AUTOMATIC DISAMBIGUATION OF TURKISH DISCOURSE CONNECTIVES BASED ON A TURKISH CONNECTIVE LEXICON

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

KEZBAN BAŞIBÜYÜK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

AUGUST 2021

Approval of the thesis:

AUTOMATIC DISAMBIGUATION OF TURKISH DISCOURSE CONNECTIVES BASED ON A TURKISH CONNECTIVE LEXICON

submitted by **KEZBAN BAŞIBÜYÜK** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. Ali Hikmet Doğru
Supervisor, **Computer Engineering, METU** _____

Prof. Dr. Deniz Zeyrek Bozşahin
Co-supervisor, **Graduate School of Informatics, METU** _____

Examining Committee Members:

Prof. Dr. Ferda Nur Alpaslan
Computer Engineering, METU _____

Prof. Dr. Ali Hikmet Doğru
Computer Engineering, METU _____

Prof. Dr. Nihan Kesim Çicekli
Computer Engineering, METU _____

Assoc. Prof. Dr. Gökhan Şengül
Computer Engineering, Atılım University _____

Assist. Prof. Dr. Selma Tekir
Computer Engineering, Izmir Institute of Technology _____

Date: 31.08.2021

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Kezban Başıbüyük

Signature :

ABSTRACT

AUTOMATIC DISAMBIGUATION OF TURKISH DISCOURSE CONNECTIVES BASED ON A TURKISH CONNECTIVE LEXICON

Başıbüyük, Kezban
Ph.D., Department of Computer Engineering
Supervisor: Prof. Dr. Ali Hikmet Doğru
Co-Supervisor: Prof. Dr. Deniz Zeyrek Bozşahin

AUGUST 2021, 148 pages

In this thesis, we developed methods for disambiguating the discourse usage and sense of connectives in a given free Turkish text. For this purpose, we firstly built a comprehensive Turkish Connective Lexicon (TCL) including all types of connectives in Turkish together with their syntactic and semantic features. This lexicon is built automatically by using the discourse relation annotations in several discourse annotated corpora developed for Turkish and follows the format of the German connective lexicon, DiMLex.

As in many other languages, Turkish has lexical connectives (referred to as single and phrasal connectives in this work), and it also includes suffixal connectives. We developed a rule-based Turkish Connective Disambiguator (TCD) in order to solve the usage ambiguity of single, phrasal and suffixal connective types. Then, we designed machine learning models to disambiguate the discourse usage and sense of connectives. We evaluated the TCD and the machine learning models by comparing their results with the human annotations in the Turkish section of the TED-Multilingual Discourse Bank and Turkish Discourse Bank 1.1. We observed that the machine learning approach

outperforms the baseline rule-based approach although both approaches yield quite good results.

Within the scope of this thesis, we developed user-friendly interfaces for the TCL and TCD programs. The TCL program lists the discourse connectives in Turkish with their features and it presents several filtering and analysis capabilities. The TCD program, on the other hand, loads the selected free Turkish text to its interface and marks the discourse and non-discourse occurrences of connectives in the text. Additionally, if the selected file has a corresponding annotation file, the program automatically evaluates the disambiguation results.

This thesis makes important contributions to Turkish discourse parsing by solving the usage ambiguity of the single and phrasal connectives as well as the suffixal connectives, which, to the best of our knowledge, has been attempted for the first time in this thesis. This thesis is also the first attempt to disambiguate the sense of all types of discourse connectives in Turkish. In this respect, it is predicted that the thesis would set baselines for future Turkish connective disambiguation works and pave the road for future researchers in the Turkish discourse parsing field.

Keywords: discourse processing, discourse relations, discourse connective disambiguation, discourse connectives, discourse connective lexicon, Turkish

ÖZ

TÜRKÇE SÖYLEM BAĞLAÇLARININ TÜRKÇE BAĞLAÇ SÖZLÜĞÜNE DAYALI OLARAK AYRIŞTIRILMASI

Başıbüyük, Kezban
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Ali Hikmet Doğru
Ortak Tez Yöneticisi: Prof. Dr. Deniz Zeyrek Bozşahin

Ağustos 2021 , 148 sayfa

Bu tezde, verilen herhangi bir Türkçe metindeki açık söylem bağlaçlarının kullanım ve anlam belirsizliğini çözmek için yöntemler geliştirdik. Bu amaçla, öncelikle Türkçe'deki tüm bağlaç türlerini sözdizimsel ve anlamsal özellikleriyle birlikte içeren kapsamlı bir Türkçe Bağlaç Sözlüğü (TCL) oluşturduk. Bu sözlük, söylem açısından işaretlenmiş çeşitli Türkçe derlemlerindeki söylem bağıntıları kullanılarak otomatik olarak oluşturulmuştur ve Almanca bağlaç sözlüğü olan DiMLex'in formatını kullanmaktadır.

Türkçe'de ilgeçler, yana sıralama bağlaçları ve zarflar (tek ya da birden fazla sözcük içeren yapılar) söylem bağlacı rolü üstlenebildiği gibi, ulaçlar da söylem bağlacı görevi yapmaktadır. Tek ya da birden fazla sözcük içeren bağlaç yapıları ve ulaç türündeki bağlaçların kullanım belirsizliğini çözmek için kural tabanlı bir Türkçe Bağlaç Ayırıcı (TCD) geliştirdik. Ardından, bağlaçların hem kullanım belirsizliğini hem de anlam belirsizliğini çözmek için çeşitli makine öğrenmesi modelleri tasarladık. TED-Multilingual Discourse Bank derleminin Türkçe bölümündeki ve Türkçe Söylem

Bankası (versiyon 1.1) derlemindeki, insan eliyle hazırlanmış söylem işaretlemeleri ile TCD ve makine öğrenmesi modellerinin sonuçlarını karşılaştırarak bir değerlendirme yaptık. Her iki yaklaşım da oldukça iyi sonuçlar vermekle birlikte, makine öğrenmesi yaklaşımının kural tabanlı yaklaşımdan daha iyi performans gösterdiğini gözlemledik.

Bu tez kapsamında TCL ve TCD programları için kullanıcı dostu arayüzler geliştirdik. TCL programı hem Türkçe'deki söylem bağlaçlarını özellikleriyle birlikte listeler hem de çeşitli filtreleme ve analiz yetenekleri sunar. TCD programı ise seçilen bir serbest Türkçe metni arayüzüne yükler ve metindeki bağlaçların söylem ve söylem dışı kullanımlarını işaretler. Ek olarak, seçilen metin dosyası için hazırlanmış bir söylem işaretleme dosyası varsa, program ayrıştırma sonuçlarını otomatik olarak değerlendirir.

Bu tez, Türkçe bağlaçların ve ulaçların söylem ve söylem dışı kullanım belirsizliğinin çözümüne önemli katkılar sağlayarak Türkçe söylem işlemlemesine katkı sağlarken, bildiğimiz kadarıyla, özellikle ulaç türündeki bağlaçların kullanım belirsizliğini çözmeye yönelik ilk denemedir. Tez, aynı zamanda Türkçe'deki her türlü söylem bağlacınının anlam belirsizliğini çözmede bir ilktir. Bu bakımdan tezdeki çalışmaların Türkçe söylem işlemleme alanındaki araştırmacılara bir referans noktası sağlayacağı ve yeni araştırma fırsatları yaratacağı öngörülmektedir.

Anahtar Kelimeler: söylem işleme, söylem bağıntıları, söylem bağlaçlarının ayrıştırılması, söylem bağlaçları, söylem bağlaç sözlüğü, Türkçe

To my husband Mustafa and
my sons Furkan and Tarık Emre

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor Prof. Dr. Ali Hikmet Doğru for his supportive, encouraging and constructive approach throughout my thesis study. I would also like to thank my co-supervisor Prof. Dr. Deniz Zeyrek Bozşahin for her guidance, continuous support, valuable comments, kindness and motivation. The ideas in present thesis are developed with the guidance of her immense knowledge of discourse. Thank you very much for being my supervisor and my co-supervisor.

I would like to thank to my thesis jury members Prof. Dr. Ferda Nur Alpaslan, Prof. Dr. Nihan Kesim Çicekli, Assoc. Prof. Dr. Gökhan Şengül, and Assist. Prof. Dr. Selma Tekir for reviewing and evaluating my thesis.

I would love to express my special thanks to my husband, Mustafa Başıbüyük and my sons for their endless patience, encouragement and support during my thesis study.

I would like to thank my lovely friend Selin Ünal Ulu for her friendship and for supporting me whenever I have difficulties.

I would like to express my appreciation for Tübitak Bilgem İltaren for providing a supportive approach for academic studies. I would also like to thank my colleagues there for presenting a friendly, instructive and innovative environment.

Finally, I wish to express my deepest gratitude to my father Hüseyin Demirtaş, my mother Münevver Demirtaş, my sister and my brother for their love and support throughout my life. With you, I am the person who I am. Thank you being my family.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xv

