ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL SCIENCE

PDTB STYLE SENTENCE LEVEL SHALLOW DISCOURSE PARSER FOR AMHARIC

Robel Arega Asfaw

A THESIS SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE N PARTIAL FULFILLMENT FOR THE DEGREE OF MASTERS OF SCIENCE IN COMPUTER SCIENCE

Addis Ababa, Ethiopia

October, 2020

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL SCIENCE

Robel Arega Asfaw

Advisor: Mulugets Libsie (PhD)

This is to certify that the thesis prepared by *Robel Arega Asfaw*, entitled: *PDTB Style Sentence Level Shallow Discourse Parser for Amharic* and submitted in partial fulfilment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee

Mulgeta Libsie (PhD)	_____	_____
Advisor	Signature	Date
_____	_____	_____
Examiner	Signature	Date
_____	_____	_____
Examiner	Signature	Date

Table of Contents

Acknowledgement i
List of Tables ii
List of Figures iii
List of Acronyms v
Abstract vi
Chapter 1: Introduction 1
1.1 Introduction 1
1.2 Motivation 2
1.3 Statement of the Problem 3
1.4 Objectives 3
1.5 Methodology 4
1.6 Scope of the Study 5
1.7 Limitations of the Study 5
1.8 Application of Results 5
1.9 Organization of the Rest of the Thesis 6
Chapter 2: Literature review 7
2.1 Discourse Overview 7
2.1.1 Discourse Coherence and Structure 7
2.1.2 Discourse Relations 8
2.1.3 Discourse Connectives 10
2.2 Theories of Discourse structure 11
2.2.1 Rhetorical Structure Theory 11
2.2.2 The Linguistic Discourse Model (LDM) 15
2.2.3 Discourse Graph Bank Theory 15
2.2.4 Intentional Discourse Model 16
2.2.5 Segmented Discourse Representation Theory (SDRT) 17
2.2.6 Discourse Lexicalized Tree-Adjoining Grammar (D-LTAG) 17
2.3 Manual Discourse Annotation 17
2.3.1 The RST Discourse Treebank 17
2.3.2 The Penn Discourse Tree Bank (PDTB) 18

2.3.3 Dependency Tree Banks 19
2.4 Automatic Discourse Parsing 19
2.5 Amharic Discourse 21
2.6 Summary 23
Chapter 3: Related Works 24
3.1 RST Style Discourse Parser 24
3.2 PDTB Style Discourse Parser 27
3.3 Summary 28
Chapter 4: Design of Discourse Parser for Amharic 29
4.1 Architecure of the System 29
4.1.1 Feature Extraction 29
4.1.2 Discourse Connective Identification 34
4.1.3 Discourse Argument Labeling 36
4.1.4 Discourse Sense Classifcation 38
4.2 Summary 40
Chapter 5: Experiement 41
5.1 Corpus Preparation 41
5.2 Analysis Results 43
5.3 Implementation 47
5.4 Evaluation 48
5.4.1 Evaluation Results of Discourse Connective detection 48
5.4.2 Evaluation Results of Argument Labeling 49
5.4.3 Evaluation Results of Sense Classification 49
5.5 Discussion 50
Chapter 6: Conclusion and Future Work 54
6.1 Conclusion 54
6.2 Recommendation 55
References 56
Appendix A 62
Appendix B 64

Acknowledgement

First of all, thanks to the Almighty God for giving me the strength to start and finalize this study. I am also very grateful to my advisor, Dr. Mulugeta Libsie for his constructive comments, support and patience during my work.

Next I would like to thank Dr. Getachew Endalamaw from Amharic Language, Literature and Folklore department in Addis Ababa University. He has provided unreserved support from the start to end of this thesis work. My deepest gratitude also goes to Dr. Fekade Getahun. He has been a great source of encouragement to complete this study. Moreover, I am very grateful for all staffs at Addis Ababa University, Department of Computer Science.

Finally, I would like to thank my colleagues Selam Tesfaye and Hilina Mesfin and my younger brother yaredo for their assistance and advice.

List of Tables

Table 2-1 Canonical Forms of Discourse Connectives 11
Table 4-1 Features used for connective Identification..... 35
Table 4-2 Features used for Argument Labeling..... 36
Table 4-3 Features used for Sense classification..... 39
Table 5-1 Lists of potential discourse Markers and their distribution list in the corpus..... 42

List of Algorithms

Algorithm 4-1 Feature Extraction Algorithms 32
Algorithm 4-2 Discourse Connective Identification Algorithm 35
Algorithm 4-3 Discourse Argument Labeling Algorithm 37
Algorithm 4-4 Discourse Relation (Sense) Labeling algorithm 39

List of Figures

Figure 2-1 The Five Schemas of RST 14
Figure 2-2 RST Diagram for Example 5 14
Figure 4-1 Full Arcitecture of the Amharic Discourse Parser 30
Figure 4-2 Sample Screen shot of Extracted Features 33
Figure 4-3 Discourse Connective Identification and Argument Lableling Component 38
Figure 4-4 Sense Labeling Component 40
Figure 5-1 Classification of selected Amharic Discourse Markers 45
Figure 5-2 PDTB Style Shallow Amharic Discourse Parser Prototype 47
Figure 5-3 Connective Detection Result 48
Figure 5-4 Argument Labeling Result 49
Figure 5-5 Sense Classification with Logistic Regression 49
Figure 5-6 Sense Classification Result with Random Forest 50
Figure 5-7 Sense Classification with Naive Bayes 50
Figure 5-8 Screenshot of correctly parsed sentense with the sense 'ተለጣጥቆ ' 51
Figure 5-9 Screenshot of correctly parsed sentense with the sense 'ማጠቃለያ ' 51
Figure 5-10 Screenshot of correctly parsed sentense with the sense 'መለየት ' 51
Figure 5-11 Screenshot of correctly parsed sentense with the sense 'ተጨማሪ ' 52
Figure 5-12 Screenshot of correctly parsed sentense with the sense 'ማነፃፀር' 52
Figure 5-13 Screenshot of correctly parsed sentense with the sense 'ምክንያት-ውጤት' 52
Figure 5-14 Screenshot of Incorrectly Parsed sentense with the sense 'መለየት' 53
Figure 5-15Screenshot of Incorrectly Parsed sentense with the sense 'ተለጣጥቆ' 53

List of Acronyms

CRF – Conditional Random Fields

DC – Discourser Connective

DCU - Discourser Constituent Units

D-LTAG – Discourse Lexicalized Tree-Adjoining Grammar

EDU - Elementary discourse unit

LDM - Linguistic Discourse Model

NLP – Natural Language Processing

PDTB – Penn Discourser Tree Bank

PTB – Penn Tree Bank

RST – Rhetorical Structure Theory

SDRT - Segmented Discourse Representation Theory

SVM – Support Vector Machine

Abstract

Research on natural language processing applications (NLP) is a very important topic in our daily life, by enabling computers to understand human languages. Such researches has come a long way in foreign languages like English, Japanese, Chinese, Portuguese and Arabic. NLP applications such as include machine translation, question answering, knowledge extraction and information retrieval are some of the fruits of such researches.

Discourse parser is one of the main components that enables the realization of such NLP applications. For foreign languages like English and Arabic, many discourse parsers are developed in different approaches. However, in the case of Amharic, there are no works done, to the best of the researcher's knowledge, on Amharic discourse parser so far.

In this study, a Penn Discourse Tree Bank (PDTB) style sentence level shallow discourse parser for Amharic is developed. We have used machine-learning algorithms to accomplish the subtasks of discourse parsing. The algorithms utilize lexical and positional features of the discourse marker and related words for segmentation and identify associated discourse relation. The parser is tested on test sentences, which are extracted from different sources. Encouraging results are observed from the experiments performed,

Keywords: NLP, Discourse, Discourse Parser, Discourse Marker, Elementary discourse unit, sense, argument, machine learning

Chapter 1: Introduction

1.1 Introduction

In linguistics and computational linguistics, the term "Discourse" refers to a text that has coherent and structured group of sentences [1]. This implies that a text is not just a simple sequence of sentences and clauses, but rather a complex structure [2]. This structure is called discourse structure. It is a hierarchical structure of the text according to discourse relations [3]. Having a discourse structure of a text is very important for many natural language processing (NLP) applications like question answering, text summarization, information extraction and machine translation. For example, in text summarization if we recognize that a unit of text further elaborates an already stated fact, then it is possible to disregard the elaborated unit and generate a concise summary based on the stated fact only [4]. In the above example, the availability of discourse parser highly facilitates the recognition of the coherence relation "elaborate" and hence the summarization task.

Discourse parsing is the task of automatically building discourse structures [3]. Hernault *et al.* [6] mentions that a number of attempts have been made to create discourse parsers using the framework of Rhetorical Structure Theory (RST). RST is the most widely used theory of text organization that formally expresses the relationship between textual units. These non-overlapping text chunks are called elementary discourse units (EDU). The process of dividing a text into EDUs is called discourse segmentation and it is the preliminary task of discourse parsing. Once a text is divided into consecutive EDUs, the rhetorical or discourse relations between these EDUs are identified and labeled with the help of pre-defined set of rhetorical relations. Finally, the discourse parser produces a tree structure as a representation of how all the EDUs relate to each other [6]. Such kind of discourse parsing that demand the establishment of a global data structure like tree or graph is called deep discourse parsing [16].

More recently, other variant of discourse parsing, shallow discourse parsing (SDP), has gained a lot of attention. Shallow discourse parsing (SDP) targets to discover the local coherence relations within text without assuming any tree or graph structure between the relations, hence the name shallow. SDP started with the the release of Penn Discourse Tree Bank (PDTB2.0) in 2008 and lately attracted more attention since it has been proposed as shared task of CoNLL 2015 and

CoNLL 2016[56,57]. In PDTB style parsing, discourse relations are viewed as binary predicate-argument, where a connective act as a predicate that takes two text spans as its arguments. The span to which the connective is syntactically attached is called Arg2, while the other is called Arg1[52,53].

Mainly two kinds of approaches are used for constructing discourse parsers: statistical and rule based. Statistical approaches are used where a large amount of discourse annotated corpus is available for training the models. Soricut and Marcu [7] used statistical approaches to build their discourse parser. Rule based discourse approaches are a valuable alternative where such corpus is not available. Marcu [8] and LeThanh [9] used such approaches to prepare discourse parsers. A discourse parser for English, Japanese and Brazilian-Portuguese is available. However, to the best of our knowledge there is no Amharic discourse parser developed so far.

It is important to understand the difference and similarities between discourse parsing and syntactic parsing or sentence parsing or simply parsing. Both parsing techniques are similar in that both techniques take text input and produce a tree structure as output. However, syntactic parser breaks down a text into its component parts of speech with an explanation of the form, function, and syntactic relationship of each part according to a given grammar while discourse parser derives a structure that depicts the semantic relationship among parts of a text according to the semantic content of the text segments.

However, it has to be noted that this does not mean a full semantic understanding is needed to derive discourse structures. Marcu [2] argues that, despite this assumption among researchers, it is possible to automatically derive discourse structures even in the absence of full semantic analyzer. Consequently, the author presented a semantic free theoretical framework that is general enough to be applicable to naturally occurring texts and facilitate algorithmic approach to discourse analysis. Therefore, even if the issue of semantic understanding is not yet addressed in Amharic, we believe it is possible to develop a discourse parsing algorithm for Amharic.

1.2 Motivation

Currently, various researches are being conducted in the area of NLP applications like question answering, text summarization, machine translation, etc. for Amharic language. Such applications require the analysis and understanding of written language higher than sentence level. This requirement shows that research interest is moving towards discourse which is essentially a text

more than a single sentence.

In contrast, the area of Amharic discourse research is untouched. In fact, there is no research conducted so far, that we know of, regarding Amharic discourse parser or its subtasks. This problem is the main motive that inspired us to investigate the nature of Amharic discourse and possibilities of developing Amharic discourse parser.

1.3 Statement of the Problem

Discourse parsing or its subtasks i.e discourse segmentation and discourse relation labeling has been addressed for other languages like English [2,3,6,7,9,11,12], German [61], Chinese [60], Spanish [62], and Portuguese [10], Arabic [20], etc. according to the special characteristics of the languages. While the early attempts employ rule based techniques using orthographical features of cue phrase and punctuation marks [2,9], recent works are mainly based on data driven approaches and use lexical and syntactical features of the specific language [7,3,11,12,60,62,10,20]. However, to the best of our knowledge, there is no such tool developed for Amharic so far.

Therefore, a discourse parser that considers the special features of Amharic language has to be developed for Amharic. In this study, we will investigate discourse parsing techniques in other languages and try to develop the design of a PDTB style sentence level shallow discourse parser for Amharic. The problem is "Given one or two Amharic sentence along with POS tags of tokens and other features to identify the discourse connective, the arguments or elementary discourse units (EDU) and the discourse relation in between."

1.4 Objectives

General Objective

The general objective of this research is to design a PDTB style sentence level shallow discourse parser for Amharic.

Specific Objectives

- Study Amharic texts, text structures, identify Amharic discourse markers and associated discourse relations.

- Review related works in other languages.
- Prepare a corpus annotated with discourse units and discourse relaions.
- Design the full discourser parser for Amharic.
- Build and evaluate the performance the component of the parser

1.5 Methodology

Literature Review

In order to understand the research topic and to be familiar with the achievements of previous efforts a number of literatures has been studied, reviewed and analyzed. The literatures reviewed include journal articles, theses, conference papers and various websites related to the subject matter. Further review has also been conducted in order to identify language dependent properties of Amharic discourse.

Data Collection

In order to study Amharic discourse and evaluate the parser developed, Amharic text corpus is required. In this regard Amharic articles written on various issues has been used as data source for this research. Therefore, documents have been collected from various Amharic newspapers, websites, and blogs history books and Amharic Bible.

Tools and Techniques

In order to achieve our objectives, specific techniques and tools has been used. Primarily we followed data driven machne learning approach to tackle the research problem.

The python crfsuite and sklearn packages are used to apply multiple machine learning algorithms, specifically Conditional Random Fields (CRF), Logistic regression, Naïve Bayes and Random Forest classifiers. In addition, Python programming language with Visual Studio Code (VSC) integrated development environment (IDE) is used as developmental tool for the prototype.

1.6 Scope of the Study

In order to meet the objective in the given period and to have a clear understanding of the works that have to be done it is essential to specify the scope of the thesis. The scope of this study is to detect discourse connective, identify the two arguments arg1 and arg2 or elementary discourse units (edus) and the discourse relationship or sense between the edus or arguments found within a sentence (Intra- Sentential) or between two consective sentences (Inter- Sentential). In this study discourse relations only signaled by explicit discourse connectives or words having a discourse function due to various morphological inflections are considered.

1.7 Limitations of the Study

This study has the following limitations:

- Texts having multiple (more than two) elementary discourse units or argumets are not included in this study. All the sentences used for training and testing only have two elementary discourse units.
- This study follows a PDTB style shallow discourse parsing approach and do not assume building of a tree or graph structure
- The size of the corpus is very small. The corpus is prepared manually for the purpose of the work.
- Implicit relations that are not explicitly marked by a discourse connective or words having a discourse function due to various morphological inflections are not included in this study.

1.8 Application of Results

Developing a discourse parser is not an end by itself. Rather, it is a means to achieve effective and efficient NLP applications. As it has been mentioned in the first section of this proposal, discourse parser can be used in many applications. In essay scoring, discourse parser helps to evaluate the coherence of the essay. In question answering, it helps to answer a series of questions in context. Therefore, the development of Amharic discourse parser highly compliments the effort of developing Amharic NLP applications.

1.9 Organization of the Rest of the Thesis

The remaining part of the thesis is organized in to five chapters. Chapter 2 covers literature review in which different definitions and theories related to the thesis are presented. Chapter 3 covers related works done by other researchers in other languages. Chapter 4 deals with the describing the architecture and components of the proposed system. It presents the general architecture of the system with its basic components and the discussion of the components and their interaction in the system. Chapter 5 focuses on details of data preparation and analysis, experiments to evaluate the proposed system and the results obtained together with their explanations. Chapter 6 presents conclusion and future work recommendations for the improvement of the system.

Chapter 2: Literature review

2.1 Discourse Overview

According to [14], discourse usually refers to a form of written text or spoken language used to communicate ideas or beliefs to be recognized by the hearer/reader. The concept of discourse deals with three dimensions of [40]: (1) language use, (2) communication of beliefs, and (3) interaction in social situations. Given these dimensions, it is not surprising that several disciplines are involved in the study of discourse including linguistics, psychology (study of beliefs), social sciences (analysis of interaction in social situations), and computational linguistics (to enhance language technology).

Discourse is not just a random sequence of sentences and clauses; rather, it is a coherent, understandable text for the reader or the hearer. Discourse studies have tended to agree on the notion that discourse has a genre-based structure which formalizes how discourse is constituted; thus, the structure of academic writing/speech differs from that of story, political, or news texts. The structure is taking into account lexical items, grammatical and morphological features, and semantic and pragmatic features such as intention and attention of propositions and the relations between them. Consequently, discourse studies in computational linguistics attempt theoretically to specify the relationships between the discourse units in a way that can be applied empirically in language applications such as text generation, summarization, argument evaluation, machine translation, speech recognition, essay scoring and question answering systems [2][41][42][43].

2.1.1 Discourse Coherence and Structure

In the introduction section we have stated that discourse has coherence and structure. According to [14] there are two aspects in which a discourse can be coherent. The first is based on the content of discourse segments. In this approach a discourse is said to be coherent if there is frequent reference to the same group of entities through “semantic congruence”, “argument overlap” or “a pattern corresponding to stereotypical situations”. The second is based on relation that exists between two or more discourse segments. The discourse segments connected by the relation could be two or more clauses, paragraphs or even chapters. These relations are called coherence relation after Hobbs [15].

On the other hand, Webber stated that there are four bases of discourse structure [16]. The first

bases are topics. When discourse is structured by topics, it will be comprised of a set of entities with concise explanation of them. Webber suggests that expository texts found in textbooks and encyclopedias are typical examples of discourse with topic structure.

Secondly, functions can also be a basis for discourse structure. News reports are good examples of discourse structured by the functions served by its elements, i.e., by their role(s) in communication. Conventionally, news reports have inverted pyramid structure comprising of three elements: lead paragraph, body and tail. Each element describes a summary, detail and less important information about the event in the report, respectively.

The third basis, eventualities, describes events and states. Discourse can be structured by these descriptions and their spatio-temporal relations. Such structure can be found in narrative texts such as accident reports.

