85%
	Hidden	759	43	716	94.33%
	Both	1020	46	974	95.49%
Ten	Independent	261	3	258	98.85%
	Hidden	759	40	719	94.72%
	Both	1020	43	977	95.78%
Eleven	Independent	261	3	258	98.85%
	Hidden	759	39	720	94.86%
	Both	1020	42	978	95.88%
Twelve	Independent	261	3	258	98.85%
	Hidden	759	37	722	95.12%
	Both	1020	40	980	96.08%
Thirteen	Independent	261	3	258	98.85%
	Hidden	759	36	723	95.26%

	Both	1020	39	984	96.47%
Fourteen	Independent	261	3	258	98.85%
	Hidden	759	31	728	95.91%
	Both	1020	34	986	96.67%
Fifteen	Independent	261	3	258	98.85%
	Hidden	759	29	730	96.17%
	Both	1020	32	988	96.86%
Sixteen	Independent	261	3	258	98.85%
	Hidden	759	27	732	96.44%
	Both	1020	30	990	97.06%
Seventeen	Independent	261	3	258	98.85%
	Hidden	759	26	733	96.57%
	Both	1020	28	992	97.25%
Eighteen	Independent	261	3	258	98.85%
	Hidden	759	23	736	96.96%
	Both	1020	26	994	97.45%
Nineteen	Independent	261	3	258	98.85%
	Hidden	759	21	738	97.23%
	Both	1020	24	996	97.65%
Twenty	Independent	261	3	258	98.85%
	Hidden	759	20	739	97.36%
	Both	1020	23	997	97.75%
Twenty-one	Independent	261	3	258	98.85%
	Hidden	759	16	743	97.89%
	Both	1020	19	1001	98.14%
Twenty-two	Independent	261	3	258	98.85%
	Hidden	759	15	744	98.02%
	Both	1020	18	1002	98.24%
Twenty-three	Independent	261	3	258	98.85%
	Hidden	759	13	746	98.28%
	Both	1020	16	1004	98.43%
Twenty-four	Independent	261	3	258	98.85%
	Hidden	759	13	746	98.28%
	Both	1020	16	1004	98.43%
Twenty-Five	Independent	261	3	258	**98.85%**
	Hidden	759	13	746	**98.28%**
	Both	1020	16	1004	**98.43%**

The above table depicts, the results of the prototype evaluation vary from form of anaphor (hidden or independent anaphor) to the location of anaphor (Intrasentential or Intersentential). The performance measure of intrasentential anaphora resolution ranges from 47.51% to 57.84%, and for the case of

Intersentential anaphora, resolution vary from 75.63% to 98.85%. Specifically, the evaluation result for hidden pronoun anaphora resolution ranges from 75.63% to 98.28% and the evaluation result for independent pronoun anaphora resolution ranges from 88.89% to 98.85%. Finally, the evaluation result for both anaphor type (hidden and independent) anaphora resolution ranges from 79.02%% to 98.43%.

In general, for intrasentential independent anaphor, the algorithm scored 47.51% and for Intrasentential, hidden anaphor the algorithm scored 57.84% and for both anaphor type in Intrasentential algorithm scored 55.20%. For Intersentential, the algorithm is evaluated in twenty-three defined scope, then independent anaphora scored 98.85% (result already scored in fifth defined scope) and hidden anaphora scored 98.28% and for both anaphor type algorithm scored 98.43%. Even if, the search scope extended to twenty-fifth the algorithm scored same result with that of twenty-three.

5.4. Discussion of results

Still, there is a lot to improve. During the experiment, the following cases are evolved that the program could not resolve.

The correct antecedent is exist as two words like full name, but the program is only able to find an antecedent that appear as one word. In the following example shows where the algorithm fails to find two antecedent.

Example,

Herodis Mootichi yommuu waan kana dhagayetti ni rifate.Yerusaalemis guutummaatti **isa** wajjiin rifatte.
[When **Herod the king** had heard these things, he was troubled, and all Jerusalem with **him**.**]**

The correct antecedent for anaphor '**isa**'/'**him**' selected by the algorithm is the word '**Mootichi**/**The King**' only, but the anaphor refers both '**Herodis'** / '**Herod'** and '**Mootichi'** / '**the king'**.

There is a case where correct antecedent and anaphor referring to it is not match in gender agreement. Since all gender mismatching candidates are dropped, so the correct antecedent might be removed as well. The example given below illustrates the mismatch of an anaphora and its antecedent.
Example,

Yerusaalemis guutummaatti isa wajjiin **rifatte**.

All Jerusalem troubled with him.

Here the verb ‘**rifatte**’ have hidden personal pronoun ‘**ishee/she**’ and it is female gender and singular in number. For this anaphor the possible candidates in this sentence is the noun ‘**Yerusaalemis**’. The candidate word is singular in number and unspecified gender as it is non- animate subject. Therefore, the prototype eliminate the proposed candidate because of gender mismatch with the anaphor in constraints rule application stage.

CHAPTER 6

CONCLUSION AND RECOMMENDATION

This chapter summarizes our approach to resolving anaphora referent for Afan Oromo personal pronouns. It also lists future works for improving the anaphora resolution system.

6.1 Conclusion

This study deals with presented paper introduces the domain of anaphora resolution and gives a detail description of the prototype implemented for the resolution of third person pronouns in Afan Oromo language for both hidden and independent personal pronouns. To implement this we have used knowledge poor approach, because it does not require semantic or deep syntactic knowledge only operates on the output of a POS tagger. The algorithm we have implemented was mainly based on the Mitikov knowledge poor approach, which was implemented for English and later implemented successfully for Arabic and Portuguese.

The prototype for Afan Oromo language Anaphora Resolution model has been implemented in Java and POS tagged dataset containing Holy Bible and fiction with 130 sentences having 270 verbs used for extracting hidden personal pronouns. In addition, 200 sentences having 165 independent personal pronouns. The tagging of the collected datasets are conducted manually with the help of linguistics due to unavailability of Afan Oromo public corpus. The major components contained in the prototype include data preprocessing, identification of anaphor, identification of candidates and resolution of the anaphor using constraints and preferences.

The foremost steps implemented in our algorithm was to identify sentences exist in the given text using delimiter ".", "?" and "!". Then it finds the third person pronoun ("inni/"he","ishee/she" and "isaan/they") in the text whether extracting from verbs for hidden pronouns or independent personal pronouns and check against lists of Afan Oromo language personal pronoun. If alike found, then it searches lists of candidates that precede the anaphor in three sentence range including anaphor sentence. The gender and number agreement filter, which is called constraints, applied to the stack of candidates. If more than one candidate remains after constraints are applied, the algorithm proceeds to apply preference factors those are customized based on language specific. Finally, the candidate with the highest aggregate score is selected as the correct antecedent.

