



JIMMA UNIVERSITY
JIMMA INSTITUTE OF TECHNOLOGY
FACULTY OF COMPUTING AND INFORMATICS

**AUTOMATIC AFAN OROMO SENTENCE IDENTIFICATION AND
SIMPLIFICATION USING RULE BASED APPROACH**

By

ABDUREHMAN MAHMUD JUNDA

ADVISOR: DR. GETACHEW MAMO (PHD)

CO-ADVISOR: MR. ADMAS ABTEW(MSc)

A Thesis Submitted to the School of Graduate Studies of Jimma Institute of Technology in Partial Fulfillment of a Master of Science Degree in Information Technology.

Jimma, Ethiopia

December, 2021

JIMMA UNIVERSITY

JIMMA INSTITUTE OF TECHNOLOGY

FACULTY OF COMPUTING AND INFORMATICS

**AUTOMATIC AFAN OROMO SENTENCE IDENTIFICATION AND
SIMPLIFICATION USING RULE BASED APPROACH**

By

ABDUREHMAN MAHMUD JUNDA

This is to certify that the thesis prepared by ABDUREHMAN MAHMUD JUNDA, titled: AUTOMATIC AFAN OROMO SENTENCE IDENTIFICATION AND SIMPLIFICATION USING RULE BASED APPROACH and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Information Technology complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Name and Sign of Examining Members


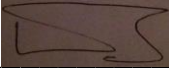


<u>List of Members</u>	<u>Signature</u>	<u>Date</u>
Chairperson: _____	_____	_____
Advisor: <u>Dr. Getachew Mamo (PhD)</u>		<u>20/12/2021</u>
Co-advisor: <u>Mr. Admsa Abtew(MSc.)</u>		<u>20/12/2021</u>
External Examiner: <u>Dr. Teklu Urgessa(PhD)</u>		<u>20/12/2021</u>
Internal Examiner: <u>Mr. Mizanu Zelalem(MSc.)</u>		<u>20/12/2021</u>

Table of Contents

Table of Contents	i
List of Tables	iii
List of Figures.....	iii
List of Algorithms	iii
Acknowledgement.....	iv
Abstract.....	v
List of Acronyms/Abbreviations	vi
CHAPTER ONE: INTRODUCTION	1
1.1 Background	1
1.2 Motivation	3
1.3 Statement of the Problem	4
1.4 Objective of the Study.....	5
1.5 Research Methodology.....	6
1.6 Scope and Limitation of the Study	9
1.7 Application of the Results	9
1.8 Organization of the Thesis	10
CHAPTER TWO: LITERATURE REVIEW	11
2.1 Introduction	11
2.2 Sentence Simplification.....	12
2.3 Approaches of Sentence Simplification	12
2.4 Method of Sentence Simplification.....	13
2.5 Evaluation of sentence simplification	16
2.6 Overview of Afan Oromo Language.....	17
2.6.1 Afan Oromo Writing system.....	18
2.6.2 Afan Oromo Word Class	18
2.6.3 Afan Oromo Clauses	22
2.6.4 Afan Oromo Sentence.....	23
CHAPTER THREE: RELATED WORK	25
3.1 Introduction	25

3.2	Sentence Simplification for English language	25
3.3	Sentence Simplification for Punjabi language	27
3.4	Sentence Simplification for French language	28
3.5	Sentence Simplification for Urdu language.	29
3.6	Summary of Related Work.....	29
CHAPTER FOUR: DESIGN AND IMPLEMENTATION		32
4.1	Introduction	32
4.2	System Architecture	32
4.3	Components of System Architecture.....	34
4.3.1	Preprocessor	34
4.3.2	Sentence Identification.....	38
4.3.3	Sentence Transformation	43
4.3.4	Postprocessor	45
CHAPTER FIVE: EVALUATION RESULT AND DISCUSSION.....		49
5.1	Data Preparation and Developmental Environment.....	49
5.2	Evaluation of the Result	50
5.3	Discussion	52
CHAPTER SIX: CONCLUSION AND FUTURE WORK		55
6.1	Conclusion.....	55
6.2	Recommendation and future work	56
References		58
Appendixes		62
Appendix A: Sample of annotated Data.....		62

List of Tables

Table 2.1 List of Afan Oromo Conjunctions	21
Table 3.1 Summary of Related Work	30
Table 4.1 The Newly Added Tagsets.....	37
Table 4.2 A summary of Afan Oromo Sentences Type Indicators	42
Table 4.3 Example of Simplified Compound Sentences.	48
Table 5.1 Evaluation Result of Sentence Identification.....	51
Table 5.2 Evaluation Result of Compound Sentence Simplification	52

List of Figures

Figure 1.1 Research Design Methodology.....	6
Figure 4.1 System Architecture	34
Figure 4.2 Input and Outputs of Sentence Identification.....	42

List of Algorithms

Algorithm 4.1 Sentence Identification and Separation Algorithm	43
Algorithm 4.2 Compound Sentence Transformation Algorithm	45
Algorithm 4.3 Sentence Rearrangement Algorithm	47

Acknowledgement

First of all, and foremost, thanks to the creature of the entire universe, almighty God, for helping me to realize this work. Thank you, Allah!

Secondly, I would like to thank my advisor, **Dr. Getachew Mamo**, whose encouragement helped me to understand from the beginning to the end of this study. He always gives me clear direction and constant guidance to follow, particularly in that he showed me different ways to approach a research problem and the need to be persistent to accomplish any goal. I have deep thanks and appreciation for his precise comments and immeasurable assistance throughout my work. Next, I would like to convey my special gratitude to my co-advisor, **Mr. Admas Abtew**, for his concerns and guidance to keep me on the right track to completing my studies.

Lastly, I would like to thank my family, friends, and everybody who directly or indirectly supported me in my important successful realization.

Abstract

In NLP, sentence identification and simplification are necessary for machine translation, parsing, question generation, information extraction, summarization, semantic role labeling, opinion mining, etc. The majority of these applications use simple sentences as preprocessing to improve their functionality, and the high coverage of sentence simplification is used for various social classes that have language difficulties, such as aphasics, children, and adults learning the language (non-native speakers).

The study provided a new automatic syntactic Afan Oromo sentence identification and simplification using a rule-based method that operates on POS tags. In this study, the main performed task can be separated into two tasks. The first task is the identification and separation of Afan Oromo declarative sentences into simple, compound, complex, and compound-complex sentences. The second task is the simplification of compound sentences into simple and self-contained sentences by preserving the meaning of the original meaning as much as possible. Sentence identification and separation were performed to improve the performance of sentence simplification.

The resursive type algorithm is developed both for sentence identification and simplification based on the syntactic structure of the sentences. To determine the syntactic structure of the sentence, the POS Tag is used as a preprocessor and then the sentence type indicators and sentence simplification features are managed. To evaluate the algorithms, a dataset containing 480 sentences was collected from the Afan Oromo textbook and annotated with the help of an expert. The performance of the sentence identification and compound sentence simplification algorithms is separately evaluated in terms of precision and recall using the result gained by the expert judgments. The expert classifies the identified and simplified sentences as correct or incorrect by comparing the system's output with the golden standard produced by the language expert. The sentence simplification evaluation criteria includes grammar and fluency of the simplified sentence and also the retainment of the original meaning. The overall performance of both sentence identification and compound sentence simplification is 90% and 84.4% F-score respectively. The evaluation result reveals that the proposed algorithm is a promising one, as it is the beginning of a less resource-intensive study.

keywords: Afan Oromo Sentences, Sentence Identification, Syntactic Sentence Simplification

List of Acronyms/Abbreviations

AI: Artificial Intelligence

LS: Lexical Simplification

MT: Machine Translation

NER: Named Entity Recognition

NLG: Natural Language Generation

NLP: Natural Language Processing

NLU: Natural Language Understanding

PBSMT: Phrase-Based Statistics Machine Translation

POST: Part of Speech Tag

SBSMT: Syntax-Based Statistics Machine Translation

SMT: Statistics Machine Translation

SS: Sentence Simplification

TS: Text Simplification

CHAPTER ONE: INTRODUCTION

1.1 Background

Natural Language Processing (NLP) is a subfield of artificial intelligence that consists of a scientifically based set of computational models for analyzing and representing natural language at different knowledge levels, dealing with a variety of activities and applications in order to achieve human-like language processing [1, 2]. NLP deals with designing and implementing computational models that analyze, understand, and generate deep features of natural languages represented in textual, speech, and image formats that can be processed at different knowledge levels such as phonetics and phonology, morphology, syntax, semantics, pragmatics, and discourse levels in their hierarchical order [3]. Nowadays, text processing is a major and crucial part of representing a natural language due to the increasing textual form of data and information on every offline and online platform, and many investigations have been performed for different activities and applications in the field of processing natural language [4].

Text simplification (TS) is among the most important NLP applications for reducing text complexity by changing the contents and structures of a text to improve readability and understandability for humans and makes it easier to process for downstream NLP applications while preserving its main idea and intently its original meaning [2, 4, 5, 6, 7] . It also represents a collaborative effort between natural language understanding (NLU), which is the comprehensive of the text to be processed, and the generation of natural language (NLG), which is the production of new structure and concepts from the input text. In the process of text simplification, the complexity of the text can be simplified at three levels: lexical-level, which deals with vocabulary simplification; sentence-level simplification, which deals with the simplification of structural and content of a sentence; and also, at document-level, which is about extracting the concept of the full document. These all types of text simplifications are a broad research area and not being investigated at a time [2, 8]. However, currently, most research in text simplification has focused on the exploration of sentence and lexical simplification. In a similar manner, our study is also focused on the area of sentence simplification.

Sentence simplification (SS) is a popular task in text simplification which modifies the content and structure of sentences by performing several rewriting transformations such as splitting, reordering, and replacement to generate easier and simpler sentences while keeping sentences'

grammar and preserving their fundamental ideas [1, 5, 9, 7]. In the task of sentences simplification process there are two approaches (syntactic and lexical simplification) that which can be which investigated separately or together [10, 2, 11] . Syntactic sentence simplification deals with the simplification of syntax structure to remove the complexity of a sentence structurally and grammatically while preserving its information and meaning as much as possible [12, 7, 2]. In sentence simplification, there are some tasks that deliver a lot of syntactic complexity. For example, splitting and transformation of sentences that have a long and complex structure like embedded clauses (independent, dependent, and relative), and discourse makers (conjunctions and ad-positions) into simple sentences, co-reference resolution, and rewriting sentences with passive voice are some tasks of syntactic sentence simplification [12, 13, 14]. The second aim of the sentence-simplification task is the simplification of sentences' content (lexical simplification). Lexical simplification is a task that changes the content of sentences by finding and substituting strong words with words that have equivalent meanings to minimize the complexity of word vocabulary in the sentence [8, 2].

In another case, in the concept of sentence simplification, the type of sentence identification and separation is the basic and necessary point that must be clearly stated, because in every language there are various types of sentences that have different structures and purposes, and they should be simplified accordingly [3, 1]. Most of the time, the categorization of sentences is generally based on two aspects, which are based on the structure of sentences and based on the purpose or types of statements (messages) they deliver. For example, Afan Oromo sentences are categorized into four types based on the types of delivered statement (Declarative, Interrogative, Imperative, and Exclamatory sentences), and in terms of their structure, again categorized into four types as simple, compound, complex and compound complex sentences [23, 21]. The detailed analysis of sentence identification and separation is the cornerstone of text simplification at sentence level, as sentence simplification is the retransformation of the constituent pattern of the sentence accordingly, and even collecting and processing datasets based on sentence type is needed to meet the necessary target [1, 3, 15].

As several studies have demonstrated, sentence simplification is extremely important in many computational modeling fields of study such as natural language processing, speech processing and query processing [16, 17, 18]. In the processing of natural language fields, sentence simplification is used as a pre-processing tool in several NLP applications,.Particularly. in

machine translation, parsing, POS tagging, information extraction, question generation, semantic role labeling, sentence-based text summarization, opinion mining, etc. In addition to the availability of different NLP fields, a broad coverage of sentence simplification is expected to be beneficial to a variety of social groups, including users with language disabilities such as dyslexia and aphasia, low-literacy users such as children, and non-native speakers [19, 5].

In area of sentence simplification different researchers have been using different methods for different purposes and categories them as rule-based, data-driven, and hybrid methods [4]. The rule-based approach is one of the most widely used and successful approach for sentence simplification [7]. The key explanation for the use of rule-based methods is the lack of annotated corpus and restricted resources for most languages, as well as the most preferable approach is the flexibility to resolve the ambiguity of the sentence structure and the efficiency of sentence simplification [20, 14]. As a result, we proposed in this study to investigate the identification and simplification of Afan Oromo sentences structurally using a rule-based method by considering the justifications of the method and the language's resources.

1.2 Motivation

Afan Oromo is one of Ethiopia's most widely spoken languages [21]. It is currently the official language of the regional state of Oromia (Ethiopia's largest regional state) and is used in offices, schools, colleges, universities, media and various written materials are published both electronically and non-electronically. The Afan Oromo language is not only spoken in Ethiopia [22]; it is also spoken in Somalia, Kenya, Uganda, Tanzania, and Djibouti. Although Afan Oromo is today spoken by such a large number of people. However, few advances have been made in computational linguistics or natural language processing in the language itself. The lack of well-studied linguistic resources has hampered computational methods to linguistic analysis in Afan Oromo. So far, the simple sentence is utilized in many natural language processing applications and directly helps social groups that have language understandability problems. Despite its enormous number of speakers, Afan Oromo is still a language for which very few computational linguistic tools have been developed, and very little has been written in the language and very little has been done in terms of making useful computer-based applications available. Thus, these gaps produce the drive and desire which motivates us to engage in this research work.

1.3 Statement of the Problem

Sentence simplification is a natural language text processing application that has been utilized as a preprocessor and postprocessor in a variety of NLP tasks to improve the efficiency of their functionality [8]. Since Afan Oromo is an infant in computational model processing and even the developed applications have been being faced the problem of being advanced computer-based applications particularly in terms of performance. For example, machine translation, sentence parsing, information extraction (NER), semantic role labeling, and grammar checking, query processing and speech processing are NLP applications that need simple sentences rather than long and grammatically complex sentences to improve their performance because a complex sentence includes several components that can be difficult to be understood by a machine. For example, structurally complex sentences contain different types of embedded clauses such as independent, dependent and relative clauses that can be simplified into simple and independent sentences, and also complex sentences contain different discourse makers such as conjunctions and coreference resolutions that can be removed (conjunctions) or replaced (coreferences) as in sentence simplification. Thus, these all-NLP applications that hinder in terms of long and structurally complex sentences can be realized [3, 5, 7].

As an Afan Oromo infant in computational model processing, it is a great challenge to initiate important high-level human-like computer applications that takes the language forward. For example, types of question generation from text, such as factual questions, multiple-choice questions, yes-or-no questions, dialogue questions, and others, are very important high-level computer-based applications, particularly in education area, but they are extremely difficult to start up due to the lack of such preprocessing sentence simplification resources [17, 24].

Sometimes different social classes that are disabled are also discouraged in naturally occurring language, particularly those who have language difficulties such as losing long and structurally complex sentence understanding (aphasia), low literacy skills (children) or adults learning languages (non-native speakers) [13, 2]. Nowadays, in the world of writing and reading different literature (books and business documents) there is also no opportunity to measure the written by authors and provide easy-to-understandable material to increase the income by satisfying the reader due to measurable text complexity [7].

However, Afan Oromo is still a language for which very few NLP applications have been developed, even for other languages where dedicated sentence simplification is hard to find and not yet commercialized due to its system complexity and restricted corpus [4]. To the best of our knowledge, Afan Oromo sentence identification and simplification is a new concept that has not been conducted yet. Thus, we proposed Afan Oromo syntactic sentence identification and simplifications in order to solve the problems justified in the above and to take up the language's resources one step.

Aside from the explanations stated above, the following research questions will be answered in this work.

- 1) How to map and build automatic sentence identification and simplification of Afan Oromo language text?
- 2) To what extent can the performance of Afan Oromo sentence identification and simplification be achieved?
- 3) What are the challenges of Afan Oromo sentence identification and simplification?

1.4 Objective of the Study

1.4.1 General Objective

The main objective of the study is to investigate and design of Afan Oromo sentence identification and simplification using a rule-based approach.

1.4.2 Specific Objectives

To achieve the main objective of the study, the following activities and procedures are necessary.