The fourth basis of discourse structures are discourse relations. Since our study is concerned with discourse structure based on these elements we will discuss them in detail in the next section.

2.1.2 Discourse Relations

Discourse relations have been studied for over the last three decades with various labels, i.e., coherence relation, rhetorical relation, and cohesion relation. In this research, we use the terms coherence relation and discourse relation interchangeably.

Hobbs [15] stated that many researchers have pointed out the existence of discourse relations in various terms and listed some of them without comprehensive theoretical basis. However, later on various studies formalized the concepts and categorized them into classes. Although, theories of discourse and text structure differ in formalizing the definition of discourse relations, they all agree that their essential purpose is to create logical or semantic association between discourse segments.

Discourse relations can be classified as explicit and implicit based on whether they are signaled by a discourse connective or not. According to Fraser [17] these discourse connectives can be conjunction (and, or, but), adverbs (because, instead, since) and prepositional phrases (in contrast). Some discourse connectives signaling explicitly various discourse relations in different location in sentences are shown in example 1 while example 2 shows a sentence with implicit " because" discourse connective.

Example 1

- Yared played tennis, **and** Hana read a book.
- **While** she is pregnant, Selam will not take a plane.
- We left late. **However**, we arrived home on time.
- We don't have to go. I will go **nevertheless**.

Example 2

> But a few funds have taken other defensive steps. Some have raised their cash positions to record levels. High cash position helps buffer a fund when the market falls.

In example 2, the discourse is still meaningful without the relation being signaled explicitly. Knott and Sanders [18] and Knott [19] argue that this is because "a discourse should be as informative as required but no more information than required." In this study discourse relations explicitly signaled with discourse connectives only will be dealt with.

On the other hand, on the basis of the message inferred from the discourse relations they are classified as intentional and informational relations [20].

Informational relations are semantic relations that relate different content or meaning of text segments. The segments could be propositions, facts, events or situation. For example, in example 3, the first sentence expresses an event, and the second expresses the writer's opinion while the last presents a fact that the color red indicates love. This discourse can be understood as that the second sentence reasons out the event in the first sentence and the third sentence elaborates the writer's opinion and conclusion in the second sentence that Daniel and Helen are in a love relationship [20].

Example 3

- Daniel gave Helen a red rose.
- He loves her so much.
- The red color often indicates love.

Intentional relations, as it can be inferred from the name, relate intentions. By that we mean a writer/speaker aims to influence the hearer/reader, for example to increase his/her belief in some proposition. In example 4 the writer tries to increase the belief of the reader in his assertion thereby creating a justification relation between the two segments [20].

Example 4

> Dr. Yohannes is serving a 7-year jail sentence for medical errors. Two nurses saw him mixing up drugs with names that sound alike.

2.1.3 Discourse Connectives

Similar to discourse relations a variety of terms are used in the literature for lexical elements that have identical function to that of discourse connectives. Fraser [17] defined discourse connectives as follows

> *"a class of lexical expressions drawn primarily from the syntactic classes of conjunctions, adverbs, and prepositional phrases. With certain exceptions, they signal a relationship between the interpretation of the segment they introduce, S2, and the prior segment, S1"* (S1 and S2 representing discourse segment 1 and 2 respectively)

Fraser also suggests from a study regarding the grammatical status of discourse connectives, that the syntactic categories conjunction, adverbial, and prepositional phrases are the ones most discourse connectives belong to. This is complementary with Asher [21] who categorized connectives in English into four main categories as follows.

- Coordinating conjunctions

  Examples: - "and", "or", "but"

  Associated discourse relation functions: Conjunction, Alternative and Contrast, respectively.

- Subordinating conjunctions

  Examples: - "because", "although" and "if"

  Associated discourse relation functions: Causal, contrast, and condition

- Adverbial connectives

  Examples:-"therefore", "then"

  Association discourse relation: causal and conditional respectively

- Prepositional phrases

  Examples:-"In contrast" and "as a result"

  Associated discourse relation function: "contrast" and "consequence"

With regard to this classification, Alsaif [20] argues that it could be generalized to other languages and the categories are not necessarily the only syntactic categories possible for connectives in all other languages.

The second issue addressed regarding discourse connectives is their position within sentences. As it is evident in the examples mentioned so far, discourse connectives can be found in different locations within a sentence. Fraser [17] suggested a range of canonical forms to formalize the position of discourse connective and its arguments A1 and A2. Discourse arguments are the two discourse segments a discourse relates. Most discourse processing studies consider arguments to be non-overlapping text spans of clauses or sentences. The forms suggested by Fraser [17] are shown in Table 2.1.

Table 2-1 Canonical Forms of Discourse Connectives

Sentence	**Canonical forms**
• **While** she is pregnant, Selam will not take a plane	<DC+ A1, A2>
• We left late. **However**, we arrived home on time.	< A1. DC, A2>
• We don't have to go. I will go **nevertheless**.	<A1.DC+A2>
• Yared played tennis, **and** Hana read a book.	<A1, DC+A2>

2.2 Theories of Discourse structure

2.2.1 Rhetorical Structure Theory

Linguists and computational linguists have proposed a number of theories to represent discourse structure over the last three decades. The theories proposed so far differ in the type of discourse, i.e., written text or dialogue, the type of organization, i.e., intentional or informational and their

background and objectives. Among these theories rhetorical structure theory (RST) is the most popular, especially within the area of computational linguists [20].

RST was developed in the 1980s by a group of researchers interested in natural language generation. It describes natural texts by characterizing their structure primarily in terms of relations that hold between parts of the text. According to RST these relations exist between two adjacent and non-overlapping text spans [22].

There are four elements defined in RST independently of the particular language and text types to which it has been applied: relation, schemas, schema applications, and structure.

Relation definitions identify particular relationships that can hold between two non-overlapping segments of a text. The portions of a text can be nuclei, the most important and essential to the writer's purpose, and satellite, the element less important to the writer's purpose. Based on this definition discourse relation in RST are divided into paratactic (multinuclear) and hypotactic (mononuclear or nucleus-satellite) relations. In paratactic relation both spans of the text are important. Examples of paratactic relations include: contrast, disjunction, and sequence. In contrast in hypotactic relation only one segment is important and the other is not. Examples of hypotactic relations include: background, circumstance, elaboration and purpose. The original RST defines 24 relations based on their analysis of texts from various genres including memos, news articles, advertisements, etc. However, the authors have explicitly suggested that the relations are not closed list and are open to extension [22]. The full list of relations including extensions can be found in Appendix A.

Schemas are defined as "*abstract patterns consisting of a small number of constituent test spans, a specification of the relations between them, and a specification of how certain spans (nuclei) are related to the whole collection*."[22] RST identifies five schemas represented by the examples shown in Figure 2.1 [22]. N and S in the figure represent nucleus and satellite respectively. The curved arrows represent relations holding between text spans and they are always drawn pointing towards the nucleus. Straight lines represent identification of the nuclear spans [22].

Schema application conventions are used to accommodate the variations of the five schemas defined. This is needed because schemas that appear in text structures are not always exact copies of the schemas as defined. The possible applications of schema are defined by three conventions

below [22].

1. Unordered spans: The schemas do not constrain the order of nucleus or satellite in the text span in which the schema is applied.

2. Optional relations: For multi relation schemas, all individual relations are optional, but at least one of the relations must hold.

3. Repeated relations: a relation that is part of a schema can be applied any number of times in the application of that schema.

The structure of an entire text is defined in terms of composition of schema applications. RST structure is simply an arrangement of regions where schemas apply such that the following four constraints hold true.

Completeness: the arrangement contains one schema application that contains a group of text spans that constitute the entire text.

Connectedness: Each span, except for the span that contains the entire text, is either a minimal unit or a constituent of another schema application.

Uniqueness: Each schema application represents unique text span, even in multi-relation schemas each relation applies for different text spans.

Adjacency: the text spans of each schema application constitute one text span.

The above definitions and constraints are sufficient to cause RST structures to be trees and to define a certain structure as RST structural analysis of a text [22]. An example text and its RST diagram is given in Figure 2.1 [22].

Example 5

Although Dioxin is toxic to certain animals, evidence is lacking that it has any serious long-term effect on human beings.

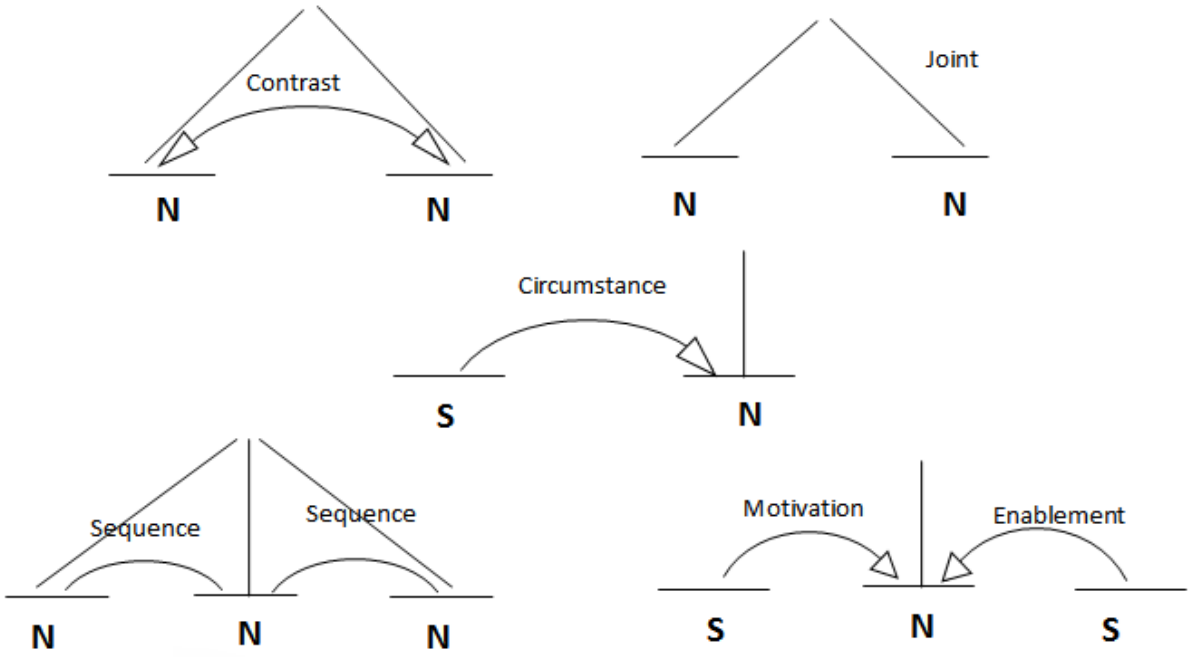


Figure 2-1 The Five Schemas of RST

In the example above the clause containing 'although' signals a Concession relation. Concession is an example of hypotactic relation defined in RST. The first step in analyzing text in RST is segmenting it to elementary units. In the original RST the size and type of elementary units is not strictly defined. However, in their analysis clauses are considered elementary units except for clausal subjects and complements and restrictive relative causes which are considered as parts of their host clause rather than as separate units. Following this convention, the segmented form of the text and its RST diagram is given as follows.

[*Although Dioxin is toxic to certain animals,*]^{1} [*evidence is lacking that it has any serious long-term effect on human beings.*]^{2}

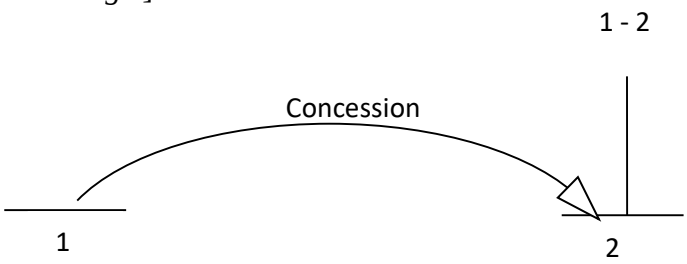


Figure 2-2 RST Diagram for Example 5

2.2.2 The Linguistic Discourse Model (LDM)

The Linguistic Discourse Model (LDM) is a theory of discourse interpretation and parsing to build a structural and semantic representation of text. The main components of LDM are discourse constituent units (DCUs- carrying propositional information such as events, facts and states), and discourse operators (DOs – carrying non-propositional information such as logical operator and connectives). The discourse parsing consists of two parts. First, the discourse units (sentences or clauses) are parsed using traditional syntactic theories. Second, these discourse units are then combined using semantic context-free relations (discourse grammar) into a tree structure. There are only three discourse grammar rules in the LDM:

- Discourse coordination is an N-ary branching rule where all RHS (Right-handsister) nodes have the same relationship to the common parent such as a list of elements and narratives.

- Discourse subordination is a binary elaboration relationship between a subordinate node (one sister) and dominant nodes (other sisters). The interpretation of the parent is the interpretation of the dominant daughter.

- Logical or rhetorical relations are derived between RHS sisters in an N-ary branching rule. The interpretation of the parent derives from the interpretation of each daughter and the relationship between them.

Polanyi and colleagues proposed an implementation of a parser based on the LDM. Nevertheless, LDM is a syntactically informed, semantically driven model, thus adopting this parser to work with other languages is a complex process [44].

2.2.3 Discourse Graph Bank Theory

Wolf and Gibson [45] introduced a “relatively shallow” and recursive graphical structure of discourse spans, with discourse relation anchoring among the siblings. Graph Bank theory, rather than analyzing a text as a binary tree structure of discourse spans built recursively via discourse relations between adjacent segments, it represents discourse as a *chain graph.*

In this approach, a text is analyzed by grouping the segments into topic and sub-topic segments, linking the non-adjacent segments or groups, if possible, using any of eleven broad classes of binary relations: Same, Condition, Attribution, Cause-Effect, Contrast, Similarity, Example, Expectation, Temporal sequence, Generalization, and Elaboration.

It appears that groupings could determine a partial hierarchical structure for parts of a text, and that grouping is a matter of constituency. But this is the only hierarchical structure in [45] approach: unlike RST and the LDM, the existence of a coherence relation between two segments does not produce a new segment that can serve as argument to another coherence relation. Moreover, this theory claims to leap upward in complexity from trees to chain graphs as a model for discourse structure.

2.2.4 Intentional Discourse Model

Grosz and Sidner's theory [46] also features recursively defined relations. In this account, discourse segments (dss) are the principal units of structure, and relations hold between these to form larger dss. However, the primitives used to define relations make reference solely to the intentions a writer has in creating a text.

Relations actually apply between discourse segment purposes (dsp); an assumption is made that a single overriding intention can be specified for each segment, and it is these intentions which are connected by relations. The fundamental metaphor is of a text embodying the execution of a plan pursued by the speaker. A point to be noted is that although Grosz and Sidner frequently use examples from task-oriented dialogues, they take care in such cases to distinguish the plan required to carry out the task from the plan required to create the text.

Using intentions in relation definitions follows quite naturally from thinking of them in the context of a recursive planning paradigm. Plans are produced to achieve user goals (or in the present case, writer goals); and do so by decomposing a principal goal into a hierarchy of sub goals.
There are only two relations in this theory: dominance and satisfaction-precedence. These are the first intentionally defined relations. Thus, dsp1 contributes to dsp2, and dsp2 dominates DSP1, when the intention dsp1 may be intended to provide part of the satisfaction of dsp2. The dominance relation invokes a partial ordering on DSPs, the dominance hierarchy. Also, dsp1 satisfaction-precedence dsp2 is true whenever dsp1 must be satisfied – recognized- before dsp2. There is no finite list of discourse purposes as there is of syntactic categories.

Another attractive feature of this theory is its account of the interaction between relations and focus. Associated with every discourse segment is a focus space, and at every point in a text a focus stack is given which models the reader's focus of attention as the discourse proceeds. The metaphor of a stack is another import from computational theories. Its pushes and pops are

determined by the dominance relations in the text: if the segment dsp2 dominates a sub-segment dsp1, then moving into dsp1 causes the focus space associated with dsp1 to be pushed onto the stack, and leaving DSP1 causes it to be popped off the stack.

2.2.5 Segmented Discourse Representation Theory (SDRT)

Rhetorical relations are also a fundamental aspect of Segmented Discourse Representation Theory (SDRT) [47]. The logical form of discourse, according to their perspective, consists of a set of labels and a mapping of those labels to *logical forms*, which can consist of rhetorical relations between the labels (arguments). A hierarchical structure is then created over the labels, allowing rhetorical relations to relate the contents of individual clauses or extended text spans. SDRT's rhetorical relations are less fine-grained than those used, for example in RST. The SDRT's Rhetorical relations must connect propositions, questions or requests. The contents of text spans can participate in more than one rhetorical relation unlike in RST.

2.2.6 Discourse Lexicalized Tree-Adjoining Grammar (D-LTAG)

Discourse Lexicalized Tree-Adjoining Grammar (D-LTAG) is a lexicalized approach to discourse relations [48][50][51]. The main belief here is that establishing relations between discourse units is based on a similar concept as establishing relations within the clause. LTAG is a tree representation of syntactic and lexical items of part of a text. However, lexicalization in D-LTAG means that each elementary tree in D-LTAG is anchored by a discourse connective which indicates a discourse relation, and links other trees for other parts of the text (arguments), using two language independent composition operations, namely substitution and adjunction. These predicate-argument trees are recursively linked to present the discourse structure. However, LTAG trees are not annotated to be linked with left and right adjacent trees, as RST does [49].

2.3 Manual Discourse Annotation

2.3.1 The RST Discourse Treebank

The RST Discourse Treebank corpus is the first publicly available, manually annotated discourse corpus grounded in the RST framework [23]. During the annotation process the annotators aimed at creating a corpus grounded in specific theory and sufficiently large to perform statistical analysis. The corpus was also limited in that it only addresses monologue textual materials, ignoring spoken language resources. The corpus primarily consisted articles from Wall Street Journal. It has been used in developing natural language applications such as summarization [2],

question answering [24], and text generation [25]. Carlson *et al.* [23] argue that the discourse structure of a text can be represented as a tree defined in terms of four aspects:

- The leaves of the tree correspond to text fragments that represent the minimal units of the discourse, called elementary discourse units.
- The internal nodes of the tree correspond to contiguous text spans.
- Each node is characterized by its nuclearity – a nucleus indicates a more essential unit of information, while a satellite indicates a supporting or background unit of information.
- Each node is characterized by a rhetorical relation that holds between two or more non-overlapping, adjacent text spans. Relations can be intentional, semantic, or textual in nature.