The performance of the prototype is measured using the success rate metric. The evaluation of the prototype is performed on 3330 sentences with 4383 words and having 261 independent pronouns plus 1174 verbs, which is used to extract hidden anaphors. Performance Measurement conducted for two different scenarios. First, the hidden intrasentential anaphor algorithm scored a success rate of 57.84% and for independent intrasentential, anaphor the algorithm scored 47.51% success rate. For both Intrasentential anaphor, the algorithm scored a success rate of 55.20%. On the other scenario, the algorithm scored success rate of 98.28% for hidden Intersentential anaphor algorithm and 98.85% for independent Intersentential algorithm. For both Intersentential anaphor, the algorithm scored success rate of 98.43%.

6.2. Recommendation

Anaphora resolution is a new study for Afan Oromo language. The task is a very complex for such under resourced languages, which requires more time and needs a different level of knowledge such as morphological, syntactic, semantic, world knowledge and another type of knowledge's which are complex to deal with. The developed Afan Oromo anaphora resolution model has parts that require further improvements and below are some of the recommendations we propose for future work.

- Implementing in case where there are two or more word candidate, like full name, are antecedent for single anaphor
- The Cataphor issue needs to be addressed in future study as it is overlooked in our work
- Efforts should be made towards incorporating Afan Oromo language POS tagger tools to make the system fully automated
- Integrating morphological analyzer within the system to extract hidden personal pronouns from verbs to make system fully automatic and increase the overall performance
- Extending this work to other pronominal and other types of anaphora, such as possessive pronoun anaphor and verb anaphor
- Incorporating of more preference factors which are specifically applied to Afan Oromo language

REFERENCE

[1] Central Statistics Agency. (2017, March 08). *National Statics Abstract Population* [online].

Available: www.csa.gov.et.

[2] Dilek Kucuk, "A Knowledge-Poor Pronoun Resolution System for Turkish," M.S. thesis, Dept. Computer Eng., Middle East Technical University, Ankara, Turkey, 2005

[3] Ruslan Mitikov, *Anaphora Resolution*, London, Britain: Pearson Education press, 2002.

[4] Tyne Liang and Dian-Song Wu, "Automatic Pronominal Anaphora Resolution in English Texts", Computational Linguistics and Chinese Language Processing, China, 2004

[5] Mohammed and Andrzej Zaborski. *Handbook of the Afan Oromo Language,* Warsaw, Poland: Polish Academy of Sciences,1990.

[6] Mylanguages.Org. (2017, March 08). *Oromo Lessons, Oromo Pronouns* [Online].

Available: www.mylanguages.org.

[7] Ruslan Mitkov. "Anaphora resolution: The state of the art". Tech. Rep. University of Wolver Hampton, UK.1999.

[8] Hobbs, J. R. "Pronoun Resolution". Res. Rep. 76-1, Department of Computer Science, City College, New York. 1976

[9] Tejaswini Deoskar, "Techniques for Anaphora Resolution: A Survey", May 2004.

[10] Massimo and Mijail, "General-Purpose, off-the-shelf Anaphora Resolution Module: Implementation and Preliminary Evaluation", Dept. Computer Science, Colchester, UK. 2004.

[11] Ruslan Mitkov, Richard Evans, and Constantin Orasan, "A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method", School of Humanities, Languages and Social Sciences, Wolverhampton, UK, 2002.

[12] Gustav Algotsson, "Automatic Pronoun Resolution for Swedish", MSc. Thesis, Dept. of Computer Science, Stockholm, Sweden, 2007.

[13] Gordana Ilic Holen, "Automatic Anaphora Resolution for Norwegian", Ph.D. dissertation, Dept. of Linguistics and Scandinavian Studies, Oslo, Norway, 2006.

[14] Pilleriin Mutso, "Knowledge-poor Anaphora Resolution System for Estonian", MSc. Thesis., Dept. Computer Science, Tartu, Estonia, 2008.

[15] Temesgen Dawit, "Design of Amharic Anaphora Resolution Model", MSc. Thesis. Dept. of Computer Science, Addis Ababa, Ethiopia, 2014

[16] Manuel Palomar and Antonio Ferrández, “An Algorithm for Anaphora Resolution in Spanish Texts”, Computational Linguistics, Volume 27 Number 4, 2001.

[17] Lucie Kucova and Zdenek Zabokrtsky, “Anaphora in Czech: Large Data and Experiments with Automatic Anaphora Resolution”, Institute of Formal and Applied Linguistics, Prague, Czech Republic, 2005.

[18] Chaves and Rino, “The Mitkov algorithm for anaphora resolution in Portuguese”, TALN, Avignon, France, 2008

[19] Shalom Lappin and Herbert J. Leass, “An Algorithm for Pronominal Anaphora Resolution,” SOAS, University of London, 1994.

[20] Yannick Versley, Simone Ponzetto, Massimo Poesio AND Vladimir Eidelman, BART: A Modular Toolkit for Coreference Resolution, Association for Computational Linguistics Human Language Technologies, 2008.

[21] Cunningham, H. *Information extraction automatic- Encyclopedia of Language and Linguistics*. Sheffield, UK: Oxford Press, 2005.

[22] Ralph Grishman, *Computational linguistics an introduction*, New York, USA: Cambridge University Press, 1986.

[23] Hamish Cunningham, Information Extraction, Department of Computer Science, University of Sheffield, Sheffield, UK, 2004

[24] Emili Sapena, Lluıs Padró and Jordi Turmo, “Coreference Resolution Survey”, TALP Research Center, Barcelona, Spain, 2008.

[25] Chaltu Fita, “Afan Oromo List, Definition and Description Question Answering System”, MSc. Thesis. Dept. of Computer Science, Addis Ababa, Ethiopia, 2016.

[26] R.Mitkov. “Robust Pronoun Resolution with limited knowledge”, In Proceedings of the 36th Annual Meeting of the Association for computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Canada: 869-875. 1998.

[27] Central Intelligence Agency (USA). (2017, April 20). *The World Fact Book* [online]. Available: https://www.cia.gov/library/publications/the-world-factbook/geos/et.html.

[28] Jurafsky and Martin, *An introduction to natural language processing computational linguistics, and speech recognition*, New Jersey, USA: Prentice-Hall Inc, 2007.

[29] Kula Kekeba Tune and Vasudeva Varma, Oromo-English Information Retrieval Experiments, IIIT-Hyderabad, India, CLEF, 2007

[30] Wakweya Olani. (2017, May 10). *Inflectional Morphology in Oromo* [online]. Available: https://www.academia.edu.

[31] Stanford University. (2017, July 05). *Chunking example* [online]. Available: http://deepdive.stanford.edu/example-chunking.