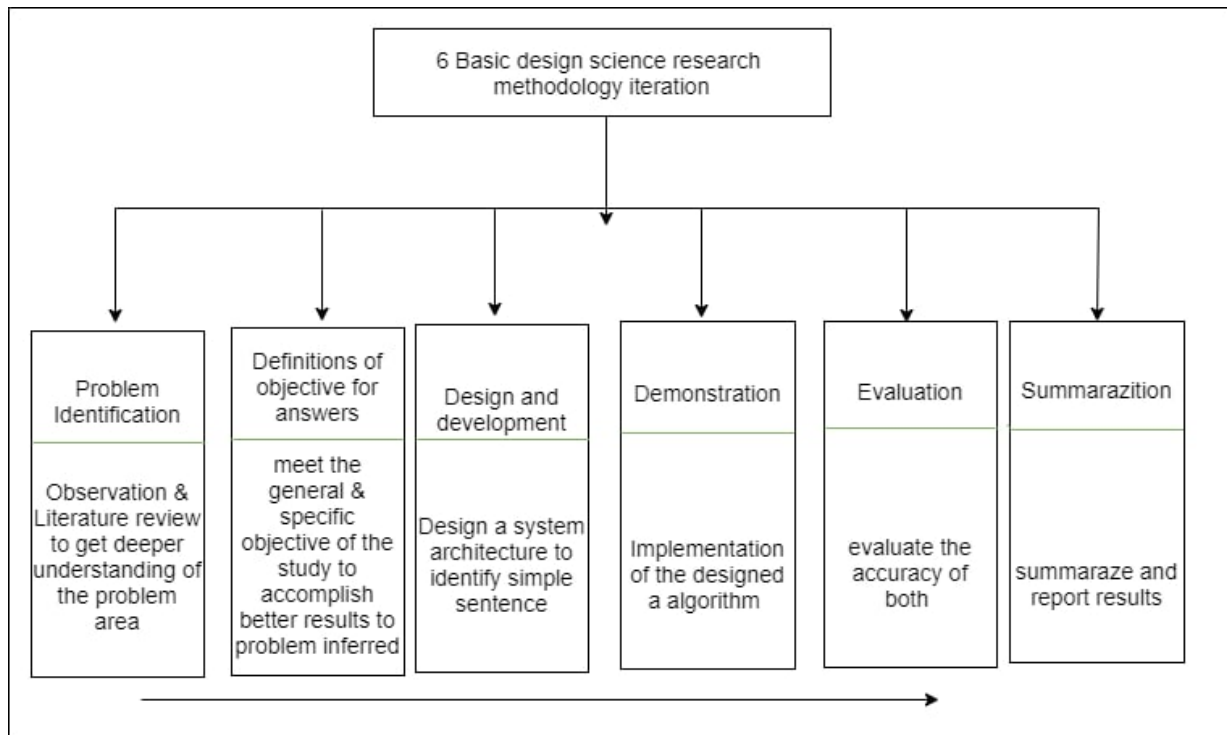
- ✚ Examine various types of literature and related works focused on sentence identification and simplification.
- ✚ Review the grammatical structures and complexity of Afan Oromo sentence.
- ✚ Review the prerequisite resources of Afan Oromo sentence identification and simplification like part of speech tagging.
- ✚ Develop a model of Afan Oromo sentence identification and simplification.
- ✚ Evaluate performance of Afan Oromo sentence identification and simplification algorithm.
- ✚ Provide reports on the final results of the experiments and limitations for future work.

1.5 Research Methodology

1.5.1 Research Design

Choosing the right methodology is a crucial step in conducting a research. As a result, a research paradigm that can produce the intended solution to the problem is employed. In order to accomplish the research aims and objectives, provide valid and reliable results, a research methodology that employs methods and techniques which are the best fit for the research is chosen. The research encompasses designing a new artifact to solve observed problems, evaluating the artifact and presenting the results. Therefore, design science research methodology is a perfect fit for this research. The methods include six basic steps as depicted on the following figure 1.1. Those are problem identification, definitions of objectives for answers, design and development, demonstration, evaluation and communication.

Figure 1.1 Research Design Methodology



➤ Problem Identification: To identify the problem, the researcher's prior observation and further investigation through literature review are applied. Journal articles, books, resources on the web are referenced and reviewed. To get a deeper understanding of the problem area, conceptual theories of research papers and books are reviewed.

- ✚ **Definition of Objective:** From the problem identified in the previous step and the solutions that were attempted by previous scholars, general and specific objectives of the solution to accomplish better results to the problem are inferred.
- ✚ **Design and Development:** The design and development process involves developing a design and architecture to detect and correct Afaan Oromo spelling errors. It is followed by designing an efficient algorithm that solves the problem defined on the first step and fulfills the objectives defined on the second step.
- ✚ **Demonstration:** It involves implementation of the designed algorithm and building a prototype. At this stage, a prototype using python programming language is implemented.
- ✚ **Evaluation:** The prototype development is followed by analyzing and evaluating the results. In order to test the performance of the model prototype, accuracy were used.
- ✚ **Summarization:** The final step involves concluding the results and presenting conclusion and recommendation from the test results

1.5.2 Literature Review and Discussion with Experts

To develop automatic sentence identification and simplification for the Afan Oromo language, recent and relevant resources like previously conducted research papers, journals, books, and other documents from the internet are accessed to explore techniques and approaches appropriate for the development of general sentence identification and simplification and to know the constituents of the sentence in the language in particular. Moreover, the syntactic structure and components of the sentence and other features that can be used in sentence simplification have been investigated. Additionally, we have discussed with experts and linguists to develop advanced and efficient sentence simplification because the approaches selected for this study require detailed knowledges of the language.

1.5.3 Data Collection and Preparation

To get fully structured sentences that are suitable for the investigation is difficult and requires a deep understanding of what makes sentences complex and how they become simple in the language. For the study, the required types of datasets were prepared by researchers with the help of experts and collected from different resources, mainly from Afan Oromo textbooks, because the sentences collected from the textbook are correctly structured and the data is manually annotated as it seemed appropriate at the time to contribute to the study in the

preparation of a standard Afan Oromo sentence corpus. As Afan Oromo structural based sentence types which includes simple, compound, complex, and compound-complex sentences are must be available for both sentence identification and simplification. To perform the both sentence identification and simplification, some preprocessing like cleaning sentences and identifying sentence structure into its parties of speech were performed because the study was focused on constituting parts of the sentences to identify and manage the sentence structures that cause sentence complexity and break them down into simple form.

1.5.4 Techniques and Tools

As the literature has shown, the task of sentence simplification has been developed by data-driven, rule based, and hybrid approaches [20]. Currently, the rule-based approach is one of the most widely used and successful approach for sentence simplification [3, 11, 20] and it is selected for this study too. The main reason for using rule-based approaches is the lack of annotated corpora and limited tools for the languages, and the approach is most preferable in terms of flexibility to overcome the complexity of sentence structure and the efficiency of sentence simplification within less dataset. Performance is very important for sentence simplification because it is used as a preprocessor and post-processor for most natural language applications and social groups. For this study, we have used a rule-based sentence simplification technique that is based on POST, and it is recommended for languages that have no annotated corpus and limited NLP tools. Therefore, the rule based Afan Oromo sentence identification and simplification algorithm has been developed using open-source tools such as the Python programming language for implementation and NLTK for text processing.

1.5.5 Evaluation method

The performance evaluation of sentence simplification is the most challenging [2]. However, there are two approaches to evaluate the outcomes of sentence simplification: intrinsic and extrinsic methods. The intrinsic method is a method that determines the accuracy of the predicted simplification based on well-known accuracy, whereas extrinsic evaluation is the evaluation of a simplified sentence by its applications [3]. The extrinsic evaluation would not be useful for our purposes because our purpose is to improve the efficiency and quality of linguistically correct sentence identifications and simplifications, not to redo the investigation of the other applications. Therefore, intrinsic evaluation is used to determine the efficiency of sentence

identification and simplification. In this study, the result is evaluated manually by comparing the system outputs with the sentences produced by an expert because there is no golden standard, and the result is described in f-score by calculating recall and precision of correctly and incorrectly processed output by using human judgments [7]. The sentence simplification evaluation criteria includes grammar and fluency of the simplified sentence and also the retainment of the original meaning.

1.6 Scope and Limitation of the Study

As stated in the introductory part, there are two basic ways of complex sentence simplification: syntactic and lexical simplification, which can be investigated either separately or together. Syntactic simplification is the process of identifying the complex structure of sentences and rewriting them into a simpler structure, whereas lexical simplification is the process of vocabulary simplifying that involves identifying a difficult word and replacing it with an easy word that is equivalent in meaning. However, the study is limited to syntactic simplification of Afan Oromo compound sentences because syntactic and lexical simplifications are different natural language processing tasks that demand different resources, techniques, and evaluation methods. As a result, due to a limitation of resources and time, the study does not offer a treatment of lexical simplification. The study involves sentence identification and separation in addition to syntactic compound sentence simplification to improve sentence simplification.

1.7 Application of the Results

Sentence simplification is important in many NLP tasks, query processing, and speech processing [2, 3]. To process tasks and increase application efficiency, sentence simplification is used as a preprocessor and postprocessor. As an example, automatic kinds of question generation such as factoids, multiple-choice questions, and yes/no questions, automatic machine translation, automatic parsing, and POS tagging, information extraction (such as entity and relation extraction), semantic role labeler, frequently asked questions, question-answering, opinion mining, title generation, automatic sentence-based summarization, grammar analysis, and other [6, 24, 17, 25]. A broad coverage of sentence simplification is also expected to be useful for different social classes, such as people with language disabilities like Aphasic, low-literacy skills such as children, adults learning the language such as non-native speakers, and users of limited channel devices [2, 5, 4] and also sentence simplification can be used as a system

(web browser) for text readers to simplify text and authors to remine their work to decide the business features. We hope that once the broad coverage of Afan Oromo text simplification is completed, we will have complete control over all accesses.

1.8 Organization of the Thesis

The thesis contains six chapters. Chapter 2 gives a comprehensive literature review in the area of text simplification. In particular, it discusses sentence simplification, including approaches, evaluation methods, and also an overview of the Afan Oromo language. Chapter 3 provides a review and summary of related works. Chapter 4 is the broad and crucial part of the research which discusses the architecture and implementation of the proposed system. Chapter 5 presents the data preparation and environment, evaluation and discussion of the results, including the challenges of Afan Oromo sentence simplifications. Chapter 6, the last chapter, presents the conclusions and recommendations based on the findings of the study and shows the points of future work to be forwarded.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

Natural Language Processing is one of the major techniques in the field of artificial intelligence and it has many applications including text simplification [2]. In past work, a lot of study in the subject of text simplification was done by considering various sorts of concepts for a number of languages in order to acquire an overview of the complexity of natural language with machines and humans, utilizing multiple methodologies and methods for different goals. Thus, after examining the previous work, we offered a literature review on text simplification, specifically on the concept of syntactic simplification at the sentence level, to strengthen the validity and understanding of the title. We provide a brief introduction of sentence simplification, approaches, procedures, and evaluation strategies in this chapter. At the end, an overview of the Afan Oromo language with its detailed structure is presented.

What is Text Simplification?

Text simplification is vital NLP tasks which modifies the content and structure of a text in order to make it easier to read and understand while retaining its main idea and approximating its original meaning [2,5]. Several researchers have used a number of independent and distinct methodologies to simplify texts in combination or separately in different ways to simplify texts in different ways. Shardlow [2] conducted a text simplification survey and discussed various approaches and tasks of text simplification, such as the lexical simplification task, which deals with identifying and substituting difficult words with understandable words while still maintaining the original meaning and content of the information. Secondly, the task of splitting and re-writing grammatically complex structures into simple structures at the sentence level is known as a syntactic simplification. Thirdly, the content reduction task, which deals with the simplification of text at the document level, and lastly, the explanation generation task, which involves finding and adding more information to complex concepts in the text to provide more understandable contexts for readers.

Some NLP tasks involve a related simplification of the text, which are different and have different aims, but they may have related operations. For example, the task can easily be confused with summarization by defining simplification [4, 5]. The goal of text summarization

is to reduce the length and content of the input, whereas simplified texts are usually shorter, and simplifying the input may be more convenient, particularly if explanations are generated while the aim of text summarization is to reduce the removal of content that is less important or redundant [10]. This is usually not explored in simplification, where the contents are normally maintained.

2.2 Sentence Simplification

Sentence simplification is one of the text simplifications tasks which is the process of simplifying complex sentences into simpler sentences in order to make them easier and more understandable [1, 17, 26, 20]. The core subject of sentence simplification is performing these procedures while keeping sentences grammatical, preserving their main concept, and creating simpler output. The simplification of sentences can be further applied through a sequence of simplification operations such as Splitting, reordering, dropping, and substitution are widely accepted as important simplification operations [7, 18, 19]. The primary activity of structural simplification is the rewriting of a single sentence into several sentences while retaining its meaning. Furthermore, the drop action eliminates unimportant sections of a sentence to make it more concise. The rearrangement process changes the order of the divided sentences or sections in one judgment. Lastly, substitute complicated words or phrases with their more common synonyms [20].

2.3 Approaches of Sentence Simplification

Syntactic and lexical simplification are two basic ways to make complicated sentences simple, which can be examined alone or combined [2, 19, 17] and covered at the following:

A) Syntactic simplification

Syntactic simplification deals with syntax components that remove complexity from the grammatical structure of the sentence while preserving its information content and meaning [19, 20]. The aim of syntactic simplification is to make the text easier grammatically to comprehend for human readers or processed by programs and for assertive technology [7, 13]. There is a lot of syntactic complexity that offers several tasks, such as analyzing and splitting long sentences into their component clauses as independent clauses, subordinate clauses, adjective (relative) clauses, adverbial clauses, and correlative clauses; sentences that use the passive voice may be rewritten; co-reference and anaphora may be resolved because inadequately composed writings

are extremely hard to draw in with. Readers may find it difficult to follow the content, lose interest sooner or later in a sentence, and inevitably quit any pretense of endeavor. On account of individuals with cognitive impairments, for example, aphasia, some language structures may even lose importance. Patients will most likely be unable to recognize the subject and article when the passive voice is utilized [7]

B) Lexical simplification

The second approach to sentence simplification is lexical simplification. Lexical simplification is the task of replacing complicated terms with simplified equivalents with the same value in a given statement. In addition to executing replacements at the single word level, lexical substitution may also be done at the phrase level, which involves some understanding of how words group into separate phrases and how these can be understood and replaced. In this participation, there is no effort to simplify a text's grammar but instead, it emphasizes on simplifying complicated aspects of vocabulary and the study does not struggle with the second means of simplifying sentences [8].

The task is widely presented as a four-step pipeline: *Identifying complicated words* (the task of determining which terms in a given sentence cannot be understood by an intended target and therefore must be simplified); *Substitution Generation* (the task of identifying terms or phrases that could replace the target word complex); *Substitution Selection* (Determine which of the candidate substitutions produced can replace the complex word in a given context without losing the sentence's grammatical or meaning.), *Substitution Ranking* (The activity of order to rank the remaining candidate substitutions by their simplicity for a given complex word). This task is generally applicable for readers with learning impairments or disorders, such as Dyslexia and Aphasia, as well as a pre-processing platform for other Natural Language Processing tasks such as text summarization, MT end etc.

2.4 Method of Sentence Simplification

Currently, the development of automatic sentence simplification has been of a great deal of interest and is becoming a popular research area [7, 4]. Sentence simplification studies developed so far for different languages are based on one of the main two methods, which are rule-based and data-driven or hybrid.

A) Rule based Method

The rule-based method has been widely implemented in the production of several natural language processing tasks. The rule-based approach seeks to simplify a sentence based on knowledge-based facts [7], i.e., grammar rules; generally based on the collective linguistic expertise of human experts in the problem area. A rule-based system is made up of two parts: a set of facts about a situation and rules for dealing with them. The inference engine repeatedly selects and implements a rule whose condition is satisfied. The rule-based SS can also be based on the syntax structure, such as structures like clauses, coordinates, punctuation, and other structures. Pattern matching applied to the output of text analysis tools such as partial parsers and POS taggers is used to trigger rules in many of the shallow pre-processing systems [14, 12]

B) Data-driven Approach

In recent research work, Sentence Simplification has remained focused on data-driven methods. In contrast to rules-based methods, data-driven approaches involve learning simplification from parallel corpora of aligned, original-simplified sentences such as English Wikipedia and simple English Wikipedia for English, among others. Data-driven is probabilistic and uses statistical models rather than deterministic rules, assuming a combination of input variables and other parameters can describe the outputs. When compared to a manual-made method, data-driven methods can simultaneously make multiple simplifications and learn very specific and complex rewriting patterns [18]. Various data-driven simplification methods are available, and different researchers use different data-driven methods for text simplification. The Scarton [16] survey provided detailed data-driven approaches and categorized them as monolingual statistical MT techniques, synchronous grammar induction, semantics-assisted, and neural sequence-to-sequence models. This section will cover the most common data-driven approaches used for sentence-level simplifications.