Based on the above framework, a group of annotators, which were all professional language analysts trained on RST theory, tagged the corpus in three phases. During the initial phase they created 100 tagged corpuses. It took about four months. The second phase was a reassessment phase, during which they measured consistency across the group and refined tagging rules. The second phase lasted one month. In the final phase (about six months) all 100 documents were re - tagged with the new approach and guidelines. The remainder of the corpus was tagged in this manner [23].

2.3.2 The Penn Discourse Tree Bank (PDTB)

The PDTB project began with the D-LTAG representations in mind. However, the annotation guidelines were subsequently made as theory independent as possible so that the corpus would be usable by a wide range of users [52] [53]. The latest version of the Penn Discourse Treebank (PDTB2) contains annotations of discourse relations and their arguments on the one million words syntactically annotated of the Wall Street Journal in the Penn Treebank. The annotation contains mostly informational discourse relations with a few pragmatic relations yielding for low-level discourse structure. The relations are mainly elementary predicate-argument relations whose predicates come mainly from discourse connectives and whose arguments come from units of discourse expressing abstract objects (AOs).

Discourse relations in the PDTB might be signaled explicitly by discourse connectives such as subordinating or coordinating conjunctions or discourse adverbials. Implicit relations are also annotated, but only between adjacent text spans. For the latter, the implicit inferable relations are annotated by inserting a so-called *implicit connective* that best expresses the inferred relation.

2.3.3 Dependency Tree Banks

The Copenhagen Dependency Treebank [54] consists of 480 annotated parallel texts in Danish and English, and 300 annotated parallel texts for German, Italian, and Spanish. Both syntactic and discourse annotation were done in the form of a tree dependency structure, linking up the top dependency node of each sentence with those of other sentences and labeling the relation between them.

The Prague Dependency Treebank, PDT 3.0, has a layer of annotation which captures discourse relations. The difference between the PDTB and the PDT is that the annotator links the mega tree of sentences (a tree structure of syntactic dependency in the PDT2) as arguments of an inter-sentential relation. For intra-sentential relations, such as clausal coordination, the syntactic annotation is already annotated in the PDT2 and should be transformed automatically into the discourse layer.

2.4 Automatic Discourse Parsing

In order to use discourse structure in developing NLP applications, it is important to develop automatic discourse parsing algorithms based on discourse structure theories. Automatic discourse parsing algorithms are used for recognizing and generating various forms of discourse structure. Mainly there are two prominent styles of parsing in the literature. This classification is based on the corpus used in the parsing. As it has been explained in the previous section, there are two discourse-annotated corpuses that are available for research purposes: RST and PDTB. While the RST corpus is based on the RST theory and demands the specification of a tree structure PDTB corpus do not demand the specification any graph or tree structure. Therfore, discourse parsing that follows RST are called Deep Discourse Parsing. In contrast, PDTB style parsing is called Shallow Discourse Parsing [55].

There are two major subtasks of discourse parsing, namely: discourse segmentation and discourse relation labeling. Discourse segmentation refers to the process of segmenting text into suitable elementary discourse units or arguments. This task shouldn’t be confused with segmenting a text

into paragraph size chunks reflecting the topic structure as it is used in Hearst [26] and Eisenstein [27]. Discourse relation labeling, based on the output of segmentation algorithm, gives the discourse realtion or sense that exists the elementary discourse units or arguments linked by local and global discourse relations.

The CoNLL 2015 and CoNLL 2016 Conference [56] [57] organized in 2015 and 2016 focused on Automatic Discourse Parsing and Argument Segmentation and labeling as the subject of its shared task. In essence, this was designed as a standardized competition geared to push the state of the art for discourse parsing and attract attention and further interest in the topic. Both competitions provided the PDTB 2.0 dataset for training and development of the models. As is standard with this dataset, Sections 2-21 of the PDTB were set aside for training only, Section 22 was marked for development and Section 23 for testing of the models. In line with standard machine learning tests, the organizers have created a separate dataset which is composed of English Wikinews. The separate dataset was annotated by two different annotators independently according to the PDTB 2.0 guidelines. The inter-annotator agreement between the two was 96% overall indicating a high degree of adherence to the PDTB 2.0 standard.

Both the 2015 and 2016 competitions were divided into two streams, the Closed and the Open tracks. The closed track was restricted on the usage of extraneous data beyond the PDTB 2.0 dataset and was only allowed to utilize phrase structure parses predicted via the Berkeley parser and dependency parses that were produced by the Stanford parser using the former dataset as input. In contrast, the open track was allowed to use any dataset or extraneous knowledge to dynamically build the model or algorithm for the task. The task also required the development of an end-to-end system that attempted full discourse parsing of explicit and implicit relations, segmenting out the arguments as well as the connectives for the explicit relations along with defining the sense of the relation.

For argument extraction, most participants resorted to casting the problem as a sequence labeling task at the token level while a few teams, in contrast, applied rule-based approaches to extract the arguments. For the sequence labeling approach, a popular choice of learning algorithm was the Conditional Random Fields (CRFs). These are types of probabilistic undirected graphical models that are discriminative and rely on the Markov Property to establish a relationship between current and historical inputs and outputs. As a result, they can account for and learn dependencies between

words of a relation that appear in a non-consecutive format thereby learning the start and end points of an argument span. For defining the sense of the relation most participants resorted to casting the problem as a multi-class classification problem. For the classification approach various classifiers, including Maximum Entropy, SVM, AdaBoost, NaiveBayes and Weka are used.

The actual scoring system was based on the F1 score which is defined as the harmonic mean of Precision and Recall. Precision focused on the ratio of true positives over all positive signals emitted by a given system whereas Recall focused on true positives over true positives and false negatives identified by the given system. The scores are very low, with the top system achieving an overall parsing score of 24.00% (F1) on the blind test set and 29.69% (F1) on theWall Street Journal (WSJ) test set. However, in terms of subtasks the tops system achieves argument labeling score of 46.37%(F1) on the blind test and 49.42%(F1) on theWall Street Journal (WSJ) test set and connective classification of 91.86%(F1) on the blind test and 94.21%(F1) on theWall Street Journal (WSJ) test set.

2.5 Amharic Discourse

Unlike the prominent western languages, there is no much work done regarding Amharic discourse except Getachew Endalamaw [34, 38]. In [34], the author studied the distribution and function of common Amharic discourse markers in Amharic narrative textual units written by college students.

The author argues that Amharic conjunctions and conjunctive phrases (*መስተፃምራን እና መስተ'ፃምራዊ ሐረጎች*) like *ስለዚህ, እና, ግን/ነገር ግን,* etc., have to be studied in relation with Amharic discourse rather than as part of grammar. The author based the argument on the suggestion made by Baye Yimam [35] and his own observation of traditional grammatical classification of Amharic words by Mersiheha'zen Weldekirkos [36] and Fantaye Teklehawaryat [37].

Consequently, the author analyzed 38 narrative textual units and studied 8 most frequently used Amharic discourse markers (*የአማርኛ ዲስኩር አመልካቾች*). In addition to their frequency of distribution, the author has explained their function and the discourse relation they are associated with.

Another work we have come across is a book by Dereje Gebre [39]. In fact this book is not about discourse but writing and composition as a whole. However, in the part of the book that discusses about cohesion/coherence in writing, the author lists words and phrases that are called connecting/co-relating phrases ("*የመሸጋገሪያ ሐረጎች*") and states that they are very helpful in creating

coherent text and lists some of them along with their role (“የመሸጋገሪያ ሐረጎች ሚና”). We have observed that the list of “የመሸጋገሪያ ሐረጎች” and “የመሸጋገሪያ ሐረጎች ሚና” in the book includes the discourse markers and associated discourse relation mentioned in [34]. From our observation of the lists and discussion with the author and other linguists, it is appropriate to consider the words and phrases listed under “የመሸጋገሪያ ሐረጎች” and “የመሸጋገሪያ ሐረጎች ሚና” as discourse markers and discourse relation respectively viewed in context of composition. Accordingly, the lists are used as input in this research work.

The objective of Getachew Endalamaw’s second work [38] is to identify the different types of Amharic discourse markers and analyze their part of speech and discourse functions. Since there is no Amharic corpus that can be used for such kinds of study, the author has prepared a mini corpus consisting of texts from ‘Addis Zemen’ and ‘Addis Admas’ Newspaprs editorial, letters written by Addis Ababa University's Human Resource Directorate and excerpts from the books ‘የስነ ጽሑፍ መሰረታውያን’, ‘’አጼ ቴዎደሮስ እና የኢትዮጵያ አንድነት’ and ‘ፍቅር እስከ መቃብር’::

Based on this corpus the author has identified 75 Amharic discourse markers with various distribution frequencies. Consequently, the discourse markers part of speech tagger is analyzed and each of the markers is categorized into words, phrases and sentences. Out of the 75 discourse markers identified, 24 (32%) are words, 29 (38.67%) are phrases and 22 (29.3 %) are sentences. Here, it is important to mention that the discourse markers categorized as ‘sentences’ are not as such sentences known in common sense. For example, the discourse marker ‘የለም’ is one of the discourse markers categorized as sentence (ሙሉ ዓረፍተ ነገር (ሙዓነ). The discourse marker ‘ስለሆነም’ is another discourse marker categorized as a sentence (ጥገኛ ዓረፍተ ነገር (ጥዓነ)). In both cases the author argues that these discourse markers can not be categorized as words or phrases by providing grammatical structure analysis of the discourse markers. The detail of the argument is not included here since it is beyond the scope of this study. It is only mentioned to avoid the confusion that might arise from the general understanding of the structure of the sentences and the context it is used in this study.

The other major contribution of this work is analysis made on the discourse functions of the identified discourse markers. According to the analysis, the author has identified 10 discourse functions namely: ምክንያት/ውጤት፣ ተጨማሪ፣ ማጠቃለል፣ መለየት፣ ማነፃፀር፣ አማራጭ፣ ማብራሪያ፣ ተለጣጥቆ፣ ገምጋሚ and ሌሎች. According to the analysis there is no one to one relation between the discourse

markers and discourse relations and a single discourse marker can exhibit multiple discourse functions. For example, the discourse marker "ግን" is found to exhibit five discourse relations.

In addition, according to literature, discourse markers can be found in the middle of a sentence (Intra sentence), between sentences (Inter sentence) at the beginning of the second discourse segment and at the end of the sentence. Similarly, the study has confirmed that Amharic discourse markers are found in the middle of sentences and between sentences at the beginning of the second discourse segment. In contrast, no occurrence is found in which Amharic discourse markers are found at the end of the second discourse segment.

2.6 Summary

In this chapter we presented an overview of discourse structure. Even if there are many aspects a discourse can be structured, most of the reviewed works consider discourse relations between arguments as a central base. Discourse relations might be signaled explicitly by discourse connectives or implicitly implicated.

RST and PDTB are the two most popular corpuses used in computational studies and applications such as text generation, automatic summarization and machine translation. Attempts also have been made to develop automatic discourse parsing algorithms in order to use discourse structures in developing NLP applications. While the discourse studies and resources discussed focused on English, the basic concepts can be generalized to other languages.

Chapter 3: Related Works

Most prominent works regarding discourse structure are based on the English language. Such algorithms developed for English language have achieved good performance and are even used in NLP applications like document summarization. The algorithms developed are also capable of analyzing discourse ranging from complex sentences to full documents. There are also attempts in Dutch, German, Brazilian-Portuguese, etc. In the following two sections related works in other languages based on the RST as well as PDTB style discourse parsing are reviewed briefly.

3.1 RST Style Discourse Parser

Soricut and Marcu [7] introduced two probabilistic models that can be used to identify elementary discourse units and build sentence level discourse parser. Their segmenter consists of two components: statistical models and segmenter. The statistical model assigns a probability to the insertion of a discourse boundary after each word in a sentence while the segmenter uses the computed probability to insert discourse boundaries. The input to the discourse parser is a lexicalized syntactic parse tree in which the discourse boundaries have been identified. Their discourse parser also consists of a parsing model and parser. Parsing model assigns a probability for every potential candidate tree while the parser finds and chooses the discourse tree that scores the highest probability. The authors used the RST-DT corpus to train both of their models.

Le Thanh *et al.* [9] argue that cue phrases are insufficient for some segmentation because only 50% of clauses contain cue phrases. Consequently, they designed an algorithm that takes a syntactic structure of a text as input and generate EDUs, in addition to the cue phrase-based process used in other works. To hypothesize rhetorical relations for sentence level discourse parsing, the authors combined syntactic information, cue phrases and WordNet. It is not indicated in the article what kind of algorithm is used for deriving the discourse structure.

Joty *et al.* [28] proposed a complete probabilistic discriminative framework for performing sentence-level discourse parsing. The authors claim that their work addresses the two key limitations found in previously proposed solutions, i.e., [7, 11]. According to the authors those limitations are: first they make string independence assumptions on the structure and the labels of the resulting discourse tree and typically model the construction of the discourse tree and the labeling of the relations separately. Second, they apply a greedy, suboptimal algorithm to build the structure of the discourse tree.

The first limitation is addressed by representing the structure and the relation of each discourse tree constituent jointly and by explicitly capturing the sequential and hierarchical dependencies between constituents of a discourse tree. The second limitation is addressed by using a bottom up parsing algorithm which is non-greedy and probably optimal. Their work is based on RST and also used the RST-DT and Instructional corpora developed in [23,11] to demonstrate their work. However, Feng and Hirst [12] claim that even if they achieved a state of the art overall accuracy in relation assignment of 55.73%, their model suffers from a high order of time complexity and thus cannot be applied in practice.

Hernault *et al.* [6] presented the first fully implemented feature-based discourse parser. Unlike the traditional single multi-class classifiers, they used two classifiers in cascade: a binary structure classifier to determine whether two adjacent text units should be merged to form a new sub-tree and a multi-class classifier to determine which discourse relation label should be assigned to the new subtree. Their work is unique in introducing the novel idea of using two cascaded classifiers and applying a variety of lexical and syntactic features. They achieved 93.8% F-score for EDU segmentation, 85.0% accuracy for structure classification, and 66.8% accuracy for 18-class relation classification. Their system was trained on the RST-DT corpus.

Feng and Hirst [12] focused on improving the performance of HILDA's tree building step by incorporating rich linguistic feature. In their study they combined linguistic features used in previous works, i.e., Hernault *et al.* [6] and Lin *et al.* [13] and contextual and other set of features of their own. With this approach they have achieved considerable improvements both in structure and relation classification. In addition to this, Feng and Thirst analyzed the difficulty of extending traditional sentence level discourse parsing to text level parsing by showing that using the same set of features, the performance of Structure and Relation classification on cross-sentence instances is consistently inferior to that of within-sentence instances. They used the RST-DT corpus for their study.

Subba and Eugenio [11] argued that rich linguistics information helps discourse parsing and that the state-of-the-art machine learning supports such approach. Consequently, they proposed a shift-reduce parser that relies on a classifier to find the appropriate relation between two text segments. Their classifier, which is based on inductive logic programming, learns first order logic rules from a large set of features including compositional semantics derived from semantic parser. They have

shown that the compositional semantics improves the classification performance. The corpus used is instructional texts annotated with compositional semantics and rhetorical relations.

Pardo and Nunes [10] presented a Brazilian Portuguese discourse parser called DiZer. DiZer uses the RST framework and is comprised of three main processes: segmentation of the text, detection of rhetorical relation, and building of the rhetorical structure. First, for the segmentation process, DiZer assigns morpho-syntactic categories to each word in the text using a POS tagger. Then the text is segmented everywhere a punctuation signal or strong cue phrase is found. DiZer makes use of discourse templates which are prepared using manually annotated corpus of discourse trees to detect rhetorical relations. It does a pattern matching process between text segments and the templates. If no relation is detected heuristics are applied. Marcu's [2] algorithm is used for building rhetorical structures with some modifications.

Li *et al.* [29] proposed a recursive neural network (RNN) based algorithm for text level discourse parsing. RNNs constitute a type of deep learning framework which was first proposed by Goller and Kucher [32]. In the beginning, the algorithm obtains the distributed representation for each of the sentences of a given document using recursive convolution based on the sentence parse tree. Then it determines the probability of two adjacent discourse units being merged to form a new subtree and selects the appropriate discourse relation label by using binary and multi-class classifier respectively. Finally, the algorithm calculates the distributed representation for the subtree so formed, gradually unifying subtrees until a single overall tree spans the entire sentence. They used the RST discourse tree bank to train the classifiers along with parameter involved in convolution. The proposed framework, evaluated based on three matrices: span, nuclearity and relation, achieves overall high performance, although it fails to score top performance on any of the parameters.

Li *et al.* [30] argued that previous researches which are based on constituency structure suffer from certain limitations and proposed to adopt the dependency structure in discourse representation. According to the authors, previous researches have explored different constituency based syntactic parsing techniques and various features for discourse parsing. However, in discourse parsing it is difficult to design a set of production rules as in syntactic parsing, since there are no determinate generative rules for the interior text spans. In addition, it is hard to develop a uniform framework to represent the different levels of discourse units, which are better represented with different

features. State of the art constituency based parsing techniques also have high time complexity and approximate parsing approaches are prone to trap in local maximum. The basic idea of dependency-based parsing is that discourse structure can be seen as a set of head-dependent links, which are labeled by functional relations. Unlike syntactic parsing the dependency relations are formed between a subordinate EDU called the *dependent* and another EDU called the *head*. This makes it possible to analyze the relations between EDUs directly, without worrying about any interior text spans. In addition, since dependency trees contain much fewer nodes and on average they are simpler than constituency-based trees, the proposed dependency parsers have a relatively low time complexity. Compared to sate of the art constituency-based parsers, i.e., [2, 6], dependency parsers exhibit higher overall performance.

Extending their work on sentence level discourse parsing Joty *et al.* [31] proposed a two-stage text level discourse parser. In this work, in addition to the two problems mentioned in their previous work, a third problem that comes with text level discourse parsing is addressed. They argued that existing discourse parsers do not discriminate between intra-sentential and multi-sentential parsing. However, distinguishing between these two conditions can result in more effective parsing. This is so because two separate parsing models could exploit the fact that discourse relation are distributed differently in intra-sententialy and multi-sententially. This also gives the models the freedom to choose their own informative features. Accordingly, the authors have developed intra-sentential and multi-sentential parsers and combined them effectively to obtain substantial improvements over existing methods. Unlike previous methods they have evaluated their parser on two very different text types: news and instructional how-to-do manuals, and successfully demonstrated the parser's consistency over various text genres.