LIST OF FIGURES xix

LIST OF ABBREVIATIONS xxii

CHAPTERS

1 INTRODUCTION 1

1.1 Motivation and Problem Definition 1

1.2 Proposed Methods 4

1.3 Contributions and Novelties 7

1.4 The Outline of the Thesis 8

2 LITERATURE REVIEW 11

2.1 Discourse Connective Lexicons 11

2.1.1 DiMLex 12

2.1.2 Lexicons Developed for Several Languages 13

2.1.3 Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives . . . 16

2.2 Discourse Relations . . . 18

2.2.1 Explicit and Implicit Discourse Relations . . . 19

2.2.2 Arguments of Discourse Relations . . . 19

2.2.3 Sense of Discourse Relations . . . 20

2.3 Discourse Parsing . . . 21

2.4 Related Work on Discourse Parsing . . . 24

3 DATA SOURCES . . . 33

3.1 PDTB . . . 33

3.2 TDB 1.0 . . . 34

3.3 TDB 1.1 . . . 37

3.4 TED-MDB . . . 40

3.5 Analysis of Data Sources . . . 42

3.5.1 Discourse Usage Distribution of Turkish Connectives . . . 42

3.5.2 Sense Distribution of Turkish Connectives . . . 44

3.5.3 Usage Frequencies of Turkish Connectives . . . 45

4 TURKISH CONNECTIVE LEXICON . . . 49

4.1 Data Sources . . . 49

4.2 Connective Categories in TCL . . . 51

4.3 The Structure of the Lexicon . . . 53

4.4 TCL Search Tool . . . 56

4.5 TCL Development Steps . . . 60

5 TURKISH CONNECTIVE DISAMBIGUATOR 71

5.1 Discourse Usage Analysis 71

5.2 Turkish Connective Disambiguator Workflow 73

5.3 Classification Rules For Single and Phrasal Connectives 75

5.4 Classification Rules For Suffixal Connectives 84

5.5 TCD Interface 89

6 MACHINE LEARNING MODELS 93

6.1 Machine Learning Methodology 93

6.2 Features for the Usage Disambiguation of Single/Phrasal Connectives 95

6.3 Features for Usage Disambiguation of Suffixal Connectives 97

6.4 Features for Sense Disambiguation 98

7 RESULTS AND EVALUATION 101

7.1 Results and Evaluation of the Turkish Connective Disambiguator . . 101

7.1.1 Error Analysis 102

7.2 Results and Evaluation of Machine Learning Models 105

7.2.1 Results of the Usage Disambiguation of Single and Phrasal Connectives 106

7.2.2 Results of the Usage Disambiguation of Suffixal Connectives . 107

7.2.3 Results of the Sense Disambiguation of Discourse Connectives 107

8 CONCLUSION 111

REFERENCES 115

A DISCOURSE USAGE ANALYSIS RESULTS 123

B DISCOURSE USAGE ANALYSIS CATEGORIES 127

C CONNECTIVE SAMPLES FROM TCL IN THE DIMLEX FORMAT . . . 129

D EXPLICIT CONNECTIVES AND THEIR SENSES 133

E CONFUSION MATRICES 139

CURRICULUM VITAE 147

LIST OF TABLES

TABLES

Table 2.1 PDTB 3.0 sense hierarchy 22

Table 2.2 Features used for connective identification (syntactic and sense) . . . 28

Table 3.1 Distribution of inter-S and intra-S explicit discourse relations in TDB 1.1 39

Table 3.2 TED talks annotated in TED-MDB 41

Table 3.3 Data sources 43

Table 3.4 DC and NDC connective counts in T-TED-MDB and TDB 1.1 . . . 44

Table 3.5 Unambiguous and ambiguous connective counts in T-TED-MDB and TDB 1.1 44

Table 4.1 Different senses of various discourse connectives obtained from different corpora 51

Table 4.2 Converbs and their allomorphs in TCL 54

Table 4.3 Different parses for noun based DCs and assigned syntactic categories 65

Table 4.4 Syntactic category mapping between Connective-Lex and TCL . . . 67

Table 5.1 A sample part of discourse usage analysis results 72

Table 6.1 Classification rules of the TCD program for solving the usage ambiguity of single/phrasal connectives and the machine learning features inspired by the rules 96

Table 6.2 Features used for the usage disambiguation of single/phrasal connectives 97

Table 6.3 Classification rules of the TCD program for solving the usage ambiguity of suffixal connectives and the machine learning features inspired by the rules 98

Table 6.4 Features for the usage disambiguation of suffixal connectives 98

Table 6.5 Features used for the sense disambiguation of single/phrasal connectives 99

Table 6.6 Features for the sense disambiguation of suffixal connectives 99

Table 7.1 Usage disambiguation results of TCD 102

Table 7.2 A nominal predicate with tense/aspect/mood suffix -dir 103

Table 7.3 A nominal predicate with a dropped tense/aspect/mood suffix 103

Table 7.4 An existential stem with inflection 104

Table 7.5 An existential stem without inflection 104

Table 7.6 A converb as a suffixal discourse connective 105

Table 7.7 A syllable mistakenly identified as a converb 105

Table 7.8 Usage disambiguation results of the machine learning models for single/phrasal connectives 106

Table 7.9 Usage disambiguation results of the machine learning models for suffixal connectives 107

Table 7.10 Sense disambiguation results of the machine learning models for single/phrasal connectives 108

Table 7.11 Sense disambiguation results of the machine learning models for suffixal connectives 108

Table A.1 Discourse Usage Analysis Results 125

Table B.1 The categories formed as a result of the discourse usage analysis of single and phrasal connectives and the list of connectives in each category 127

Table E.1 The confusion matrix for the sense classification of single/phrasal connectives - T-TED-MDB fold1 139

Table E.2 The confusion matrix for the sense classification of single/phrasal connectives - T-TED-MDB fold2 139

Table E.3 The confusion matrix for the sense classification of single/phrasal connectives - T-TED-MDB fold3 140

Table E.4 The confusion matrix for the sense classification of single/phrasal connectives - T-TED-MDB fold4 140

Table E.5 The confusion matrix for the sense classification of single/phrasal connectives - T-TED-MDB fold5 140

Table E.6 The confusion matrix for the sense classification of single/phrasal connectives - TDB 1.1 fold1 141

Table E.7 The confusion matrix for the sense classification of single/phrasal connectives - TDB 1.1 fold2 141

Table E.8 The confusion matrix for the sense classification of single/phrasal connectives - TDB 1.1 fold3 141

Table E.9 The confusion matrix for the sense classification of single/phrasal connectives - TDB 1.1 fold4 142

Table E.10 The confusion matrix for the sense classification of single/phrasal connectives - TDB 1.1 fold5 142

Table E.11 The confusion matrix for the sense classification of single/phrasal connectives - total 142

Table E.12 The confusion matrix for the sense classification of suffixal connectives - T-TED-MDB fold1 143

Table E.13 The confusion matrix for the sense classification of suffixal connectives - T-TED-MDB fold2 143

Table E.14 The confusion matrix for the sense classification of suffixal connectives - T-TED-MDB fold3 143

Table E.15 The confusion matrix for the sense classification of suffixal connectives - T-TED-MDB fold4 144

Table E.16 The confusion matrix for the sense classification of suffixal connectives - T-TED-MDB fold5 144

Table E.17 The confusion matrix for the sense classification of suffixal connectives - TDB 1.1 fold1 144

Table E.18 The confusion matrix for the sense classification of suffixal connectives - TDB 1.1 fold2 145

Table E.19 The confusion matrix for the sense classification of suffixal connectives - TDB 1.1 fold3 145

Table E.20 The confusion matrix for the sense classification of suffixal connectives - TDB 1.1 fold4 145

Table E.21 The confusion matrix for the sense classification of suffixal connectives - TDB 1.1 fold5 146

Table E.22 The confusion matrix for the sense classification of suffixal connectives - total 146

LIST OF FIGURES

FIGURES

Figure 2.1 Main steps of discourse parsing 23

Figure 3.1 A sample relation annotated in TDB 1.1 file 00013112 38

Figure 3.2 Usage frequencies of all types of connectives (single, phrasal and suffixal) in T-TED-MDB and TDB 1.1 45

Figure 3.3 Usage frequencies of single and phrasal connectives in T-TED-MDB and TDB 1.1 46

Figure 3.4 Usage frequencies of suffixal connectives in T-TED-MDB and TDB 1.1 46

Figure 4.1 Distribution of connectives in data sources 50

Figure 4.2 A sample discourse connective *gibi* 'like' in TCL 54

Figure 4.3 Variants of the single discourse connective *ve* 'and' 55

Figure 4.4 Variants of the suffixal discourse connective *-(y)ArAk* 'by means of' 55

Figure 4.5 Interface of the TCL search tool 57

Figure 4.6 First tab of the TCL search tool result screen 58

Figure 4.7 Second tab of the TCL search tool result screen 58

Figure 4.8 Third tab of the TCL search tool result screen 59

Figure 4.9 Discourse use analysis results screen 60

Figure 4.10 A sample annotation file in XML file format from TDB1.1 . . . 61

Figure 4.11 A sample annotation file in pipe-delimited file format from T-TED-MDB dataset 61

Figure 4.12 TCL Flowchart 62

Figure 4.13 Connective *mesela* 'for example' in TCL 68

Figure 4.14 Connective *mesela* 'for example' in DiMLex format 68

Figure 5.1 Turkish Connective Disambiguator workflow 74

Figure 5.2 Decision tree for the usage disambiguation of single and phrasal connectives 76

Figure 5.3 Decision tree for the usage disambiguation of suffixal connectives 86

Figure 5.4 Turkish Connective Disambiguator interface 90

Figure 6.1 CRISP-DM Process Model 94

Figure C.1 A sample suffixal connective -(y)ArAk 'by means of/ and' in the DiMLex format 129

Figure C.2 A sample suffixal connective -AlI 'since' in the DiMLex format . 130

Figure C.3 A sample single connective *çünkü* 'because' in the DiMLex format 130

Figure C.4 A sample single connective *için* 'in order to' in the DiMLex format 131

Figure C.5 A sample phrasal continuous connective *gene de* 'nevertheless' in the DiMLex format 131

Figure C.6 A sample phrasal continuous connective *bu sebeple* 'for this reason' in the DiMLex format 132

Figure C.7 A sample phrasal discontinuous connective *önce ... sonra* 'first ... then' in the DiMLex format 132

Figure C.8 A sample phrasal discontinuous connective *tıpkı ... gibi* ‘as if’ in the DiMLex format 132

LIST OF ABBREVIATIONS

ADV	Adverb
ADP	Adposition
AltLex	Alternative Lexicalization
Arg1	Argument 1
Arg2	Argument 2
CCONJ	Coordinating Conjunction
DC	Discourse Connective
EntRel	Entity Relation
ML	Machine Learning
NDC	Non-Discourse Connective
NLP	Natural Language Processing
NoRel	No Relation
PDTB	Penn Discourse Treebank
PoS	Part of Speech
RST	Rhetorical Structure Theory
TCD	Turkish Connective Disambiguator
TCL	Turkish Connective Lexicon
TDB	Turkish Discourse Bank
TED-MDB	TED Multilingual Discourse Bank
T-TED-MDB	Turkish section of TED-Multilingual Discourse Bank
XML	Extensible Markup Language

CHAPTER 1

INTRODUCTION

In this first chapter, we present an introduction to the main topics of the thesis. First of all, we start with the motivation and the problem definition. Then, we explain the methods proposed to tackle the main object of investigation, and our contributions to the literature. We conclude the chapter by presenting the outline of the thesis.

1.1 Motivation and Problem Definition

Researchers have treated a text as a set of words or a set of sentences and have benefited from this approach for a long time in Natural Language Processing (NLP). The position, order, adjacency and context of words or sentences in a text can actually play an important role in many NLP tasks. The discourse structure of a text, i.e. the semantic (and pragmatic) relations that hold between text segments, can give strong clues about this type of information [1].