I. Monolingual Statistical Machine Translation

Different methods consider SS as a monolingual MT task; SMT has been performed as original and simplified respectively as source language and target language [27]. The objective of an SMT model is to generate a simplification in the target language, provided the sentence is in the source language. This framework is based on a translation model and a language model. Furthermore, a decoder is responsible for producing the most likely translation given a sentence.

The language model is single-lingual and thus easier to construct [16]. Different researchers have used different translation model approaches for the simplification of the sentence as SMT. For example:

a) *Phrase-Based Approaches.* Phrases (word sequences) are the main unit of translation behind Phrase-Based SMT (PBSMT). The translation model, therefore, depends on the standardized number of times each possible phrase-pair occurs. Phrase-Based SMT (PBSMT) is based on the idea of using phrases (word sequences) As a fundamental component of translation. As a result, the translation model is based on a normalized count of how many times each possible phrase pair appears. These figures are derived from comparable corpora and automatic phrase alignments, which are derived from word alignments. Decoding is the problem of searching and also seeing sentences that optimize translation and resolve it with the best first search algorithm [21]. From the point of view of transformation capabilities, PBSMT-based simplification models are capable of performing substitutions, short distance reordering, and deletions, but do not learn more complex operations (e.g., splitting) that may require more knowledge about the structure of the sentences and the relationships between their components.

b) *Syntax-Based Approaches:* The basic unit for translation in the Syntax-Based SMT (SBSMT) is not just a phrase but a syntactic component in parse trees. In PBSMT, the language model and phrase alignments are used as the characteristics that tell the models how likely the simplification can be made, and also, unlike in PBSMT, in SBSMT it is possible to extract information on the basis of parallel [10]. The Tree-based Simplification Model suggested by Zhu [27] can execute four text transformations: splitting, dropping, reordering, and substitution as the simplification operation is required.

II. Grammar Induction

The SS is represented as a tree-to-tree rewriting method in this approach. Typically, approaches are implemented in two steps: (1) extract a set of tree simplification rules from parallel corpora of linked, original-simplified sentence pairs; and (2) learn how to apply the rule (s) to unseen sentences to produce the best-simplified sentence. This is similar to how an SBSMT approach works in that the principles are the features, and the decoder uses the learned model to decide how to apply those rules.

III. Neural sequence-to - sequence models

In this method, SS is characterized as a sequential problem that is typically solved with an attention-based encoder-decoder architecture [18]. The encoder converts the source sentence into a set of non-stop vector representations that the decoder uses to create the target sentence. This method has the advantage of allowing end-to-end fashions to be trained without the need to extract capabilities or estimate particular version components, such as the language model. Furthermore, rather than creating distinct processes like in other research, all simplification transformations can be learned at the same time.

2.5 Evaluation of sentence simplification

Several evaluation metrics have been used for performance scoring of automatic sentence simplification systems to enhance the readability and comprehensibility of both programs and users, irrespective of the methodology used to simplify the sentences. Various metrics have been used in various research, but no measure is considered as a standard. Different authors perform experiments on different datasets. Evaluations in the literature have tended to be on a small scale, at the level of a sentence, and evaluated either automatically or manually by expert judgments.

A) Human Judgement

Asking human judges to assess the quality of a simplification is the most effective way to define evaluation criteria for system performance, such as the correctness and fluency of the simplification [7]. The evaluator was asked to rate correctly and incorrectly simplified sentences by the system to compare them with the original, simplified sentence version. Different authors conduct experiments on different datasets under the human judgment of sentence simplification and use various measures to report experiment outcomes, such as precision and recall, F1 ratings [28]

Precision

Precision is the percentage of accurately simplified sentences. For example, if A is the number of correctly simplified sentences and B is the number of incorrectly simplified sentences, then

$$\text{Precision} = A / (A+B)$$

Recall

The recall of the sentence simplification system shows the coverage of the system. For example, if C is the number of correctly simplified sentences and D is the number of sentences that are not simplified by our system, then

$$\text{Recall} = C / (C+D)$$

The F1 score is the harmonic mean between precision and recall that measures the overall score of the system:

$$F1 = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

B) Automatic Evaluation

Even though human evaluation is the most reliable, it is very time-consuming and boring if done repeatedly. Therefore, with some compromise of quality, automatic evaluation metrics are useful for testing the system quickly and comparing various architectures frequently. Since personal biases do not play a part, they may also be called more impartial than humans. However, the metrics used in the SS analysis are unreliable and most of these metrics are taken from the literature of the MT since SS can be seen as translating a text from complex to simple [26]. There are different kinds of automatic evaluation metrics that have been commonly used to evaluate the performance of sentence simplification, such as BLEU, SARI, Flesch metrics, etc. BLEU calculates the output by counting N-gram matches to the reference and it was found to be reliable only for evaluation in MT but not for other tasks of natural language generation, but it is not suitable for most SS [25]. SARI compares the output with both the reference and the input sentence but is limited to lexical simplifications and short-distance reordering. Despite the possibility of more text transformations and the evaluation of full documents and not sentences, Flesch metrics were produced.

2.6 Overview of Afan Oromo Language

Afan Oromo is one of the Ethiopian languages spoken by the largest ethnic group of the Oromo people, and Ethiopia's Statistical Population Report and Housing Census (2007) found that 34.5% of the ethnic distribution of the Ethiopian population was Oromo. Currently, Afan Oromo is the official language of the regional state of Oromia and has been used as a working language in offices and as an educational language. At the national level, in Ethiopia, some public universities are offering up to Ph.D. programs majoring in the Afan Oromo language. Afan Oromo is also broadly used as a written and spoken language in some other Ethiopian regional states and neighboring countries, such as Kenya and Somalia [27]. This section presents the structure of Afan Oromo word groups, clauses, and sentences in detail.

2.6.1 Afan Oromo Writing system

Afan Oromo uses the 'Qubee' alphabet, which is the Latin alphabet for writing purposes and Afan Oromo's writing system is almost phonetic as it is written the way it is spoken, i.e. one letter corresponds to one sound with consonants and vowels sounds. There are 33 consonants in Afan Oromo, of which seven are consonant letters combined: 'CH', 'Ph', 'SH', 'TS', 'DH', 'NY', and 'ZH' The combined letters together are known as 'Qubee Dachaa'. Afan Oromo has five vowels that are short and five long. Capital letters and small letters like the English language distinguish the Afan Oromo alphabet [29] As stated in [27], every natural language has its standard word order that is used to construct a grammatical and meaningful sentence. For example, English and Afan Oromo have differences in their syntactic structure. Particularly, the syntax of English is Subject-Verb-Object (SVO) whereas the Afan Oromo sentence is Subject-Object-Verb (SOV) in a declarative sentence. For example, in the Afaan Oromo sentence “Caalaan kitaaba dubbisa”, “chaalaan“ is a subject, “kitaaba” is an object and “dubbisa” is a verb. Therefore, it has an SOV structure, whereas in English, "Chala reads a book," which has an SVO structure.

Furthermore, in English, the adjective comes before the noun or pronoun it modifies, whereas in Afan Oromo, nouns or pronouns come before adjectives. For example, in Afan Oromo, "waraqaa adii". "waraqaa" is a noun, "adii" is an adjective, Therefore, the noun precedes the adjective, whereas in English "white paper," the adjective precedes the noun. The other difference between English and Afan Oromo is the formation of adverbs. In Afan Oromo, the adverb precedes the verb, whereas in English it is preceded by a verb. For example, in Afan Oromo, "dabalaan kalleessa dhufe". "Dabalaan" is a noun, "kalleessa" is an adverb, and "dhufe" is a verb. Therefore, it shows the adverb precedes the verb. In English, "Dabela came yesterday", an adverb preceded by a verb. In contrast to English, Afan Oromo does not need articles that occur before nouns. There is also a difference in the formation of articles. In Afaan Oromo, the noun's final vowel is lost, and suffixes (-icha, -ittii, -attii) are added instead of a definite article to show definiteness. For example, "The boy" (masculine) is "muccicha" to mean English certainty.

2.6.2 Afan Oromo Word Class

In this section, we discuss the word classes that are in use in the Afan Oromo language and have contributions to our study. Words are the basic unit of a given language and the combination of

these words on the basis of the language gives us phrases, clauses, and sentences. However, the extent to which a certain word impacts the meaning of a sentence, on the other hand, is determined by that word's contribution. In the Afan Oromo language, based on their context and formation in the sentences. However, different scholars hold opposing viewpoints on the Afan Oromo word categories. In recent works [29, 22, 30] Afan Oromo words are categorized into five major groups. These categories are noun, adjective, adverb, verb, ad-position, and each of these categories can again be divided into other sub-classes. For example, the noun category includes all nouns and pronouns, whereas the ad-position includes conjunction and other prepositions and postpositions,

Noun (Maqaa)

Afan Oromo nouns are names that are used to identify any group of people, animals, places, things, or abstract ideas. The number of nouns can be indicated by using numerals and quantifiers. Usually, numerals and quantifiers, for example, indicate numbers in countable nouns. In sa'a lama (two cows), lama (two) indicates the number of cows, and in saawwan baay' ee (many cows), baay' ee (many) is the quantifier to reflect plurality. Most of the time in Afan Oromo, a sentence begins with a noun that begins with a capital letter and uses a noun as a subject and also the noun which is the subject of a sentence often optionally follows the direct object and indirect object [21]. For instance

✚ “Abdiin mana barumsaa deeme” (‘Abdi went to school’)

In this sentence, both “Abdiin/Abdi” and “Mana barumsaa/School” are nouns acting as subject and object respectively. Afan Oromo noun has three major types of cases: nominative (subjective), objective, and possessive. The common nominative case markers in Afan Oromo are –n, -ni, and –i. These markers inflect the noun and finally form the subjective case.

✚ chaaltuun deemuufi (chaltu will be leaving)

✚ Barattooni dhufan (students were come)

Pronoun (Maq-dhaala)

Words that can be used in place of nouns are known as pronouns. Pronouns, like nouns, have a number and a gender [29]. Such as, **isa** which means “he” is masculine (singular) and **ishee** **which** means “she” is feminine (singular) and also **isaan** which means 'they' is plural can be masculine or feminine. Pronouns can also be classified according to their functions and meanings

in a sentence. Personal pronouns, possessive pronouns, demonstrative pronouns, relative pronouns, and reciprocal pronouns are examples [29, 31].

Verbs (Xumura)

Verbs are words in Afan Oromo that are used to indicate an action or event that occurs within time boundaries [32] and that can be transitive, intransitive, modal, or auxiliary in nature. Transitive verbs are those that transfer a message to a complement or object, whereas intransitive verbs do not do so and thus do not have a complement or object. Afan Oromo verbs are found at the end of sentences. The examples below demonstrate this point.

- ✚ Abdiin kitaab **dubbise**. (Abdi reads a book)
- ✚ Boontuun boru **dhufti** (Bontu will come tomorrow)

In these sentences, verbs are found at the end of the sentences, and the verbs "dubbise/read" and "dhufti/come" are transitive and intrusive verbs, respectively.

Adjective (Ibsa Maqaa)

A noun or a pronoun is modified by an adjective by describing, identifying, or quantifying words. In Afan Oromo, an adjective is usually used after the noun or pronoun that it modifies. An adjective can also be classified according to its functions and meanings in a sentence. These are descriptive adjectives (dheeraa, Adii, furdaa), possessive adjectives (kiyya, kee), quantitative adjectives (hedduu baayee, hundaa), number and rank (tokko, sadaffaa), and adjectives or quaternary adjectives (maalii). The following are some of Afan Oromo's adjectives:

- ✚ Milkeessaan dheeraa dha (milkesa is a tall)
- ✚ Kitaabich kiyya (the book is main)
- ✚ Abdiin kitaaba tokko qaba (abdii has two books)
- ✚ Mucaa kam? (Which boy?)

Adverb (Ibsa Xumuraa)

An adverb is words that change the meaning of a verb, an adjective, another adverb, or a clause. Adverbs frequently appear before the verbs they modify or describe in Afan Oromo sentences. An adverb provides an answer to a question by indicating the time, method, place, cause, or degree. In the following example, each of the bold words is an adverb.

- ✚ Abdiisaan **kalleessa** deeme. (Abdisa went yesterday)
- ✚ Guddattuun **daftee** dufte (Gudatu came quickly)

✚ Magarsann **yeroo hunda** fiiga (Magarsa runs every time)

Ad-positions (Dur-Duubee)

Prepositions and postpositions are words that have full meaning only when combined with other words, such as nouns, adjectives, or verbs. akka 'as,' eegasii' since,' hamma 'till,' gara 'to,' gadi 'below,' irra /gubbaa 'on,' and so on are some prepositions in Afaan Oromo. Prepositions and postpositions can be used before or after the category to which they add syntactic meaning. In the following, the words "irra" and "gara" are postposition and preposition respectively.

✚ Biifaan adurree teessoo **jala** ke'e (Bifa put the cat under the chair)

✚ Biiftuun **gara** mana barumsaa deemte (Biftu went to school)

Conjunction (wal-qabsiiftuu)

A conjunction is a word that joins words, phrases, clauses, or sentences together. In Afan Oromo conjunctions, there are coordinating and subordinating conjunctions. Coordinating conjunctions are used to join main clauses that are given equal emphasis by the user. Subordinating conjunctions are those conjunctions that are used to join the main clause with the subordinate clause. A clause usually follows a subordinating conjunction. Table 2.1 illustrates some of the basic Afan Oromo conjunctions.

Table 2.1 List of Afan Oromo Conjunctions

Coordinate Conjunctions	English Equivalent	Subordinate Conjunctions	English Equivalent
fi	and	yoo/osoo	If
Garuu	but	wayta/ yeenna	when
Yookiin/moo	or	Hamma	Until
Kanaaf/kanaafuu	So/therefore	Erga	After
Haata'uu malee	However,	dursa	Before
Ta'us/Ta'ullee	Though/Even though	akka waan	As if
Sababiin isaa	Because		

Let us see them in the following sentences.

✚ Boontuun daa'ima **garuu** waan baaayee beekti (Bontu is a child but she knows everything)

✚ Suuta deema **sababiinsa** nan dadhabe (I walk slowly because I am tired.)

✚ Caalaan baayee cimaadha **kanaafuu** tokkoffa baha (chala is very clever so he ranked first)

✚ **Yoo** baayee shakaleta ni mo'atta (if you train more you will win)

✚ **Wayta** chalaan galu ani rafa jira. (When chala was come I was sleeping)

In the first three sentences, the bolded words are used as coordinative conjunctions to connect the independent clauses, whereas in the last two sentences, they are used as subordinate conjunctions to connect dependent and independent clauses by preceding the dependent clauses.

2.6.3 Afan Oromo Clauses

A clause is a group of words that contain subject and verb in their structure. The Afan Oromo clause has at least a verb and subject [28]. In this study, the knowledge of clause is the main important to break down complex sentences into simple accordingly. As stated by Hailu [22], In Afan Oromo there are two main types of Afan Oromo clauses: Independent clause (Main clause) and Dependent clause (subordinate clause).

Independent Clause

An independent clause is a group of words that has both a subject and a verb and gives a meaningful and complete message. An independent clause is a clause that can stand-alone, which is a simple sentence. For instance,

✚ Abdiin kitaaba bite (Abdi bought a book)

✚ Boontuun mana barumsaa deemuufi (Bontu will go to school)

These clauses are independent clauses or a simple sentence that contain subject and verb to express a complete and meaningful message.

Dependent clause

A dependent (subordinate) clause is a clause that depends on another clause to make a complete sentence. Subordinate clauses contain both a subject and a verb, but do not express a complete message and it is attached to another independent clause by subordinate conjunction to give a complete message. For Instance,

✚ Osoo Abdiin kitaaba bitee”(if Abdi bought a book)

✚ “Erga inni dhufe” (after he came)

In this example, both clauses cannot deliver a complete message; but the independent clauses complete them so that a complete message could be delivered. In the following sentences, the bolded are independent clauses that completed the incomplete dependent clauses.