3.2 PDTB Style Discourse Parser

Lin [56] investigated a natural language problem of parsing a free text into its discourse structure according to Penn Discourse Treebank representation in a fully data-driven approach. The author proposed a classifier-based parser that makes use of contextual features, word-pairs, and constituent and dependency parse features. Then, designed a parsing algorithm and implement it into a full parser in a pipeline. In addition to that the author has given a comprehensive evaluation on the parser from both component-wise and error-cascading perspectives. This research work is the first work to develop a parser that performs end-to-end discourse parsing in the PDTB style.

Ghosh [58] employs a data driven approach to identify arguments of explicit discourse connectives. The author designed the argument segmentation task as a cascade of decisions based on conditional random fields (CRFs). The CRFs is trained on lexical, syntactic and semantic features extracted from the Penn Discourse Treebank and evaluate feature combinations on the commonly used test split. According to the results, the best combination of features includes syntactic and semantic features. The evaluation results on the standard test set of the PDTB has achieved a rebalancing of precision and recall with improved F-measures across the board.

3.3 Summary

In this chapter we reviewed related works on automatic discourse parsing. The related works are classified, according to the discourse corpus and associated theory they rely on, in to two: RST and PDTB style discourse parser. The availability of large discourse annotated corpus has inspired researchers to conduct various studies and helped so much in achieving near human standard results. However, compared to sentence level discourse parsers, still there is a lot to improve in document level discourse parsers.

To the best of our knowledge, there is no study conducted on automatic discourse processing of Amharic labguage and there are no corpora and tools to be used as basis for the study. Therefore, in this study a PDTB style sentelnce level discours parser for Amharic is proposed.

Chapter 4: Design of Discourse Parser for Amharic

This chapter is dedicated to explain the architecture of the proposed PDTB style sentence level Amharic discourse parser. In the first section, explanation of full architecture of the system is given along with the specific components it consists of. Then, the following subsections elaborate on each components and algorithms. Finally, a summary of the chapter is given in the second section.

4.1 Architecure of the System

In this section an overview of the architecture of the parser is discussed in detail. The architecture of the parser follows a PDTB style approach in which the arguments, discourse connective and discourse relations between the arguments is given as output. Unlike RST style discourse parsing this style do not assume a tree of graph structure. The parser consists of three main components: connective identifier, argument labeler and sense classifier. In each component, different machine learning approaches along with lexico syntactic features generated by a preprocessor are applied. The purpose of connective identfier is to decide wether an explicit discourse marker or morphologicaly inflected word in the given text is functioning as discourse connective. Argumnet labeling is to identify both argumens of the input text using a statistical model and features generated from the prerpocessor. Sense classification classifies the discurse relation (sense) between the arguments based on the identified discourse connective and other features fom the preprocessor. Full architecture of the Amharic discourse parser is showm in figure 3-2. In the rest of the sections, each component and the relevant features used will be discoussed in details.

4.1.1 Feature Extraction

As shown in figure 3-2 the first component of the Amharic discourse parser is feature extractor. This component is used ot extract the lexical and mporphological features required for the training of the three models. It is composed of seven algorithms namely: tagger, stemFetcher, affixFetcher, romanFormFetcher, argumentFeature, connectiveFeature and senseFeature. The tagger algorithm is used to generate word-PoS tag tuples using the Habit Amharic tree tagger. The stemFetcher, romanFormFetcher and affixFetcher algorithms extract stem, roman form and affix repectively for each input word using the HornMorpho morphological analyser. The rest three algorithms combine the results of the above mentioned algorithms and for the purposes of the connective indentifiction,

argument labeling and sense labeling tasks. The algorithms used for feature extraction are listed ing Algorithm 4-1 and 3-5. A sample output of extracted features is shown in figure 4-2.

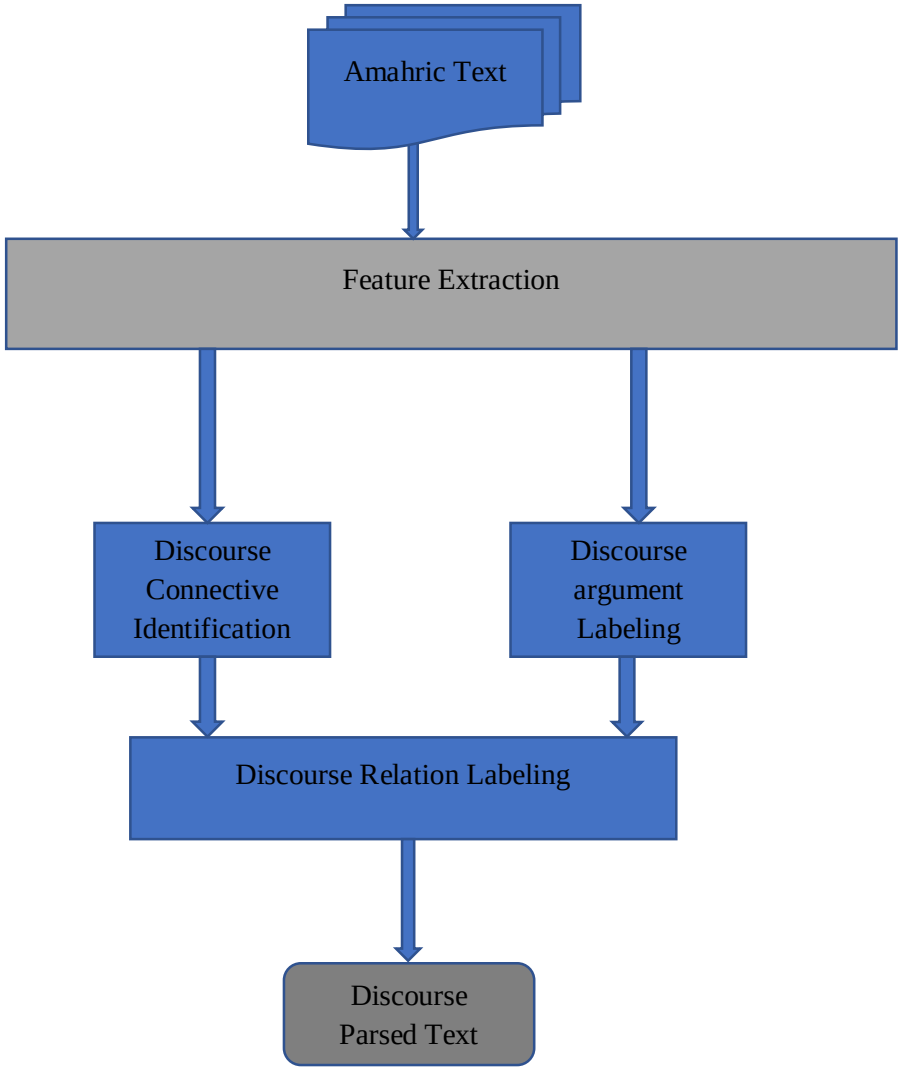


Figure 4-1 Full Arcitecture of the Amharic Discourse Parser

```
###Feature Extraction Algorithms
function tagger (Word):
    read Word
    Call the PoS tagger (Habit Tagger)
    Open filePOS
    For each Word
        Write Word, PoStag tuple into filePOS
    Close filePOS
return Word, PoStag tuple

function stemFetcher(Word,tag):
    read Word,tag
    Call Morphological Analyzer (Horn Morpho)
    For each Word
        check for stem output
            If found
                return stem
            else
                return none

function affixFetcher(Word):
    read Word
    Call Morphological Analyzer (Horn Morpho)
    For each Word
        Check for prefix of Word
            If found
                affix=prefix
            else
                return none
        Check for suffix of Word
            If found
                affix = affix+suffix
            else
                return none
return affix

function romanFormFetcher(Word):
    read Word
    Call Morphological Analyzer
    For each Word
        extract romanForm
return romanForm
```

```
function argumentFeature(Word,tag)
   Call affixFetcher
   Call romanFormFetcher
   Features = [affix,romanForm,Word,PoS]
return Features

##Input - A nested array which contains a list of sentences where each sentence is a list of word-pos tag tuples
function ConnectiveFeatures(Sentence)
   For each sentence
      for each tuple
         call affixFetcher
         call romanFormFetcher
         If Word is not the initial word
            return previous word and previous PoS tag
         If Word is not the final word
            return next word and next PoS tag

         Feature = [affix, romanForm,Word,PoS,previous word+Word,previous PoS,previous PoS+PoS,Word+next word,next PoS,PoS + next PoS]
return Feature

##Input - A nested array which contains a list of sentences where each sentence is a list of word-pos tag tuples
function senseFeatures(Sentence)
   For each sentence
      for each tuple
            If Word is not the initial word
            return previous word and previous PoS tag
         If Word is not the final word
            return next word and next PoS tag

         Feature = [Word, PoS,previous word+Word]
return Feature
```

Algorithm 4-1 Feature Extraction Algorithms

Figure 4-2 Sample Screen shot of Extracted Features

4.1.2 Discourse Connective Identification

There are 75 types of discourse connectives identified in Amharic [38]. However, for the corpus prepared for this study we are considering only thirteen (13) discourse connectives. As per the corpus analysis made the occurrence of a connective do not necessarily dicate a discourse function and the connective may have sentential function. Therfore, given a connective occurrence the parser needs to disambiguate the connective i.e wether the connective has a discourse function or not, before using it in argument labeling and sense classification tasks. In addition, in a situation where discourse relation is not indicated by explicit discourse marker but words having a discourse function due to various morphological inflections, it is necessary to detect such words. Hence, both identifying the function of an occurence of a discourse marker and detection of a word having a discourse function is formulated as a sequence labeling problem.

In order to resolve this sequence labeling problem Conditonal Random Field (CRF) is used in this study. CRF is a type of probabilistic graphical model that can be used to model sequential data, such as labels of words in a sentence. In CRF, a set of features are extracted for each word in a sentence. During model training, CRF will try to determine the weights of different feature functions that will maximise the likelihood of the labels in the training data. In this study, a manually labeled data set is used for training the CRF model.

In statistical models, selecting appropriate feature is the primary determining factor of the performance of the model. Therefore, it is necessary to extract as many features as possible and train the model with various combinations to achieve an optimal performance. Various studies show that lexical and syntactic fatures are very useful in disambiguating discourse connectives. [13] has said that lexical features like the connectives context and POS tags give a strong indication of its discourse usage. [59] on the other hand has showed that syntactic features are a crucial role in disambiguating discourse connectives. In this study, since there is no Amharic syntax parser only the lexical features suggested by [13] are used. Table 4-1 shows the features and their description and algorithm 4-2 shows the connective identification algorithm.

Table 4-1 Features used for connective Identification

	Features	Desciption
1	C POS	Connectives Part of Speech
2	Prev + C	Connective and the preceeding token
3	prev POS	Part of Speech of the previous word
4	Prev POS and C POS	Part of Speech of the connective and preceding token
5	C + next	Connective and the next token
6	Next POS	Part of Speech of the next word
7	C POS +Next POS	Part of Speech of the connective and next token

##Discourse connective Identification Algorithm

##Input-Marked sentences with N and D for Non-discourse and discourse markers respectively
###Output- Discourse Connective dentification Model
For each sentence in marked sentences

Append a word, marker 2-tuple array of the sentence to 'sentence' array

For each 2-tuple sentence in 'sentence' array

call tagger function to fetch PoS tag for each word

append a word, PoS,marker 3-tuple array of a sentence to 'data' array

For each 3-tuple sentence in 'data' array

call ConnectiveFeatures function to fetch 'Features'

Separate marker information and append to FullSentenceMarkerData array

append 'Features' to FullSentenceFeatureData array

Split the two arrays into train and test sentence
train and test the model
save the final mode.

Algorithm 4-2 Discourse Connective Identification Algorithm

4.1.3 Discourse Argument Labeling

The next component of the discourse parser is argument labeler. Argument labeling is concerned with correctly identifying the position of the arguments. In this study, argument labeling is also formulated as a sequence-labeling problem. The problem is to assign each word in the input text an observation category c ∈ {B, I, E}, where B denotes a beginning of arguments, I denotes a continuation of argument and E an end of argument. In order to resolve this sequence labeling problem Conditonal Random Field (CRF) are used in this study. CRF is a type of probabilistic graphical model that can be used to model sequential data, such as labels of words in a sentence. In CRF, a set of features are extracted for each word in a sentence. During model training, CRF will try to determine the weights of different feature functions that will maximise the likelihood of the labels in the training data. In this study, a manually labeled data set is used for training the CRF model. Tabel 4-2 shows the features used for argument labeling and the algorithm used for argument labeling is shown in algorithm 4-3.

Table 4-2 Features used for Argument Labeling

	Features	Desciption
1	Token T	Each token in the give text
2	T POS	Part of Speech of the each token in the given text
3	T Stem	Stem of each toke (where applicable)
4	T Prefix	Prefix of each token
5	T suffix	Suffix of each token

```
##Argument labeling algorithm

##Input labeled sentences with Arg1-N,  Arg1-I, Arg1-E,Arg2-B,Arg2-I,Arg2-E to indicate
the begining ,Inside and Ending of
###Output – Argument Labeling Model
Arg1 and Arg2

for each sentence in labeled sentences

    append a word-label 2-tuple array of the sentence to 'sentence' array

for each 2-tupled sentence in a 'sentence' array

    call tagger function to fetch PoS tag for each word

    append a word, PoS, Label 3-tuple array of a sentence to 'data' array

for each 3-tupled sentence in 'data' array

    call argumentFeature function to fetch 'Features'

    separate label information and append to FullSentenceLabel array

    append 'Features' to FullSentenceFeatureData array

split the two arrays for train and test set
train and test model
save the final model
```

Algorithm 4-3 Discourse Argument Labeling Algorithm

Figure 4-3 describes the discourse connective identification and argument labeling components. The two tasks are casted as a sequence labeling problem and the process is similar inboth cases. During the training stage a discourser marked and argument labeled corpus is given to the preprocessing component. The preprocessing component extracts feature relevant for each tasks using the algorithms described in section 3.3.1. Using the input corpus and relevant features a CRF model is trained and saved. During the testing stage unlabeled text is given to the preprocessor to extract relevant features. Then, using the features extracted and the previously trained and saved model the labeling for the input text is predicted.

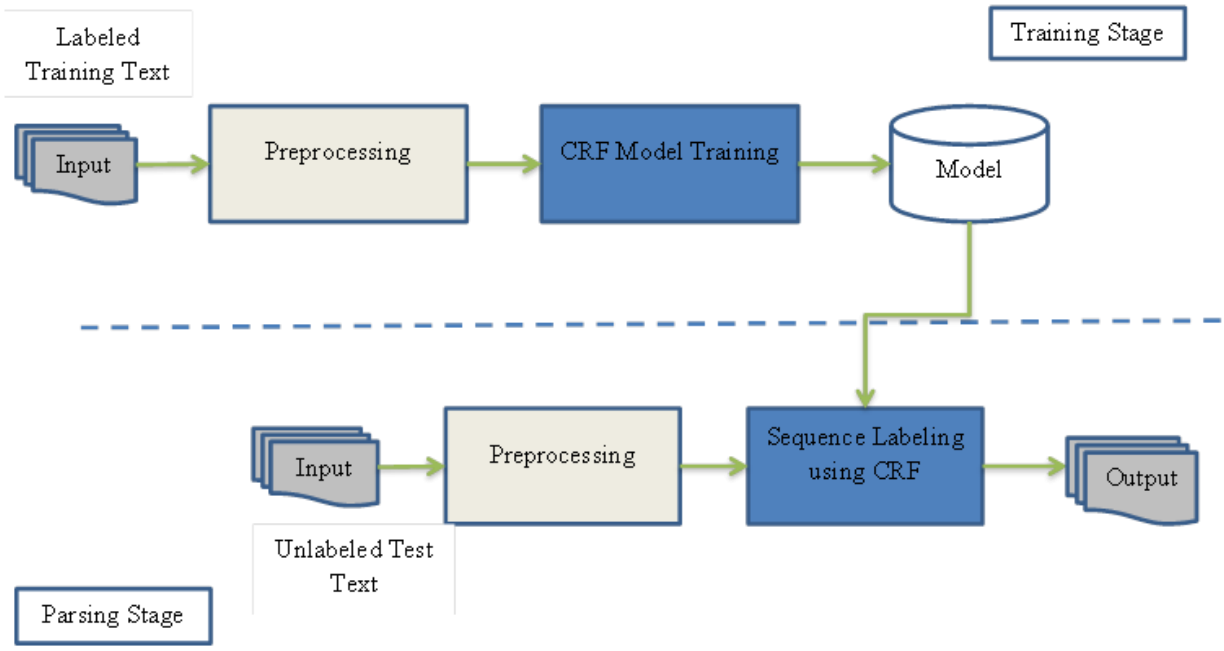


Figure 4-3 Discourse Connective Identification and Argument Lableing Component

In the test and training corpus used for this study discourse relations appears at two positions: in the same sentence (SS) and in two consecutive sentences (2CS). Therefore, our system focus on identifying these kinds of discourse relations by building a model for recognizing discourse relations in the same sentence (**SS**) as well as two consecutive sentences (**2CS**).

4.1.4 Discourse Sense Classifcation

In this study, the discourse sense classification is formulated as a multi-category classification problem. The purpose of the sense classifier is to determine the discourse relation between the the discourse arguments, given both of the discourse arguments and the discourse connective. This is important because a given connective may indicate multiple relation sense under different circumstance. The sense classifier has a training stage and testing stage. In the training stage a classifier is trained with a text with its arguments labeled, discourse connective identified and sense labeled is and relevant features. In the test phase, a text with its arguments labeled, discourse connective identified but sense not labeled give to the model after the relevant features are extracted. Based on the features and given input the model labels the sense from the known list of

senses identified in the corpus analysis. As the discourse connective is a strong feature by itself, we have trained the classifier using the following three features shown in table 4-3.