Discourse structures can be defined as *patterns* seen in multi-sentence or multi-clause in a text [2]. In order to properly interpret the information in the text, it is important to understand these patterns together with the elements that compose them. These elements can be *topics*, *functions*, *eventualities* or *discourse relations*.

- *Topics* consist of a set of entities and sentences about these entities summarizing the concepts in the text. The topic structure usually occurs in informative texts, such as texts in textbooks and encyclopedias.
- *Functions* are the role(s) of the text parts in the communication and discourse of a text can be structured by the function(s) of its elements. For example, the genre

of the news reports can be a good sample. The functions in a news report can be explained as (1) a summary in the *first paragraph* about who was in the incident, what happened, when and where it happened, why and how it happened etc. (2) through paragraphs that give detailed information about these topics in the *body*, (3) with less important information in the *final paragraph*.

- *Eventualities* can be defined as the descriptions of events and states and their spatio-temporal relations. These structures are often found in all kinds of texts such as news reports, in the methods section of a scientific paper, in narrative-type texts.

- *Discourse relations* show the relation between two semantic contents in a text or between a speech act and a semantic content. The semantic content can be composed of one or more clauses or sentences and it can be defined as an abstract object (Asher [3]) – a proposition, a fact, an event, a situation, etc. Discourse relations can be realized explicitly by using discourse connectives or implicitly through utterance adjacency.

Discourse structure plays an important role in Language Technologies and contributes to several NLP tasks such as text summarization, scoring student essays, information extraction, opinion mining, sentiment detection and statistical machine translation. For example, the discourse structure can help in selecting important sentences for text summarization by excluding the subordinate discourse parts with the content of explanation, comment or example from the summary [2].

In this thesis, we deal with discourse relations, which are an essential aspect of discourse structure. Two popular computational approaches for discourse relation modelling are Rhetorical Structure Theory (RST) [4] and the Penn Discourse TreeBank (PDTB) framework [5]. RST models a discourse relation with a tree structure in which the dominance of a node shows the relation between the child nodes and the precedence between nodes shows the order in the text. In the PDTB framework, a discourse relation is annotated with a discourse connective, its arguments and sense(s).

Today, several corpora have been annotated with discourse relations following the principles of the RST or the PDTB framework in many languages [2]. These languages

include Modern Standard Arabic [6], Chinese [7] [8], Czech [9], Danish and Italian [10], Dutch [11], English [12], German [13] [14], Hindi [15] and Turkish [16].

Many corpora developed so far have been annotated with discourse relations manually. Manual discourse annotation is a laborious and time-consuming task. So it is crucial to develop automatic discourse annotation methods to facilitate the discourse annotation task. There are various studies about automatic discourse annotation in different languages and the related work about this subject is provided in subsection 2.4 of this thesis.

Although Turkish is widely used as a native language in wide geography, NLP studies in the Turkish language have started late. Tools and resources for the Turkish language have been developed recently, approximately in the last two or three decades. The main reason for this situation is that the Turkish language offers specific challenges in many natural language processing tasks such as tag set design, statistical language modelling, syntactic modelling and statistical machine translation etc. Turkish is an agglutinative language and most of the language processing challenges arise from the complex morphology of Turkish [17].

Considering the Turkish NLP studies, it is observed that there are still very few studies about the discourse structure and in this respect, Turkish studies fall behind the other languages. One of the main reasons for this situation is that there are few resources developed for Turkish such as discourse annotated corpora and discourse connective lexicon.

Discourse relations can be expressed through explicit discourse connectives but languages vary in expressing them [18]. For example, while English and German express discourse relations lexically, Turkish conveys the discourse relations through morphological suffixes in addition to lexical connectives. The number of discourse connectives that languages use to express a discourse relation can be also different. Finally, some of the discourse connectives are polysemous and may signal to more than one sense. For such reasons, there is a need for a resource beyond a classic dictionary of discourse connectives. Given the important role of discourse connectives in discourse structure and discourse processing, in this thesis, we aimed to develop a Turkish Connective Lexicon (TCL) to bring all Turkish discourse connectives together and provide detailed

information about their syntactic and semantic features.

After developing the TCL, we wanted to develop a Turkish Connective Disambiguator (TCD) to detect discourse connectives in a raw text and assign the relations [sense(s)] of the connectives. We hope that the research community will benefit from the pipelines of the TCL and TCD developed within the scope of this thesis, which are also hoped to set the basis of future discourse parsing research.

1.2 Proposed Methods

A piece of text is often not understood on its own but it should be considered within the greater context, that is, with other pieces of text around it such as clauses, sentences or even paragraphs. A text span can be related to another text span in a way such as cause-effect relationship, contrast relationship etc. which are called *discourse relations*. *Discourse parsing* is the process of extracting such discourse relations between the text spans by understanding the internal structure of the text [19].

For representing the discourse relations between two text spans, generally explicit discourse connectives are used. Some of the connectives are always used with a discourse role and they signal only one relation sense, in other words, they are unambiguous in terms of discourse usage and relation sense. However, this is not the case for most connectives. A connective can play a discourse role or non-discourse role depending on the context it is used. For example, the connective *ve* 'and' in sentence 1 plays a discourse role relating two clauses while the connective *ve* 'and' in sentence 2 plays a non-discourse role relating two nouns. This type of ambiguity is called *usage ambiguity*.

1. Her cumartesi en sevdiği kafeye gider ve arkadaşlarıyla buluşur.
   'She goes to her favourite cafe every Saturday and meets with her friends.'

2. Çay ve kahve içmeyi sever.
   'She likes to drink tea and coffee.'

A discourse connective can signal different relation senses in different contexts. For

example, the connective *kadar* 'as much as' used in sentence 3 shows a comparison relation while the connective *kadar* 'until' used in sentence 4 shows a temporal relation. This type of ambiguity is called *sense ambiguity*.

3. Okuduğun kadar öğrenirsin.
   'You learn as much as you read.'

4. Okullar açılana kadar yazlığa gitti.
   'She went to the summer house until the schools opened.'

Another ambiguity related to discourse connectives is *PoS ambiguity*. In PoS ambiguity, the sense of a relationship changes according to the PoS of the connective. Webber et al. [20] state that the PoS of the connectives *since*, *before* and *however* in English can affect the senses of the connectives. Similarly, Zeyrek and Kurfalı [21] mention the relationship between the PoS tags and the senses of the connectives *ardından* 'then', *sonra* 'after' and *önce* 'before' in Turkish. These connectives are mostly presented through the *precedence* sense with the PoS tag *adverb*, and *succession* sense with the PoS tag *postposition*.

For example, in sentence 5, the connective *sonra* 'then' has the PoS tag *adverb* and has the relation sense *precedence* while in sentence 6, the connective *sonra* 'after' has the PoS tag *postposition* and carries the relation sense *succession*.

5. Ressam, önce tuvalini ve fırçasını buldu sonra yağlı boya yapmaya başladı.
   'The painter first found her canvas and brush then started to paint in oil.'
   [Temporal:Asynchronous:Precedence]
   (adverb)

6. Kahvaltısını yaptıktan sonra yürüyüşe çıktı.
   'After she has breakfast, she went for a walk.'
   [Temporal:Asynchronous:Succession]
   (postposition)

We propose methods for solving the usage ambiguity and the sense ambiguity of the Turkish explicit discourse connectives. We benefited from the PoS ambiguity concept

in determining the sense of the discourse connectives by using the PoS feature as a distinguishing feature in the machine learning models. For this purpose, we followed the steps below:

- **Turkish Connective Lexicon:** To detect discourse connectives in a text, a list of connectives specific to language is needed to mark the discourse usages. In Turkish, we could not come across such a comprehensive connective list, in other words, a connective lexicon. For this reason, we developed a lexicon both to use in our studies within the scope of this thesis and to offer it to future researchers.

  TCL was built automatically by using discourse annotations of the Turkish Discourse Bank (TDB) 1.0 [22], TDB 1.1 [23] and Turkish section of TED-Multilingual Discourse Bank (T-TED-MDB) [24]. Also, syntactic and semantic properties of discourse connectives are computed automatically using these corpora.

  TCL uses the XML format of DiMLex, a German connective lexicon, and contains approximately 180 connectives in single, phrasal and suffixal types.

  We have designed a user-friendly interface for the TCL using the C# programming language in the Visual Studio environment to provide users with several displaying and filtering capabilities.

- **Turkish Connective Disambiguator:** After completing the TCL, we developed the TCD program for marking all of the occurrences of connectives in a given Turkish text and distinguishing the discourse usages from the non-discourse ones. TCD makes use of the connective list developed in the present thesis containing single, phrasal and suffixal types, to mark the connective occurrences in the text. TCD program enables two rule-based decision trees for usage disambiguation of connectives, one for the single/phrasal type connectives and one for the suffixal type connectives. We derived the classification rules of decision trees by considering the Turkish grammar rules related to Turkish discourse structure and analyzing the usage scenarios of discourse connectives in sample sentences of TDB 1.1 and T-TED-MDB datasets.

  We designed a user-friendly interface for the TCD program by using the C#

programming language in the Visual Studio environment. By using this program, a user can select and load any Turkish text file to the program interface and after the disambiguation process, he/she can display all discourse and non-discourse occurrences of connectives marked with different colours in the text.

- **Machine Learning Models for Usage Disambiguation:** In addition to the algorithmic approach used in the TCD program, we tried to solve the usage ambiguity problem with the machine learning approach. For this purpose, we designed features for the usage disambiguation of Turkish connectives. While designing these features, we were inspired by the classification rules and the features used for the usage disambiguation of connectives in other languages. Since the discourse structure of suffixal connectives in Turkish is quite different from single and phrasal connectives, a different feature set is designed for the suffixal type.

- **Machine Learning Models for Sense Disambiguation:** Another method that we propose in this thesis is designing machine learning features for the sense disambiguation of Turkish discourse connectives.

  There are 3 levels of sense in the PDTB sense hierarchy, as shown in Table 2.1. In this thesis, we aimed to disambiguate the senses at all levels of the PDTB 3.0 sense hierarchy and for all types of Turkish discourse connectives, namely single, phrasal and suffixal types. While designing the machine learning features for the sense disambiguation of Turkish discourse connectives, we referenced the features designed for sense disambiguation in other languages.

Two papers were written reflecting the studies carried out in the context of this thesis. Zeyrek and Başıbüyük [18] describes the TCL and Başıbüyük and Zeyrek (in submission) [25] focuses on the TCD and the machine learning models.