[32] Allen James, Natural Language Understanding, University of Rochester, The Benjamin/Cummings Publishing Company Inc, California,USA,1995

[33] Imran Q. Sayed. 2003. Issues in Anaphora Resolution. Project Report. Stanford University, USA

[34] Ruslan Mitkov, *The Oxford Handbook of Computational Linguistics*, New York, USA: Oxford University Press, 2003.

[35] Michel Denber. "Automatic Resolution of Anaphora in English", Technical report, Imaging Science Division, Kodak. 1998.

[36] Ruslan Mitkov. "Outstanding issues in anaphora resolution (invited talk)", In Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing. Mexico City, Mexico 2001.

[37] Ruslan Mitkov. "Factors in anaphora resolution: They are not the only things that matter. A case study based on two different approaches", ANARESOLUTION '97 Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts. pp 14-21, 1997.

[38] Ruslan Mitkov and Lamia Belguith. "Multilingual robust anaphora resolution", 1998.

[39] Brennan, S.E., Friedman, M.W. & Pollard, C.J. (1987). 'A Centering approach to pronouns.'

Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, Stanford, Calif., 155-162.

[40] Niyu Ge, Jon Hale and Eugene Charniak. "A statistical Approach to Anaphora Reoslution", 1998.

[41] Getachew Mamo, "Part-Of-Speech Tagging for Afaan Oromo Language", MSc. Thesis. Dept. of Computer Science, Addis Ababa, Ethiopia, 2009.

[42] Mohammed Hussein, "Part Of Speech Tagger for Afaan Oromo Language Using Transformational Error Driven Learning (Tel) Approach", MSc. Thesis. Dept. of Computer Science, Addis Ababa, Ethiopia, 2010.

[43] Michael Gasser. "HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya", Conference on Human Language Technology for Development, Alexandria, Egypt, 2011.

[44] Vaclav Nemcik, "Anaphora Resolution", MSc. Thesis. Faculty of Informatics, Brno, Czech Republic, 2006

[45] Kennedy, C. and B. Boguraev. "Anaphora for everyone: Pronominal anaphora resolution without a parser". In The Proceedings of the 16th Inter-national Conference on Computational Linguistics (COLING'96), Copenhagen, Denmark, pp. 113–118, 1996

[46] Gezehagn Gutema, "Afaan Oromo Text Retrieval System", MSc. Thesis. School of Information Science, Addis Ababa, Ethiopia, 2012.

[47] Vicedo J.L., Ferrández A. “Applying Anaphora Resolution to Question Answering and Information Retrieval Systems”. In: Lu H., Zhou A. (eds) Web-Age Information Management. WAIM 2000. Lecture Notes in Computer Science, vol 1846. Springer, Berlin, Heidelberg

APPENDICES

Appendix A: Sample independent pronouns and referring nouns in Intrasentential

1. **Herodis** yommuu akka beettonni sun **isa** gowwoomsan hubatetti akka malee aare.
2. Garuu inni hanga **ishiini** ilma hangafa deettutti **ishiitti** hin buune**.**
3. Akkuma **Abbaan** keessani **inni** samii irraa mudaas hin qabne sana isinis warra mudaa hin qabne taaa.
4. **Yesuusis** hanga namootaa ufirraa geegessutti akka barattoonni bidiruu yaabbatanii **isa** dura dabranii yeruma sana gama ceaniif isaan ajaje.
5. **Sinboon** baayee bareedduu fi bifa **isheen** abbaa isheetti kan baatedha.
6. Obboo **Angaasuu fi Aadde Seenaan** mucaan **isaanii** akkaa yaadan qabamtee waan argaaniif akka mana barumsaa seentuu maritan.
7. **Yesuus** otoo achii kaee deemaa jiruu jaamonni lama **isa** faana deemanii Yaa Ilma Daawiti nu maari jedhanii iyyan.
8. **Namoonni** kunneen gara gandootaa dhaqanii akka nyaata bitataniif **isaan** ergi jedhaniin.
9. Yesuus yommuu warra **kudhaniilamaan** kanneen ergetti akkana jedhee **isaan** ajaje.
10. **Namoonnis** yeroo sana sababii Zakkaariyaas yeroo akkas dheeratu mana qulqullummaa keessatti tureef dinqifachaa **isa** eegaa turan.
11. **Abbaan** kee **inni** waan dhossaadhaan hojjatamu argu sun muldhinnaan gatii sii kenna.
12. Isaanis akkana jedhanii deebisaniif kutaa biyya Yihuudaa magaalaa Beetaliheemiitti dhalata kunis dubbii raagichi Yaa **Beetaliheem ishii** biyya Yihuudaa ati matumaan bulchitoota Yihuudaatii gadii miti.
13. Yesuusis akkana jedhee deebiseef **Keessummoonni** cidhaa fuudhaa fi heerumaa yeroo dhiirsii misirroo **isaan** bira jirutti gadduu ni dandawuu.
14. Innis luboota **hangafootaa fi barsiistota** seeraa hunda saba keessaa walitti qabee Kiristoos eessatti akka dhalatu **isaan** gaafate.
15. Innis harka **isii** tuqe dhukkubni nafa nama gubu sunis **isii** dhiise.

Appendix B: Sample Hidden pronouns and referring nouns in Intrasentential

1. **Herodis** Mootichi yommuu waan kana dhagayetti ni **rifate**.
2. **Yerusaalemis** guutummaatti isa wajjiin **rifatte**.
3. **Yooseefis** hirriibaa dammaqee akkuma maleekkaani gooftaa isa ajajetti Maariyaamin niitummaadhaan mana isaatti **fudhate.**
4. **Abbaan** kee inni waan dhossaadhaan hojjatamu argu sun muldhiinnaan gatii sii **kenna**.
5. Yaa **dargaggoota** isinis akkasuma maanguddootaaf **ajajamaa**.
6. **Namichi** kunis abbaa Aleksaandiroosiitii fi Ruufoos **ture**.
7. **Yesuus** hoboloo tokko yaabbatee galaana ceee gara magaalaas isaa **dhufe**.
8. **Katabbiin** waan inni ittiin himatamees mootii Yihudootaa jedha **ture**.
9. **Haati** mana isaanii Seenaan hara egaa deessee jia jaha guute **jirtii**.
10. **Harmeen** ishee yeroo gabaa Ooltee dhuftu buna danfistee hanga humna ishee hojjettee kan Eegduu **turte**.
11. **Seenaan** dhugaati akkaa Araqee Farsoo fi kkkf baastee **gurgurti**.
12. Yeroo kanatti garuu **bidiruun** sun walakkaa galaanaa **turte**
13. **Inni** otoo addunyaan hin uumaminii dura **filatame**.
14. **Ilmi** namaa garuu bakka itti mataa isaa irkifatullee hin qabu **jedhe**.
15. **Jinniiwwan** sunis Yoo nu baafte gara hoomaa booyyee sanaatti nu ergi jedhanii isa **kadhatan**.
16. **Saaraanis** Abrahaamiin gooftaa jettee waamaa isaaf ajajamaa **turte**.