Table 4-3 Features used for Sense classification

	Features	Desciption
1	C	Connective
2	C POS	Part of Speech of the connective
3	C + prev	Connective and the previous token

```
### Discourse Relation Labeling
### Input - A sentence with Arguments Arg1 Arg2, Connective and Sense Identified
###Output – Relaion Labeling Model
function senseClassifier (sentence)
   For each sentence
      separate sense and append to 'sense' array
      call senseFeature and get connective POS and connective and previous word and append
into 'FullSenseFeature'
      append the rest of the input and 'SenseFeature' into 'text' array
   Assign 'sense' and 'text' arrays to 'sensedf' and 'textdf' data frame
   Split the data frame into train and test set
   vectorize the 'textdf' dataframe train and tes the model using classifier1, classifier2 and
classifier3
   save the final models
```

Algorithm 4-4 Discourse Relation (Sense) Labeling algorithm

Figure 3-10 and figure 3-11 describes the discourse relation labeling component and the algorithm used for discourse labeling. This tasks is casted as a multi class classification problem. During the training stage a sentence corpus labeled with argument, connective and sense is given to the preprocessing component. The preprocessing component extracts feature relevant for sense labeling task using the algorithms described in section 3.3.1. Using the input corpus and relevant features multiple (three) classifier models are trained and saved. During the testing stage

arguments and connective are extracted from the previous models are given to the preprocessor to extract relevant features. Then, using the features extracted and the previously trained and saved model the labeling for the input text is predicted.

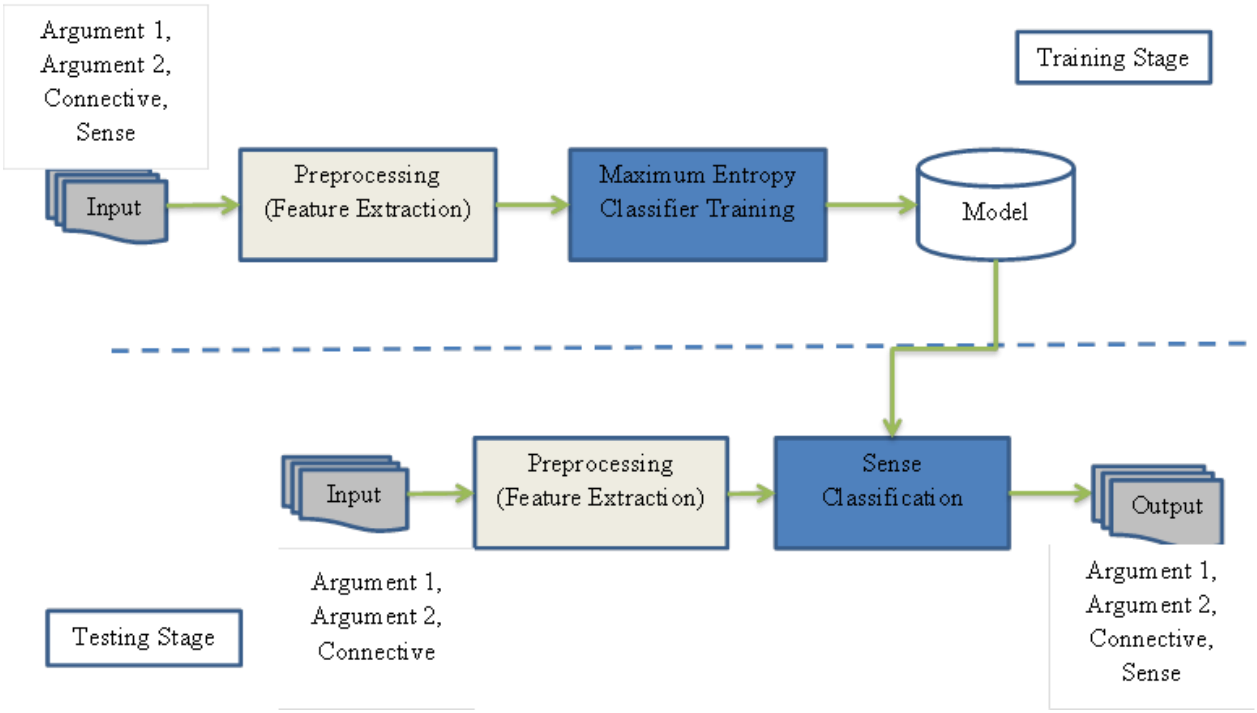


Figure 4-4 Sense Labeling Component

4.2 Summary

In this chapter full architecture of the proposed PDTB style sentence level discourse parser along with the three components of the Amharic discourse parser and the approaches used in these components is explained in detail. In addition, the algorithms used in each componets are presented and explained in detail. In the next chapter, the experiements conducted on the prototype will be discussed.

Chapter 5: Experiement

In this chapter the experiemnts conducted on the proposed system is presented. The first section explains the preparation process of the data used for the experiement. Section two presents the analysis reslutls obtained from the corpus preparation. In section 3 the implementation of the prototype and the development tools used are described. The evaluation metrics and results are given in section 4. Finally, a discussion on evaluation results is provided.

5.1 Corpus Preparation

The corpus preparation for this study is performed in collaboration with two linguists, both PhD candidates (During the course of this study) in Addis Ababa University and currently working at Debrebirhan University, on a News corpus taken from Ethiopian News Agency (ENA). The ENA corpus was classified into 12 genres namely: Accident, Health, Sports, Politics, Social, Science and Technology, Economy, Education, Environment Preservation and weather condition, Law and Justice, Foreign relation, and Culture and Tourism.

Taking the list of words and phrases labeled as cohesive devices by Dereje Gebre [39] and Amharic discourse markers used in his MSc thesis by Getachew Endalamaw [38] we collected 70 potential words and phrases that can be used as discourse markers.

First we recorded the distribution and occurrences of each potential discourse marker in each genre of the ENA corpus using a concordance tool. From this record we selected the potential discourse markers to use for our purpose based on two requirements. The first requirement is that the total number of occurrences of the potential discourse markers should be more than 1000. The second requirement is that each potential discourse marker should have at least one occurrence in each genre. In addition to this we have included potential discourse markers which exhibited occurrences lower than 1000 but have distribution in each genre. Accordingly, we selected 24 words and phrases that can be used as discourse markers. Table 4.1 shows the list of words and phrases selected for the study.

Consequently, we automatically extracted text fragments containing each potential discourse markers selected based on the previous requirements. We extracted as small as 6 text fragments for rarely used potential discourse markers and as big as 30 text fragments for frequently used ones. Totally, we extracted 503 text fragments for this purpose.

Table 5-1 Lists of potential discourse Markers and their distribution list in the corpus

No.	*Words and Phrases*	*Accident*	*Culture and tourism*	*Economy*	*Education*	*Environment preservation and weather condition*	*foreign relation*	*Health*	*Law and Justice*	*Politics*	*Science and Technology*	*Social*	*Sport*	*Total*
1	እንዲሁም	426	1527	3596	1363	583	876	1911	887	4108	356	1759	881	18273
2	ደግሞ	491	1053	2911	960	330	479	1206	808	2650	228	1108	1142	13366
3	እና	180	751	878	667	314	823	719	1042	2803	273	934	649	10033
4	በኋላ	224	548	699	458	150	583	503	583	2069	148	463	518	6946
5	በተጨማሪ/ከዚህበተጨማሪ	140	468	1372	640	155	204	850	192	859	126	535	196	5737
6	ግን	136	495	785	169	192	301	714	395	1518	103	466	229	5503
7	በፊት	137	468	932	338	167	299	750	259	1265	157	434	266	5472
8	አሁን	45	496	861	316	143	231	634	138	1360	113	454	104	4895
9	ቀደምሲል	59	209	1102	445	110	110	749	237	763	112	402	51	4349
10	ሁለተኛ	22	174	243	1082	42	59	105	154	426	30	167	607	3111
11	በመሆኑም	68	247	339	191	129	95	256	198	672	9	295	69	2568
12	በአሁኑጊዜ	35	71	693	252	43	58	269	36	280	35	171	8	1951
13	በሌላበኩል	61	101	323	120	74	38	181	108	449	20	92	88	1655
14	በወቅቱ	59	82	358	44	58	81	112	120	362	23	134	43	1476
15	ስለዚህ	2	51	54	8	126	16	32	86	315	6	30	6	732
16	እንደገና	14	54	91	40	29	67	44	41	249	13	61	21	724
17	እዚህ	8	121	98	15	50	47	37	48	195	5	47	28	699
18	በተከታታይ	22	33	83	124	17	35	68	22	162	10	44	33	653
19	በመጀመሪያ	3	26	60	230	18	19	46	17	108	11	37	8	583
20	ቀጥሎ	11	44	36	7	14	15	22	19	86	8	15	111	388
21	በመጨረሻም	5	50	19	13	3	12	20	22	172	2	31	16	365
22	ስለሆነም	5	36	28	14	19	24	28	47	131	4	22	6	364
23	በተጓዳኝ	1	20	81	60	16	8	35	6	25	4	36	10	302
24	ቢሆንም	10	21	39	18	19	9	26	15	50	8	16	10	241

Next, the text fragments were given for the linguists for analysis. The linguist job was

1. To identify whether the potential discourse markers have discourse function or not in the given text fragments
2. If the potential discourse markers have discourse function,
    a. to identify what discourse relation, the DMs are implying, and
    b. to identify which text spans are related by the discourse relations
3. Finally, to draw a classification of the discourse markers according to their associated relation

5.2 Analysis Results

According to the analysis, out of the 24 words and phrases considered in this study, nine of them are found to have no discourse function and should be studied in relation to their grammatical usage. These words and phrases are ቀደም ሲል, አሁን, እዚህ, እንደገና, በፊት ,በወቅቱ, በአሁኑ ጊዜ , በተጓዳኝ, and በ ተከታታይ. In addition to this, the phrase (በሌላ በኩል) is removed from the list of potential discourse markers because the linguists concluded that the text fragments are not properly constructed to do the analysis. The rest fourteen phrases and words have exhibited single or multiple discourse functions. The linguists have also prepared a classification of the discourse markers that exhibited a discourse function based on Getachew Endalamaw's [38] work. According to their classification, the selected Amharic discourse markers are categorized into three major discourse relations and five specific discourse relations. Their classification is shown in Table 5-2 and Figure 5-1.

Samples of the text fragments that are identified to show a discourse usage of the discourse markers listed in Table 5-1 are shown in Appendix B. The linguists have identified 88 text fragments to show discourse usage of the discourse markers listed in Table 5.2. The linguists were unable to provide a conclusive suggestion on whether the discourse marker’s usage is sentential or discourse on the other text fragments having the same discourse markers except the three discourse markers (ሁለተኛ፣ በኋላ፣ በመጀመሪያ) in which they concluded that the usage of these phrases in the rest of the fragments is clearly sentential.

Table 5-2 Classification of selected Amharic Discourse Markers

ተ/ቁ	መለየት	ተጨማሪ	ተለጣጥቆ	ማጠቃለያ	ምክንያት-ውጤት	ማነፃፀር
1	ግን	ደግሞ	እና	በመጨረሻም	በመሆኑም	ግን
2	ቢሆንም	በተጨማሪ	በመጨረሻም	ስለዚህ	ስለዚህ	ደግሞ
3		እንዲሁም	በመጀመሪያ	ስለሆነም	ስለሆነም	
4		ግን	በኋላ			
5			ቀጥሎ			
6			ሁለተኛ			

Eventually, the analysis results of the linguists were given to an expert on the subject area to be corrected and verified. According to the verification one additional discourse marker, i.e., እና is identified by the linguist from the text fragments given do not actually show discourse usage and recommended to drop the results of the analysis. The expert recommended using well-written texts, for example, History and Fiction Books, inorder to appropriately show the discourse usage of the mentioned discourse markers. Therefore, the total number of remaining verified discourse markers used in this study is thirteen. A closer look at the fragments shows that only three discourse markers (ሁለተኛ፣ በኋላ፣ በመጀመሪያ') out of 13 exhibit a sentential usage in addition to a discourse usage. The other discourse markers only exhibit a discourse usage. Accordingly, additional corpus that consists of texts from history book and bible is added to improve the corpus. Finally, a corpus having 250 sentence is prepared.

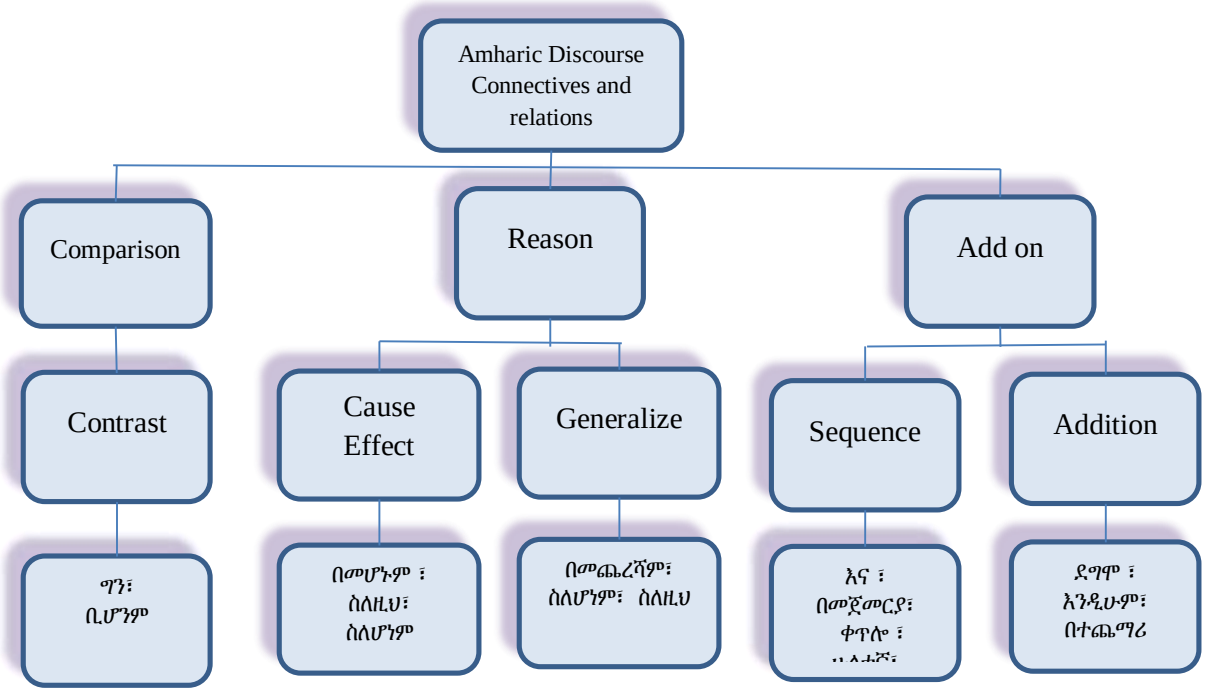


Figure 5-1 Classification of selected Amharic Discourse Markers

Consequently, the final corpus is manually tagged for three computational purposes as follows.

- For Connective classification

As per the experts' analysis the discourse connectives in the text were tagged with two labels D and N; D for discourse function and N for non discourse function.

Example -1

መንግስት N ካለበት N የአቅም N ውስንነት N አንፃር N እነዚህን N ሁሉ N ተግባራት N ሊወጣ N የማይችልበት N ሁኔታ N አለ N ስለሆነም **D** ዛሬም N ቢሆን N የግብረ N ሰናይ N ድርጅቶች N ድጋፍና N አስተዋጽኦ N ይፈለጋል N

Example 2

አበባየሆይ N የሚባለው N ሕዝባዊ N ዜማ N አሁንም N ቀለሙን N ያልለቀቀ N ቢሆንም **D** ቃላት N ላይ N ግን N መለያየት N መኖሩን N ትናገራለች

Example 3

በተጨማሪም N ከአሰራርና N አደረጃጀት N ጋር N በተያያዘም N ከቴክኖሎጂ N አንጻር N ፍርድ N ቤቶቹ N ትልቅ N ማሻሻያ N እንዳደረጉ N ከሪፖርቱ N መረዳት N መቻሉንም N ገልጸዋል N በተለይም N በመጀመሪያ **D**

ደረጃ N ፍርድ N ቤት N የህጻናትና N የሴቶች N ጉዳይ N ልዩ N ትኩረት N ተሰጥቷቸው N እንዲታዩ N መደረጋቸው N የሚበረታታ N ተግባር N ነው N ብለዋል N

Example 4

የፓርቲው N አመራር N ዶክተር N መራራ N ጉዲና N በዚሁ N ወቅት N እንደተናገሩት N መድረክ N ቢመረጥ N መሬት N እንዲሸጥ N እንዲለወጥ N ያደርጋል N በመሆኑም **D** አባላቱና N ደጋፊዎቹ N በምርጫዉ N በንቃት N በመሳተፍ N ድምፃቸዉን N እንዲሰጡ N ጠይቀዋል N

- For Argument Labeling

The sentences in the training set were tagged using –B, -I, -E label set, represent Beginning of Argumement, Inside of Argument and End of Argument, along with the prefix Arg1 and Arg2 as shown in the following two examples.

Example -1

መንግስት Arg1-B ካለበት Arg1-I የአቅም Arg1-I ውስንነት Arg1-I አንፃር Arg1-I እነዚህን Arg1-I ሁሉ Arg1-I ተግባራት Arg1-I ሊወጣ Arg1-I የማይችልበት Arg1-I ሁኔታ Arg1-I አለ Arg1-E ስለሆነም Arg2-B ዛሬም Arg2-I ቢሆን Arg2-I የግብረ Arg2-I ሰናይ Arg2-I ድርጅቶች Arg2-I ድጋፍና Arg2-I አስተዋጽኦ Arg2-I ይፈለጋል Arg2-E

Example 2

አበባየሆይ Arg1-B የሚባለው Arg1-I ሕዝባዊ Arg1-I ዜማ Arg1-I አሁንም Arg1-I ቀለሙን Arg1-I ያልለቀቀ Arg1-I ቢሆንም Arg1-E ቃላት Arg2-B ላይ Arg2-I ግን Arg2-I መለያየት Arg2-I መኖሩን Arg2-I ትናገራለች Arg2-E

- For Sense Classification

Example -1

መንግስት Arg1-B ካለበት Arg1-I የአቅም Arg1-I ውስንነት Arg1-I አንፃር Arg1-I እነዚህን Arg1-I ሁሉ Arg1-I ተግባራት Arg1-I ሊወጣ Arg1-I የማይችልበት Arg1-I ሁኔታ Arg1-I አለ Arg1-E ስለሆነም Arg2-B ዛሬም Arg2-I ቢሆን Arg2-I የግብረ Arg2-I ሰናይ Arg2-I ድርጅቶች Arg2-I ድጋፍና Arg2-I አስተዋጽኦ Arg2-I ይፈለጋል Arg2-E [ማጠቃለያ]

Example 2

አበባየሆይ Arg1-B የሚባለው Arg1-I ሕዝባዊ Arg1-I ዜማ Arg1-I አሁንም Arg1-I ቀለሙን Arg1-I ያልለቀቀ Arg1-I ቢሆንም Arg1-E ቃላት Arg2-B ላይ Arg2-I ግን Arg2-I መለያየት Arg2-I መኖሩን Arg2-I ትናገራለች Arg2-E [መለየት]

5.3 Implementation

The proposed amharic discourse parser is developed using the python programming language. Python programming language is a widely used programming language in used in the development of natural language processing applications. In particular, we have used Python version 3 and the sklearn python package which is a python library for machinle learning offered free of charge [63]. In addition to this, we have used two previously developed amharic NLP tools: namely the Habit Amharic POS tagger and HornMorpho version 2.5. The IDE used for the development is visual studio code which is also a free software. The physical machine used during implementation is Laptop installed with Windows 10 Operating system.