1.3 Contributions and Novelties

The main contributions of this thesis can be summarized as follows:

- We developed the TCL, the first comprehensive lexicon covering all types of

discourse connectives in Turkish, namely single, phrasal and suffixal types. For each discourse connective, we provided the orthographic variants, syntactic category, semantic sense(s) and sample sentences. Additionally, we provided the possible allomorphs of the suffixal connectives for the first time in the literature. We plan to submit the output of the TCL to the connective-lex [1] website which presents a web-based multilingual lexical resource for connectives. Thus, we will make the TCL available to the research community.

- Turkish suffixal discourse connectives are suffixed to the *verbs* and they have different variations depending on the vowel harmony and consonant harmony. Such variants of a suffixal connective are called *allomorphs* of the connective. We provided all possible allomorphs of suffixal connectives, hence we simplify parsing the suffixal connectives in a text.

- We designed a set of classification rules for usage disambiguation of single/phrasal type connectives in Turkish and a different set of classification rules for the usage disambiguation of the suffixal type.

- We discovered machine learning features to be used for the usage disambiguation of the single/phrasal and suffixal type connectives in Turkish. As far as we know, there are very few attempts for solving the usage ambiguity of single/phrasal type connectives in Turkish while there is no attempt for suffixal type.

- We attempted to solve the sense ambiguity of single/phrasal and suffixal type explicit discourse connectives for three levels of PDTB 3.0 sense hierarchy for the first time in the literature to the best of our knowledge.

- User-friendly interfaces were developed for the TCL and TCD programs.

1.4 The Outline of the Thesis

The rest of the thesis is organized as follows. Chapter 2 provides the literature review on discourse connective lexicons, discourse relations, and discourse parsing. The literature review covers both conceptual explanations about these topics and the related

[1] http://www.connective-lex.info

work. In Chapter 3, we provide detailed information about the data sources we used within the scope of this thesis. In Chapter 4, we present the content of the TCL, the TCL program interface and describe how we developed the lexicon. In Chapter 5, we explain the usage disambiguation approach employed in the TCD program. Chapter 6 presents the machine learning models developed for solving the usage ambiguity and the sense ambiguity of the connectives in Turkish. Chapter 7 provides the results and evaluation, and Chapter 8 concludes the thesis.

CHAPTER 2

LITERATURE REVIEW

In this section, we firstly present information about discourse connective lexicons and the related work on this subject. Then we describe the notion of discourse relation by providing information and samples about explicit and implicit discourse relations, arguments of discourse relations and the sense(s) of discourse relations. We continue with the concept of discourse parsing and conclude the chapter with a summary of work on discourse parsing.

2.1 Discourse Connective Lexicons

A lexicon of discourse connectives can be defined as a dictionary giving the list of discourse connectives in a language along with their syntactic and semantic properties. Discourse connective lexicons have been developed in many languages and are used especially in the field of discourse parsing. One of the first initiatives is the German discourse connective lexicon, DiMLex [26], which offers an XML schema that can be used by other languages. Many languages later used the DiMLex XML format in their lexicon development studies.

In this section, we first describe the format of the DiMLex and then give detailed information about the discourse connective lexicons developed for several languages. Finally, we mention the connective-lex website which aims to present multilingual lexical resources for connectives.

2.1.1 DiMLex

DiMLex is a lexicon of German connectives which presents an XML schema followed by many other languages later. Using the DiMLex XML structure, the following information about a lexicon entry can be provided as described in the paper [27]:

- orthographic variants
- ambiguity of the connective (whether the connective can be used as non-discourse)
- examples for non-discourse usage
- syntactic category
- semantic senses
- examples for semantic senses

Stede and Umbach [26], the creators of the DiMLex, believe that a dictionary dedicated to discourse connectives (discourse markers), carrying syntactic, semantic and pragmatic information about them, would work well in the natural language processing field especially in text generation and text understanding. In text generation, the selection of the discourse markers to be used in a sentence is a part of the sentence planning job. In text understanding, discourse markers are important cues in inferring the semantic relations in a text. DiMLex can be used as a declarative resource in both tasks. To identify the items in the lexicon, information from grammars, dictionaries, and the linguistics research literature is used. The lexical features defined for lexicon items include the part of speech of a marker and the discourse relations expressed by a marker.

The content of DiMLex has been enhanced later by Scheffler and Stede [27] to include all known German discourse connectives currently in use. The lexicon contains 275 items and the information described for each item includes orthographic variants, ambiguity information (whether the connective has non-connective readings), examples of non-connective readings, syntactic category and relational information. In defining the relational information of lexicon items, PDTB 3.0 [28] sense hierarchy was used.

For obtaining the relation(s) which a discourse connective can carry, four different methods can be used according to the authors [27]. The first method is the usage of the annotated corpus. The relation(s) assigned to a connective in a corpus is derived. The second method is using a lexicon in a different language. In this method, a connective is translated and the relation(s) of the translated connective is found in its lexicon. As a third method, grammar or other literature can be used to find semantic information about a connective. Finally, instances of a connective can be found in a raw text and it can be annotated using annotation guidelines. For DiMLex, all methods were used to collect comprehensive information about connective relations, but since it is a costly method, the fourth method was used in limited numbers.

2.1.2 Lexicons Developed for Several Languages

After the German discourse connective lexicon, many other languages developed their own discourse connective lexicons.

LEXCONN [29] is the first French lexicon with 328 discourse connectives and contains syntactic categories and discourse relations of connectives. This lexicon was created manually, relational information was presented in SDRT [30] format and FRANTEXT corpus was used as an example source. The authors state that ambiguity information about discourse and non-discourse usage must be added to LEXCONN.

Keskes et al. [31] created an Arabic lexicon manually containing 174 connectives to use in their discourse parsing studies. For each connective, they provided various information such as its type, signalling force (strong or weak), possible parts of speech and the possible set of discourse relations.

The lexicon for Italian discourse connectives, LICO [32], was developed in two releases: the first release in 2016 and the second release in 2018. The first release of LICO includes about 170 discourse connectives with their orthographical variants, part of speech(es), semantic relation(s), and a number of usage examples. The semantic relations are aligned with the PDTB 3.0 sense hierarchy. Feltracco et al. [32] inspired by DiMLex and they used the XML format proposed by DiMLex. For each discourse connective, LICO specifies:

- lexical variants and orthographic variants
- whether the connective is a single token or phrasal token
- whether the connective is continuous or discontinuous
- syntactic category
- semantic relation(s)
- alignments with German lexicon
- examples of semantic relations

To compile the list of discourse connectives, Feltracco et al. [32] used a number of grammatical and lexical resources for Italian and they used DiMLex for finding the equivalent connectives in Italian. The second release of LICO [33] enriched the existing resource by using real corpus data, basically Contrast-Ita Bank. The study was limited to the connectives signalling contrast relations in text. For each semantic relation, they specified examples taken from several corpora.

CzeDLex [34] is the discourse connective lexicon for the Czech language based on the discourse annotations in the Prague Discourse Treebank 2.0. It has an XML-based structure largely inspired by the DiMLex. In addition to primary connectives, it also includes secondary connectives. The lexicon provides the morphosyntactic and semantic properties of the connectives and gives non-connective usages.

One of the recent lexicons, LDM-PT [35], is a lexicon developed for European Portuguese and contains 252 discourse marker/rhetorical sense pairs. Each discourse marker is assigned its possible rhetorical sense(s) (discourse relations or coherence relations) according to the PDTB 3.0 sense hierarchy. Mendes et al. state that this lexicon can provide data for annotation of discourse relations and also it can be useful for parsing, text processing and summarization. The list of discourse markers are obtained from three different sources: frequency lists extracted from a corpus of contemporary Portuguese, a multilingual discourse treebank (TED-MDB) and contrastive analysis with English discourse markers. The lexicon gives the following information about discourse markers:

- rhetorical sense
- whether the discourse marker is continuous or discontinuous
- whether the discourse marker is a single token or a phrasal token
- type (primary connective, secondary connective or alternative lexicalization)
- syntactic category
- restrictions on the context
- English near-synonym(s)
- corpus example

After gathering information about discourse markers, the lexicon was presented in the DiMLex format.

DiMLex-Eng [36] is a lexicon created in the format of DiMLex, containing 149 English connectives. This lexicon was created by compiling the annotations of the Penn Discourse Tree Bank 2.0 and RST Signaling Corpus and using an additional relation signal list. Developed to be part of the connective-lex database, this lexicon includes the following features related to the connectives; possible orthographic variants, syntactic category, non-connective usage and a set of discourse relations with example usages.

Bourgonje et al. [37] created a discourse connective lexicon for Dutch containing 207 entries of the categories; adjective, subordinating conjunction, coordinating conjunction, preposition and other. The lexicon is publicly available and it presents both human- and machine-readable formats. The lexicon has been prepared in DiMLex format since it has been used by many languages in recent years and it offers a language-neutral XML structure at a sufficient level. The lexicon was generated in two steps. In the first step, they determined candidate connectives using two parallel German-Dutch corpora and a German connective lexicon (DiMLex) and in the second step, they evaluated the connective candidates using Dutch resources. The authors state that such a lexicon can be a useful resource for language learners as it provides information about semantic relations for each connective and also it can be used by studies such as machine translation and automatic discourse parsing.

2.1.3 Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives

Connective-Lex [38] is a multilingual connective database that puts together discourse connective lexicons developed in multiple languages. It has a web-based interface [1] which provides connective information in these languages and allows querying of this information.

Currently, connective-lex includes the lexicons of the following languages: Ukrainian, Arabic [31], Bangla, Czech [34], German [39], Dutch[37], English [36], Portuguese [35], French [29] and Italian [32].

The authors discuss the definition of the connectives that are covered in this source and they provide the syntactic, semantic/pragmatic and lexicographic information about the connectives. They argue that the integration of lexicons developed in many languages provides additional advantages to researchers such as cross-linguistic comparison and linking of connectives in different languages. The authors also state that connective-lex has an infrastructure suitable for developing both in breadth (by adding different languages) and in-depth (by adding more exact and more comprehensive information about connectives).

Connective definition and delimitation: The first important decision to make when creating a specific lexicon is which items to include in the lexicon. It is important to define what a discourse connective means, because the definition of a connective may vary according to the researchers or language characteristics. The authors specify four main criteria for discourse connective definition [38] following the definition of the German DiMLex lexicon:

"A lexical item or phrase x is a connective if and only if:
1. x cannot be inflected or modified;
2. the meaning of x is a two-place relation;
3. the arguments of this relation are abstract objects (propositions, events, states, or processes);
4. these arguments are typically expressed as clauses or sentences (but they can also

[1] http://www.connective-lex.info

be expressed as groups of sentences, or as noun phrases that denote abstract entities)."