✚ “Osoo Abdiin kitaaba bitee **ani ni dubbisa**” (if Abdi bought a book I read the book)

✚ “Erga ani dhufe **mana barumsaa deeme**” (after I came he went to school)

2.6.4 Afan Oromo Sentence

Afan Oromo sentence is made from a combination of one or more clauses that give a meaningful message. The Afan Oromo language follows a subject-object-verb (SOV) grammar order, unlike English, which follows a subject-verb-object sequence of words, for instance, "Abdiin mana barumsaa deeme" is written as "Abdi went to school" The subject and verb are the most important components of Afan Oromo sentences. The subject of a sentence is a noun phrase that always comes before the verb. Afan Oromo sentences are categorized based on two main aspects [29, 22]: According to the purpose they achieve, sentences are of four types: declarative, interrogative, imperative, and exclamatory sentences, and based on their structure, Afan Oromo sentences are categorized into four simple, compound, complex and compound-complex sentences. Since our work is focused on the syntax and structure of the declarative sentence category.

Simple sentence

A simple sentence is a sentence with a single independent clause, or it is a sentence that consists of only a single verb in its structure. For instance,

✚ “Abdiin mana barumsaa deeme” (Abdi went to school)

This is a simple sentence or an independent clause. It contains a subject (Abdiin/Abdi) an object (“mana barumsaa/school”) and a verb (deeme/went), and it expresses a complete thought.

Compound sentence

A compound sentence in Afan Oromo is a sentence comprising at least two simple sentences or independent clauses that are combined to create a compound sentence. Each main clause has its own subject and predicate and they might share the subject, the object, or the verb of the main clause. These clauses are combined by coordinating conjunctions, semicolons, and adding markers like {-e} to a main verb. For example, the following are examples of independent clauses that can be joined together by coordinate conjunctions.

✚ Abdiin Boontuu waamaa jira **garuu** boontuun mana hin jirtu (Abdi is calling Bontu but Bontu is not at home)

✚ Dabalaan mana barumsaatii dhufee Rafe (Dabela came from school and he had slept)

These sentences are compound sentences that are constructed from two simple sentences (independent clauses) joined by coordinate conjunction (garuu) and another feature (-e).

Complex sentence

A complex sentence is made from the composition of independent and dependent clauses. It must contain at least one independent clause and one or a number of dependent clauses. Dependent and independent clauses are joined by using subordinate conjunctions. The thing that makes this sentence complex is the existence of a dependent clause. For instance:

✚ “Yoo dhufte, yoo haftes naatti himi (tell me either you came or not)”

The sentence on the above example is made up of two dependent clauses and one independent clause.

Compound-Complex Sentence

A compound-complex sentence is created by combining compound and complex types of sentences. There are at least two independent clauses and one subordinate clause. This means that there are two or even more independent clauses and a single or more dependent clauses in a single sentence. For example, the sentence below contains two dependent clauses and three independent clauses.

✚ “Yommuu dhaqes, yommuu gales, natti goree na gaafatee darbe” (He asked me when he had gone and come).

CHAPTER THREE: RELATED WORK

3.1 Introduction

Sentence simplification is one of the most common areas of research in the field of natural language processing due to the rapid growth of text [4]. In recent years, it has taken a great deal of interest in natural language processing and social communities [4]. Even though the early simplification of the text has been investigated, there are fewer SS works. However, some conducted research in different languages is based on various approaches and many of them carried out in English and some other foreign languages. To the best of our knowledge, no research has been done on Afan Oromo sentence identification and simplification directly to get out the problem of the previous work even for local languages. In this section, the overview is mainly focused on the previous work related to the syntactic level of sentence simplification but does not include the task of lexical simplification.

3.2 Sentence Simplification for English language

The work of [1], has proposed that simple sentences be generated from complex and compound sentences. A novel algorithm was developed in the study that uses dependency parsing of input text to generate simple sentences from complex and compound sentences. The internal part of sentence simplification, such as different types of clauses and co-reference resolution, and also the accesses of the simplified sentence in some other NLP applications, were discussed and justified as simple sentences improve their accuracy. Text summarization based on sentences, information extraction and machine translation, etc. are used as preprocessors and postprocessors. To implement the desired algorithm, they used the available ‘Stanford Parser1’ and ‘Stanford CoreNLP2’ along with the help of the Stanford Typed Dependency. They completed a few preliminary tasks, such as detecting and distinguishing simple sentences from others, before applying the algorithm to the dependency parsing of the input corpus.

The performance of the proposed system was tested using data collected from different openly available online resources. Five linguistic experts judged the correctness of a simplified sentence, and the last value of the result was taken from the average of their judgments. To calculate the performance of the proposed system, data was collected from two Wikipedia pages. From the collected sentences 92 were given to the system while the others were separated as

simple sentences and the end judgmental average accuracy of the proposed system is 91.102%. As a drawback, they stated that the system generates an incomplete sentence due to the incorrectness of the sentence structure, but using the technique of dependence parsing by itself results in an error as the sentence becomes long and complex, especially for the Afan Oromo language that has dedicated NLP applications.

In the paper [16], a rule-based method for simplifying sentences has been proposed to improve the machine translation system from English to Tamil. They presented the simplification in the English language of a complex sentence and then translated the sentence into the Tamil language without altering the meaning of the sentence, but before the simplification, the system lacked accuracy due to sentence complexity. They suggested a method to solve the difficulty that first identifies the compound and complex sentences and then simply translates them into a simple sentence. In order to obtain simple sentences for machine translation, the strategy was based on relations like relative pronouns, coordinating, and subordinating conjunctions. In the paper, sentence simplification, sentence segmentation, POS tag, and machine translation have been used.

Different scholars have been focusing on different machine learning methods to simplify complex sentences, particularly for the languages that have the largest and openly available corpus [20]. In the study [11], an unsupervised method of SS was proposed using a multi-stage encoder based on a sequence-to-sequence model to solve the problem of the single-stage encoder (conventional Seq2seq) model which was developed by [18]. To improve SS over the conventional Seq2seq model, the proposed model was implemented in three stages. The first stage is the N-gram reading stage, which catches N-gram feature embedding for other stages. The second stage is the glance-over, in which local and global information about the source sentence is obtained. The final stage is the encoding stage, which uses the information gathered in the previous two phases to better encode the source text. To experiment with the proposed model, three public datasets, WikiSmall, WikiLarge, and Newsela, have been demonstrated.

Signs of syntactic complexity have been studied by [7] for rule-based sentence simplification of the English language. In the study, two approaches were used to minimize the number of double clauses and nominally attached relative clauses. The first part is a sign tagger that classifies signs automatically according to the annotation scheme used to annotate the corpus. The second part

is an iterative sentence conversion tool based on rules. The sentence transformation method automatically rewrites long sentences containing compound clauses and nominally binding relative clauses as sequences of shorter single-clause sentences by using the sign tagger along with other NLP components. In rewriting sentences containing compound clauses, evaluation of the various components reveals reasonable efficiency, but less accuracy when rewriting sentences containing nominally bound relative clauses.

For semantic function labeling, K. Vickrey [33] used sentence simplification. By retaining all the arguments for a verb, they condensed the sentence by eliminating a bit of detail beyond arguments and the verb. Zhu et al. [12] proposed a Tree-based Simplification Model (TSM) to generate the parse tree of a complex sentence by applying [34] to generate the parse tree of a complex sentence by applying operations of splitting, reordering, substitution, and falling. They carried out sentence simplification using tree transformation techniques based on statistical machine translation techniques.

The Simplified Factual Statement Extractor was created by Heilman and Smith [17] to extract multiple condensed sentences from a source sentence. They demonstrated that the addition of text simplification improved the results of automatic question generation. A Sentence Based on an Entity Miwa et al. proposed a technique for simplifying relationship extraction. The target sentence's meaning is less important than preserving the true value of the relationship in order to simplify the sentence in relation to extraction. Two rules, clause selection rules, and entity phrase rules were therefore established. The rule of clause selection is used before and after the relevant clause to eliminate noisy information. Whereas the rule of the entity phrase is used to simplify an entity holding region without altering the relation's true value.

3.3 Sentence Simplification for Punjabi language

The research article of [28] has proposed a novel approach for the conversion of participial types of complex sentences in the Punjabi language. The author carried out lexical, syntactic and content simplification. The producer used morphological characteristics along with part of speech (POS) tagging for syntactic simplification. In order to distinguish dependent and independent clauses, more clause boundary knowledge is used, and then dependent clauses are transformed into independent clauses in all possible situations. Lexical simplification was performed based on the frequency of words using support vector machine (SVM) and content

reduction was performed. The performance of the proposed algorithm was tested in terms of Recall and Precision within a large corpus and the overall gained accuracy is more than 93.79% F score. The developed system is suggested for Aphasic and Dyslexia readers as well as for machine translation systems, summarization systems, and other Natural Language Processing applications.

In the study [15], the initial implementation of complex sentence identification and separation of the corpus of the Punjabi Language has been developed based on the manually developed pattern by using the characteristics of a complex sentence in the language. They introduced the identification mark (conjunction) that helps to separate predicate-bound and non-predicate-bound types of complex sentences. The accuracy has been tested on the Punjabi corpus taken randomly from different internet sites by grouping into A and B samples. Sample A and B contain 2400 and 3100 sentences respectively. Sample A has an accuracy of 85 percent for predicate bound sentences and 81 percent for non-predicated bound sentences, whereas simple B has an accuracy of 82 percent for predicate bound sentences and 80 percent for non-predicated bound sentences.

The paper of [3] proposed an algorithm for the identification of different types of sentences using a specific feature of the sentences in the Punjabi language. such as the identification of types and the number of clauses in the sentences. For example, in order to identify complex sentences, they identify the dependent clause. A dependent clause is identified using a special characteristic of the dependent clause such as the presence of a finite verb in the dependent clause, the presence of a specific postposition after the verb's root form, etc. Similarly, compound sentences identify the presence of multiple independent clauses in a sentence. Likewise, for compound sentences, they identify the presence of several independent clauses within a sentence. These multiple independent clauses are expressed by having several verbs in the sentence. If there is only one stand-alone clause, it is known as simple sentence structurally. The proposed system was measured in terms of precision and recall, and 90% accuracy was gained.

3.4 Sentence Simplification for French language

The paperwork of [12] has presented a method of syntactic simplification for French texts to make texts easier to understand by simplifying complex syntactic structures. The investigation aimed the data-driven approach, which is based on two parallel corpora of encyclopedia articles

and tales. The goal is to establish and organize the linguistic phenomena involved in the manual simplification of French texts within a typology. To produce simpler sentences, they proposed a syntactic simplification scheme that relies on typology. The module begins by generating all possible variants before the best subset is selected. The assessment shows that about 80% of the system's simplified sentences are correct.

Violeta Seretan [12] developed the corpus-based acquisition of syntactic simplification set of rules for the French language. The work introduced the improvement of a complex sentence in the language that centers syntactic complexity decrease and manages the assignment of how that the data-driven manual acquisition of simplification rules can be supplemented by the semi-automatic detection of syntactic constructions requiring simplification. In particular, they are interested in syntactic changes that alter the macro-structure of a complex sentence, i.e., splitting clauses, extracting clauses, and dis-embedding non-finite clauses, while leaving most of the internal structure of the clauses intact.

3.5 Sentence Simplification for Urdu language.

The study has developed a clause boundary identification algorithm using classifier and clause markers in the Urdu language. As for the classification strategy and the clause labels, they used Conditional Random Field. The markers of the clause perform the role of distinguishing the form of the subordinate clause with or beyond the main clause. After checking with different sentences if there is some misclassification, then more rules are found to get high recall and accuracy. The findings show that the method effectively defines the form and boundary of the subordinate clause. The POS and chunked tagged corpus have been considered as input data. The approach to machine learning is initially applied, in which linguistic rules are used.

3.6 Summary of Related Work

In this chapter, we presented some earlier work connected to our work throughout the world which has been examined by various studies using various methodologies for different languages. We found three types of SS approaches in the review: rule-based methods, which are handmade rules or heuristics; data-driven methods, which are based on the utilization of training data; and hybrid methods. The technique of evaluation has also been considered. The effectiveness of simplifications has been assessed using automated evaluation metrics (i.e.,

BLUE, SARI) that can be carried out by a machine, whereas human evaluation is carried out manually by human judgments [4, 18].

Because of annotated data and tools are not publicly available, most SS are performed using rule-based approaches in previous works, and better performance is obtained [1, 28]. Moreover, the rule-based method has been used because of its flexibility to overcome the complex structure in spite of different languages. However, currently, different research has been developing data-oriented simplification methods which mean simplification based on machine learning algorithms, and also better performances are obtained [16, 25, 26], but data-driven approaches require processed datasets of more than ten thousand (i.e., small Wikipedia and large Wikipedia, Newsela, etc.) and prerequisite tools for sentence simplification (i.e., POS tagger, parser, tree-bank, etc.) are required. As a result of the gained knowledge, for Afan Oromo language, sentence simplification is a novel notion which needs deep knowledge and resources, but the language’s resources are limited, Therefore, we suggest a rule-based approach to the Afan Oromo Sentence simplification in light of this challenge and the results of the review of related publications. And also, sentence identification and separation is the main task that should be discovered before simplification to enhance the performance of complex sentence conversion into simple [3]. As indicated in Table 3.1, we offer an overview of related research that emphasizes approaches, mechanisms, used datasets, and the outcomes produced.

Table 3.1 Summary of Related Work

Author, year	Title & language	Approach category	Techniques And tools	Dataset’s source and size	Overall results & Evaluation techniques
Das et al., 2018	A Novel System for generating Simple Sentences from complex and compound Sentences	Rule based.	<ul style="list-style-type: none"> ✓ Dependency Parser from Stanford CoreNLP2’ and Stanford Parser1 ✓ Co-reference resolution ✓ Algorithm works on Sanford parser to identify & generate simple sentence 	162 sentences are collected from Wikipedia pages	F-measure ~91.102% by Human evaluation
Jindal 2019	Simplification of Punjabi Sentences: Converting Complex Participial Sentences	Hybrid	<ul style="list-style-type: none"> ✓ LS using word frequencies (SVM) ✓ Syntactic using Morphological features & POST 	5876 sentences from deferent online resources	F-measure~ 93.79%

	into Simple Sentences		✓ Content reduction		
Evans 2018	Identifying signs of syntactic complexity for rule-based sentence simplification	Hybrid	<ul style="list-style-type: none"> ✓ Classify sign automatically using sign tagger ✓ SS using iterative rule-based sentence transformation 	Collected from online resource	stated as promised
Lemin Zhang, 2019	Sentence Simplification Based on Multi-Stage Encoder Model	Data-drive	SS in multi-stage Seq2qen model using RNN encoder and decode algorithm	<p>WikiSmall (89,042 sentence pairs)</p> <p>WikiLarge (296,402 sentence pairs) Newsela (94,208 sentence pairs)</p>	<p>30.74 % SARI (WikiSmall)</p> <p>94.60% BLEU (WikiLarge)</p> <p>30.02% SARI (Newsela)</p>
Brouwes 2014	Syntactic sentence simplification for French	Rule based	Typology based simplification	parallel corpora (encyclopedia articles and tales)	80%
Dhanalakshmi 2011	Rule based Sentence Simplification for English to Tamil Machine Translation System	Rule based	<ul style="list-style-type: none"> ✓ POS tagger (Sanford parser) ✓ Splitting and Simplification. 	Not defined	Better translation accuracy.
Chandni 2014	Identification and Separation of Simple, Compound & Complex Sentences in Punjabi Language	Rule based Algorithm	<ul style="list-style-type: none"> ✓ POS tag sets (manual) ✓ Identify based on number and types of verb and clauses 	1100 Sentences	<p>F-score 90%</p> <p>Human evaluation</p>
Navneet Kaur, 2013	Identification and Separation of Complex Sentences from Punjabi Language	Rule based Algorithms	Developing pattern based on the sentence grammar	Collected from internet as sample A (2400) and sample B (3100) sentences	85% (predicate bounded) and 82 (non-predicate bound sentences)

CHAPTER FOUR: DESIGN AND IMPLEMENTATION

4.1 Introduction

In this chapter, we discuss the overall design, processes and implementation of the proposed automatic sentence identification and simplification of Afan Oromo text. First, we illustrate the general architecture of the proposed system from the perspective of the system's phases and flow of each operations. Then we present a detailed explanation of these phases along with the subcomponents within the implementation algorithms.