The prototype is implemented as an interactive application that can be launched from the windows operating system command line. It prompts the user to enter a text and return the parsed output on the same command line after processing th input amharic text. Figure 5-2 shows the screen shot of the prototype prompting a user for input text

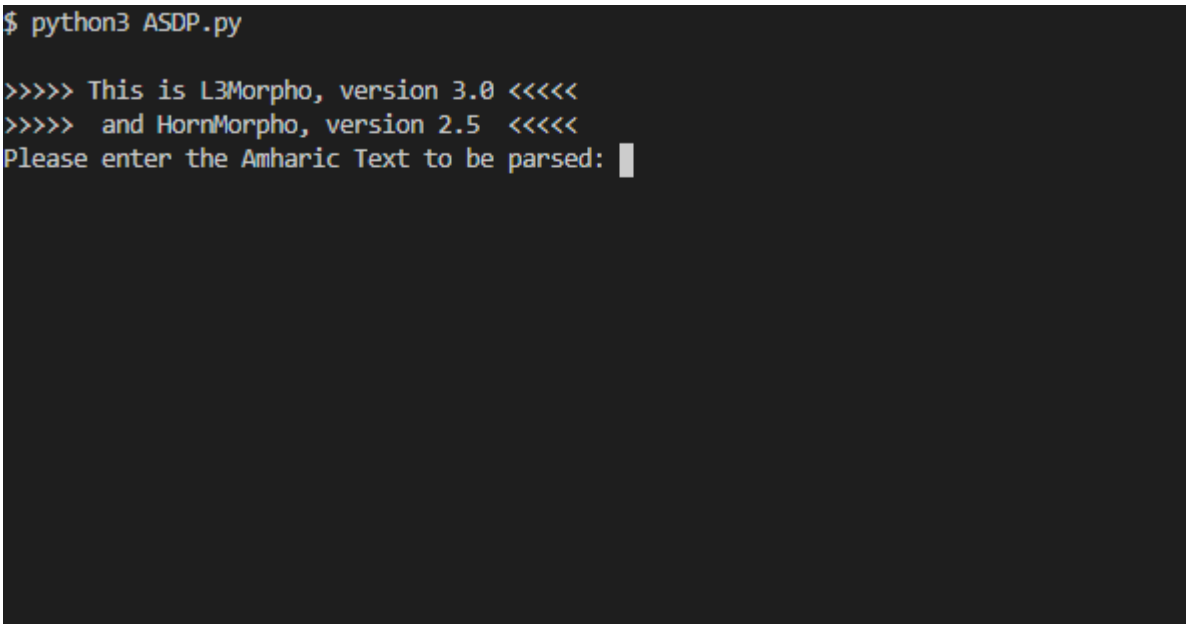


Figure 5-2 PDTB Style Shallow Amharic Discourse Parser Prototype

5.4 Evaluation

The evaluation of the system is given using the three well known measures: precision, recall and F1 measures. To compute the precision and recall of the system we have used the scemes suggested by [57]: exact and overlap. In the exact scoring scheme, an argument extracted by the system is counted as correct if its extent exactly coincides with one in the manulay-annotated corpus. In the overlap scheme, an expression is counted as correctly detected if it overlaps with an expression in the manulay-annotated corpus, i.e. if their intersection is nonempty.

Therfore,

$$Precision = \frac{\text{True Posetive}}{\text{True Posetive} + \text{False Posetive}}$$

$$Recall = \frac{\text{True Posetive}}{\text{True Posetive} + \text{False Negative}}$$

$$F1 = 2 \text{ x } \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where,

Precision – the proportion of predicted or classifed posetives tht are are truly positive

Recall – the proportion of actual positives that are predicted or classified correctly

F1 – the harmonic mean of precision and recall

Accordingly, for each component the result obtained using the training dataset is presented below.

5.4.1 Evaluation Results of Discourse Connective detection

The connective detection component is measured for the labels D and N which represent given token has a discourse connective function or not respectively. Accordingly, it has achieved 85% precision, recall and f1-score for label D and 99 % precision, recall and f1-score for label N.

	precision	recall	f1-score
N	0.99	0.99	0.99
D	0.85	0.85	0.85
accuracy			0.99

Figure 5-3 Connective Detection Result

5.4.2 Evaluation Results of Argument Labeling

The argument labeling componets is measured for each label used i.e Arg1-B, Arg1-I, Arg1-E, Arg2-B, Arg2-I and Arg2-E which represent wether each word is found at the beginning, inside or end of argument 1 or 2.

	precision	recall	f1-score
Arg1-B	1.00	0.88	0.94
Arg1-I	0.99	0.97	0.98
Arg1-E	0.91	0.80	0.85
Arg2-B	0.91	0.80	0.85
Arg2-I	0.93	0.99	0.96
Arg2-E	1.00	1.00	1.00
accuracy			0.96

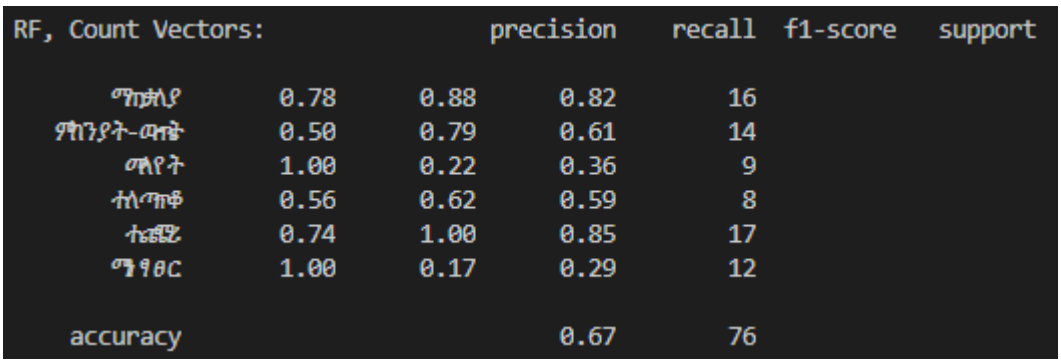
Figure 5-4 Argument Labeling Result

5.4.3 Evaluation Results of Sense Classification

The sense classification component is measure interms of the the six senses identified መለየት, ተጨማሪ, ተለጣጥቆ, ማጠቃለያ, ምክንያት-ውጤት and ማነፃፀር. A comparison of three classifiers namely Naïve-Bayes, Logistic Regression and Random Forest is presented here. According to the result the Logistic Regression classifier performs best by achieving 71% of accuracy. The Naïve Bayes and Random Forest classifier has achieved 62% and 67% accuracy respectively.

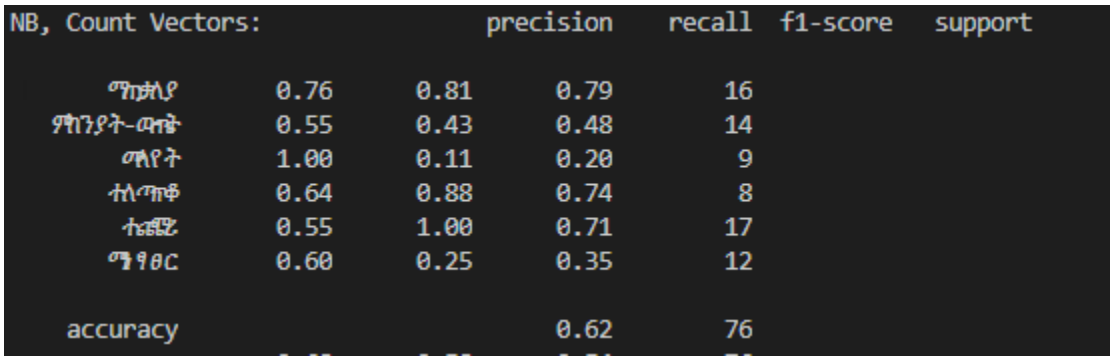
LR, Count Vectors:		precision	recall	f1-score	support
ማጠቃለያ	0.74	0.88	0.80	16	
ምክንያት-ውጤት	0.55	0.79	0.65	14	
መለየት	0.80	0.44	0.57	9	
ተለጣጥቆ	0.58	0.88	0.70	8	
ተጨማሪ	0.89	0.94	0.91	17	
ማነፃፀር	1.00	0.17	0.29	12	
accuracy			0.71	76	

Figure 5-5 Sense Classification with Logistic Regression



RF, Count Vectors:	precision	recall	f1-score	support
ማክለያ	0.78	0.88	0.82	16
ምክንያት-ውጤት	0.50	0.79	0.61	14
መለየት	1.00	0.22	0.36	9
ተለጣቀ	0.56	0.62	0.59	8
ተጨማሪ	0.74	1.00	0.85	17
ግዥፀር	1.00	0.17	0.29	12
accuracy			0.67	76

Figure 5-6 Sense Classification Result with Random Forest



NB, Count Vectors:	precision	recall	f1-score	support
ማክለያ	0.76	0.81	0.79	16
ምክንያት-ውጤት	0.55	0.43	0.48	14
መለየት	1.00	0.11	0.20	9
ተለጣቀ	0.64	0.88	0.74	8
ተጨማሪ	0.55	1.00	0.71	17
ግዥፀር	0.60	0.25	0.35	12
accuracy			0.62	76

Figure 5-7 Sense Classification with Naive Bayes

5.5 Discussion

The following screenshots show sample of texts that are correctly identified using the two trained CRF models and the Logistic Regression classifier. The samples consist of correctly parsed sentenece i.e the arguments, connective and relation is correctly identified and classified. In addition, the time it took for each parsing is give using linux time command. Based on the length of the word count a time between 1 minute and 14 seconds and 3 minute and 41 seconds is registered for the sample sentences.

Example-1

አጼ ቴዎድሮስ ከእንግሊዞች ጋር ጦርነት ገጥመው እጄን አልሰጥም ብለው የራሳቸውን ራስ በራሳቸው ሽጉጥ ከጨለጡዋት በኋላ የሥልጣን ተተኪያቸው የሆኑት አጼ ተክለ ጊዮርጊስ ነበሩ

```
አጼ Arg1-B ቴዎድሮስ Arg1-I ከእንግሊዞች Arg1-I ጋር Arg1-I ጦርነት Arg1-I ገጥመው Arg1-I እጄን Arg1-I አ
ልሰጥም Arg1-I ብለው Arg1-I የራሳቸዉ Arg1-I ራስ Arg1-I በራሳቸው Arg1-I ሽጉጥ Arg1-I ከጨለጡዋት Arg1-I
 በኋላ Arg1-E የሥልጣን Arg2-B ተተኪያቸው Arg2-I የሆኑት Arg2-I አጼ Arg2-I ተክለ Arg2-I ጊዮርጊስ Arg2-I
 ነበሩ Arg2-E ,በኋላ

['ተለጣጥቆ']

real    2m30.315s
```

Figure 5-8 Screenshot of correctly parsed sentense with the sense 'ተለጣጥቆ '

Example 2

ተኩሱ አንድ ሰዓት ያህል እንደ ቆየ ቀስ እያለ እየቀነሰና እየራቀ ሄደ በመጨረሻም ከናካቴው ቆመ

```
ተኩሱ Arg1-B አንድ Arg1-I ሰዓት Arg1-I ያህል Arg1-I እንደ Arg1-I ቆየ Arg1-I ቀስ Arg1-I እያለ
Arg1-I እየቀነሰና Arg1-I እየራቀ Arg1-I ሄደ Arg1-E በመጨረሻም Arg2-B ከናካቴው Arg2-I ቆመ Arg2-E
 ,በመጨረሻም

['ማጠቃለያ']

real    1m14.860s
```

Figure 5-9 Screenshot of correctly parsed sentense with the sense 'ማጠቃለያ '

Example 3

በፍቅር የተገነባ ቤታችን ልጅ እሚጮህበት ባለመሆኑ ደስታችንን አደብዝዞት ቢቆይም ብዙ ከዘገየን በኋላ ሁለት ወንድ ልጆች ለማፍራት ችለናል

```
በፍቅር Arg1-B የተገነባ Arg1-I ቤታችን Arg1-I ልጅ Arg1-I እሚጮህበት Arg1-I ባለመሆኑ Arg1-I ደስታችንን
Arg1-I አደብዝዞት Arg1-I ቢቆይም Arg1-E ብዙ Arg2-B ከዘገየን Arg2-I በኋላ Arg2-I ሁለት Arg2-I ወንድ
 Arg2-I ልጆች Arg2-I ለማፍራት Arg2-I ችለናል Arg2-E ,ቢቆይም

['መለየት']

real    1m31.502s
```

Figure 5-10 Screenshot of correctly parsed sentense with the sense 'መለየት '

Example 4

ፍትሃዊ ነፃ ሰላማዊ ዴሞክራሲያዊና በሕዝቡ ዘንድ ተዓማኒነት ያለው ምርጫ እንዲካሄድ ሙያዊ ግዴታችንን በአግባቡ እንወጣለን ብለዋል እንዲሁም ከፖለቲካ ፓርቲዎች ወገንተኝነት በፀዳ መልኩ ትክክለኛ ሚዛናዊና ግልፅነት ያለው መረጃ ለጎበረተሰቡ እናደርሳለን ሲሉም ባለሙያዎቹ በአቋም መግለጫቸው ገልፀዋል

```
ፍትሃዊ Arg1-B ነፃ Arg1-I ሰላማዊ Arg1-I ደማክራሲያዊና Arg1-I በሕዝቡ Arg1-I ዘንድ Arg1-I ተጓጓኔነት Arg1-I ያለ
ው Arg1-I ምርጫ Arg1-I እንዲከሄድ Arg1-I ማዋዊ Arg1-I ግዴቻችንን Arg1-I በአግባቡ Arg1-I እንወጣን Arg1-I ብለዋ
ል Arg1-E እንዲሁም Arg2-B ከፖለቲካ Arg2-I ፓርቲዎች Arg2-I ወንጮኝነት Arg2-I በፀዳ Arg2-I መልኩ Arg2-I ትከከለ
ኛ Arg2-I ማዘናዊና Arg2-I ግልፅነት Arg2-I ያለው Arg2-I ሚኛ Arg2-I ለጎበረቱሰቡ Arg2-I እናደርሳለን Arg2-I ሰሉ
ም Arg2-I ባለመዋምቹ Arg2-I በእቁም Arg2-I መለሳቹው Arg2-I ገልፀዋል Arg2-E ,እንዲሁም

['ተጨማሪ']

real    3m41.526s
```

Figure 5-11 Screenshot of correctly parsed sentense with the sense '*ተጨማሪ* '

Example 5

እንደዚህ ያለውን የብልሆችንና የትጉሆችን ኑሮ ሲያዩት መፈጠርን ያስመሰግናል ባንድ ወገን ደግሞ የሰነፎችንና የንዝህላሎችን ሁናቴ ሲመለከቱት ተፈጥሮን ያስጠላል

```
እንደዚህ Arg1-B ያለውን Arg1-I የብልሆችንና Arg1-I የትጉሆችን Arg1-I ኑሮ Arg1-I ሲያዩት Arg1-I መፈጠርን Arg1-I
ያስመሰግናል Arg1-I ባንድ Arg1-E ወገን Arg2-B ደግሞ Arg2-I የሰነፎችንና Arg2-I የንዝህላሎችን Arg2-I ሁናቴ Arg2-I
 ሲመለከቱት Arg2-I ተፈጥሮን Arg2-I ያስጠላል Arg2-E ,ደግሞ

['ማነፃፀር']

real    1m36.172s
```

Figure 5-12 Screenshot of correctly parsed sentense with the sense '*ማነፃፀር*'

Example 6

በደርግ የአገዛዝ ሥርዓት የሃይማኖት ነፃነት አልነበረም በመሆኑም ዜጎች እምነታቸውን ከደው የእሱን ርዕዮት-ዓለም እንዲከተሉ በኃይል የተገደዱበት አጋጣሚ የቅርብ ጊዜ ታሪካችን ነው

```
በደርግ Arg1-B የአገዛዝ Arg1-I ሥርዓት Arg1-I የሃይማኖት Arg1-I ነፃነት Arg1-I አልነበረም Arg1-E በመሆኑም Arg2-B
ዜጎች Arg2-I እምነታቸውን Arg2-I ከደው Arg2-I የእሱን Arg2-I ርዕዮት-ዓለም Arg2-I እንዲከተሉ Arg2-I በኃይል Arg2-I
 የተገደዱበት Arg2-I አጋጣሚ Arg2-I የቅርብ Arg2-I ጊዜ Arg2-I ታሪካችን Arg2-I ነው Arg2-E ,በመሆኑም

['ምክንያት-ውጤት']

real    1m33.370s
```

Figure 5-13 Screenshot of correctly parsed sentense with the sense '*ምክንያት-ውጤት*'

Unlike the above samples we have also encountered incorrectly parsed sentences. Incorrectly parsed texts are text in which the argument is incorrectly identified or connective is incorrectly identified or sense is classified incorrectly. Certain samples of incorrectly parsed sentences is give below

Example 1

አዝዞ መከበር የለም መከበር የሚቻለው ሠርቶ ነው ስለዚህ ሁሉም ሰው ለመበልፀግና ለመከበር ተግቶ መሥራት አለበት የሚል ፅኑ እምነት አለኝ

```
አዝዞ Arg1-B መከበር Arg1-I የለም Arg1-I መከበር Arg1-I የሚቻለው Arg1-I ሠርቶ Arg1-I ነው Arg1-E ስለዚህ
 Arg2-B ሁሉም Arg2-I ሰው Arg2-I ለመበልፀግና Arg2-I ለመከበር Arg2-I ተግቶ Arg2-I መሥራት Arg2-I አለበት
 Arg2-I የሚል Arg2-I ፅኑ Arg2-I እምነት Arg2-I አለኝ Arg2-E ,ስለዚህ

['መለየት']

real    1m56.411s
```

Figure 5-14 Screenshot of Incorrectly Parsed sentense with the sense '*መለየት*'

In figure 4-12 even if the arguments and discourse connective are correctly identified, the sense is classified incorrectly as '*መለየት*' while the correct sense is '*ማጠቃለያ*'.