In addition to single words, phrases with the above characteristics can also be defined as connectives. This definition is similar to the discourse connective definitions in many other lexicons and it is also compatible with the definition of *primary connectives* in the Danlos et al. [40]. However, the authors ask whether a language-independent and well-established discourse connective definition is possible.

According to the English PDTB [41], discourse connectives are "discourse-level predicates that take two abstract objects such as events, states, and propositions (Asher, 1993) as their arguments". PDTB [12] makes several syntactic restrictions for items to be classified as connectives. The first type of constraint is related to the part of speech tag and requires the connective to be in one of three classes: a subordinating or coordinating conjunction, or an adverbial. The other limitation is related to the arguments of the connective and requires that the arguments can be expressed as clauses.

Lexicon entries: The information given about the connective-lex items can be summarized as follows:

- **Syntax** : The major syntactic categories of connectives are coordinating and subordinating conjunctions, adverbials, and adpositions. The connectives can be grouped by their syntactic categories using the search facilities offered by the connective-lex.
- **Semantics/pragmatics** : The semantic features refers to the possible discourse relations that a connective may signal. Since a connective can have more than one syntactic category, semantic features are given under these syntactic categories. PDTB 3.0 sense hierarchy [28] was chosen for the representation of discourse relations.
- **Lexicographic information** : Lexical items in a language, including connectives, display specific features. These features such as the definition, examples, punctuation, position in the sentence, translation into other languages and possible modifiers may vary from language to language. For this reason, such information about connectives is left optional to languages.

Linking the languages: Individual connective lexicons are very useful in discourse and NLP studies such as automatic identification of connectives [42] [43] and discourse processing [19] [44]. However, linking these resources offers additional benefits to researchers such as machine translation [45] [46], manual translation, language learning and teaching in CALL (Computer-Assisted Language Learning) systems [47]. Conn-Lex gives users the possibility to query connectives that have the same syntactic category or the same senses in multiple languages.

2.2 Discourse Relations

Discourse is a language unit above the sentence level and represents the relations between the clauses in a sentence or between the sentences in a text. Theories of discourse theory state that when the sentences in a text are somehow related to each other, the text is coherent [48] [2]. For example, the sentences in example 7 are in a cause-effect relationship: the boy was very happy because his friend gave him a nice toy. However, in example 8, there is no relationship between the boy's being happy and the flowers' being grown. Therefore, in the first example, the text is coherent while in the second example the text is incoherent.

7. Çocuk çok mutluydu. Arkadaşı ona güzel bir oyuncak hediye etmişti.
   'The boy was very happy. His friend gave him a nice toy.'

8. Çocuk çok mutluydu. Bahçedeki çicekler ne kadar da büyümüş!
   'The boy was very happy. How grown the flowers in the garden!'

The relationships between the clauses or the sentences in a text are called coherence relations, conjunctive relations, rhetorical relations or discourse relations. In this thesis, we prefer the term *discourse relations* and we present sample sentences related to discourse relations throughout the thesis. For the sample sentences selected from Turkish discourse-annotated corpora, we provide the related corpus name and file name along with the sample sentence. We also provide the English translations of the sample sentences. The TED Multilingual Discourse Bank (TED-MDB) dataset contains English files corresponding to the Turkish files. However, for the samples

given from the Turkish section of TED-MDB dataset, we do not use the English section of TED-MDB dataset for the translations. Instead, we prefer to use back-translation for glosses to keep the discourse and non-discourse connective usages in the sentences.

2.2.1 Explicit and Implicit Discourse Relations

Discourse relations in a text can be realized explicitly by using an explicit *discourse connective* or implicitly without using any connective [12]. In explicit relations, the relation between clauses or sentences is clearly highlighted with a connective, while in the case of implicit relations, the relation is interpreted by the reader using the context of the text. For instance, in example 9, the relation between the sentence parts is expressed by an explicit discourse connective *için* 'because', while in example 10, the user understands the relation between the sentences from the context.

9. Çiçekleri sevdiği için balkonu çiçeklerle donattı.
   'She decorated the balcony with flowers because she loves the flowers.'

10. Çiçekleri seviyor. (implicit=Bu yüzden) Balkonu çiçeklerle donattı.
    'She loves the flowers. (implicit=So) She decorated the balcony with flowers.'

In the PDTB annotation style (which is the annotation style used by the corpora used in the present thesis), while annotating implicit relations, a connective that best expresses the meaning of the relation is placed in the text, which is called an *Implicit* connective. In the PDTB, to disinguish implicit connectives from connectives used in the texts, the term explicit connective used. These terms are also used in the present thesis.

2.2.2 Arguments of Discourse Relations

In explicit and implicit relations, discourse connectives have two arguments (i.e. text spans that constitute the discourse relation), simply labelled as Arg1 and Arg2. According to the definition of the PDTB [12], in explicit relations, Arg2 is the argument to which the connective is syntactically bound, while Arg1 is the other argument. In relations between adjacent sentences, arguments follow the linear order as Arg1 comes

before Arg2. The arguments of explicit connectives do not have to be a single clause or a single sentence, they can consist of multiple clauses or sentences. However, the argument is expected to consist of minimal parts that is needed to infer the meaning of the discourse relation which is called as *minimality principle*. In the representation of discourse relations in PDTB, discourse connectives are underlined, Arg1 is shown in italics and Arg2 is written in bold letters as shown in example 11. We use the same convention in the present thesis. Additionally, we show non-discourse connectives with NDC subscript.

11. **Çok çalıştığı** için *bir kahve molası verdi.*
    He took a coffee break because **he worked hard.**
    [Contingency:Cause:Reason]

The arguments of the implicit connectives can be the successive sentences in a paragraph or the parts of a sentence separated by a semi-colon (";") or colon (":") [12]. In example 12, the first consecutive sentence is labelled as Arg1 while the second one is labelled as Arg2.

12. *Sabahtan bu yana çok çalışmıştı.* (implicit=Bu yüzden) **Bir kahve molası verdi.**
    He had worked hard since the morning. (implicit=So) **He took a coffee break.**
    [Contingency:Cause:Result]

Example 13 below shows an implicit relation in a sentence where the first part of the sentence is the Arg1 of the implicit relation while the second part is the Arg2.

13. *Her tür spora ilgi duyuyor;* (implicit=mesela) **basketbol oynamaya bayılıyor.**
    He is interested in all kinds of sports; (implicit=for example) **loves to play basketball.**
    [Expansion:Instantiation]

2.2.3 Sense of Discourse Relations

In discourse relations, sense tags of connectives define the relationship between the arguments such as Temporal, Contingency, Comparison or Expansion [12]. For

instance, the sense of the explicit discourse connective *için* 'because' in sample 11 is *Contingency:Cause:Reason* because the first clause of the sentence (Arg2) is the cause of the second clause (Arg1). On the other hand, the sense of the implicit discourse connective *Bu yüzden* 'So' in sample 12 is *Contingency:Cause:Result* because Arg2 is the result of Arg1.

In PDTB, semantic senses are defined hierarchically, with three levels: class, type and subtype as shown in Table 2.1 (Webber et al. [49]). The top-level *classes* represent the main semantic classes which are Temporal, Contingency, Comparison and Expansion. A set of *types* is defined under each *class* to refine the senses. The third level, *subtype*, defines the semantic contribution of the arguments to the relationship.

Discourse connectives can have a single sense, or they can have multiple senses [41]. In example 14, Arg1 and Arg2 represent two consecutive events and also Arg2 is the result of Arg1. So, the discourse connective *ve* 'and' carries both Temporal:Asynchronous sense and Contingency:Cause:Result sense in the sentence.

14. *Kadın hediye paketini açtı* ve **sevinçle havaya zıpladı.**
    The woman opened the gift package and **jumped into the air with joy.**
    [Temporal:Asynchronous | Contingency:Cause:Result]

2.3 Discourse Parsing

In NLP, *discourse parsing* can be defined as understanding the internal structure of a text and determining the discourse relations in that text. For discourse analysis and parsing, researchers have proposed several frameworks from different perspectives. One of the frameworks developed for this purpose is PDTB [12] which has discourse-level annotations. PDTB annotates explicit and implicit relations, argument spans of these relations, the sense of each relation, and their attribution spans if any. There are various challenges related to discourse parsing in PDTB. These are separating discourse usages from non-discourse usages, determining the argument spans of the relation, classifying the relation sense of the relation and marking the attribution spans. All of these steps must be completed to make a PDTB-style fully automatic

Table 2.1: PDTB 3.0 sense hierarchy

Level-1	**Level-2**	**Level-3**
TEMPORAL	SYNCHRONOUS	-
	ASYNCHRONOUS	[PRECEDENCE, SUCCESSION]
CONTINGENCY	CAUSE	REASON
		RESULT
		NEGRESULT
	CAUSE+BELIEF	REASON+BELIEF
		RESULT+BELIEF
	CAUSE+SPEECHACT	REASON+SPEECHACT
		RESULT+SPEECHACT
	CONDITION	[ARG1,ARG2]-AS-COND
	CONDITION+SPEECHACT	-
	NEGATIVE-CONDITION	[ARG1,ARG2]-AS-NEGCOND
	NEGATIVE-CONDITION+SPEECHACT	-
	PURPOSE	[ARG1,ARG2]-AS-GOAL
COMPARISON	CONCESSION	[ARG1,ARG2]-AS-DENIER
	CONCESSION+SPEECHACT	ARG2-AS-DENIER+SPEECHACT
	CONTRAST	-
	SIMILARITY	-
EXPANSION	CONJUNCTION	-
	DISJUNCTION	-
	EQUIVALENCE	-
	EXCEPTION	[ARG1,ARG2]-AS-EXCPT
	INSTANTIATION	[ARG1,ARG2]-AS-INSTANCE
	LEVEL-OF-DETAIL	[ARG1,ARG2]-AS-DETAIL
	MANNER	[ARG1,ARG2]-AS-MANNER
	SUBSTITUTION	[ARG1,ARG2]-AS-SUBST

end-to-end discourse parser [19]. Figure 2.1 shows the main steps of the discourse parsing procedure.