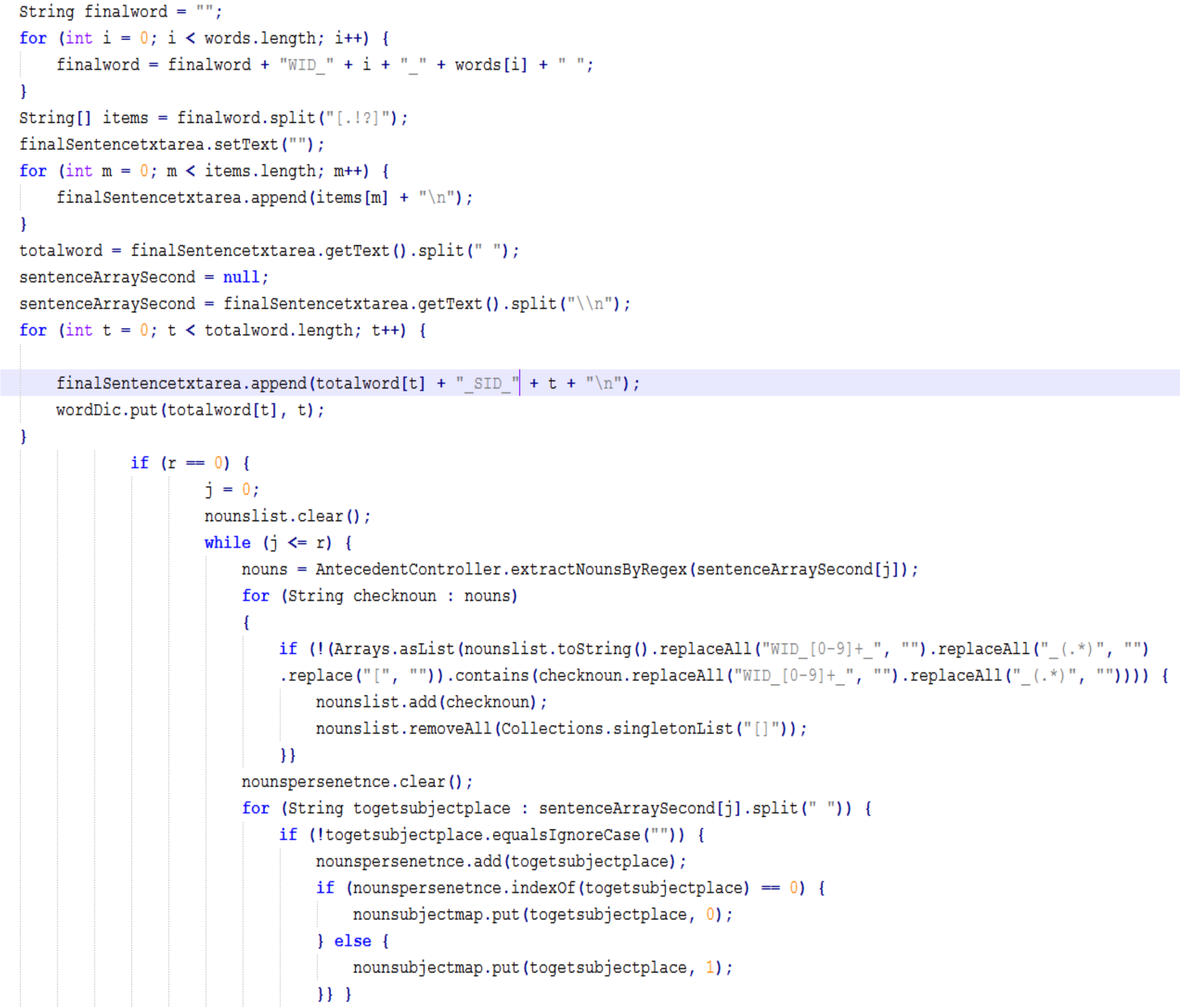
Appendix C: Sample Independent pronouns and referring nouns in Intersentential

1. **Yooseefis** hirriibaa dammaqee akkuma maleekkaani gooftaa isa ajajetti Maariyaamin niitummaadhaan mana isaatti fudhate. Garuu **inni** hanga ishiini ilma hangafa deettutti ishiitti hin buune.
2. **Herodis** Mootichi yommuu waan kana dhagayetti ni rifate. Yerusaalemis guutummaatti **isa** wajjiin rifatte.
3. **Herodis** yommuu akka beettonni sun isa gowwoomsan hubatetti akka malee aare. **Innis** akka yeroo beettota sana irraa bareetti ijoollee dhiiraa warra umuriini isaanii waggaa lamaa fi hagasii gadi tae kanneen Beetaliheemii fi naannoo isii mara keessa jiraatan nama itti ergee ficcisiise.
4. Haati isaa **Maariyaam** kaadhimaa Yooseefi turte isiinis otoo Yooseef wajjiin walbira hin gayin hafuura Qulqulluun ulfooftee argamte. Yooseef Kaadhimaan **ishii** sunis waan nama qajeelaa tureef dhossaadhaan ishii dhiisuu murteesse malee uumata duratti ishii qaanessuu hin barbaanne.
5. Kana booddee **Herodis** beettota sana dhossaatti waamee yeroo itti urjiini sun muldhate isaani irraa hubate. **Innis** dhaqaatii jabeessaa mucicha barbaadaa yeroo argitanitti immoo akka anis dhufee isaaf sagaduuf koottaa natti himaa jedhee Beetaliheemitti isaan erge.
6. **Yesuus** garuu Warri duan duaa isaanii haa awwaalatanii ati immoo na faan koottu jedheen. Yommu **inni** bidiruu yaabbatettis barattoonni isaa isa hordofan.
7. **Jarri** garuu achii baanii kutaa biyya sanaa mara keessatti waayee isaa odeessan. Otoo **isaan** gadi bawuutti jiranuu jarri tokko arrabdidaan jinniidhaan qabame tokko Yesuusitti fidan.
8. **Abbaa** mana ishee immoo yeroo boqonnaa isaa mana baaburaa deemee kan ishee barbaachisu hundaa isheef daaksisaa ture. Mucaan **isaanii** sinboon haraa guddattee waggaa kudhan guutte jirti sinboon akkumaa maqaa ishee sinboo qabeettii qalbii qabeettii fi bareedduu turtee.
9. **Innis** harka isii tuqe dhukkubni nafa nama gubu sunis isii dhiise. Isiinis kaatee **isa** tajaajiluuf jalqabde.
10. **Abrahaamis** akkuma Sanyiin kee akkas ni taa jedhametti otoo abdiin hin jiraatin abdiin amanee abbaasaboota baayee tahe. **Innis** waan umriin isaa gara waggaa dhibbaa tureef nafni isaa akkuma waan duee tauu isaati fi Saaraanis dhabduu tauu ishii yommu argettillee amantii isaatti hin laafne.