4.2 System Architecture

The proposed automatic Afan Oromo sentence identification and simplification is the first study of the Afan Oromo language and the investigation of the study is implemented based on a rule-based approach. The main decision to use a rule-based approach is mainly due to the public unavailability of sentence-level annotated corpora and other prerequisite resources or tools such as sentence identification and separation, clause detection, POS-tagging, Co-reference resolution, discourse maker detection, etc. are the necessary resources in the sentence identification and simplification, but they are not completed and available publicly for research purposes rather than educational fulfilment. Also, many scholars recommend the approach for such Afan Oromo that has fewer resources, particularly to handle the complexity of sentence structure and to get the promised results [7, 29]. In another case, sentence simplification is one of the high-level NLP applications that passes through different applications that are typically used as preprocessor resources.

The resources that are used to get the best final result of sentence simplification include detection and separation of sentence types, detection of clause and clause boundary typically, detection of discourse maker or connectors (conjunctions), coreference resolution, and other syntactic knowledge level correspondence such as fluency, grammar, even semantic analyzer, etc. So, it is difficult to run through all the applications in a machine learning approach at once, and it can fail under very dangerous errors, even may not possible to talk about the results within a small-scale sample of datasets. In a sense, Afan Oromo has had computational infancy because none of its features are available to the research community except for academicians' trials for academic exercises. As a result, it is the worst to have an appropriate dataset for all types of

sentences at once but gradually. Consequently, the proposed investigation of automating the identification and simplification of sentences for Afan Oromo text is actualized based on a hand-made method that operates on POS-tagged.

The Architecture of the proposed system has four major components or phases. Those are: preprocessor phase, Sentence Identification and Separation phase, Sentence Transformation phase, and Postprocessor phase. In the Preprocessor Phase at the beginning, the input of Afan Oromo text is cleaned and the only types of declarative sentences are selected from the input text because the study is particularly focused on the identification and simplification of declarative sentences than other types. Then the selected declarative sentence is tagged into an appropriate POS tag. The second phase, Sentence Identification phase determines types of declarative Afan Oromo sentences structurally based on POS tags and separates the determined sentences accordingly for the next further simplifications to make the processes easier. The third phase is Sentence Transformation. The sentence Transformation phase simplifies or splits the compound sentence into its simple sentences or independent clauses based on the syntactic structure of the language. Lastly, the Postprocessor phase rearranges the simplified or split sentences to make a complete and self-contained sentence. The flow of processes and relationships between the components of the proposed system architecture are shown as the following 4.1 figure.

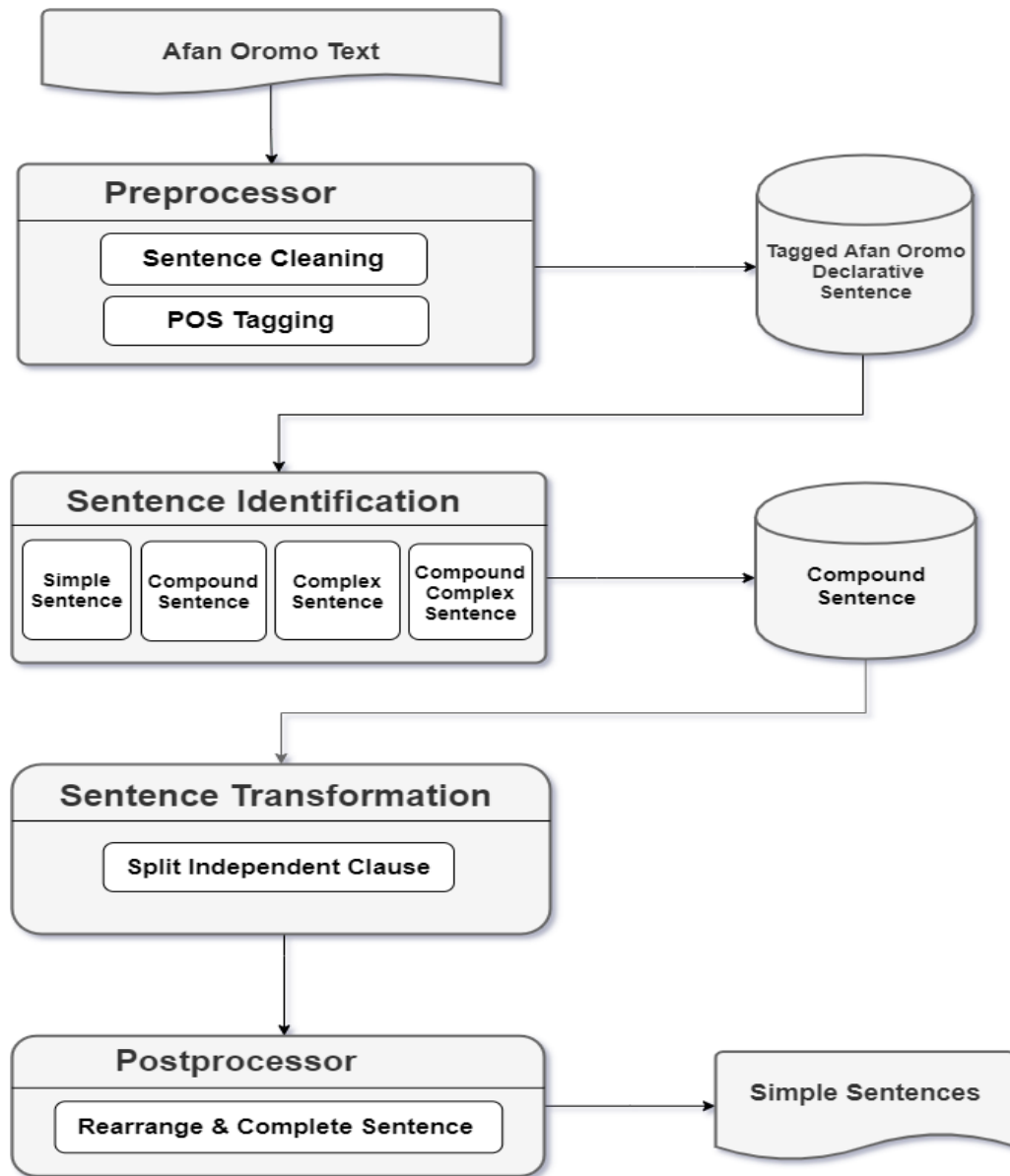


Figure 4.1 System Architecture

4.3 Components of System Architecture

The detailed description of components of the system architecture is discussed in the following sections.

4.3.1 Preprocessor

The proposed system accepts properly cleaned and tagged sentences as an input and outputs a set of structurally identified and simplified sentences. Therefore, the preprocessor phase makes that the input text ready for further processing. In the preprocessor phase, two main tasks are

performed to make the sentences ready for the next consecutive steps. Those two main tasks are sentence cleaning and POS tagging, and they are discussed in detail in the following sections.

4.3.1.1 Sentence Cleaning

The preprocessor phase starts by cleaning and filtering out well-formed declarative types of sentences from the input text. Because of that, the study is only limited to investigation of declarative sentence types both for sentence identification and simplification too. In the Afan Oromo language, sentences are categorized based on two basic aspects: Depending on the purpose or type of statement conveyed Afan Oromo sentences are of four types: declarative, interrogative, imperative, and exclamatory and also based on the grammatical structures again Afan Oromo sentences are categorized into simple, compound, complex, and compound-complex sentences [29]. Therefore, before the discussion of sentence identification and simplification, it is important to look for all types of Afan Oromo sentences with their syntactic features that enable us to identify and simplify them appropriately. Thus, the second aspect, the structural-based classification of the Afan Oromo sentence, is discussed under the 4.3.2 topic. Purpose-based sentence categorizations are discussed here because the sentence cleaning task of the preprocessor phase is only about getting declarative sentences and they are detailed as the following with appropriate examples.

Declarative sentences: declarative sentences are purpose-based Afan oromo sentence types that used to convey information. In Afan oromo declarative sentence is always ended with the fullstop punctuation mark (.) which is equivalent to in English. For example

- ✚ Falmataan kitaaba seenaa dubbisa (Falmeta reads a history book)
- ✚ Baratootni gara mana barumsaa deeman (Students have gone to school)
- ✚ Boonaan mana barumsaatii dhufee rafe (Bona came from school and he has slept)
- ✚ Abdiin Boontuu waamaa jira garuu boontuun mana hin jirtu (Abdi is calling bontu but bontu is not at home)
- ✚ Dabalaan yoo dhufe ni deemna (If dable come, we will go)

Interrogative sentences are sentences that used to ask a question. In Afan Oromo, these types of sentences are always ended with a question mark (?).

Example:

- ✚ Falmantaan kitaba dubbisaa? (Does Falmata read a book?)

- ✚ Baratooni eessaa deeman? (Where did students go?)
- ✚ Abdiin eesssa dhufee rafe? (Where did Abdi come and slept)
- ✚ Boontuun mana hinjirtuu eenyutu boontuu waama jira? (Bontuu is not at home who is calling bontu?)
- ✚ Dabalan ni dhufaa? (Will dabela come?)

Imperative sentences mostly used to give a command or make a request and it ends with a period (‘.’). for example:

- ✚ Kitaabicha naaf kenni (Give me the book.)
- ✚ Ati mana barumsaa deemi (You go to school.)
- ✚ Boontuu ciree nyaadhu (Boontu have a breakfast.)

Exclamatory sentences express emotions, pleasure, sorrow, or suddenly happening situations. In Afan Oromo, there are some words (introjections) that are used in an exclamatory sentence like in many other languages. For instance, “Ishoo” for happiness “wayyoo” for sadness ah for silent event or situation, and this type of sentence ends with an exclamatory mark (!). for example:

- ✚ Ishoo baga dhufete! (Wow! well come.)
- ✚ Wayyoo hin dhufe! (Oh! He came.)

As a result , the sentence cleaning task performs the task of filtering out the declarative sentences of the input text based on the delimiters (‘.’, ‘?’, ‘!’). In the Afan Oromo text, the existence of those delimiters is used to indicate purpose based types of sentences and determine the end of the sentences. Declarative types of Afan Oromo sentences are ended and filtered out by the fullstop delimiter (‘.’) and non-declarative sentences can also be detected easily by other delimiters. Therefore, we used those delimiters to tokenize the input text into sentences and filter out the required declarative sentence for the next process.

4.3.1.2 POS Tagging

After the declarative sentences are filtered out from other types of Afan Oromo sentences of an input text, the next sub-task of the preprocessor is POS Tagging. In the POS Tagging task, each declarative sentence is tokenized into words and tagged with their appropriate part-of-speech to provide the information of words of the sentences. In this study, POS tagging is needed for two main reasons. The first reason is that the study is particularly illustrated on the syntactic sentence

identification and simplification, and the second reason is the method preferred for the study is a hand-crafted method which most of the implementation processes are manual. So, POS tagging must be required to get information of the syntax structure or constituent parts that the sentence is made from. Even though, there is no publicly available Afan Oromo POS tagging tools, for the study, each word of the sentences is annotated manually by the researcher with the help of the linguistic expert. To annotate the prepared dataset, we used the Afan Oromo tag sets developed by Getachew [35], and Abraham [36] scholars are used in combination.

However, as the already developed tag set is not fully completed it does not cover every word as needed for this study. For example, in Afan Oromo, conjunction is one part of speech that has a vital role in sentence identification and simplification as their type (coordinate and subordinate conjunctions) to detect the clauses boundary but the scholars ignored them as stop words and/or tagged all types by the same tag set instead of identifying them typically. So, we added about eight new tag sets to an already developed tag sets to control the structure of the sentences. Then the sentence structure is controlled as needed after the new tag sets. The newly added tag sets are described in the following 4.1 table.

Table 4.1 The Newly Added Tagsets

SN.	Tag set	Definition of the tag	Examples
1.	SCC	Sentence level coordinate conjunction	Abdiin boontuu fi caalaa waamaa jira garuu/SCC isaan mana hin jiran
2.	SSC	Sentence level subordinate conjunction.	Yommuu/SSC mana barumsaa deemtu na waami.
3.	VVC	Main verb that used as verb and coordinate conjunction	Abdiisaan mana barumsaatii dhufee/VVC rafe.
4.	VVS	The main verb that used as verb and subordinate conjunction	Isheen waan sirritti hojatuuf/VVS dadhabdi.
5.	PNC	Punctuation mark used as punctuation and coordinate conjunction	Rabbiin galanni haa ga'u ./PNC jalqaba barakanaa irratti tokkummaa tokko argannee jirra
6.	NNS	Noun used as Subject (s) in the sentence	Abdiin/NNS Boontuu waamaa jira garuu Boontuun/NNS mana hin jirtu.
7.	NNO	Noun used as Object(s) in the sentences	Abdiin Boontuu/NNO waamaa jira garuu Boontuun mana hin jirtu.

8.	NNSO	Noun used both as subject and object	Abdiin Boontuu/NNSO waamaa jira garuu mana hin jirtu.
----	------	--------------------------------------	--

At the end, the preprocessor phase must fit the following specifications for the next phase:

- ✚ The text should be split into sentences.
- ✚ The declarative sentences should be filtered out.
- ✚ Declarative sentence should be part-of-speech tagged at word level
- ✚ The subject and object should be marked-up and annotated.

4.3.2 Sentence Identification

After the declarative sentence is filtered out from the text and each word of the sentence is tagged into appropriate parts of speech, the second phase of the system architecture, sentence identification task is followed. Sentence identification is the process of detecting and separating types of sentences either structurally or based on their purposes by using different techniques and methods [1, 7]. As described under 4.3.1.1, Afan Oromo sentences are generally categorized into declarative, interrogative, imperative, and exclamatory sentences based on their purposes, in addition to the structural-based categorization of each of them. In this study, the sentence identification phase performs the task of detecting and separating the declarative types of sentences structurally into simple, compound, complex and compound-complex based on the content parts of a sentence by applying a set of syntactic structures of the language. Basically, the information about syntactic structure that enables us to identify sentence types is gained from the coined tag sets and some other morphological features. Shortly, the sentence identification phase performs the task of detection and separation of Afan Oromo declarative sentences based on the POS tag sets. That is what the study is limited to in the case of sentence identification. The sentence type identification is performed to simplify the task of sentence simplification and also to improve the performance of sentence simplification, even though sentence identification by itself needs further studies as the Afan Oromo sentences are very informative and complex in nature to make complete the sentence detection and separation tasks like in other languages. However, more or less, the study looked at sentence identification separately to achieve higher accuracy in further simplification of Afan Oromo sentences.

Before the identification of declarative sentence types, it is necessary to review the four types of declarative sentences and their constituent parts, which enables us to identify and separate them typically and they discussed as the following

Simple Sentence: a sentence that has only a single independent clause which is known as a "simple sentence," but it has no dependent clauses.

Compound sentence: is a sentence that contains at least two independent clauses or simple sentences that have been combined by sentence level coordinate conjunctions.

Complex sentence: is one that contains only one independent clause and one or even more dependent clauses that cannot stand alone and must be combined by a subordinate conjunction with an independent clause to be informative.

Compound-Complex Sentence: A compound-complex sentence combines the qualities and parts of both compound and complex sentences into one sentence. It has at least two independent clauses as well as dependent clauses. Independent clauses are joined together by coordinate conjunctions, while dependent clauses are joined to independent clauses by subordinate conjunctions.

The process of sentence identification is started by checking the sentence structures (constitute parts of the sentences at the word level) based on POS tag sets. The developed sentence identification algorithm recognizes sentence types by determining the types of verbs (main and auxiliary), the number of verbs, the types of sentence-level conjunctions (coordinate and subordinate), punctuation marks (semicolon and comma) and some other morphological features. Hence, the developed iterative algorithm detects and separates types of declarative Afan Oromo sentences using the following facts:

If the sentence contains exactly a single independent clause which has a single main verb and has no other sentence-level connectors such as coordinate and subordinate conjunctions, punctuation marks (comma & semicolon) and other morphological features in the sentence then the sentence will be identified as a simple sentence. Still, in a simple sentence, more than one subject can be shared by a single verb, but the number of subjects or Noun phrases in the sentences is not matter for this study. For example,

✚ Abdiin mana barumsaa **deeme** (Abdi went to a school)

✚ Abdii fi caalaan kallessa **dhufan** (Abdi and Chala came yesterday)

If the sentence contains more than one independent clause with different main verbs and the clauses are combined by sentence-level coordinate conjunctions but the sentence has no dependent clauses means, there are no sentence-level subordinate conjunctions in the sentence then the sentence is identified as compound sentences. For example,

- ✚ Abdiin Boontuu **waamaa** jira *garuu* Isheen mana hin **jirtu** (Abdi is calling Bontu but she is not at home).
- ✚ Sorrettin ni **qu'atti** Caaltuun *immoo* ni **rafti** (Soreti is studying but Chaltu is sleeping)
- ✚ Nyaatan **barbaada sababiinsa** nan **beela'e** (I need food because I am hungry).
- ✚ Rabbiin galanin haa **ga'u** , jalqaba bara kanaa irratti tokkummaa tokko **argannee ture** (Thanks to God, at the beginning of this year we have got a unity).