A connective can have a discourse use or non-discourse use in its context. This is an ambiguous situation and called *usage ambiguity* [20]. In example 15, the connective *ve* 'and' connects two logical clauses in the sentence and plays a discourse role while in example 16, the connective *ve* 'and' connects two nouns in the sentence and used as non-discourse.

15. *Ali kütüphaneden bir kitap aldı* ve **onu okumaya başladı**.

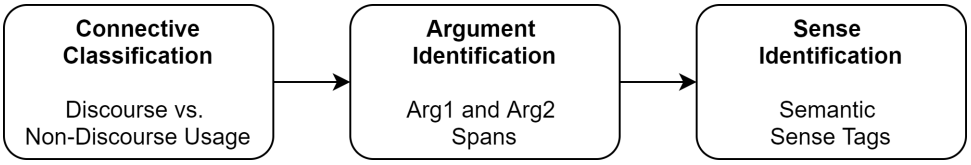

Figure 2.1: Main steps of discourse parsing

'*Ali bought a book from the library* and **started reading it**.'

16. Ali ve $_{\text{NDC}}$ Ayşe kütüphanede kitap okuyor.
    'Ali and $_{\text{NDC}}$ Ayşe are reading books in the library.'

After identifying that a connective is used as discourse connective, the next step is identifying the arguments of the connective. In the identification of arguments, argument spans must be determined, which includes both the location and extent of the arguments. Prasad et al. [12] state that since Arg2 is syntactically bound to the discourse connective, Arg2 is easier to find than Arg1.

The next step in discourse parsing is finding the sense(s) of the discourse relation. A discourse connective of a discourse relation may have different senses depending on its context, which is called *sense ambiguity* as mentioned in the paper of Webber et al. [20]. Some connectives are always used in a single sense, while others are quite ambiguous in terms of sense. For example, the connective *ayrıca* 'in addition' in Turkish is always used in a single sense, i.e Expansion:Conjunction while the connective *ve* 'and' may be used in many different senses. For example, in sentence 17 it is used with *Expansion:Conjunction* sense, while in sentence 18 it is used with *Temporal:Synchronous* sense.

17. *Yıldızın önüne işte böyle geçecek* ve **harika bir gölge yaratacak**.
    '*This is how it will pass in front of the star* and **create a wonderful shadow**.'
    [Expansion:Conjunction]
    (T-TED-MDB file 1976)

18. *Orada, aynı yerde durur* ve **onu bekler**.
    '*He stands there in the same place* and **waits for him**.'

[Temporal:Synchronous]
(TDB 1.1 file 00002213)

Another issue about the senses of connectives is that a connective may have more than one sense simultaneously in the same relationship. In such cases, each sense tag which the connective carries is provided [12]. In example 19, the connective *ve* 'and' signals to both *Expansion:Conjunction* sense and *Contingency:Cause:Result* sense, and therefore both senses are given in the discourse annotation.

19. *Bunun ona daha çok kararlılık verdiğini gördüm* ve **tekrar tekrar denedi**.
'*I saw that it gave him more determination* and **tried again and again**.'
[Expansion:Conjunction | Contingency:Cause:Result]
(T-TED-MDB file 1978)

2.4 Related Work on Discourse Parsing

Many studies about discourse parsing have been conducted in the literature. One of the first studies on this subject was carried out by Pitler and Nenkova [50]. The authors state that two types of ambiguity problems should be solved during the discourse parsing process, which is the usage ambiguity and the sense ambiguity. They showed that the connective itself and the various syntactic features extracted from the parse trees can be very useful in solving these ambiguity problems. These features are as follows with their descriptions from the article [50]:

- **Self Category** "The highest node in the tree which dominates the words in the connective but nothing else."
- **Parent Category** "The category of the immediate parent of the Self Category."
- **Left Sibling Category** "The syntactic category of the sibling immediately to the left of the Self Category."
- **Right Sibling Category** "The syntactic category of the sibling immediately to the right of the Self Category."

In the disambiguation of discourse vs. non-discourse usage of connectives, they used the connective itself and the syntactic features separately/together in different combinations hence they got high accuracy results. Using connective and syntactic features together performed better than using these features separately. The set with the highest performance in the study includes the pairwise interaction of the connective and the syntactic features and the pairwise interaction between the syntactic features.

The authors state that most of the connectives are mostly used in a single sense, but some connectives are quite ambiguous. For example, the connective *since* may signal to the Temporal sense or the Contingency sense. After the authors developed syntactic features to solve the usage ambiguity problem, they tried the same feature set to solve the sense ambiguity problem. The connectives were classified according to the four top-levels in the PDTB 3.0 sense hierarchy. Using only the connective itself in the sense classification has given quite good results and adding syntactic features has increased the success rate. The results in the sense classification are almost the same as those obtained with the inter-annotator agreement, indicating that the success is very high.

Another important study on discourse parsing was carried out by Lin et al. [19]. They developed a PDTB-styled end-to-end discourse parser in English to parse the free texts. The input of the parser is a free text and the output is the discourse structure of the given text. The parser has 3 main steps. In the first step, explicit connectives, their arguments and the relation sense are identified. In the second step, the non-explicit relations are found which are Implicit, AltLex, EntRel and NoRel. Finally, the third step includes finding attribution spans.

In their paper, Lin et al. [19] state that nearly 30 percent of the connective occurrences used in the PDTB corpus are used as discourse, while 70 percent are used as non-discourse. So before labelling the arguments and sense of a connective, it is critical to correctly identify whether this connective is discourse or non-discourse. The authors observed that the context and the PoS of the connective give strong clues about its discourse use. The syntactic parse path from the connective itself to the root of the parse tree also shows the functionality of the connective in the sentence. Based on these observations, the authors used a set of lexico-syntactic features and path features

of the connective for the discourse usage identification. For a connective *C*, its previous word *prev* and its next word *next*, the features used for the connective classification are listed in the paper as follows:

- "C POS"
- "prev + C"
- "prev POS"
- "prev POS + C POS"
- "C + next"
- "next POS"
- "C POS + next POS"
- "path of C's parent -> root "
- "compressed path of C's parent -> root"

The first seven features above reflect the context and PoS of the connective, while the last two features show the parse tree from the connective to the root.

After determining that a connective is used as a discourse connective, the relation which the connective signals must be determined since a connective can be used in different semantic senses depending on its context. In the sense classification, the connective itself is a pretty good feature since only a few of the connectives are ambiguous in terms of sense, and those that are ambiguous tend to be used in certain senses. The authors [19] used the following three features for this purpose:

- "C string"
- "C's POS"
- "C + prev"

For training, development and testing purposes, they used different sections of PDTB in their experiments and made evaluations for the performance of each component

individually as well as all components together. While the individual component accuracy results are quite well, the success of the whole system together decreases due to the error propagation issue.

Laali et al. [51] developed the CLaC discourse parser as part of the CoNLL-2016 shared task on shallow discourse parsing. Shallow discourse parsing involves finding the relations, the arguments, and the sense(s) in a text. Within the scope of the CoNLL shared task, four types of discourse relations are annotated. These relation types are explained in the paper as follows:

- **Explicit Discourse Relations :** Explicit discourse relations are marked explicitly by using discourse connectives such as *and, but, so* in the text.
- **Implicit Discourse Relations :** In Implicit discourse relations, no connective is used in the text. The reader infers the relation from context by inserting a connective best expresses the relation. These connectives are called implicit discourse connectives.
- **AltLex Discourse Relations :** Similar to implicit discourse relations, AltLex relations do not have an explicit connective. However, the relation is alternatively lexicalized with some non-connective expressions. Therefore, there is no need to use an implicit connective.
- **EntRel Relations :** In EntRel relations, there is an entity-based coherence relation between arguments.

For connective classification, the authors used the C4.5 decision tree binary classifier and a set of features which are listed as follows in the paper[51]:

- "1. The discourse connective text in lowercase."
- "2. The categorization of the case of the connective: all lowercase or initial uppercase."
- "3. The highest node (called the SelfCat node) in the parse tree that covers the connective words but nothing more."

- "4–6. The parent, the left sibling and the right sibling of the SelfCat."
- "7–10. The left and the right word of discourse connective and their parts of speech."

After the connective classification, the same algorithm and the same feature set was used for sense labelling. The parser achieved high success with (F1=0.9020) in the connective classification and (F1=0.7622) in the sense labelling.

Gopalan and Devi [52] developed a system for automatic identification of the explicit connectives, their arguments and sense(s) in biomedical texts. They used the Conditional Random Fields algorithm with the syntactic features given in Table 2.2 for the connective classification and the sense labelling [52].

Table 2.2: Features used for connective identification (syntactic and sense)

S.No	Features for Connectives	Examples
1	Word	previous word, current word, next word
2	PoS (P)	PoS of previous word, PoS of current word, PoS of next word
3	Combination of PoS and Chunk	PoS and Chunk of current word
4	Combination of 1, 2, 3	Current word, PoS of current word, Chunk of current word
5	Connective	Connective is an exceptional feature for sense identification

They obtained high accuracy results with these features and they also developed several post-processing rules by analysing the errors to improve the results. They reported 86.42% F-score for connective identification using 10 fold cross-validation and CRFs algorithm.

Another automatic discourse connective detection study conducted for biomedical texts is the study of Polepalli Ramesh et al. [53]. They state that discourse connective (word or phrase) detection is a first step of discourse relation identification in a text. In this work, they developed supervised machine learning methods to automatically detect the discourse connectives in biomedical texts. For this purpose, they used two machine learning models; support vector machines and conditional random fields. The

authors trained the machine learning models on the Biomedical Discourse Relation Bank dataset and discovered novel dataset adaptation techniques. As a result, they stated that the discourse connectives in biomedical texts can be detected by machine learning techniques and reported the highest performance as 0.761 F-score. The authors performed an error analysis on the best performing CRF classifiers and observed that most of the errors are due to the data sparsity of some connectives. In general, as the frequency of connectives used in the dataset increases, the success in detecting those connectives also increases. This is an expected result because the increase in training data provides a better classification.

With the increase of corpora annotated at the discourse level, discourse parsing studies have also increased in these languages. Various shared tasks organized with the participation of researchers from many languages show that there is an ongoing interest in discourse parsing. Some of these shared tasks are CoNLL-2015 [54], CoNLL-2016 [55] and DISRPT 2019 [56]. Such shared tasks bring researchers and resources from the discourse parsing field together and contribute to the progress of research about this field.