Appendix D: Sample Hidden pronouns and referring nouns in Intersentential

1. Innis harka **isii** tuqe dhukkubni nafa nama gubu sunis isii dhiise. Isiinis kaatee isa tajaajiluuf **jalqabde**.
2. **Innis** jecha tokkoon hafuurota sana baase. Dhukkubsattoota hundumaas ni **fayyise**.
3. **Innis** eegii namoota of irraa galchee booddee Waaqa kadhachuuf jedhee kophaa isaa gaaratti ol bahe. Yommuu lafti galgalaaes kophuma isaa achi **ture**.
4. Guyyaa tokkoo galgala keessa abbaa ishee **Angaasuu** bira fuula ishee gudunfitee deemtee abbaa koo Nuti maaliif akkaa warra Beenya faa Harree Loon Hoolaa dhabnee? Jetten abbaan ishee gaddaa guddan itti dhagaamee waan dubbatu **wallaale**.
5. **Yooseefis** kahee mucichaa fi haadha mucichaa fudhatee halkaniin gara Gibxitti sokke. Hanga dua Herodisiittis achi **jiraate**.

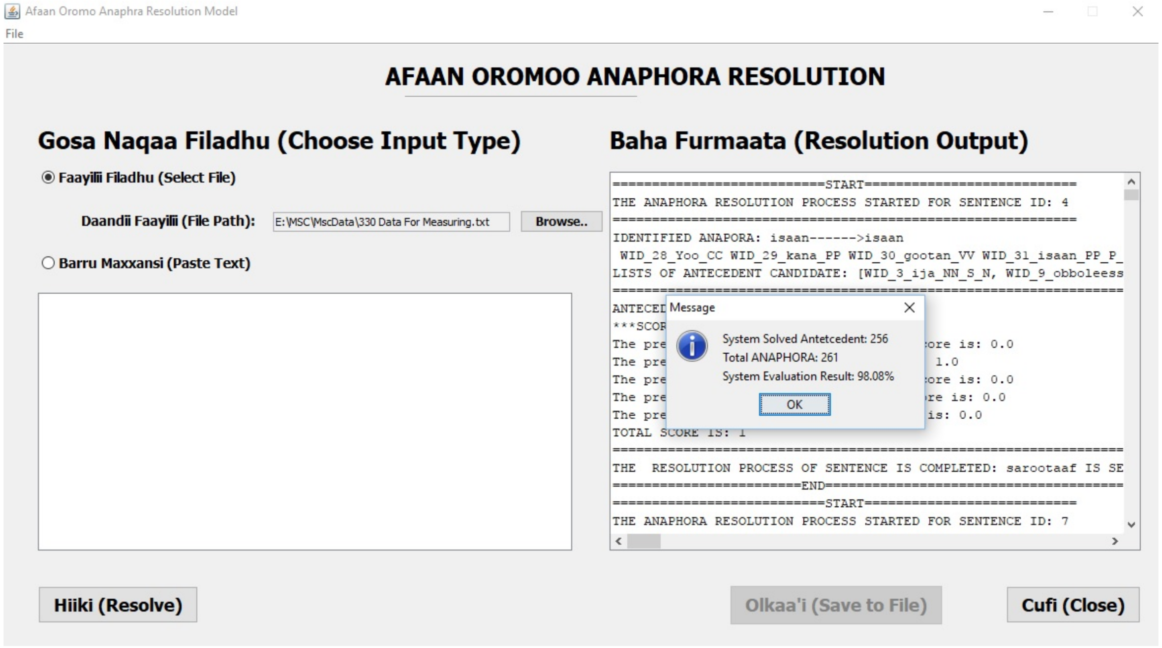
Appendix E: Sample Code



```
String finalword = "";
for (int i = 0; i < words.length; i++) {
    finalword = finalword + "WID_" + i + "_" + words[i] + " ";
}
String[] items = finalword.split("[.!?]");
finalSentencetxtarea.setText("");
for (int m = 0; m < items.length; m++) {
    finalSentencetxtarea.append(items[m] + "\n");
}
totalword = finalSentencetxtarea.getText().split(" ");
sentenceArraySecond = null;
sentenceArraySecond = finalSentencetxtarea.getText().split("\\n");
for (int t = 0; t < totalword.length; t++) {

    finalSentencetxtarea.append(totalword[t] + "_SID_" + t + "\n");
    wordDic.put(totalword[t], t);
}
            if (r == 0) {
                    j = 0;
                    nounslist.clear();
                    while (j <= r) {
                        nouns = AntecedentController.extractNounsByRegex(sentenceArraySecond[j]);
                        for (String checknoun : nouns)
                        {
                            if (!(Arrays.asList(nounslist.toString().replaceAll("WID_[0-9]+_", "").replaceAll("_(.*)", "")
                            .replace("[", "")).contains(checknoun.replaceAll("WID_[0-9]+_", "").replaceAll("_(.*)", "")))) {
                                nounslist.add(checknoun);
                                nounslist.removeAll(Collections.singletonList("[]"));
                            }}
                        nounspersenetnce.clear();
                        for (String togetsubjectplace : sentenceArraySecond[j].split(" ")) {
                            if (!togetsubjectplace.equalsIgnoreCase("")) {
                                nounspersenetnce.add(togetsubjectplace);
                                if (nounspersenetnce.indexOf(togetsubjectplace) == 0) {
                                    nounsubjectmap.put(togetsubjectplace, 0);
                                } else {
                                    nounsubjectmap.put(togetsubjectplace, 1);
                                }} }
```

Appendix F: Sample Output


Afaan Oromo Anaphra Resolution Model
File
AFAAN OROMOO ANAPHORA RESOLUTION
Gosa Naqaa Filadhu (Choose Input Type)
● Faayilii Filadhu (Select File)
Daandii Faayilii (File Path):
Browse..
○ Barru Maxxansi (Paste Text)
Baha Furmaata (Resolution Output)
Hiiki (Resolve)
Olkaa'i (Save to File)
Cufi (Close)