In the above sentences, the bolded words are main verbs, they are more than one in every sentence and the bolded and italic words are sentence-level connectors (coordinate conjunctions and punctuation mark (,) in the last sentence).

Again, in a compound sentence, a single subject or a noun phrase in the sentence can be shared by two or more main verbs and interconnected by adding the long sound like {'-e'}, {'-i'} at the end of the main verb without coordinate conjunctions. In the dataset, that verb is tagged by VVC that used as both main verbs and sentence level coordinate conjunction to feature the compound sentence. For example,

- ✚ Dabalaan mana barumsaatii dhufe**e** Rafe (Dabela came from school and slept)
- ✚ Caalaan boontuu waame**e** dhufe (Chala has come and called Bontu.)
- ✚ Hawwii fi chaaltuun dhufani**i** deeman (Hawi and chaltu have come and gone)

If a sentence includes a single independent clause and one or more dependent clauses with at least one main verb, and every dependent clause is joined to each other and the independent clause by sentence-level subordinate conjunctions, but the sentence has no sentence-level coordinate conjunctions which reveal the existence of more than one independent clause, then the sentence will be identified as a complex sentence. For example:

- ✚ Boontuun **Yoo dhufte** naatti **himti** (Bontu would have told me if she came)”
- ✚ **Yoo** haalaan **hojjette**, qormaata ni **dabarta** (If you work hard, you will pass the exam.)

- ✚ ***Yoo sooromtes, yoo hiyyoomtes*** gorsa abbaa kee hin **dagatin**. (whether you become rich or poor, don't forget your father's advice.)
- ✚ dheeressaan yeroo hundaa **yommuu** nyaata **nyaatu nisirba**. (Deressa always sings a song while eats food.)

In the above sentences, there is at least one subordinate conjunction (bold and italic) and at least two main verbs (bold). Again, in the complex sentence, some verbs are used as both main verbs and connectors of the dependent clause to the independent clause without a subordinate conjunction. Those verbs are tagged by VVS in the data to handle such types of complexity. For example:

- ✚ Isheen waan sirritti ***hojatuuf*** dadhabdi (Because of hard work, she is tired)

If a sentence includes more than two independent clauses and one or more dependent clauses with the main verbs of every clause in the sentence, and independent clauses are joined to each other by sentence-level connectors (coordinative conjunctions and punctuation marks), and independent clauses are joined to each other or independent clauses by sentence-level subordinate conjunctions, then the sentence will be identified as a compound-complex sentence, i.e., as the name indicates, compound-complex sentences share the characteristics of both compound and complex sentences in common. For example:

- ✚ Dabalaan yommuu dhaqes, yommuu gales, natti goree na gaafatee darbe. (Debela asked me when he had gone and come).

In the above sentence there are two dependent clauses ('Dabalaan yommuu dhaqes', 'Dabalaan yommuu gales') and two independent clauses ('Dabalaan natti goree darbe', 'Dabalaan na gaafatee darbe'). Hence, the sentence combines elements of compound sentences (two or more independent clauses) and complex sentences (at least one dependent clause with an independent clause).

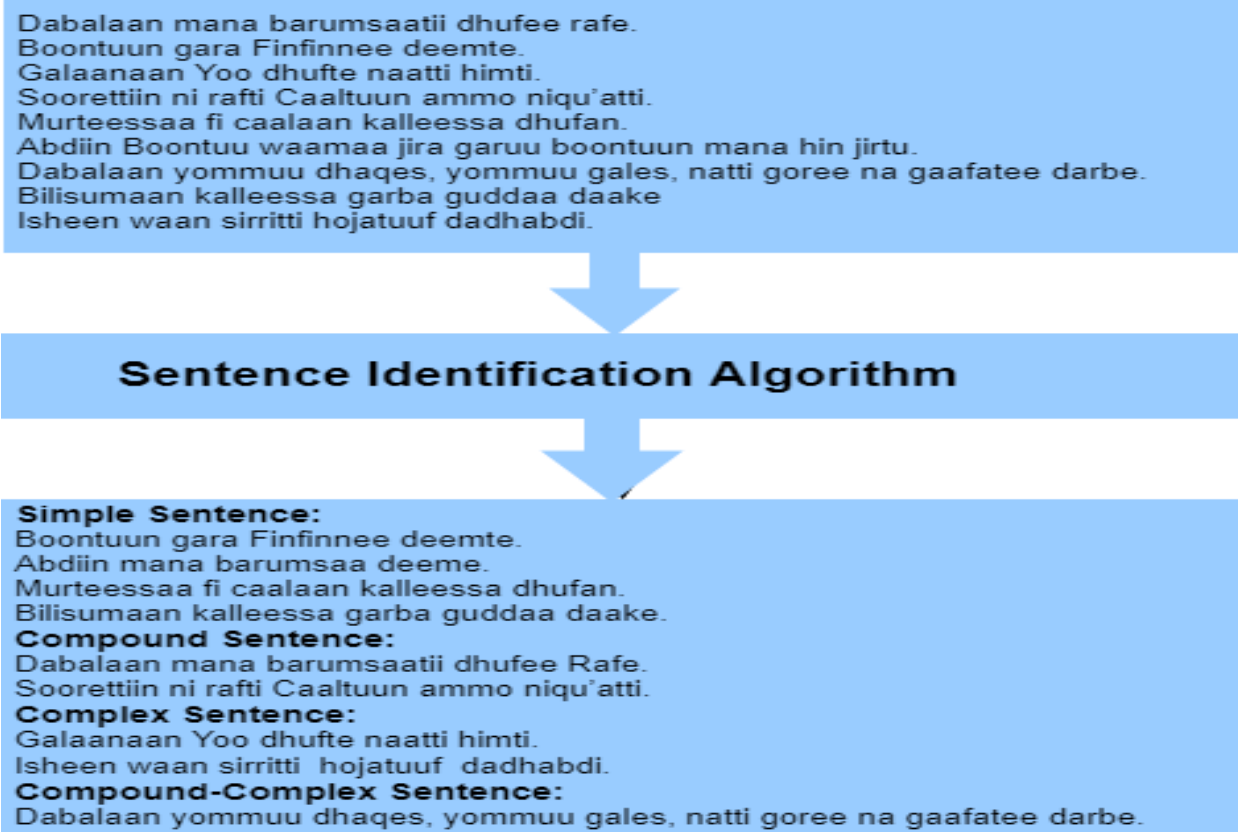
Generally, the syntactic content and structure of Afan Oromo sentences that enable us to identify and separate declarative sentences into their types, i.e., into simple, compound, complex, and compound-complex sentences, is summarized in the following 4.2 table.

Table 4.2 A summary of Afan Oromo Sentences Type Indicators

Types of a sentence	Numbers and types of clause in a sentence		Sentence/Clause Type Indicators		
	Independent Clause	Dependent Clause	Main verb	Sentence level Coordinate conjunction (SCC, VVC, PNC)	Sentence level subordinate conjunction (SSC, VVS)
Simple	1	0	1	0	0
Compound	≥ 2	0	≥ 2	≥ 1	0
Complex	1	≥ 1	≥ 2	0	≥ 1
Compound-complex	≥ 2	≥ 1	≥ 3	≥ 1	≥ 1

Altogether, figure 4.2 shows the input and output of the sentence identification subsystem of the Afan Oromo sentence simplification distinctly.

Figure 4.2 Input and Outputs of Sentence Identification



Proposed Algorithm:

To implement the identification and separation of declarative sentences types the following 4.1 algorithm is proposed.

Algorithm 4.1 Sentence Identification and Separation Algorithm

Require: *POS tagged Afan Oromo declarative sentence : Sent*

Ensure: *List of simple, compound, complex and compound-complex sentence*

MV \leftarrow *Number of main verbs in Sent*

SCC \leftarrow *Number of sentence coordinate conjunctions in Sent*

SSC \leftarrow *Number of sentence subordinate conjunctions in Sent*

while *MV* \neq 0 **do**

if *MV* is only **one** and *SCC* or *SSC* are **zero** in *Sent* **then**
 add Sent into simple sentence list

else if *MV* \geq 2 or *SCC* \geq 1 or *SSC* == 0 in *Sent* **then**
 add Sent into compound sentence list

else if *MV* \geq 2 or *SCC* == 0 or *SSC* \geq 1 in *Sent* **then**
 add Sent into complex sentence list

else if *MV* \geq 3 or *SCC* \geq 1 or *SSC* \geq 1 in *Sent* **then**
 add Sent into compound-complex sentence list

else
 add Sent into not identified sentence list

end if

Output *list of sentences type separately*

end while

4.3.3 Sentence Transformation

In the process of Afan Oromo sentence identification and simplification, the sentence transformation phase is followed after the types of declarative sentences are detected and separated by the sentence identification phase. Sentence transformation is a subtask of sentence simplification that splits or breaks down structurally complex sentences into their simple parts (clauses). Here, the study is mainly focused on the simplification of the syntactic complexity of compound sentences from other types of declarative sentences without losing the original meaning as much as possible because a simple sentence is easier for humans to understand and can be processed by a machine than a structurally complex compound sentence, it is preferred. Hence, the sentence transformation phase task is to split or break down compound sentences into simple structures or clauses by detecting a set of syntactic information that enables us to simplify

them into their appropriate parts. The transformed parts (clauses) of the sentence might need further processing to be complete and self-contained simple sentences that can be passed into the next postprocessor phase. A set of syntactic information used to recognize and split the sentences is formulated based on various features of Afan Oromo sentences that are identified during part of speech tagging.

In this case, a declarative Afan Oromo compound sentence is formed from two or more independent clauses (simple sentences). A compound sentence's independent clauses or simple sentences are connected to one another by sentence-level coordinate connectors such as coordinate conjunctions (SCC), punctuation marks (PNC), and main verbs that are used as verbs and connectors (VVC). So, in this study, the process of compound sentence transformation or splitting of the compound sentence into independent clauses (simple sentences) is performed by detecting the locations of those sentence-level coordinate connectors. To split independent clauses in a compound sentence, first the sentence-level coordinate connectors that join the clauses are detected, and then the sentence is split at the connectors. The developed algorithm accepts POS tagged as input and returns a list of simple parts or clauses. The algorithm first looks for sentence-level connectors (SCC, PNC, and VVC) from left to right. If a sentence-level connector is detected, the sentence is split there and the first clause is added to the clause list, and the remaining part of the sentence is processed in the same procedure. If the connector is not available, the remaining clause is added to the clause list. For example, the following sentences are transformed into simple parts as:

✚ Abdiin/NNS Boontuu//NN waamaa/VV jira/AX garuu/SCC Boontuun/NNS mana/AD hin/NG jirtu/VV

- Abdiin Boontuu waamaa jira.
- Boontuun mana hin jirtu.

✚ Oromiyaan/NNS saba/NN baa'ee/JJ qabdi/VV akkasumas/SCC qileensaa/NN gaarii/JJ qabdi/VV ammas//SCC baayee/JJ guddoodha/VV

- Oromiyaan saba baa'ee qabdi
- qileensaa gaarii qabdi
- baayee guddoodha.

✚ Jaalanneen/NNS kaleessa/AD mana/NNO barumsarra/JJ hafte/VV ta'ullee/SCC hojii/NNO manaa/JJ hojjatee/VV jira/AV

- Jaalanneen kaleessa mana barumsarra hafte.
- hojii manaa hojjatee jirti.
- ✚ Abdiin/NNS hin/NG baranne/VV garuu/SCC waan/PP baay'ee/JJ beeka/VV
 - Abdiin/NNS hin/NG baranne/VV
 - waan/PP baay'ee/JJ beeka/VV

Algorithm 4.2 shows the implementation of compound sentence transformation that splits or breaks down the syntactic complex of compound sentences into simple structures (clauses).

Algorithm 4.2 Compound Sentence Transformation Algorithm

Require: *POS tagged Afan Oromo compound sentence* : Comp_Sent

Ensure: *List of split clauses*

```

Connector_index ← Find connector's index tagged as SCC, VVC and PNC in Comp_Sent
for each sentence S in Comp_Sent do
  if Len(Connector_index) == 1 then
    split S into two at the Connector_index and add them into clause_list
  else if Len(Connector_index) ≥ 1 then
    for each Connector_index C in S do
      Substring S at C and add into Clause_list
    end for
  else
    The sentence is not compound sentence
  end if
  Output list of clauses
end for

```

4.3.4 Postprocessor

After the sentence has been transformed into a simple structure, the postprocessor phase has been implemented. The simplified structure of the compound sentences may not be self-contained and arranged mannerly during transformation. Hence, to make all the simplified parts of the compound sentence self-contained and correctly arranged, the next postprocessor phase is required. The preprocessor phase is the process of making newly generated sentences complete, self-contained, and rearrangeable in the order of SOV forms while the sentence preserves the original meaning as much as possible.

The rearranging task starts by checking if the transformed parts are either completed or not, because all of the newly transformed parts (independent clauses) of the compound sentence are not self-contained and mannerly organized. As described under 4.3.2, in a compound sentence,

a single subject or noun phrase can be shared by two or more main verbs, and the sentence may have different contents that make it difficult to get an independent clause from the sentence. For example, in the following sentences, "Oromiyaan" and "Abdiin" are subjects that are shared in common by all the clauses of the sentence and they must be transformed with the subject to be self-contained simple sentences as the following.

✚ Oromiyaan/NNS saba/NN baa'ee/JJ qabdi/VV akkasumas/SCC qileensaa/NN gaarii/JJ qabdi/VV ammas//SCC baayee/JJ guddoodha/VV

- Oromiyaan saba baa'ee qabdi.
- Oromiyaan qileensaa gaarii qabdi.
- Oromiyaan baayee guddoodha.

✚ Abdiin/NNS hin/NG baranne/VV garuu/SCC waan/PP baay'ee/JJ beeka/VV

- Abdiin hin baranne.
- Abdiin waan baay'ee beeka.

In a compound sentence, an independent clause can be joined without conjunction by the main verb that is tagged as 'VVC' in the data. The previous phase, sentence transformation, only splits the sentences at the main verb tagged as "VVC" including the verb, but the postprocessor phase changes the verb form by removing postfixes (-e) and likes of the main verb that is tagged as "VVC" and checking the subject of the newly transformed clause. If it is not available, get it from the preceding clause because the subject (s) is common for all verbs in the sentence. For example:

✚ Oromiyaan/NNS mana/NNO seentee/VVC boorsaa/NNO fudhattee/VVC baate/VV

- Oromiyaan mana seente
- Oromiyaan boorsaa fudhatte
- Oromiyaan baate

And also, the postprocessor performs the task of rearranging the order of transformed sentences in SOV form because in a compound sentence, sometimes the subject of the followed clause is found in the preceded clause, which was not incorporated during the transformation phase. For example, in this sentence "Sorrettin/NNS nirafti/VV Caaltuun/NNS ammoo/SCC niq'u'atti/VV" Both "Sorrettin/NNS" and "Caaltuun/NNS" are subjects that are found before the conjunction

"ammoo/SCC" and both subjects are split into the first clause during the transformation phase as:

- “Sorrettin/NNS nirafti/VV Caaltuun/NNS” and
- “ammoo/SCC niqu’atti/VV”

However, the postprocessor takes the appropriate subject for the remaining clause from the clause that contains the subjects and adds it to the remaining clause that has no subjects to make a self-contained sentence as:

- “Sorrettin/NNS nirafti/VV” and
- “Caaltuun/NNS ammoo/SCC niqu’atti/VV”

Algorithm 4.3 shows the implementation of the postprocessor phase of sentence simplification that rearranges the simplified compound sentences to obtain correct and independent simple sentences.

Algorithm 4.3 Sentence Rearrangement Algorithm

Require: *POS tagged Afan Oromo clauses* : Clause_List

Ensure: *List of complete and self-contained simple sentences*

for *each clause C in Clause_List* **do**

Check Subject of C

if *C has no Subject* **then**

Set Subject by the subject found before the verb of the preceded clause
 end if

if *Verb in C is tagged as VVC* **then**

Remove suffix from the verb
 end if

Output list of Simple Sentences

end for

The following table 4.3 shows examples of the overall correct outputs (transformation and postprocessor phases) of the compound sentence simplification algorithms.

Table 4.3 Example of Simplified Compound Sentences.