CoNLL-2015 [54] is a shared task about Shallow Discourse Parsing, in which 16 countries from 3 continents participate. The goal of this shared test is to identify the discourse relations in a given English newswire text. To identify discourse relations, explicit discourse connectives, their arguments and sense should be found by the participating systems. PDTB was selected as the representation framework since it provides a shallow representation of discourse structure and it is the largest dataset annotated with discourse relations currently. The participating teams were asked to deploy their systems to run on the test set, rather than sending their own results. The F1 score of the best system is 24.0%, showing that there is an error propagation problem between the components of an end-to-end system.

CoNLL-2016 shared task [55] is a follow-on to the CoNLL-2015, this time it is about Multilingual Shallow Discourse Parsing. 23 countries from three continents participated in this task. The aim is to find the discourse relations in a natural language text, similar to CoNLL-2015, but in addition to English, the Chinese language is also included. The best system in English achieved a 27.77% F1 score, while the best

system in Chinese achieved a 26.90% F1 score.

Another shared task DISRPT 2019 [57] follows a series of events about discourse relation studies that take place every two years. These events, especially focusing on RST theory, were held first in Brazil (2007, 2009, 2011, 2013) and then in Spain (2015 and 2017). The 2019 workshop aimed to expand the scope by including participants from different discourse theories (especially, RST, SDRT and PDTB). 14 theoretical and applied papers and 4 shared task papers were published under this shared task.

DISRPT 2019 [56] is about elementary discourse unit segmentation and connective detection across different formalism. The DISRPT dataset contains 15 datasets from 10 languages, including both annotated and plain-tokenized data. For Turkish, three systems try the Turkish connective detection and the GumDrop system reports the highest performance with a 79.19 F-score on treebanked data and 79.17 F-score on plain data. The authors state that there is still more work to do in the field of discourse parsing, especially in lower resource and less studied languages. We briefly summarize the four shared task papers below.

Yu et al. [58] developed the GumDrop system, which is the contribution of Georgetown University to the DISPRT 2019 shared task on automatic discourse unit segmentation and connective detection. The system consists of three trainable parts, which are sentence splitting, discourse unit segmentation, and connective detection. The system accepts two types of inputs; raw text or gold syntax text containing sentence splits and PoS tags. When raw text is given as input, the text is first split into sentences and then parsed using UDPipe. Then for texts with predicted or gold syntax, discourse unit segmentation and connective detection are performed.

Bourgonje and Schafer [59] conducted various experiments on datasets of different languages and types. These datasets are annotated with coherence relations according to different theoretical frameworks. In particular, they searched for a method for discourse segmentation with a unified (theory-neutral) approach. For this purpose, they tried a RandomForest classifier and an LSTM classifier on 14 datasets from 9 languages. The results show that the performance of the methods varies quite a lot depending on the language and genre. In segmentation performance, genre rather than language seems to be a more reliable indicator. Therefore, the authors propose that

future work can be considered about genre differences in discourse segmentation.

Another group participating in the DISPRT 2019 workshop is Iruskieta et al. [60]. The authors emphasize that Elementary Discourse Units (EDUs) are quite similar in different theories and segmentation is one of the first steps in the rhetorical annotation. In this shared task, they use pre-trained word embeddings and a neural networks approach for segmentation on six languages.

Muller et al. [61] state that segmentation can be defined as finding the minimum text spans linked by the discourse relation. The authors develop a generic system, named ToNY to conduct the EDU segmentation and connective detection for multiple discourse representation schemes, languages and domains. They used straightforward sequence prediction architecture with pre-trained contextual embeddings, separately trained on each corpus and achieved a level of performance comparable to current systems.

There have been important improvements in the field of Turkish NLP in the last two decades, but most of these developments are at the sentence level [62]. The studies of Yüksel and Bozşahin [63], Yıldırım et al. [64], Tüfekçi et al. [65] Kurfalı et. al [66] are among the very few studies in the field of Turkish discourse structure. One of the main reasons why the studies on Turkish discourse structure are less than those in other languages is the scarcity of various Turkish resources such as discourse annotated corpora and discourse connective lexicon that can be used in these studies.

To summarize, this chapter provided the background for discourse connective lexicons, discourse relations, and discourse parsing. In the next chapter, we describe the data sources used in the thesis.

CHAPTER 3

DATA SOURCES

In this thesis, three main data sources are used, namely TDB 1.0 [22], TDB 1.1 [23] and T-TED-MDB [24]. All of these corpora are annotated in the style of PDTB [12]. For this reason, this section firstly gives information about PDTB, then continues with the discourse annotated Turkish corpora used in this thesis and concludes with some analysis showing the characteristics of the data sources.

3.1 PDTB

PDTB [12] is a large-scale resource annotated with discourse relations and their two abstract object arguments over the 1 million word Wall Street Journal (WSJ) corpus. Various versions of PDTB was published up to now which are PDTB-1.0, PDTB-2.0 and PDTB-3.0.

According to the PDTB [41], discourse connectives are "discourse-level predicates that take two abstract objects such as events, states, and propositions (Asher, 1993) as their arguments". Five types of discourse relations are annotated in the PDTB corpus. The definitions of these discourse relation types are given below [41]:

Explicit : In Explicit discourse relations, the relation between sentences, clauses, and rarely phrases are realized explicitly by using a connective such as *and, but, so* which are drawn from a predefined connective list.

Implicit : In Implicit discourse relations, no connective is used between the sentence pairs in a paragraph but the relation is inferred by the reader from the context. A discourse connective best expresses this relation is inserted between sentences which

are called *implicit discourse connective*. If there is more than one relation between sentence pairs, more than one implicit discourse connective can be inserted.

AltLex : Similar to Implicit discourse relations, AltLex discourse relations do not have an explicit connective. However, the relations are alternatively lexicalized with some non-connective expression. So, there is no need to insert an implicit connective.

EntRel : In EntRel relations, there is no relation between sentence pairs but the second sentence explains an entity in the first sentence.

NoRel : In NoRel relations, there is no relation or entity-based coherence relation between sentences.

The sense of Explicit, Implicit and AltLex relations are annotated but the sense of EntRel and NoRel relations are not annotated since there is no relation in these types. In sense annotation, the senses are selected from a three-level sense hierarchy.

The PDTB 2.0 version has been used widely by researchers in the NLP community and by researchers in psycholinguistics, as well. Additionally, PDTB 2.0 has encouraged the development of similar resources in other languages (such as Chinese, Czech, Hindi, Modern Standard Arabic, Turkish and most recently, French) and domains (such as biomedical texts and conversational dialogues).

After PDTB 2.0, PDTB 3.0 has been published [49] to enhance the PDTB 2.0 annotations and make up for the deficiencies observed in it. PDTB 3.0 includes more discourse relation annotations than PDTB 2.0. Also in PDTB 3.0, the annotation decisions of certain senses were standardized, some senses that were difficult to annotate were removed and some new senses were added. Thus, PDTB 3.0 has been made both a bigger and a better resource. The PDTB 3.0 sense hierarchy is given in Table 2.1.

3.2 TDB 1.0

The Turkish Discourse Bank (TDB) 1.0 is the first publicly available discourse annotated corpus developed for Turkish [22]. It is a large-scale corpus containing nearly 400,000-word in modern written Turkish with various genres. The discourse relations

in this corpus were annotated following the principles of PDTB.

In TDB 1.0, only explicit discourse connectives and phrasal expressions (such as *bundan dolayı* 'for this reason') in discourse relations were annotated with their two arguments. Implicit connectives and sense annotations were left for later versions. Annotation is done manually by the annotators using a tool [67] developed for the annotation of discourse structures in Turkish.

Discourse connectives in Turkish are drawn from three main syntactic categories and five classes. In the annotation of discourse connectives in TDB 1.0, these syntactic classes, excluding simplex subordinators, were used as the basis to prepare an initial list of connectives. In argument annotation, it is ensured that the smallest text spans, which give the meaning of discourse relation, are chosen as arguments by following the *minimality principle* [12].

Discourse connectives in Turkish are determined based on the following syntactic classes [68]:

- Coordinating conjunctions
  - Simple coordinating conjunctions
  - Paired coordinating conjunctions
- Subordinators
  - Simplex subordinators
  - Complex subordinators
- Anaphoric connectives

Simple coordinating conjunctions : Simple coordinating conjunctions combine two main clauses with the same syntactic type in Turkish like in English. These kinds of conjunctions generally link clauses in a sentence, but they can also link clauses in adjacent sentences. This type includes connectives such as *ve* 'and', *veya* 'or', *fakat* 'but' etc. The connective *çünkü* 'because' in sentence 20 is an example of a simple coordinating conjunction.

20. *Yapılarını kerpiçten yapıyorlar, ama sonra taşı kullanmayı öğreniyorlar. Mimarlık açısından çok önemli,* çünkü **bu yapı malzemesini baska bir malzemeyle beraber kullanmayı, ilk defa burada görüyoruz.**
    '*They constructed their buildings first from mud bricks but then they learnt to use the stone. Architecturally, this is very important* because **we see the use of this construction material with another one at this site for the first time.**'
    (Zeyrek and Webber [68])

Paired coordinating conjunctions : Paired coordinating conjunctions consist of two lexical items, usually, the second item is a duplicate of the first item. Two items together form a single discourse relation. This type includes connectives such as *ya ... ya* 'either... or', *ne ... ne* 'neither ... nor', *hem ... hem* 'both ... and' etc. The connective *ya ... ya* 'either ... or' in sentence 21 is an example of a paired coordinating conjunction.

21. *Birilerinin* ya *isi vardır, aceleyle yürürler,* ya **kosarlar.**
    '*Some people are* either *busy and walk hurriedly,* or **they run.**'
    (Zeyrek and Webber [68])

Simplex subordinators : Simplex subordinators are a combination of a non-finite verb and a suffix, called converb, such as *–sA*, 'if', *-(y)ArAk* 'by means of/ and' etc. There is a limited number of these connectives in Turkish and it is a small set. The converb *-erek* 'by means of/and' in sentence 22 is an example of a simplex subordinator.

22. Kafiye Hanım beni kucakladı, **yanagını yanagıma sürt**erek *iyi yolculuklar diledi.*
    'Kafiye hugged me and **by rubbing her cheek against mine,** *she wished me a good trip.*'
    (Zeyrek and Webber [68])

Complex subordinators : Complex subordinators usually consist of two parts; a postposition and a suffix on the (non-finite) verb of the subordinate clause. This type

of connectives has a larger set than simplex subordinators. In the following example, the lexical part of the connective is underlined while the suffix part is typed with small caps. The connective *için* 'since' in sentence 23 is an example of a complex subordinator.