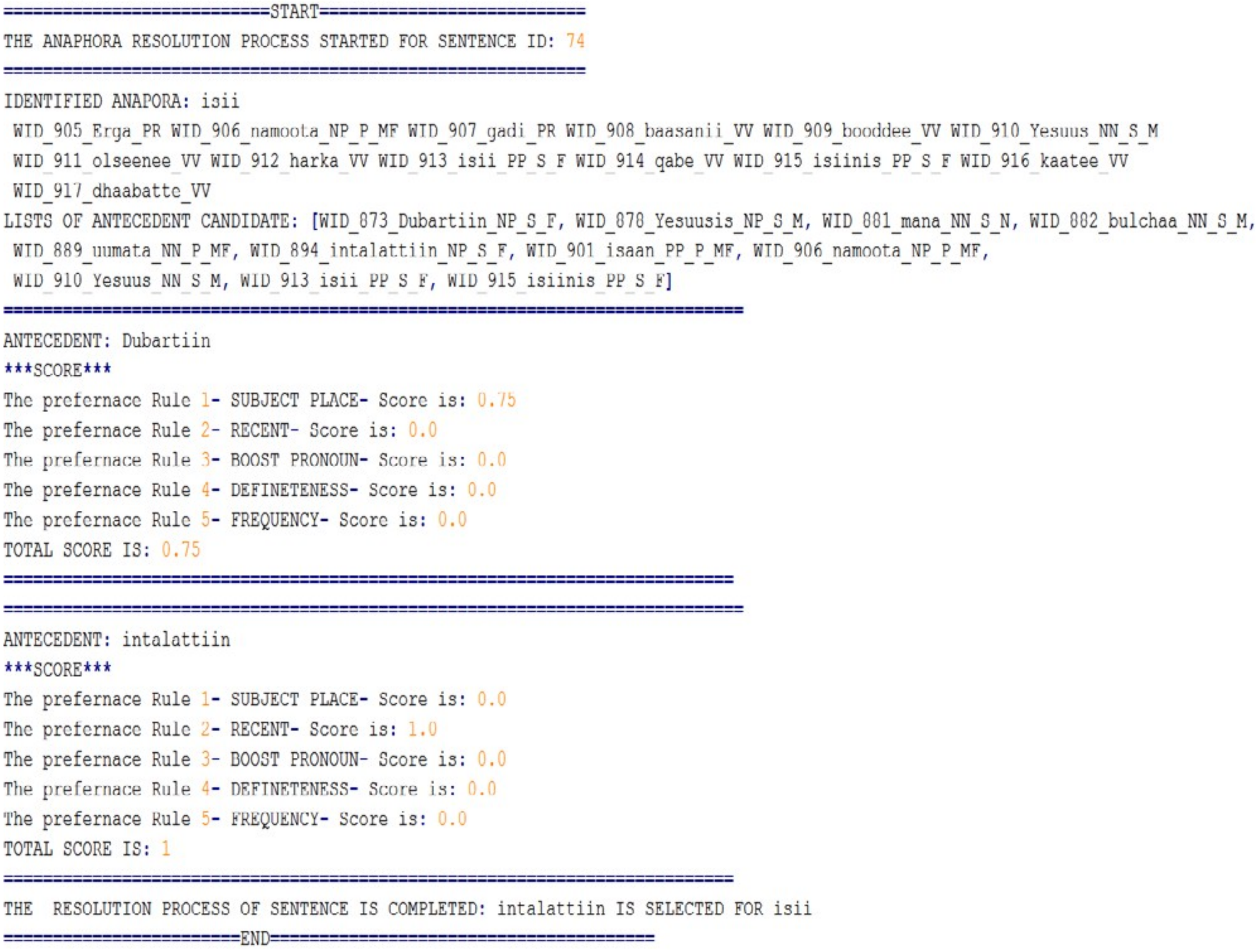
Daandii Faayilii (File Path): E:\MSC\MscData\330 Data For Measuring.txt

```
==========================START===========================
THE ANAPHORA RESOLUTION PROCESS STARTED FOR SENTENCE ID: 4
==========================================================
IDENTIFIED ANAPORA: isaan------>isaan
 WID_28_Yoo_CC WID_29_kana_PP WID_30_gootan_VV WID_31_isaan_PP_P_
LISTS OF ANTECEDENT CANDIDATE: [WID_3_ija_NN_S_N, WID_9_obboleess
=================================================================
ANTECED
***SCOR
The pre                                   ore is: 0.0
The pre                                     1.0
The pre                                   ore is: 0.0
The pre                                   re is: 0.0
The pre                                   is: 0.0
TOTAL SCORE IS: 1
=================================================================
THE  RESOLUTION PROCESS OF SENTENCE IS COMPLETED: sarootaaf IS SE
========================END======================================
==========================START===========================
THE ANAPHORA RESOLUTION PROCESS STARTED FOR SENTENCE ID: 7
```

Message

System Solved Antetcedent: 256
Total ANAPHORA: 261
System Evaluation Result: 98.08%