SN.	Original sentences	Simplified sentences
1.	Abdiin/NNS Boontuu/NNO waamaa/VV jira/AX garuu/SCC isheen/PPS mana/AD hin/NG jirtu/VV	1. Abdiin Boontuu waamaa jira. 2. Isheen mana hin jirtu.
2.	Kitaabicha/NNS qarshii/NN dhibbaan/JN bitee/VVC hiriyaasaf/NNO kenne/VV garruu/SCC maalgodha/PR innuu/PP hin/NG dubbisuu/VV	1. Kitaabicha qarshii dhibbaan bite 2. Kitaabicha hiriyaasaf kenne 3. Kitaabicha maalgodha innuu hin dubbisuu
3.	Oromiyaan/NNS mana/NNO seentee/VVC boorsaa/NNO fudhattee/VVC baate/VV	1. Oromiyaan mana seente 2. Oromiyaan boorsaa fudhatte 3. Oromiyaan baate
4.	Tulluun/NNS hojii/NNO manaa/NN boontuuf/PP hojjate/VV garuu/SCC isheen/PPS sirrii/JJ miti/NG jettee/VVC mana-barumassa/NNO irraa/PR hafte/VV	1. Tulluun hojii manaa boontuuf hojate 2. Isheen sirrii miti jette 3. Isheen mana-barumassa irraa hafte
5.	Jaalanneen/NNS kaleessa/AD mana/NNO barumsarraa/JJ hafte/VV ta'ullee/SCC hojii/NNO manaa/JJ hojjatee/VV jirti/AV	1. Jaalanneen kaleessa mana barumsarraa hafte. 2. Jaalanneen hojii manaa hojjatee jirti
6.	Sorrettin/NNS nirafti/VV Caaltuun/NNS ammoo/SCC niq'u'atti/VV	1. Sorrettin nirafti 2. Caaltuun niq'u'atti
7.	Abdiin/NNS hin/NG baranne/VV garruu/SCC waan/PP baay'ee/JJ beeka/VV	1. Abdiin hin baranne 2. Abdiin waan baay'ee beeka
8.	Isheen/PPS qall'oo/JJ dha/VV haata'uuyyumalee/SCC jabduu/JJ dha/VV	1. Isheen qall'oo dha 2. Isheen jabduu dha

CHAPTER FIVE: EVALUATION RESULT AND DISCUSSION

The evaluation of automatic sentence simplification systems is not straightforward, it needs techniques and methods. Hence, this chapter deals with details about data preparation and annotation, as well as evaluation techniques based on the method and algorithms developed in chapter four. Furthermore, the results and analysis conducted are discussed.

5.1 Data Preparation and Developmental Environment

5.1.1 Data Preparation

In chapter four, we have discussed all the Afan Oromo sentences required for the study in detail, both as purpose based (declarative, interrogative, imperative, and exclamatory) and structural based (simple, compound, complex, and compound-complex) types of sentences. However, the study focused on the declarative types of sentences both for sentence identification and simplification. As a result, the study required all types of Afan Oromo declarative sentences, which are simple, compound, complex, and compound-complex sentences, but there were no available appropriate and annotated sentences that were appropriate for the study. Thus, the only solution is to start the preparation of the data from scratch, including the collection and construction of the sentences according to their types, organizing and annotating data (part of speech tagging) and error corrections.

Indeed, the required types of datasets were prepared by researchers with the help of experts and it is collected from different resources such as Afan Oromo textbooks and documents developed by different scholars because the sentences collected from them are grammatically correct when compared to other resources and they also seem to have common understandability for all users. Because of the public unavailability of the Afan Oromo POS tagger, the researcher manually annotated all data using a tag set which was developed by different researchers [27, 28] with a dependency information tag set that was newly added to control the sentence complexity, as it is described under chapter four in the 4.1 table. The researcher annotated data with the assistance of the language's linguistics as it seems to contribute to the study in the preparation of a standard Afan Oromo sentence corpus. However, it took a long time because reviewing different literature, undermining the sentence structure, and bringing it into the study were the necessary steps, as the study's approach was based on the knowledge of the language.

Generally, in this study, the dataset used for evaluation consists of 480 POS-tagged African declarative sentences that consist of 200 simple sentences, 120 compound sentences, 80 complex sentences, and 80 compound-complex sentences.

5.1.2 Development of the Environment

To implement and test the proposed system, we have used a Toshiba laptop with an Intel Core i3 CPU at 2.5 GHz speed, 4 GB of RAM, a 500 GB hard disk, and the Windows 10 operating system. Draw.io to draw the diagram, Jupiter Notebook editor with Python 3.8 programming language to edit code, NLTK for implementation purposes, and MS office 2019 to write the document.

5.2 Evaluation of the Result

In order to evaluate the performance of text simplification systems and compare it to the performance of similar systems or human judgments, several evaluation metrics have been defined. The most popular is the confusion metric, which consists of precision, recall, and F-measure. Precision and recall are two generally used mathematical classifications. In a sentence identification and simplification scenario, the result is classified as correctly and incorrectly identified and simplified sentences. Therefore, precision is defined as the number of relevant sentences recognized by the system divided by the total number of sentences recognized by the system, and Recall is defined as the number of relevant sentences recognized by the system divided by the total number of existing relevant sentences which should have been identified and simplified.

Where precision is the percentage of correctly identified and simplified sentences found by the system. It can be expressed that:

$$\text{Precision} = \frac{\text{Number of correctly simplified sentence found by the system}}{\text{Number of simplified sentences found by the system}}$$

and recall is the percentage of simple sentences existing in the data and which were found by the system. It can be expressed that:

$$\text{Recall} = \frac{\text{Number of simplified sentences found by the system}}{\text{Total number of simple sentences in the data}}$$

A mixed metric exists, F-score, defined that:

$$F\text{-measure} = \frac{2(P * R)}{P + R}$$

In this study, the evaluation briefly focuses on two tasks. The first task is the evaluation of sentence identification and separation, and the other task is the overall evaluation of compound sentence simplification. To evaluate both the identification and simplification of the sentence, the researcher classifies the output of the system into correctly and incorrectly identified and simplified sentences by comparing them with the gold standard produced by the linguistic model. The experiment results are described in terms of precision and recall.

5.2.1 Evaluation of Sentence Identification

The evaluation of the sentence identification algorithm is performed separately because the performance of the sentence identification algorithm has a significant effect on the accuracy of the sentence simplification. The sentence identification algorithm is tested using 480 declarative types of Afan Oromo sentences. From the provided sentences, 200 sentences are simple sentences and the algorithm identifies 189 sentences correctly; 120 sentences are compound sentences and 107 sentences are correctly identified; 80 sentences are complex and compound-complex sentences and the algorithm correctly identifies 68 and 67 sentences respectively. In general, the sentence identification algorithm was tested with 480 declarative types of Afan Oromo sentences and it achieved 90% of the overall accuracy in terms of F-score which is the promised result. The following 5.1 table shows the summary evaluation results of the declarative Afan Oromo sentence identification and separation algorithm.

Table 5.1 Evaluation Result of Sentence Identification

Sr. No .	Corpus Type	Total number of sentences	Correctly identified sentences	In-correctly identified sentences	Not Identified	Precision %	Recall %	F-score %
1	Simple Sentences	200	189	7	4	96	94.5	95
2	Compound sentences	120	107	8	5	93	89	91
3	Complex Sentences	80	68	7	5	90	85	87
4	Compound -Complex Sentences	80	67	8	5	89	84	86

Overall results	92	88	90
------------------------	-----------	-----------	-----------

5.2.2 Evaluation of Compound Sentence Simplification

The evaluation of compound sentence simplification shows the overall performance of the Afan Oromo compound sentence simplification algorithm in terms of precision and recall. It is the combined outcome of the sentence transformation and rearranging (postprocessor) of the system phases. To determine the performance, the system outputs or simplified sentences are classified into correct and incorrect sentence by human judgments. The compound sentence simplification algorithm is tested using 120 compound sentences that contain 268 simple sentences. The result of the evaluation showed that out of 268 simple sentences in the dataset, 223 of them were correctly simplified. The algorithm achieved a performance of 84.4% f-score as summarized in the 5.1 table.

Table 5.2 Evaluation Result of Compound Sentence Simplification

Total number of compound sentences	Number of simple sentences in corpus	Correctly simplified sentences	In-correctly Simplified sentences	Not Simplified	Precision %	Recall %	F-Score
120	268	223	38	7	85.44	83.21	84.4 %

5.3 Discussion

In this section, the basic concepts of the study are discussed in detail. The discussion includes what tasks are performed and how they are performed; the results and evaluation processes; challenges of performing the tasks; and, meanwhile, the consideration shows whether the investigation answered the provided research questions or not.

The study is focused on syntactic or structurally Afan Oromo sentence simplification and identification by using the handcrafted rule method. The rule is coined based on the identified constituent parts of the sentences at the word level based on the language's parts of speech. The identified constituent parts of the sentences are the types and number of main verbs, coordinated and subordinate conjunctions, punctuation marks (comma and semicolon), and some morphological features like the long sound {e} and {i} with main verbs. We developed iterative types of algorithms that can monitor all features of a given sentence to identify and simplify

types of sentences according to the syntactic structure of Afan Oromo sentences. Before sentence simplification, the types of sentences are identified and separated because sentence identification plays a great role in improving the accuracy of sentence simplification. After the sentences are correctly identified and separated, the sentence simplification algorithm simplifies compound sentences into clauses by determining the bounding box of every clause in the sentence. Finally, the split or determined clauses are rearranged to make them complete and self-contained simple sentences.

The performance of the sentence's simplification and identification are determined separately. As shown in table 5.1, the overall performance of sentence identification is 90% F-score. However, In the sentence's identification, as the types of sentence structure become more complex, i.e., from a simple sentence to a compound-complex sentence, the accuracy of sentence identification decreases. From the types of identified Afan Oromo declarative sentences, only a compound sentence is focused on being simplified due to resources and time as the area is a large and complex concept. As shown in table 5.2, the accuracy of Afan Oromo compound sentence simplification is 84.4% f-score, which is good but it was evaluated by small size data.

As for the limitations of our system, in particular, the found result (a simplified sentence) sometimes has a limit in sentence completeness, coherence, and semantic problems when compared with the original sentence because of several factors such as incorporated and incompleteness of resources like POST, phrase chunking and coreference resolution, malformed sentences. For instance, POS tagging is mandatory to identify constituent parts of the sentences, and it has been investigated by many researchers but still not completed, particularly in the area of conjunction types. They are considered as a removal word or tagged as a single set, but they play a lion's share in the sentence simplification to identify sentences and the boundary of clauses to split or simplify the sentences. Thus, as showed in table 4.1 we added about eight new tag sets to an already developed tag sets somewhat to control the structure of the sentence but we couldn't able to the whole control of the noun phrase replacement at rearrangement of the transformed sentences so it needs the incorporation of phrase analyzer too. The other factor is that of the method used for implementation. A rule-based method is used due to resource scarcity and flexibility to control sentence structure. However, it takes a long time to come up with a rule for each and every sentence. So, limited coverage of the rules used to rewrite sentences and an

inability to discriminate between various subtypes of clause coordination have a great impact on sentence simplification performance.

The evaluation of the results was somewhat difficult due to the lack of common and standard criteria that state if such and such conditions are met. If so, the simplification is correct; otherwise, the simplification is incorrect because of naturally occurring language ambiguity. However, the simplified sentence is tried to be classified into correct and incorrect with the help of the language's linguistics to determine the system's effectiveness by comparing the system's output to the gold standard as determined by the language expert. Furthermore, it was tough to find specialists who could evaluate the correctness of the simplified sentences. Another difficulty with sentence simplification is the intended audience; it tends to mean that the simplified sentence can be used for different people with different problems such as social classes that have a problem with long sentences' understandability (like aphasic) and language learners (children and foreigners). Moreover, a simplified sentence is utilized in different NLP applications such as machine translations, question generation, sentence parsing, information extraction, and sentence-based text summarization are some NLP applications that access a simplified sentence for initiation and/or better performance, however both social classes, and NLP applications may not have simple sentence availability in common rather than subjective.

The other main challenge outside the correspondent study is that Afan Oromo has suffered from computational infancy because there are no deep features available to the research community except for academic exercises. In fact, it takes much time trying to understand and bring each and every structure of the sentences to the study rather than what should be expected from the researcher in the subjective work. Furthermore, the selected area of text simplification is very young. In particular, sentence simplification is new for the Afan Oromo language, which needs so many investigations into different aspects and approaches that is seriously recommended for future work.

CHAPTER SIX: CONCLUSION AND FUTURE WORK

This chapter focuses on summaries that indicate the whole picture of the study in the conclusion, based on the findings of the experiment and recommendations that the researcher has suggested for future work.

6.1 Conclusion

Text simplification is a completely critical NLP application which changes the content and structure of a textual content to enhance clarity and understandability for distinct social instructions and to decrease complexity to make it a less complicated method for downstream NLP programs and tools at the same time as maintaining its fundamental concept and relativity to its original meaning. Fundamentally, textual content simplification may be a syntactic simplification, which simplifies the complexity of sentences structurally on the sentence degree, and lexical simplification replaces a strong word with an easy one on the word degree or the combination of each simplification.

The primary objective of the study is the syntactic Afan Oromo sentence identification and simplification by using rules-based approaches and it was accomplished by governing the constituent parts of the language's sentences at word level based on part of speech tagged. The work presented in the study can be divided into two major tasks. The first and most important task in this study is to simplify syntactic complex sentences into simple and self-contained sentences by preserving the meaning as much of the original sentence as possible. From the types of structurally complex Afan Oromo sentences, only compound sentence simplification was performed because of its very broad area and also resources and time limitations. In the other task, the types of sentences are identified and separated as simple, compound, complex and compound-complex to improve the achievement of sentence simplification.

The sentence simplification and identification performance are separately determined and assessed manually by comparing the system's output to the golden standard assessed by the language expert. The overall performance of sentence identification is 90% F-score and compound sentence simplification is also 84.4% F-score. The obtained result is a promised result as it is the groundwork of less resource-intensive study. However, from the results, we can conclude that sometimes simplified sentence has a limit in sentence completeness, coherence,

and semantic problems when compared with the original sentence particularly due to unincorporating of phases and semantic analyzer.

Simple sentences are used for different natural language processing applications to enhance the achievement of their functions particularly in machine translation, question generation, sentence paring, information extraction, semantic role labeling, sentence-based text summarization, opinion mining and other language processing applications like query processing, and speech processing. Besides naturally occurring language processing, the high coverage of sentence simplification is utilized by different social classes that have different problems, like complex and long-sentence understandability problems (aphasic), low-literacy (children) and adults learning a language (non-native speakers).

Our Contributions

This work has several contributions to Afan Oromo language processing. We have contributed:

- ✚ An annotated dataset of 480 Afan Oromo declarative sentences.
- ✚ A design model procedure for automatically identifying and simplifying sentences
- ✚ An Afan Oromo sentence identification and separation algorithm.
- ✚ The Afan Oromo sentence simplification algorithm which breaks down compound sentences into simple parts and rearranges them to produce simple and self-contained sentences.

6.2 Recommendation and future work

As discussed above, the investigation of text simplification for Afan Oromo is in its infant stage. Hence, there are several areas of research for Afan Oromo and other Ethiopian languages, in particular, that should be highly recommended for future researchers in the area of sentence identificiation and simplification. The following categories of recommendation and future work are made based on the findings and limitations of the current work:

- ✚ Sentence identification and simplification is more advanced and higher NLP application that incorporates several applications to be efficient and complete applications. The scarcity and incompleteness of resources and tools such as speech tagging, clause boulder detectors, phrase chunking (noun phrase), coreference resolution and the preparation of standard sentence corpora have clearly impacted the performance of this work. Particularly, the found result (simplified sentence) has a limit in sentence completeness, coherence and semantic problems. Thus, it is necessary to consider all of them to improve the quality and

performance of sentence simplification. This is due to the reality that sentence simplification is used in many NLP applications and social classes, so the best results are required to further robust the system.

- ✚ Regarding the coverage of the Afan Oromo sentence simplification area, only syntactical compound sentence simplification was conducted and small dataset was used for evaluation. Therefore, another research that investigates the generic Afan Oromo sentences' simplification using a large dataset is desirable and the sentence simplification is not only limited to syntactical simplification but also lexical simplification, or integration of both approaches together which propels the study forward. Syntactic simplification is structurally simplifying a complex sentence into a simple sentence using various techniques, whereas Lexical simplification is identifying strong words and replacing them with common one by scanning a text and picking out the words.
- ✚ The other future work is focused on the approach of implementing sentence simplification and identification. Regarding this, there are two basic approaches, rule-based and driven-based approaches or hybrid. In this study, the rule-based approaches were used to simplify sentences and the simplification of compound sentences due to the scarcity of resources, and the promised result was achieved. However, it takes a long time and is difficult to finish the rule of every natural language, so we highly recommend other researchers conduct the study by using data-driven approaches such as Memory-Augmented Neural Networks and Deep Reinforcement Learning with large corpora.
- ✚ In this study, the evaluation of both sentence identification and simplification was assessed manually by comparing the output with the golden standard produced by the language experts who were difficult to compute for precision, recall, and f-measure. Additionally, it is difficult to set the evaluation criteria and they vary from person to person. Therefore, the researcher recommends a tool that can automatically evaluate the sentence simplification system.