Example 2

ከፊት የቀደሙት ቃፈር ፈረሰኞች ነበሩ ቀጥሎ ብዛት ያለው እግረኛ ጦር ተከተለ

```
ከፊት Arg1-B የቀደሙት Arg1-I ቃፈር Arg1-I ፈረሰኞች Arg1-I ነበሩ Arg1-I ቀጥሎ Arg1-E ብዛት Arg2-B ያለው
 Arg2-I እግረኛ Arg2-I ጦር Arg2-I ተከተለ Arg2-E ,ቀጥሎ

['ተለጣጥቆ']

real    0m55.164s
```

Figure 5-15Screenshot of Incorrectly Parsed sentense with the sense '*ተለጣጥቆ*'

In figure 4-13 while the sense is classified correctly as '*ተለጣጥቆ*' and the discourse connective is identified correctly the span of the first argument is incorrectly identified. The correct span of the first argument ends at the word '*ነበሩ*'.

In this chapter the evaluation metrics used for evaluation is introduced and results are given. In addition, sample outputs of correctly and incorrectly parsed sentence area discussed. Even if the results are encouraging, the incorrectly parsed sentences clearly show that there is a lot of room for improvement in the future.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

In this study, the design and development of Amharic shallow discourse parser is described. The thesis began with a brief discussion on the concept and applications of discourse, discourse structure and discourse parsing. In this discussion, it is indicated text is not just a sequence of sentences but an entity that has a complex structure having a vrious meaningful relations called discourse relation. In addition, it is shown that knowledge and understanding of this structure is very useful for the development of various NLP applications.

Consequently, a literature review is conducted on various theoretical frame works of discourse that are used as a basis for the manual annotation of discourse tree banks and development of automatic discourser parsers. In addition, the prominent styles of automatic discourse parsing and the approaches used are discussed. Finally, research works on Amharic discourse are reviewed in consideration with identifying language specific features of Amharic to be used in designing the and development of the parser.

Next, the preparation and analysis of the corpus used for the study is described in detail. Due to the absence of a discourse annotated corpus of Amharic, the corpus is prepared by the researcher in collaboration of linguistic experts. This discourse annotated corpus is used as an input for the parser. The thesis has also presented the model and approach used to design and develop the Amharic shallow discourse parser. The visual studio development platform with sklearn python package and Habit Tagger with HornMorpho is used as a development tool.

Experiments were conducted on the parser by splitting the corpus in to training and testing dataset by 70/30 ratio. Based on the results from the experiemtns the discourse parser has given encouraging results. However, there is a lot that has to be done to improve the discourse parser.

6.2 Contribution of this work

In this thesis, two important contributions are made to the field of Amharic discourse parser. In chapter 4, the architecture of a PDTB style sentence level Amharic discourse parser is given along with components and their specific role. In addition, three algorithms that are used for discourse connective identification, argument labeling and sense labeling are developed. To the best of our

knowledge there is no discourse parser designed or algorithms developed specifically for Amharic so far.

In Chapter 5, we have explained the preparation of the discourse corpus used for experiment purposes in which a corpus consisting of 250 texts is prepared. This corpus can be used in future works to improve the proposed system or develop a new one buy applying alternative approaches. Therefore, we believe design of Amharic discourse parser architecture and algorithm and preparation of Amharic discourse corpus are the major contributions of this study.

6.3 Recommendation

Discourse parsing is a difficult task, which requires more time and needs more linguistic resources to make it full-fledged. Hence, further improvements and modifications are required. Below, additional features that can be added to increase the performance of the system and future research directions are listed.

- Improving the current discourse parser by including syntactic information: The discourse parser proposed in this study can be improved by taking advantage of syntactic information gathered from syntactic parsers.

- Replicating the work in other local languages: Since Ethiopia is a multilingual country, implementing the sentence level discourse parser for languages other than Amharic is a potential future work.

- Preparation of manually annotated Amharic syntactic as well as discourse tree bank corpus: Even if the research in computational linguistics in Amharic language is improving from time to time, there is still a big gap that needs to be filled in relation to manually annotated corpus that can be used in natural language processing studies. Therefore, preparing such corpus can be a future work.

- Text level discourse parser: The discourse parser presented here only works for discourse relations within a sentence or between two sentences. However, discourse structure is not limited to sentence level but spans paragraphs and the whole text. Therefore, developing a text level discourse parser for Amharic is also another future work.

References

[1] D. Jurafsky and J. Martin, “An Introduction to Language Processing, Computational Linguistics and Speech Processing”, 2007, pp. 783-784.

[2] D. Marcu, “The Theory and Practice of Discourse Parsing and Summarization”, London, England: The MIT Press, 2000, pp. 1-2.

[3] V. Feng, “RST-Style Discourse Parsing and its Applications in Discourse Analysis”, Unpublished PhD Dissertation, University of Toronto, Canada, 2015.

[4] H. Hernault, D. Bollegala, and M. Ishizuka, “A Sequential Model for Discourse Segmentation”, in Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics, 2010, pp.315-326.

[5] W. Mann and S. Thompson, “Rhetorical Structure Theory: Toward a Functional Theory of Text Organization”, Text 8, 1988, pp. 243–281.

[6] H. Hernault, H. Predinger, D. duVerle, and M. Ishizuka, “HILDA: A Discourse Parser Using Support Vector Machine Classification”, Dialogue and Discourse, Vol. 1, No. 3, 2010, pp. 1-33.

[7] R. Soricut and D. Marcu, “Sentence Level Discourse Parsing Using Syntactic and Lexical Information”, in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003, pp. 149-156.

[8] D. Marcu, “Building Up Rhetorical Structure Trees”, in Proceedings of the National Conference on Artificial Intelligence, 1996, pp 1069–1074.

[9] H. Le Thanh, G. Abeysinghe, and C. Huyck, “Generating Discourse Structures for Written Texts,” in Proceedings of the 20th International Conference on Computational Linguistics, 2004, pp 329–335.

[10] T. Pardo and M. Nunes, “On the Development and Evaluation of a Brazilian Portuguese Discourse Parser”, Latin America Learning Technology Journal, Vol. 15, No. 2, 2010, pp. 43-64.

[11] R. Subba and B. Eugenio, “An Effective Discourse Parser That Uses Rich Linguistic Information”, in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009, pp. 566–574.

[12] V. Feng and G. Hirst, “Text-level Discourse Parsing with Rich Linguistic Features”, in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2012), 2012, pp. 60–68.

[13] Z. Lin, H. Ng, and M. Kan, “A PDTB-styled end-to-end Discourse Parser”, Natural Language Engineering, Vol. 1, No.1, 2012, pp. 1–35.

[14] T. Sanders, "Toward a Taxonomy of Coherence Relations", Discourse Processes, Vol. 15, No. 1, 1992, pp.1-35.

[15] J. Hobbs, “On the Coherence and Structure of Discourse”, Center for the Study of Language and Information (CSLI), Stanford University,1985.

[16] B. Webber, M. Egg, and V. Kordoni, “Discourse Structure and Language Technology”, Natural Language Engineering, Vol. 18, No. 4, 2012, pp. 47-490.

[17] B. Fraser, “What are Discourse Markers?”, Journal of Pragmatics, Vol. 31, No. 7, 1999, pp. 931-952.

[18] A. Knott and T. Sanders, “The Classification of Coherence Relations and their Linguistic Markers: An Exploration of Two Languages”, Journal of Pragmatics, Vol. 30, No. 2, 1998, pp. 135-175.

[19] A. Knott. “A Data-Driven Methodology for Motivating a Set of Coherence Relations”, Unpublished PhD Thesis, University of Edinburgh, Edinburgh, 1996.

[20] A. Alsaif, “Human and Automatic Annotation of Discourse Relations for Arabic”, Unpublished PhD Thesis, University of Leeds, School of Computing, Leeds, 2012.

[21] N. Asher, “Reference to Abstract Objects in Discourse”, Boston, MA, Kluwer Academic Publishers, 1993.

[22] W. Mann and A. Thompson, “Rhetorical Structure Theory: A Theory of Text Organization”, Technical Report, ISI/RS- Information Sciences Institute, 1988.

[23] L. Carlson, D. Marcu, and M. Okurowski, "Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory", In Proceedings of the Second SIGdial Workshop on Discourse and Dialogue, 2001.

[24] R. Girju, "Automatic Detection of Causal Relations for Question Answering", In Proceedings of the ACL 2003, Workshop on Multilingual Summarization and Question Answering - Machine Learning and Beyond , Sapporo, Japan, 2003, pp. 76-83.

[25] S. Williams and E. Reiter, “A Corpus Analysis of Discourse Relations for Natural Language Generation”, In Proceedings of Corpus Linguistics, Lancaster University, 2003, pp. 899-908.

[26] M. Hearst, “Text Tiling: Segmenting Text into Multi-paragraph Subtopic Passages”, Computational Linguistics, Vol. 23, No. 1, 1997, pp. 33-64.

[27] J. Einstein, “Hierarchical Text Segmentation from Multi-scale Lexical Cohesion”, In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009, pp. 353-361.

[28] S. Joty, G. Carenini, and R. Ng, “A Novel Discriminative Framework for Sentence-Level Discourse Analysis”, in Proceedings of the Conference on Empirical Methods in Natural Language Processing and the Conference on Natural Language Learning, 2012, pp. 904-915.

[29] S. Li, R. Li, and E. Hovy, “Recursive Deep Models for Discourse Parsing”, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 2061–2069.

[30] S. Li, L. Wang, Z. Cao, and W. Li, “Text-level Discourse Dependency Parsing”, in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1 , 2014, pp. 25–35.

[31] S. Joty, G. Carenini, and R. Ng, “Combining Intra- and Multi-Sentential Rhetorical Parsing for Document-Level Discourse Analysis”, in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013, pp. 486-496.

[32] C. Goller and A. Kuchler, “Learning Task-Dependent Distributed Representations by Back Propagation Through Structure”, in Proceedings of IEEE International Conference on Neural Networks, Vol. 15, 1996, pp. 347–352.

[33] L. Carlson and D. Marcu, “Discourse Tagging Reference Manual”, ISI Technical Report, ISI-TR-545, 2001.

[34] ጌታቸው እንዳላማው፡፡ (1987)፡፡ የአማርኛ ዲስኩር አመልካቾች ስርጭትና ተግባር በተማሪዎች በተጻፉ ትረካዊ ውሁድ አሃዶች፡፡ አዲስ አበባ፤ አዲስ አበባ ዩንቨርስቲ፡፡

[35] ባዬይማም፡፡ (1986)፡፡ አማርኛ ሰዋሰው፡፡ አዲስ አበባ፤ ት.መ.ማ.ድ

[36] መርስኤጎዘን ወልደቂርቆስ፡፡ (1948)፡፡ ያማርኛ ሰዋሰው፡፡ አዲስአ በባ፤ አርቲስቲክ ማተሚያ ድርጀት፡፡

[37] ተክለማርያም ፋንታዬ ፡፡ (1944)፡፡ ሆህተፅሁፍ ዘስነጥበብ፡፡ አዲስ አበባ፤ ትንሳኤ ዘጉባኤ ማተሚያ ቤት፡፡

[38] ጌታቸው እንዳላማው፡፡ (2007)፡፡ የአማርኛ ጽሁፋዊ ዲስኩር አመልካቾች ሰዋሰዋዊ ከፍልና ዲስኩራዊ ትንተና፡፡ አዲስ አበባ፤ አዲስ አበባ ዩንቨርስቲ፡፡

[39] ደረጀ ገብሬ፡፡ (1996)፡፡ ተግባራዊ የጽሕፈት ክሂል መማሪያ፡፡ አዲስ አበባ፤ ንግድ ማተሚያ ድርጀት፡፡

[40] M. Halliday and R. Hassan, “Cohesion in English”, London, England: Longman Press, 1976.

[41] M. Taboda and W. Mann, "Rhetorical Structure Theory: Looking Back and Moving Ahead", Discourse Studies, Vol 8. 2006.

[42] D. Marcu, L. Carlson and M, Watanabee "The automatic translation of discourse structures", in the proceedings of Association for Computational Linguistics, Vol 1, 2000.

[43] E. Hovy, "Autamated Discurse Generation Using Strycture Relations", Artificials Intelligence, Vol 63, Issue 1-2, pp 341-386, 1993.

[44] L. Polanyi,C. Culy,V.D Berg,M.A, Thione,D Lorrenzo, D. AHN "A Rule Based Approach to Discourse Parsing", in the proceedings of Global Workshop on Discourse and Dialogus, Vol 5, pp 108-117, 2004.

[45] F. Wolf,E. Gibson "Representing discourse coherence:A corpus based study", Computational Linguistics, Vol 31, issue 2, pp 24-287, 2005.

[46] B.J. Grosz, C.L. Sidner "Attention, Intention and the Structure of Discourse", Computational Linguistics, Vol 12, Issue 3, pp 175 -204, 1986.

[47] A. Lascarides, N. Asher "Segemented Discourse Representation Theory", Computing Meaning , pp 87 -124, 2003.

[48] B. Webber "D-LTAG:extending lexicalized TAG to Discourse", Cognitive Science, Vol 28, Issue 5, pp 751-779,2004.

[49] B. Webber "Acounting for Discourse Relations: Constituency and Dependency", Intelligent Linguistic Architectures, CSLI Publications, pp 339-360, 2006.

[50] B. Webber, M. Stone, A. Joshi and A.Knott "Anaphora and Discourse Structure", Computational Linguistics, Vol 29, Issue 4, pp. 545-587, 2003

[51] K,Forbes-Riley, B. Webber and A.Joshi "Computing Discourse Semantics:The Predicate-Argument Semantics of Discourse Connectives in D-LTAG", Journal of Semantics, Vol 23, Issue 1, pp. 55-106, 2006.

[52] R. Prasad, N. Dinesh,A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, B. Webber "The Penn Discourse TreeBank 2.0",in the proceedings of the international Conference on Language Resources and Evaluation,2008.

[53] B. Webber and R.E.A. Prasad "The Penn Discourse TreeBank 1.0 Annotation Manual",University of Pennsylvania:Institute for reserach in Cognitive science,2006.

[54] M. Buch-Kromann,I.Korzen "The Unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks",in the proceedings of Linguistic Annotation, Vol 4, pp 127-131, 2010.

[55] C. Chiarcos and N. Schenk " A Minimalist Approach to Shallow Discourse Parsing and Implicit Relation Recognition", in the proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task, pp 42–49, 2015.

[56] N. Xue, H. Tou Ng S. Pradhan, R. Prasad, C. Bryant, and A. T. Rutherford “The CoNLL-2015 Shared Task on Shallow Discourse Parsing”, in the proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task, pp 1–16, 2015.

[57] N. Xue, H. Tou Ng S. Pradhan, B. Webber, C. Wang, H. Wang, and A. T. Rutherford “The CoNLL-2015 Shared Task on Shallow Discourse Parsing”, in the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp 1–19, 2016.

[58] S. Ghosh “End-to-End Discourse Parse using Cascaded Structured Prediction”, University of Trento, Italy.

[59] E. Pitler, A. Louis, A. Nenkova “Automatic sense prediction for implicit discourse relations in text”, In Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore, 2009, pp. 683-691.

[60] H. Chen, W. Liao, H. Huang and H, Chenn “Fine grained Chinese Discourse Relation Labeling”, National Taiwan University, Taiwan.

[61] U. Sidarennka, A. Peldsus and M, Stede “ Discourse segmentationof German Texts”, Journal

of Languages and Computational Linguistics,Vol. 30, Issue 1, pp. 71-98.

[62] I. Cunha, E. SanJuan, J. Torres-Moreno, M, Cabre, and G. Sierra "A symbolic Approach for Automatic Detection of Nuclearity and Rhetorical Relations among Intra-sentence Discourse Segments in Spanish", International Conference on Computational Linguistics and Intelligent Text Processing, pp.462-474,2012.

[63] F. Pedregosa et al, "Scikit-learn: Machine Learning in Python" , Journal of Machine Learning Research, Vol. 12, pp. 2825-2830, 2011.

Appendix A

A total of 53 mononuclear and 25 multinuclear rhetorical relations are listed here. Table 1 below is a complete listing of all the relations, arranged alphabetically by mononuclear relation. Mononuclear relations are listed in Column 1 if the satellite is the unit that characterizes the relation name. For example, in a BACKGROUND relation, the satellite provides background information for the situation presented in the nucleus. Mononuclear relations listed in Column 2 are those in which the nucleus characterizes the relation name. For example, in a CAUSE relation, the nucleus is the cause of the situation presented in the satellite. Column 3 lists the multinuclear relations. Corresponding mono and multinuclear relations are shown across a single row. (In some cases, this results in the multinuclear relations appearing out of alphabetical order.)