OK



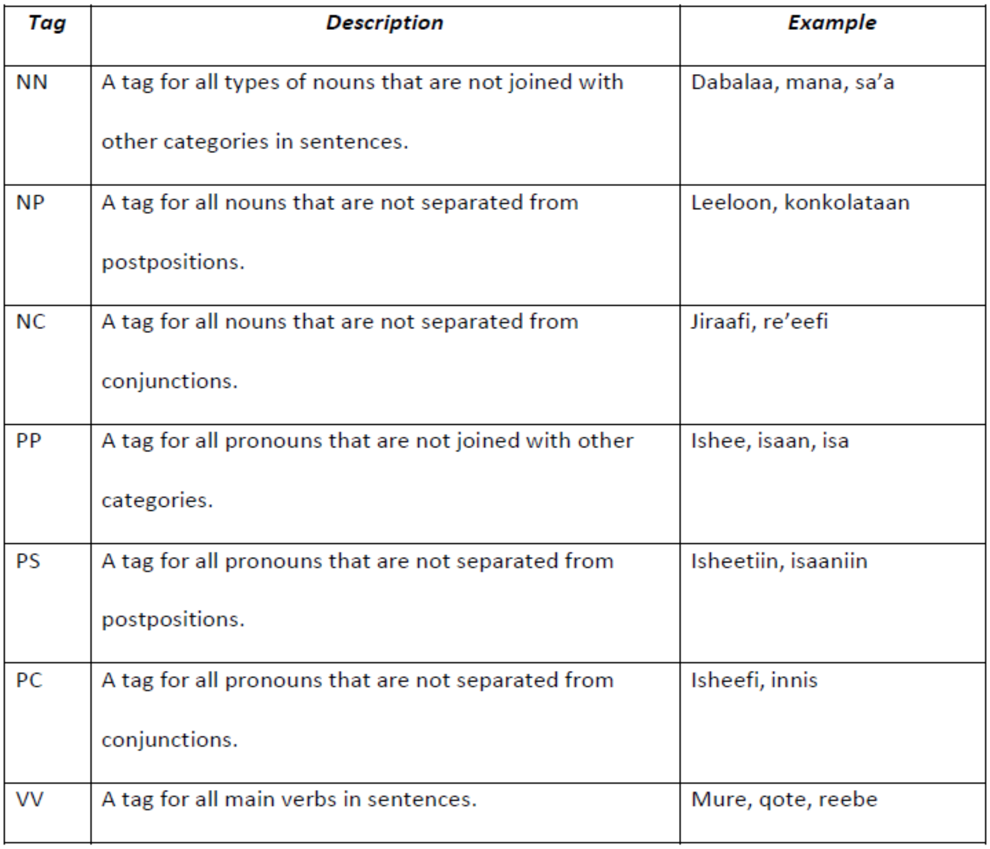
```
===========================START==========================
THE ANAPHORA RESOLUTION PROCESS STARTED FOR SENTENCE ID: 74
==========================================================
IDENTIFIED ANAPORA: isii
 WID_905_Erga_PR WID_906_namoota_NP_P_MF WID_907_gadi_PR WID_908_baasanii_VV WID_909_booddee_VV WID_910_Yesuus_NN_S_M
 WID_911_olseenee_VV WID_912_harka_VV WID_913_isii_PP_S_F WID_914_qabe_VV WID_915_isiinis_PP_S_F WID_916_kaatee_VV
 WID_917_dhaabatte_VV
LISTS OF ANTECEDENT CANDIDATE: [WID_873_Dubartiin_NP_S_F, WID_878_Yesuusis_NP_S_M, WID_881_mana_NN_S_N, WID_882_bulchaa_NN_S_M,
 WID_889_uumata_NN_P_MF, WID_894_intalattiin_NP_S_F, WID_901_isaan_PP_P_MF, WID_906_namoota_NP_P_MF,
 WID_910_Yesuus_NN_S_M, WID_913_isii_PP_S_F, WID_915_isiinis_PP_S_F]
==========================================================================
ANTECEDENT: Dubartiin
***SCORE***
The prefernace Rule 1- SUBJECT PLACE- Score is: 0.75
The preferance Rule 2- RECENT- Score is: 0.0
The prefernace Rule 3- BOOST PRONOUN- Score is: 0.0
The preferance Rule 4- DEFINETENESS- Score is: 0.0
The prefernace Rule 5- FREQUENCY- Score is: 0.0
TOTAL SCORE IS: 0.75
=========================================================================
==========================================================================
ANTECEDENT: intalattiin
***SCORE***
The preferance Rule 1- SUBJECT PLACE- Score is: 0.0
The prefernace Rule 2- RECENT- Score is: 1.0
The preferance Rule 3- BOOST PRONOUN- Score is: 0.0
The prefernace Rule 4- DEFINETENESS- Score is: 0.0
The preferance Rule 5- FREQUENCY- Score is: 0.0
TOTAL SCORE IS: 1
=========================================================================
THE  RESOLUTION PROCESS OF SENTENCE IS COMPLETED: intalattiin IS SELECTED FOR isii
======================END===================================
```

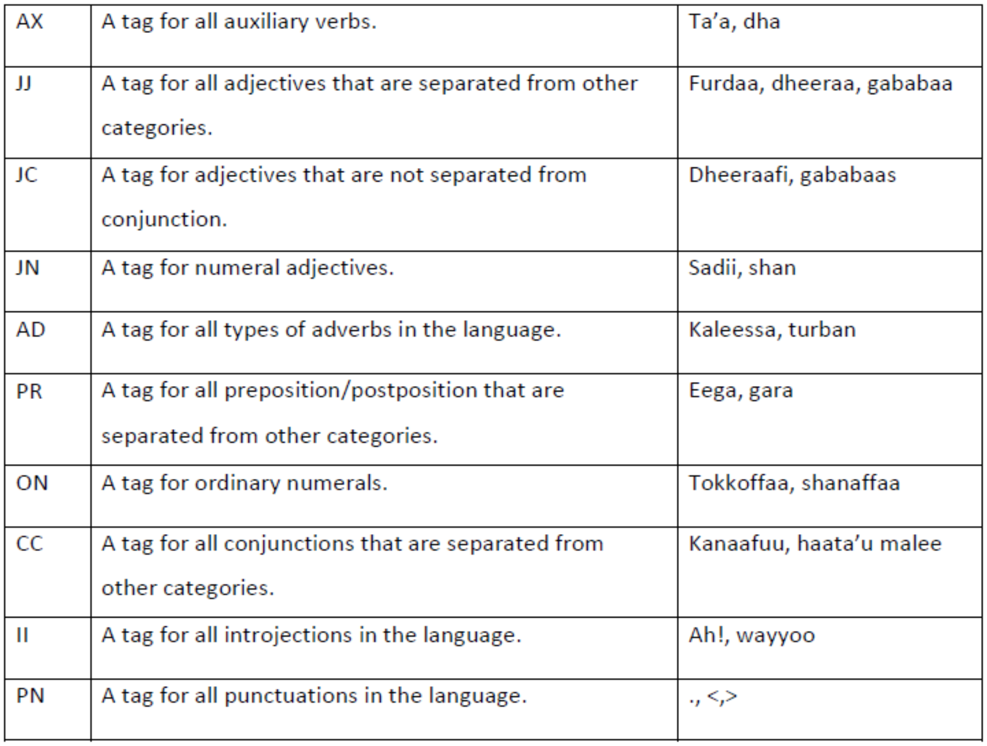
Appendix G: Manually prepared sample tagged data