References

- [1] M. M. P. Das, "A Novel System for Generating Simple Sentences from Complex and Compound Sentences," *International Journal of Modern Education and Computer Scienc*, pp. 1, 57-64, 2018.
- [2] M. Shardlow, "A Survey of Automated Text Simplification," (*IJACSA*) *International Journal of Advanced Computer Science and Applications, Special Issue on Natural Language Processing*, pp. 56-70, 2014.
- [3] R. N. S. K. S. Chandni, "Identification and Separation of Simple, Compound and Complex Sentences in Punjabi Language," *International Journal of Computer Applications & Information Technology*, pp. 123-128, 2014.
- [4] Association for Computational Linguistics, Carolina Scarton, Lucia Specia, "Data-Driven Sentence Simplification: Survey and Benchmark," *Association for Computational Linguistics*, vol. Volume 46, no. Computational Linguistics, pp. 136-183, 2019.
- [5] R. R. Finegan-Dollak, "Sentence Simplification, Compression, and Disaggregation for Summarization of Sophisticated Documents," *JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY*, ••(••);, pp. 2437-2453, 2016.
- [6] O. A. A. R. Elier Sulem, "Simple and Effective Text Simplification Using Semantic and Neural Methods," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, p. 162–173, 2018.
- [7] RICHARD EVANS and CONSTANTIN ORASAN ~, "Identifying signs of syntactic complexity for rule-based sentence simplification*," p. 69–119, 17 september 2018.
- [8] S. H. Paetzold, "A Survey on Lexical Simplification," *Journal of Artificial Intelligence Research* 60, pp. 549-593, 2017.
- [9] L. Zhang, "Sentence Simplification with Deep Reinforcement Learning," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 7–11, 584–594 September 2017.
- [10] Sai Surya† Abhijit Mishra‡ Anirban Laha‡ Parag Jain‡ Karthik Sankaranarayanan, "Unsupervised Neural Text Simplification," *IBM researchh*, 21 Agu 2019.
- [11] B. H. T. M. Y. Tu Vu, "Sentence Simplification with Memory-Augmented Neural Networks," *Proceedings of NAACL-HLT*, pp. 76-85, 1-6 June 2018.

- [12] D. B.-L. L. F. Brouwers, "Syntactic Sentence Simplification for French," *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014*, p. 47–56, 26-30 April 2014.
- [13] S. Sharma, "Sentence Reduction for Syntactic Analysis of Compound Sentences in Punjabi Language," *EAI Endorsed Transactions on*, 30 January 2019.
- [14] D. B. A.-L. L. F. Brouwers, "Syntactic Sentence Simplification for French," *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014*, pp. 47–56,, april 26-30 2014.
- [15] Navneet Kaur, Kamaldeep Garg, Sanjeev Kumar Sharma, "Identification and Separation of Complex Sentences from Punjabi Language," *International Journal of Computer Applications (0975 – 8887)*, Vols. Volume 69– No.13,, pp. 22-24, 2013.
- [16] Poornima C, Dhanalakshmi V, Anand Kumar M, Soman K P, "Rule based Sentence Simplification for English to Tamil Machine Translation System," *International Journal of Computer Applications (0975 – 8887) Volume 25– No.8*, pp. 38-42, 2011.
- [17] Heilman, Smith, "Extracting Simplified Statements for Factual Question Generation," 29 march 2010.
- [18] Yaoyuan Zhang, "A Constrained Sequence-to-Sequence Neural Model for Sentence Simplification," *arXiv:1704.02312v1 [cs.CL]*, 7 april 2017.
- [19] Y. Peng, "A Sentence Simplification System for Biomedical Text," *IEEE International Conference on Bioinformatics and Biomedicine*, pp. 211-216, 2012.
- [20] D. V. K. M. S. K. P. Poornima C, "Rule based Sentence Simplification for English to Tamil Machine Translation System," *International Journal of Computer Applications (0975 – 8887) Volume 25– No.8*, pp. 38-42, 2011.
- [21] A. S. Genemo, "AFAAN OROMO NAMED ENTITY RECOGNITION USING HYBRID APPROACH," ADDIS ABABA UNIVERSITY, SCHOOL OF GRADUATE STUDIES, COLLEGE OF NATURAL SCIENCES, DEPARTMENT OF COMPUTER SCIENCE, ADDIS ABABA, 2015.
- [22] G. H. DABALO, "AUTOMATIC SYNTACTIC PARSER FOR AFAAN OROMO COMPLEX SENTENCE USING CONTEXT FREE GRAMMAR," ADDIS ABABA UNIVERSITY, COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES, SCHOOL OF INFORMATION SCIENCE, ADDIS ABABA, 2016.

- [23] R. S. Y. M. T. Makoto Miwa, "Entity-Focused Sentence Simplification for Relation Extraction," *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 788-796, 2010.
- [24] Heilman, A. Smith, "Question Generation via Overgenerating transformations and Ranking," 2009.
- [25] A. S. Manisha Divat, "Automatic question generation approaches and evaluation techniques," Manisha Divat, Ambuja Salgaonkar, Mumbai, 2017.
- [26] LEMIN ZHANG AND HUIFANG DENG, "Sentence Simplification Based on Multi-Stage Encoder Model," *IEEEAccess*, Vols. VOLUME 7, 2019, 16 December 2019.
- [27] Zhemin Zhu, Delphine Bernhard, Iryna Gurevych, "A Monolingual Tree-based Translation Model for Sentence Simplification," *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, p. 1353–1361, 1253 2010.
- [28] R.M. Jindal^{1,*}, V. Rana¹ and S.K. Sharma², "Simplification of Punjabi Sentences: Converting Complex Participial Sentences into Simple Sentences," *EAI Endorsed Transactions on Scalable Information System*, pp. 1-11, 27 February 2020.
- [29] H. Beshada, "DESIGN AND DEVELOP SENTENCE PARSER FOR AFAN OROMO LANGUAGE USING TOP-DOWN CHART PARSING ALGORITHM," MSc.Thesis, BAHIR DAR INSTITUTE OF TECHNOLOGY, SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES, FACULTY OF COMPUTING, Bahir der, 2017.
- [30] A. S. Genemo, "AFAAN OROMO NAMED ENTITY RECOGNITION USING HYBRID APPROACH," ADDIS ABABA UNIVERSITY, SCHOOL OF GRADUATE STUDIES, COLLEGE OF NATURAL SCIENCES ,DEPARTMENT OF COMPUTER SCIENCE, ADDIS ABABA, 2015.
- [31] M.-H. ABUBEKER, "PART OF SPEECH TAGGER FOR AFAAN OROMO LANGUAGE USING TRANSFORMATIONAL ERROR DRIVEN LEARNING (TEL) APPROACH," ADDIS ABABA, Addis Abebe, 2010.
- [32] ABRAHAM, "MPROVING BRILL'S TAGGER LEXICAL AND TRANSFORMATION RULE FOR AFAAN OROMO LANGUAGE," AAU, ADDIS ABEBA, 2013.
- [33] K. Vickrey, "Sentence Simplification for Semantic Role Labeling," *Proceedings of ACL-08: HLT*, p. 344–352, 2008.

- [34] Z. Zhu and I. G. Delphine Bernhard, "A Monolingual Tree-based Translation Model for Sentence Simplification," *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, p. 1353–1361, August 2010.
- [35] Getachew Mamo, Million Meshesha, "Parts of Speech Tagging for Afaan Oromo," *International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*, pp. 1-5, 2011.
- [36] A. G. AYANA, "IMPROVING BRILL'S TAGGER LEXICAL AND TRANSFORMATION RULE FOR AFAAN OROMO LANGUAGE," AAU, ADDIS ABEBA, 2013.
- [37] D. K. Dan Feblowitz, "Sentence Simplification as Tree Transduction," *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages, pp. 1-10, 4-9 August 2013.
- [38] R. Chandrasekar and B. Srinivas, "Automatic induction of rules for text simplification. Knowledge-Based Systems," 1997.
- [39] A. Siddharthan, "Syntactic simplification and text cohesion," University of Cambridge, United Kingdom, 2004.
- [40] A. v. d. B. E. K. Sander Wubben, "Sentence Simplification by Monolingual Machine Translation," *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, p. 1015–1024, 8-14 July 2012.

Appendixes

Appendix A: Sample of annotated Data

Badhaasaan/NN harreetti/NP okaa/NN fe'e/VV
Mataa/NN olnu/PS qabachiise/VV
Boontuun/NN gara/PR Finfinnee/NN deemte/VV
Bilisumaan/NN kalleessa/AD garba/NN guddaa/JJ daake/VV
Barattooni/NN iyaata/NN kalleessaa/AD dhiisanii/VV jiru/AX
kana/PP dhoksaadhaan/AD hojecha/VV ture/AX
Inni/PP facaafne/JJ nuuf/PS biqile/VV
Galmeen/NN barbaachisoon/JJ badanii/VV jiru/AX
Galama/NN kanbiraan/PR kiisii/NN xarapheezaa/JJ keessaa/AD arge/VV
Tolaan/NN mana/AD jira/VV
Inni/PP konkolaataa/NN irra/PR jira/VV
Inni/PP ariitin/AD figa/VV
Calaan/NN baay'ee/AD cimaa/JJ dha/VV
caaltuun/NN gara/PR mana/NN barumsaa/JJ deemte/VV
Toolaan/NNP konkolaataa/NN keessa/PR jira/VV
Bariisoon/NN konkolaataa/NN lama/JN qaba/VV
Roobeenn/NNP kutaa/NN isiitii/PS tokkoffaa/ON baate/VV
Baratootni/NNP gara/PR mana/NN barumsaa/JJ deeman/VV
Ani/PP kitaaba/NN bite/VV
Mucayyoon/NN barattuu/NN cimtuu/JJ dha/VV
Toolan/NN leenca/NN ajjese/VV
Biqilaan/NN boru/AD dhufa/VV
yoosan/NN mana/NN Magarsaa/JJ deeme/VV
Abbabaan/NNP kitaabicha/NN Caalaaf/NP kenne/VV
Jaalanneen/NN kaleessa/AD mana/NN barumsarra/JJ hafte/VV garuu/SCC hojii/NN mana/JJ hojjatee/VV jirti/AV
Abdiin/NN hin/NG baranne/VV garuu/SCC waan/PP baay'ee/JJ beeka/VV
Isheen/PP qall'oo/JJ dha/VV haata'uuyyuumalee/SCC jabduu/JJ dha/VV
Sorrettin/NP nirafti/VV Caaltuun/NN ammoo/SCC niq'u'atti/VV
Abdiin/NN Boontuu/NN waamaa/VV jira/AX garuu/SCC boontuun/NN mana/AD hin/NG jirtu/VV

Rabbiin/NN galanni/NN haa/PR ga'u/VV ./PNC jalqaba/PR bara/AD kanaa/PP irratti/PR tokkummaa/NN tokko/JN argannee/VV jirra/
 AX
 Galmoota/NN wallitti/PR fuunaanee/VVC kenne/VV
 Shamarree/NN Daai'amaa/JJ waggaa/JN 14/JJ haadha/NN jalaa/PR fuudhaanii/VVC waraabeesaa/NN darban/VV
 Namni/NN mana/NN ijaareefi/VVC namni/NN barumsa/NN barate/VV gaariidha/VV
 Oromiyaan/NN saba/NN baa'ee/JJ qabdi/VV akkasumas/SCC qileensaa/NN gaarii/JJ qabdi/VV
 Nyaatan/NN barbaada/VV sababiinsa/SCC nan/PP beela'e/VV
 Ani/PP kochee/NN nyaadhe/VV kanaafuu/SCC garaa/NN kaasan/JJ qaba/VV
 Israa'eelii/NN fi/CC Liibaanoos/NN wal/PP waraansa/JJ eegalan/VVC cimsanii/AD itti/PR fufaniiru/VV
 Bala/NN lolaa/JJ dirree/NN dawaa/JJ midhaa/NN gudaa/JJ qaqaabsiise/VV haata'uuyyuumalee/SCC ummani/NN qaamolee/NN desgra
 saa/JJ addadaattin/JP baraaramanii/VV jiru/AX
 Ibbiddi/NN godinaalee/NN sadi/JN keessatti/AD ka'ee/VVC ture/AX too'atame/VV
 shaampiyoonni/NN atileetiksii/NN afrikaa/JJ sportii/NN baayee//AD guddaadha/VV akkasumas/SCC baayee/AD nama/NN hawwwata
 /VV
 Wayta/SSC inni/PP dhufu/VV ./PNC ani/PP barressaa/VV jira/AX
 Baratotni/NN hundinu/JJ qoruumsa/NN darbuuf/VC sirritti/AD qo'atan/VV
 Yoo/SSC haalaan/JJ hojjetta/VV ./PNC qormaata/NN ni/PR dabarta/VV
 Yommuu/SSC gabaa/NN dhaqxu/VV na/PP waami/VV
 Akka/SSC argite/VV hin/NG dubatiin/VV
 Yoo/SSC sooromtes/VC ./PNC yoo/SSC hiyyoomtes/VV gorsa/NN abbaa/NN kee/PP hin/NG dagatin/VV
 Yoo/SSC bokkaan/NN robe/VV malee/SSC margi/NN hin/NG margu/VV
 Waan/SSC isheen/NN hamtuu/JJ taateef/VV namni/NN ishee/NN hin/NG jaalatu/VV
 Busaa/NN ittisuuf/VC sochiin/NN cimaan/JJ taasifamaa/VV jira/AX
 yommuu/SSC nyaata/NN nyaatu/VV hundaa/AD dheeressaan/NN nisirba/VV
 erga/SSC nyaata/VV booda/PR ./PNC kummarran/NN kutaa/NN isaa/PP kessatti/AD taphata/VV
 yoo/SSC isiin/PP sirbite/VV anillee/PR nisirba/VV
 isheen/PP waan/PR sirritti/AD hojatuuf/VC huqatte/VV
 waan/SSC barfadheef/VV baayee/AD aare/VV
 yoo/SSC ijjooleen/NN gitaara/NN rukutan/VV maatiin/NN hin/NG jaalatan/VV
 Yommuu/SSC dhaqes/VV, yommuu/SSC gales/VV natti/PS goree/VVC na/PP gaafatee/VVC darbe/VV
 Gaaddisaan/NN yommuu/SSC kutaa/NN isaa/PP keessatti/AD spoortii/NN hojatu/VV nisirba/VV akkasumas/SCC caaltuun/NN yemm
 uu/SSC figdu/VV nisirbiti/VV.
 yommuu/SSC inni/PP kolfu/VV isiin/PP nibootti/VV garuu/SCC taphataa/VV jiru/AX
 Yoo/SSC ijjoollen/NN kubbaa/NN taphatan/VV ./PNC yoo/SSC ijjooleen/NN gitaara/NN rukutan/VV maatiin/NN hin/NG jaalatan/VV
 akkasumas/SCC warri/NN olaaa/JJ niaaru/VV
 sirritti/AD hojachuun/VV barbaachisaadha/JJ yookiin/SCC barnoota/NN sirritti/AD barachuudha/VV yoo/SSC milkaayuu/VV feete/A
 X
 osoo/SSC ijjoollen/NN hin/NG galin/VV ./PNC maatiin/NN nyaata/NN qoppeesan/VV garuu/SCC manatti/ hin/NG eegne/VV
 Yoo/SSC humna/NN cimaa/JJ qabaachuu/VV baatan/AX ./PNC waan/PP barbaaddan/JJ sana/PP argachuu/VV fi/SCC bakkee/NN barb
 aadan/JJ sana/PP ga'uun/VV baay'ee/AD rakkisaa/AD ta'a/AX
 Kun/PP olola/NN oofuu/VV otuu/SSC hin/NG taane/VV ./PNC waan/PP irra/PR deddeebi'amee/AD dhaga'ame/VV argame/VV ./PN
 C dhugaa/NN lafa/JJ jiruu/VV dha/AX