Mononucleus (Satellite)	**Mononucleus (Nucleus)**	**Multinuclear**
analogy		Analogy
antithesis		Contrast
attribution	attribution-n	
background		
	Cause	Cause-Result
circumstance		
comparison		Comparison
comment		
		Comment-Topic
concession		
conclusion		Conclusion
condition		
consequence-s	consequence-n	Consequence
contingency		
		Contrast (see antithesis)
definition		
		Disjunction
elaboration-additional		
elaboration-set-member		
elaboration-part-whole		
elaboration-process-step		
elaboration-object-attribute		
elaboration-general-specific		
enablement		

Mononucleus (Satellite)	**Mononucleus (Nucleus)**	**Multinuclear**
evaluation-s	evaluation-n	Evaluation
evidence		
example		
explanation-argumentative		
hypothetical		
interpretation-s	interpretation-n	Interpretation
		Inverted-Sequence
		List
manner		
means		
otherwise		Otherwise
preference		
problem-solution-s	problem-solution-n	Problem-Solution
		Proportion
purpose		
question-answer-s	question-answer-n	Question-Answer
reason		Reason
restatement		
	result	Cause-Result
rhetorical-question		
		Same-Unit
		Sequence
statement-response-s	statement-response-n	Statement-Response
summary-s	summary-n	
	temporal-before	
temporal-same-time-s	temporal-same-time-n	Temporal-Same-Time
	temporal-after	
		TextualOrganization
		Topic-Comment
topic-drift		topic-drift
topic-shift		topic-shift

Appendix B

1. ወንዶቹ ብቻቸውን ማደን ሲችሉ እናታቸው ታባርራቸዋለች ሴቶቹ ግን ረዘም ላለ ጊዜ ከእናታቸው ጋር አብረው ይቆያሉ
2. አልፎ አልፎ እንደማንኛችንም ይታመማል ግን እስከዚህ ሊዘልቅ የሚችል ቋሚ ሕመም አላውቅበትም
3. ቡና አይታየኝም ቡና ግን ይሸተኛል
4. ባዶ የሆነ ሰው ስለራሱ ብዙ ጉራዎችን መንዛቱ መታወቂያው ነው ስለራሳችን ጉራ ስንነዛ ግን ባዶነታችንን ትልቅ ማድረጋችን እንደሆነ እናስተውል
5. ያጣሁትን ነገር ያጣሁት በድክመቴ ነው ብዬ ነው የማምነው ስለዚህ ከባለቤቴ ጋር በተያያዘ የሚሰጥ መሰል አስተያየት ብዙም አልጎዳኝም አይጎዳኝም
6. እኛ ደግሞ ልጆቹን ትምህርት አስጀምረን አቋርጡ ማለት አንችልም ስለዚህ አሁን የምንፈልገው የአራት ዓመቱን አንዴ እንዲሰጡንና እርፍ ብለን እንድናስተምር ነው
7. አሁን መለስ አርፏል ተብሎ ተነግሮናል ስለዚህ የሞተን ሰው የምንወቅስበት ጊዜ አይደለም
8. ሆስፒታሉን ስንመለከት የእናቶች የጤና አገልግሎት አይሰጥም ስለዚህ አዋላጅ ነርሶቻችን የተግባር ሥልጠና እንዲያደርጉ ስንፈልግ የምንወስደው የእናቶችና የሕፃናት ሕክምና አገልግሎት ወደሚሰጡ ሆስፒታሎች ነው
9. ይህ ፋብሪካ ጂፕሰም ቦርድ በማምረት በአገራችን የመጀመሪያው ነው ስለዚህ ለማስተዋወቅና ለማስለመድ ብዙ ሥራ ይፈልጋል
10. አዲሱ ባል አሮጌውን ባል እጅግ አድርጉ ይፈራዋል ስለዚህ በዋለበት አይውልም
11. በዕድሜና በሰውነት ከሁሉም ትንሽ እኔ ስለነበርኩ የሚሰጠኝ ልብስ እየሰፋኝ እቸገር ነበር ስለዚህ በቀበቶ ታስሮ እንዲጠብ ይደረጋል
12. ማህበረሰቡ አሸባሪ አለመሆናችንን ያውቃል ስለዚህ ምንም የሚፈጠር ችግር የለም
13. መድኃኒቱ ሲያልቅባት ሆስፒታል ሄዳ አልወሰደችም ስለዚህ ይጥላት ጀመር
14. ውሻህ የሚያሳክክ ቅንቅን አለበት ስለዚህ ቪታሚን ገዝተህ ስጠው
15. ትምህርቴ ዘመናዊው ላይ ቢሆንም ለመሥራት የምፈልገው የኢትዮጵያን የባህል ልብስ ነው
16. የዕርዳታ ድርጅቶች የሚያደርጉት እገዛ ጥሩ ቢሆንም ዕርዳታ ለማግኘት የሚጠቀሙበት መንገድ የዕርዳታን መጥፎ ገጽታ ያሳያል
17. ምንም እንኳን የዘመናዊ ባንክ ሥርዓት የ20ኛው ክፍለ ዘመን ውጤት ቢሆንም ኢትዮጵያ የዘመናዊ ገንዘብ ግብይት ሥርዓትን የጀመረችው እጅግ ቀደም ብላ ነበር
18. ምንም እንኳን ወታደሮቻቸው ሕይወታቸውን እየሰው ቢሆንም ፕሬዚዳንት አል አሳድ የመጨረሻው መፍትሔ በጦርነት አውድማው ላይ ይገኛል ሲሉ ጦርነቱ የሚቀጥል መሆኑን በቀጥታ ተናግረዋል
19. እናትና አባቴ በልጅነቴ ስለተለያዩ ከሁለቱም ጋር በመፈራረቅ መኖር ግዴታ ሆኖብኝ ነበር
20. ቀዝቅዞና የመከራ ጭጋግ ሸፍኖት ወደ ነበረው መኖሪያ ቤቴ ስደርስ ልጆቼ አድገው ባለቤቴም ምንም እንኳን ብቻዋን ልጆች ይዛ ያሳለፈችው ዘመን ከባድ ሆኖባት እንደቆየ ብረዳም ጤንነቷና ሁኔታዋ ደህና ሆኖ ስለጠበቀኝ በህይወት የመኖር ተስፋዬ በራ

21. ዛሬ የሚካሄደው የኢህአዴግ ስራ አስፈፃሚ ኮሚቴ ስብሰባ ከሹምሽሩ ጋር የተያያዘ እንደሆነ ቢታሰብም የሰብስባው ዋና ርዕሰ ጉዳይ አመታዊ ግምገማና አመታዊ እቅድ እንደሆነ ተነግሯል
22. መንግሥት ለረጅም ዘመን ሙስናን የመዋጋት ቁርጠኝነትም ቢያሳይም በዚያው ልክ ግን ችግሩ በአገራችን አሳሳቢ ደረጃ ላይ አልደረሰም ሲል ኖሯል
23. በእኛ በኩል ካልሆነ ከእግዚአብሔር ጋር መገናኘት አትችሉም የሚሉ ሰዎች ቢኖሩም ማናቸውንም ነገር በእናንተ በኩል እንድናደርግ ስላልተነገረን ዋሾዎች ናችሁ እንላለን
24. የገጠር ወጣቶቹ በትምህርት ቤት ቆይታቸው በግብርና ሙያ የተወሰነ እውቀት ቢኖራቸዉም ለእርሻ ሥራቸው የእርሻ በሬዎች የአርሻ ማሣሪዎችና የምርጥ ዘር ያስፈልጋቸዋል
25. መንግሥት ከሀገራዊ መፍትሔ በተጨማሪም ዓለም አቀፍ መፍትሔዎችንም እንዲተገብሩ እያደረገ ነው
26. በቤት አካባቢ ወይም ከዚያው ውጭ ልሠራው ያቀድኩት ምንም ነገር አልነበረም ስለዚህ ተኝቼ መጽሐፍ በማንበብና ሹራብና ዳንቴል በመሥራት የእረፍቴን ቀን አጋመስኩት
27. በቤታቸው ውስጥ ሌላ ማንም ረዳት ስለሌላቸው ማስታመምና ምግባቸውን ማብሰል የእኔው ግዴታ ሆነ ስለዚህ ሁኔታውን ለወላጆቼ አማክሬአቸው ይህኑኑ ግዴታዬን ለመወጣት ከእሳቸው ጋር መኖር ጀመርኩ
28. ቁስለኞቹ በጉዞ ላይ ብዙ ስለቆዩና የደእሰባቸውም አደጋ ቀላል ስላልነበረ ከመድከማቸው የተነሳ እንደ ልብ መነጋገር አልቻሉም ስለዚህ የምጠይቃቸውን ሁሉ አቅማቸው በፈቀደው መጠን ከመመለስ ወደ ኋላ ባይሉም ማስቸገሬ ሳይሰማኝ አልቀረም
29. ንጉሠ ነገሥቱም ፊት ለፊት በመግጠም የሚደርሰውን ጉዳት በማመዛዘን አደጋ እየጣሉ ብቻ እንጂ ፊት ለፊት ገጥመው እንዳይዋጉ አስፈላጊውም በቶሎ እንዲደርስላቸው መታሰቡን እየገለጽጡ መለሱላቸው ስለዚህ ወኔው የሚፈነቅለው ጀግና ሁሉ ሌት ተቀን ሳይል አደጋ እየጣለ የጀብዱ ሥራ ለመሥራት ይሽቀዳደም ጀመር
30. ከጥቂት ጊዜ በኋላ ግን ተኳሹ ሳይታወቅ በኛ ሠራዊት መካከል ከፀት የማይበልጡ የጠመንጃ ተኩሶች ተሰሙ ስለዚህ ወታደሩ ለጠብ አኩብኩቦ በሚጓዝበት ጊዜ እነዚህ ተኩሶች መሰማት አበረገገው
31. የትግራይን ምድር ጣሊያን በቦምብ አቃጥሎት ዛሬ እንኩዋንስ ሰው ወፍ የሚሸሸግ ጫካና ቁጥቁዋጦ አይገኝም ስለዚህ ጦሩ አንድ ስራ ሳይሰራ በሰፈረበት ባይሮፕላን ተደብድቦ ሳያልቅ ቀድሞ ባደጉናን ለመምታት ማሰቡ ማለፊያ አሳብ ነው
32. ሚበዛው የኢትዮጵያ ህዝብ ስላልተማረ መንግስትን ከሱ የተለዬ እሱን እንዲገዛ እግዚአብሔር በላዩ የጫነበት ባለስልጣን አድርጎ ነው የሚያየው ስለዚህ የኢትዮጵያ መንግስት ቢገዛው የጣሊያን መንግስት ቢገዛው ልዩነት ያለው መስሎ አይታዬውም
33. በእርግጥም አሉላ ያን እጅግ አስፈላጊ የነበረ ምሽግ ከያዙ እንግሊዝ በግብፅ ላይ ታዝላ ይዛው የነበረውን ጥቅምና ክብር ወደ ማጣቱ ልትደርስ ሆነ ስለዚህ ከዚያ ውድቀት ላይ ከመድረሱዋ በፊት ነገሩን በዲፕሎማሲ ማለዘቡን ሁለቱም ወገኖች ስለ አመኑበት የሱዳን ጠቅላይ አገረ ገዥ የነበረው ጐርደን ከራስ አሉላ ጋር ለመነጋገር እንዲችል ፈቃድ የሚጠይቅ ደብዳቤ ጻፈላቸው

34. የእንግሊዝ እጅ ግብፅን ይዞ ከአጼ ዮሐንስና ከራስ አሉላ ዘንድ ከስምምነት ለመድረስ የጀመረውን ስልት የፈረንሣይ ዐይን ደግሞ በቅናት ይመለከተው ጀመር ስለዚህ እንግሊዞች ከአጼ ዮሐንስ ቤተ መንግሥት ቀርቶ ከራስ አሉላ በር እንኳ እንዳይደርሱ ለመከላከል የፈረንሳይ መንግሥት ጥረት ጀመረ
35. የግብጽ ወታደሮች ሰሕጢ ላይ እንደ ሠፈሩ እንዲቆዩ አጼ ዮሐንስ ከአድሚራል ሒዊት ጋር የተስማሙት ከአሥመራ ወደ ምፅዋ የሚመላለሱት ነጋዴዎች በወስላቶች እንዳይቸገሩ ለመከላከልና ለማስጠበቅ ሲሉ ነበር ስለዚህ ግብጾች ቀበሌዎቹን ከለቀቁላቸውና አሉላ ራሳቸው ጦር አሥፍረው ለማስጠበቅ እንደማይቸገሩ የተረዱት እንግሊዞች በቀበሌው ኢጣሊያኖች ቢሠፍሩ ራስ አሉላ በዝምታ እንደማይመለከቱት ተረዱት
36. በምሥራቅ ሱዳን የተነሣውን የማሕዲስቶች ንቅናቄ እጅግ የፈሩት እንግሊዞች ያለ አሉላ ከንድ የሚገታው የለም» ብለው ወሰኑ ስለዚህ አሉላ ፈጥነው እንዲዘምቱ ውትወታቸውን ቀጠሉበትና ከስድስት መቶ በላይ የሚሆን ጠመንጃ አስይዘው ማርኮፖሎ ቤይ የተባለውን ወዳጃቸውን ለአማላጅነት ከምፅዋ ወደ አሥመራ ሰደዱት
37. ይህን ሁሉ ሀብት አከማችተን የምንሠራበት ጉዳይ ነው የቸገረን ስለዚህ ያከማቸነውን ሀብት ከእናንተ ጋር ለመካፈል ነው የመጣነው
38. ለወታደርም ለባላገርም ለነጋዴም ቢሆን ትልቅ ኩራት ያገሩ ነፃነት መሆኑ የማይጠረጠር ነው ስለዚህ ከመካከላችሁ ጠብና ተንኮል እንዲጠፋ ፍቅርና አንድነት እንዲሰፋ ሎሌ ለጌታው ጭፍራ ላለቃው መታዘዝን እንዲያውቅ በሚቻላችሁ ሁሉ ማስረዳት ነው
39. ደጃዝማች ኃይለ ሥላሴ ጉግሣ ከጂኔራል ሳንቲኒ ጋር ተገናኝቶ ለኢጣልያ ታማኝነቱን ካስረዳ በኋላ ከጠቅላይ አዝማቹ ከጀኔራል ደቦኖ ጋር እንዲነጋገር ጠየቀ ስለዚህ በማግሥቱ ፲ ሰዓት ከ፴፫ ደቂቃ ሲሆን በኩአቲት ከጀኔራል ዲቦኖ ጋር ተገናኘ
40. እኔ የኢጣልያ ወዳጅ ነኝ ስለዚህ የሥልጣኔ አምጭ የሆነውን የኢጣልያን ጦር ሠራዊት ሥራ በማናቸውም ረገድ አልቃወምም
41. እናንተ ተማሪዎች ብዙዎቻችሁ እንግዲህ የትምህርት ቤት ኑሮው በቅቶአቸው ከአስተማሪዎችና ከፈተናዎች ነፃ ልትወጡ ስትሆኑ ጥቂቶቻችሁ ግን ትምህርታችሁን በከፍተኞች ትምህርት ቤቶች ልትቀጥሉ ናችሁ ስለዚህ ይህ ቀን ለአብዛኛዎቻችሁ የመጨረሻው የተማሪነት ቀን ይሆን ይሆናል
42. እማማ የልጅ ልጅ ለማየት በጣም እንደምትጓጓ አውቃለሁ ስለዚህ የልጅ ልጅ ስታገኝ ፍቅርዋ አስተሳሰቧ በከፊል ወደርሱ እንደሚሳብ እርግጠኛ ነኝ

43. አንድ ኢትዮጵያዊ በጣም ወደ ራቁ ሀገሮች ቀርቶ ቅርብ ጎረቤቶቻችን ወደ ሆኑት ወደ ሱዳን ወደ ኬንያ ወደ ሱማሊያ ቢሔድ የሚታወቀው በኢትዮጵያዊነቱ እንጂ በአንድ ቀበሌ ባለቹው ትንሽ ርስቱ አይደለም ስለዚህ አንድ ሰው ለሀገሩ ጠቅላላ ደኅንነት ራሱ ሳይነካ የሚቀርብለትን ጠቃሚና አዳዲስ ሐሳብ ሁሉ የልማድ ሰንሰለቱን በጣጥሶ እየጣለ ሊቀበል ይገባዋል
44. አዲሷ ኢትዮጵያ የምትራመድበት ቅን ጐዳና በጣት የሚቈጠሩ ሰዎችን የሀብት ሥልጣን በጠቅላላው የሁሉ ነገር ጌቶች አድርጎ አብዛኛውን ሕዝቧን የውርደትና የችግር ሁሉ ተሸካሚ የሚያደርግ ሳይሆን ሀገሪቱ ለሚኖራት ለማንኛውም ነገር ጠቅላላው ሕዝብ ሙሉ ተሳታፊነት እንዲኖረው የሚያደርግ መሆኑ የተረጋገጠ ነው ስለዚህ እያንዳንዱ

ኢትዮጵያዊ ሀገሩ ለምትጠቀምበትም ሆነ ለምትጎዳበት ነገር ሙሉ ኃላፊነት ያለበት መሆኑን አምኖ ለግል ቤቱ መሟላት የሚያደርገውን ልባዊ ጥረት ለናት ሀገሩ ደኅንነትም ማድረግ ብሔራዊ ተግባሩ ነው

45. ፖሊሶች ወንበዴዎችን ከሰላም ሰዎች ለይቶ ለመምታት መብራት ስለሚያስፈልግና ተጨማሪ ረዳት ቢያስፈልግ ከጣቢያ ለመጥራት ስልክ ስለሚያስፈልግ እነዚህ በተቆረጡባቸው ሰፈሮች መግባት አይወዱም ነበር ስለዚህ ጩኸቱን ቢሰሙም ያልሰሙ የሚመስሉበት ጊዜ ይበዛ ነበር
46. የድሮ አለቃው አማር ቶሪጆስ ምክንያቱ ባልታወቀ ሁኔ በአውሮኘላን አደጋ ከሞተ በኋላ ለሁለት ዓመት ያህል መንገዱን ሲያደላድል ቆይቶ በኦገስት 1983 ሚስተር ኖሬጋ እንደ ጀኔራልና ዴ-ፋክቶ አምባገነን በመሆን የፖናማን መንግሥት ያዘ
47. በሕይወት እንድንኖር ወደዚያ ወርዳችሁ እህል ግዙልን አለዚያ ማለቃችን ነው'' አላቸው ስለሆነም አሥሩ የዮሴፍ ወንድሞች እህል ለመግዛት ወደ ግብፅ ወረዱ
48. ያዕቆብ የዚያን ቀን ምሽት ከእርሻ ሲመለስ ሊያ ወጥታ ተቀበለችውና ልጇ ባመጣው ፍሬ ስለተከራየሁህ ዛሬ የምትተኛው ከእኔ ጋር ነው አለችው ስለሆነም በዚያ ሌሊት ከእሷ ጋር አደረ
49. አጋፋሪ አሉላም ከራሳቸው አሽከሮች ሌላ በርካታ ወታደሮች ተጨምረውላቸው በአንድ አቅጣጫ አሰልፈው ሲዋጉ ዋሉ በመጨረሻም በለስ ቀናቸውና ራሳቸውን አጼ ተክለ ጊዮርጊስን ማርከው ለደጃዝማች ካሣ አቀረቡ
50. ተኩሱ አንድ ሰዓት ያህል እንደ ቆየ ቀስ እያለ እየቀነሰና እየራቀ ሄደ በመጨረሻም ከናካቴው ቆመ

Declaration

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

Declared by:

Name: _____

Signature: _____

Date: _____

Confirmed by advisor:

Name: _____

Signature: _____

Date: _____