1. Dhalachuuni_NP Yesuus_NN_S_M Kiristoosi_NN_S_M akkana_AD ture_VV.
2. Haati_NN_S_F isaa_PP_S_M Maariyaam_NN_S_F kaadhimaa_NN Yooseefi_NN_S_M turte_VV.
3. Isiinis_PP_S_F otoo_CC Yooseef_NN_S_M wajjiin_PR walbira_PR hin_AX gayin_VV hafuura_NN Qulqulluun_NN ulfooftee_VV argamte_VV.
4. Yooseef_NN_S_M Kaadhimaan_NP ishii_PP_S_F sunis_PP_S_N waan_NN nama_NN_S_MF qajeelaa_VV tureef_VV dhossaadhaan_NP ishii_PP_S_F dhiisuu_VV murteesse_VV malee_PR uumata_NN duratti_JJ ishii_PP_S_F qaanessuu_VV hin_AX barbaanne_VV.
5. Otoo_CC inni_PP_S_M waan_PR kana_PP_S_N yaadaa_VV jiruu_VV Maleekkaani_NN_S_M gooftaa_NN_S_M abjuun_NP itti_PR muldhatee_VV akkana_AD jedheen_VV.
6. Yaa_AX Yooseef_NN_S_M ilma_NN_S_M Daawiti_NN_S_M kaadhimni_NP_S_F tee_PP_S_M_F Maariyaam_NN_S_F Hafuura_NN Qulqulluudhaan_NP waan_PR ulfoofteef_VV isii_PP_S_F fuudhuu_VV hin_AX sodaatin_VV.
7. Ishiin_PP_S_F ilma_NN_S_M deetti_VV atis_PP_S_M_F maqaa_NN isaa_PP_S_M Yesuus_NN_S_M jettee_VV moggaafta_NN inni_PP_S_M saba_NN isaa_PP cubbuu_NN isaanii_PP irraa_PR ni_PR fayyisaatii_VV.
8. Yooseefis_NP_S_M hirriibaa_NN dammaqee_VV akkuma_PR maleekkaani_NN_S_M gooftaa_NN_S_M isa_PP_S_M ajajetti_VV Maariyaamin_NP_S_F niitummaadhaan_NP_S_N mana_NN_S_N isaatti_PP_S_M fudhate_VV.
9. Garuu_CC inni_PP_S_M hanga_PR ishiini_PP_S_F ilma_NN_S_M hangafa_JJ deettutti_VV ishiitti_PP_S_F hin_AX buune_VV.
10. Dhufaatii_VV Beettotaa_NP_P_MF Bara_NN_S_N mooticha_NP_S_M Herodisi_NN_S_M keessa_PR erga_PR Yesuus_NN_S_M kutaa_NN_S_N biyyaa_NN_S_N Yihuudaa_NN_S_N magaala_NN Beetaliiheemitti_NP_S_F dhalatee_VV booddee_PR beettonni_JJ tokko_VV baa_NN_S_N biiftuutii_NP_S_N gara_PR Yerusaalemi_NN_S_MF dhufan_VV.
11. Isaanis_PP_P_MF inni_PP_S_M mootii_NN_S_M Yihuudotaa_NP_P_MF tauuf_VV dhalate_VV sun_PP_S_N eessa_PP_S_N jira_VV.
12. Herodis_NN_S_M Mootichi_NP_S_M yommuu_AD waan_PR kana_PP_S_N dhagayeeti_VV ni_PR rifate_VV.
13. Yerusaalemis_NP_S_F guutummaatti_JJ isa_PP_S_M wajjiin_PR rifatte_VV.
14. Innis_PP_S_M luboota_NP_P_MF hangafootaa_NP_P_MF fi_CC barsiistota_NP_P_MF seeraa_NN_S_N hunda_AD saba_NN_S_N keessaa_PR walitti_PR qabee_VV Kiristoos_NN eessatti_PP_S_N akka_PP_S_N dhalatu_VV isaan_PP_P_MF gaafate_VV.

15. Isaanis_PP_P_MF akkana_AD jedhanii_VV deebisaniif_VV kutaa_NN_S_N biyya_NN_S_N Yihuudaa_NN_S_N magaalaa_NN_S_N Beetaliheemiitti_NP_S_N dhalata_VV kunis_PP_S_N dubbii_NN_S_N raagichi_NP_S_M Yaa_AX Beetaliheem_NN_S_F ishii_PP_S_F biyya_NN_S_N Yihuudaa_NN_S_N ati_PP_S_MF matumaan_NP_S_N bulchitoota_NP_P_N Yihuudaatii_NP_S_N gadii_PR miti_AX.
16. Saba_NN_S_N koo_PP_S_N Israaelin_NP_S_N kan_PR bulchu_VV si_PP_S_MF keessaa_PR ni_PR baaatii_VV jedhee_VV barreesse_VV sanaa_PP_S_MF dha_AX.
17. Kana_PR booddee_PR Herodis_NN_S_M beettota_NP_P_MF sana_PP_S_N dhossaatti_JJ waamee_VV yeroo_AD itti_PR urjiini_NP_S_N sun_PP_S_N muldhate_VV isaani_PP_P_MF irraa_PR hubate_VV.
18. Innis_PP_S_M dhaqaatii_VV jabeessaa_VV mucicha_NP_S_M barbaadaa_VV yeroo_AD argitanitti_CC immoo_CC akka_PR anis_PP_S_N dhufee_VV isaaf_PP_S_M sagaduuf_VV koottaa_VV natti_PP_S_MF himaa_VV jedhee_VV Beetaliheemitti_NP_S_N isaan_PP_P_MF erge_VV.
19. Isaanis_PP_P_MF eegii_PR waan_PR mootichi_NP_S_M jedhe_VV sana_PP_S_N dhagahanii_VV booddee_PR adeemsa_NN isaanii_PP_P_MF itti_PR fufan_VV urjiini_NP_S_N isaani_PP_P_MF bahaa_NN_S_N biiftuutti_NP_S_N argan_VV sunis_PP_S_N hanga_AD bakka_NN_S_N mucichi_NP_S_M ture_VV gahee_VV dhaabatutti_CC isaani_PP_P_MF dura_JJ deeme_VV.
20. Yooseefis_NP_S_M kahee_VV mucichaa_NP_S_M fi_CC haadha_NN_S_F mucichaa_NP_S_M fudhatee_VV halkaniin_AD gara_PR Gibxitti_NP_S_N sokke_VV.
21. Hanga_AD duha_NN Herodisiittis_NP achi_AD jiraate_VV.

Appendix H: Tagset for Afaan Oromo Getachew [41]





Tag	***Description***	***Example***
NN	A tag for all types of nouns that are not joined with other categories in sentences.	Dabalaa, mana, sa'a
NP	A tag for all nouns that are not separated from postpositions.	Leeloon, konkolataan
NC	A tag for all nouns that are not separated from conjunctions.	Jiraafi, re'eefi
PP	A tag for all pronouns that are not joined with other categories.	Ishee, isaan, isa
PS	A tag for all pronouns that are not separated from postpositions.	Isheetiin, isaaniin
PC	A tag for all pronouns that are not separated from conjunctions.	Isheefi, innis
VV	A tag for all main verbs in sentences.	Mure, qote, reebe
AX	A tag for all auxiliary verbs.	Ta'a, dha
JJ	A tag for all adjectives that are separated from other categories.	Furdaa, dheeraa, gababaa
JC	A tag for adjectives that are not separated from conjunction.	Dheeraafi, gababaas
JN	A tag for numeral adjectives.	Sadii, shan
AD	A tag for all types of adverbs in the language.	Kaleessa, turban
PR	A tag for all preposition/postposition that are separated from other categories.	Eega, gara
ON	A tag for ordinary numerals.	Tokkoffaa, shanaffaa
CC	A tag for all conjunctions that are separated from other categories.	Kanaafuu, haata'u malee
II	A tag for all introjections in the language.	Ah!, wayyoo
PN	A tag for all punctuations in the language.	., <,>