23. **Herkes çoktan pazara çık**TIĞI için *kentin o dar, egri bügrü arka sokaklarını bosalmıs ve sessiz bulurduk.*
    'Since **everyone has gone to the bazaar a long time ago,** *we would find the narrow and curved back streets of the town empty and quiet.*'
    (Zeyrek and Webber [68])

Anaphoric connectives : Anaphoric connectives require only one argument syntactically and they take the second argument from the previous discourse. This type includes connectives such as *sonuç olarak* 'consequently', *ne var ki* 'however', *yoksa* 'or else' etc. The connective *sonuç olarak* 'consequently' in sentence 24 is an example of an anaphoric connective.

24. *Ali hiç spor yapmaz.* Sonuç olarak çok istedigi halde **kilo veremiyor.**
    '*Ali never exercises.* Consequently, **he can't lose weight**' although he wants to very much.'
    (Zeyrek and Webber [68])

In addition to these syntactic classes mentioned above, discourse connectives can also be drawn from the discourse adverbials such as *oysa* 'however', *öte yandan* 'on the other hand', *ayrıca* 'in addition/separately', *ne var ki* 'even so' etc. [22]

3.3 TDB 1.1

Turkish Discourse Bank (TDB) 1.1 is a subcorpus of TDB 1.0 enriched with the annotations of all PDTB-style relation types (Explicit, Implicit, AltLex, EntRel and NoRel) and the senses of explicit relations, implicit relations, and AltLexes [23]. It contains 20 files from various genres and has approximately 40K-word. This enrichment started by adding the senses of the previously annotated explicit discourse

connectives, then continued with the new annotations for Implicit relations, AltLexes and EntRels.

In sense annotation of the discourse relations in TDB 1.1, the PDTB 2.0 annotation manual and the PDTB 2.0 sense hierarchy were used. During the sense annotation process, three Level-2 senses were added to the sense list to cover the senses encountered in Turkish. These added senses are COMPARISON: Degree, EXPANSION: Manner and EXPANSION: Correction.

TDB 1.1 was annotated manually using the annotation tool (DATT) developed for TDB [67], as in TDB 1.0. An example annotation file created using this tool is given in figure 3.1.

```
<Relation note="" sense="COMPARISON: Concession: Arg2-as-denier"
    sense2="" type="EXPLICIT" type2="intra_conj">
    <Conn>
        <Span>
            <Text>ama</Text>
            <BeginOffset>311</BeginOffset>
            <EndOffset>314</EndOffset>
        </Span>
    </Conn>
    <Mod/>
    <Arg1>
        <Span>
            <Text>Avcı bir toplum, tarım henüz başlamış</Text>
            <BeginOffset>272</BeginOffset>
            <EndOffset>309</EndOffset>
        </Span>
    </Arg1>
    <Arg2>
        <Span>
            <Text>son derece gelişmiş bir toplumsal örgütlenme var</Text>
            <BeginOffset>315</BeginOffset>
            <EndOffset>363</EndOffset>
        </Span>
    </Arg2>
    <Supp1/>
    <Supp2/>
    <Shared/>
    <Supp_Shared/>
</Relation>
```

Figure 3.1: A sample relation annotated in TDB 1.1 file 00013112

In a later study, Zeyrek and Kurfalı [21] annotated a new set of intra-sentential explicit discourse connectives, i.e. converbs, in TDB 1.1. An example converb *-sA* 'if' is given in sentence 25.

25. **İzini bulursanız,** *bu numaraya haber verin.*

'If **you find his trail,** *call this number.*'
[Contingency:Condition]
(Zeyrek and Kurfalı [21])

One of the most challenging problems encountered during PDTB-style discourse annotation is to distinguish whether connectives are used as discourse or non-discourse. After this distinction, only discourse connectives were annotated in TDB 1.1. As with other connectives, converbs can be used as discourse or non-discourse. In case a converb is used as discourse, we call it as a suffixal discourse connective. Abstract object criterion works well in distinguishing discourse connectives from non-discourse ones.

Zeyrek and Kurfalı (2018) made several analyses on TDB 1.1. One of these analyses is the frequency comparison of inter-sentential and intra-sentential discourse connective usages. It was seen that explicit inter-sentential discourse connective use is more frequent than intra-sentential discourse connective use. The results of this analysis are given in Table 3.1 from Zeyrek and Kurfalı [21].

Table 3.1: Distribution of inter-S and intra-S explicit discourse relations in TDB 1.1

TYPE	Frequency	Ratio
Inter-S Explicit DR	688	79.2%
Intra-S Explicit DR	180	20.8%

Another analysis on TDB 1.1 corpus is about the relation between the inter-S or intra-S usage of a discourse connective and its sense. It has been observed that three explicit discourse connectives in Turkish, namely *ardından* 'then', *sonra* 'after' *önce* 'before', tend to give Asynchronous: precedence or Asynchronous: succession sense depending on their inter-S or intra-S usage. The examples below show how the sense of the connective *sonra* 'after' changes depending on its position:

26. *Durun, oturun biraz. Anlatın.* **Gidersiniz** sonra.
'*Wait, have a sit for a while. Tell us.* **You can leave** later.
[Temporal:Asynchronous:precedence]

[adverb]
(Zeyrek and Kurfalı [21])

27. **İyice kendine geldikten** sonra *getirilen kuru elbiseleri giydi.*
'After **regaining his consciousness,** *he wore the dry clothes they brought.*
[Temporal:Asynchronous:succession]'
[postposition]
(Zeyrek and Kurfalı [21])

Part of Speech (PoS) tags of these connectives, i.e. *ardından* 'then', *sonra* 'after' *önce* 'before', change when they are used inter-sententially or intra-sententially. Therefore, this situation shows that there is also a relationship between PoS tags and the senses of such connectives. In this thesis, this topic is addressed under the PoS ambiguity issue and in the machine learning experiments we used the PoS of the connective as a distinguishing feature in discourse connective sense classification.

3.4 TED-MDB

TED-MDB is a multilingual corpus created by annotating TED talks in English and translations into six languages (German, Polish, European Portuguese, Russian, Turkish and Lithuanian) at the discourse level following the principles of the PDTB [69]. While discourse annotated corpora developed for a single language are more common, multilingual discourse annotated corpora are still rare. However, such a resource would be very useful in cross-linguistic discourse analysis as well as other natural language processing applications such as discourse parsing, text understanding and text generation. This resource, developed for filling the gap in the multilingual discourse annotated corpora, has several important purposes. It will establish a basis for comparing the discourse structure of different languages and understand the similarities and differences. Also, it will be an important resource for developing discourse parsers for languages other than English.

TED talks are previously prepared monologues narrated to live audiences. TED website provides the transcripts of TED talks along with their translations into other

languages. Zeyrek et al. [69] give the list of TED talks annotated in TED-MDB as follows:

Table 3.2: TED talks annotated in TED-MDB

ID	Author	Title
1927	Chris McKnett	The investment logic for sustainability
1971	David Sengeh	The sore problem of prosthetic limbs
1976	Jeremy Kasdin	The flower-shaped starshade that might help us detect Earth-like planets
1978	Sarah Lewis	Embrace the near win
2009	Kitra Cahana	A glimpse of life on the road
2150	Dave Troy	Social maps that reveal a city's intersections and separations

Discourse connectives are assumed to be discourse-level predicates with two arguments. These arguments are abstract objects such as eventualities, facts, propositions as defined in Asher [3]. Explicit discourse relations were annotated when an explicit discourse connective was used inter-sententially or intra-sententially while implicit relations were annotated only inter-sententially. In implicit relation annotation, an implicit connective that expresses the sense of the relation is inserted. In addition to explicit and implicit relations, AltLex, EntRel and NoRel relationship types are also annotated. In this thesis, the explicit relation annotations in Turkish were used.

In argument annotation of discourse relations, PDTB's minimality principle was followed in selecting the text spans of arguments. For sense annotation, PDTB 3.0 sense hierarchy given in Table 2.1 was used.

TED talks have a specific genre where the purpose of the talker is to convince the audience about the talk's importance. In these talks, speakers usually ask some questions and then answer the questions themselves. These Q/R pairs aim to attract the attention of the audience to some topics. To meet the sense of such Q/R pairs, a new sense *hypophora* was added to the top level of the PDTB 3.0 sense hierarchy.

Zeyrek et al. [69] state in their paper that discourse relations across languages in TED-MDB were analyzed and various observations were made. The usage percentage of explicit connectives are close across the languages and in between 37% and 44%.

However, the distribution of other types of relations varies from language to language.

3.5 Analysis of Data Sources

In the present thesis, we first created the TCL, which is a comprehensive lexicon listing the single, phrasal and suffixal connectives in Turkish with their syntactic and semantic features. The TCL entries are compiled automatically from three discourse-annotated corpora, namely TDB 1.0, TDB 1.1, and T-TED-MDB.

Secondly, we tried to solve the usage ambiguity and sense ambiguity of connectives. For this purpose, we developed the TCD program to solve the usage ambiguity and several machine learning models to solve both the usage and the sense ambiguity of connectives. Both the disambiguator and the machine learning models were evaluated on the TDB 1.1 and T-TED-MDB datasets. In Table 3.3 below, we present summary information about these three corpora [25].

When a raw Turkish file is loaded into the TCD program, all candidate connectives in that file are automatically marked by the program. During the thesis studies, we loaded the raw files in the TDB 1.1 and T-TED-MDB datasets to the TCD program and listed the candidate connectives in the files. Among these candidate connectives, we noticed that some connectives are actually used as discourse but are not annotated by human annotators. So we conducted an annotation update study to add missing annotations to the annotation files. In Table 3.3, the number of annotations in the updated datasets and original datasets are given. In the thesis, we used the updated annotation files for evaluation purposes.

3.5.1 Discourse Usage Distribution of Turkish Connectives

Since one of the main purposes of the studies carried out in this thesis is to disambiguate the discourse usage of connectives from non-discourse uses, we analyzed discourse vs. non-discourse distribution of Turkish connectives in the T-TED-MDB and TDB 1.1 datasets.

We used the capabilities of the TCD program and the annotation files of the T-TED-

Table 3.3: Data sources

	Number of words	Annotations	Corpus characteristics	Number of explicit discourse connectives and phrasal expressions	Data source for
TDB 1.0	400000	Explicit relations, Phrasal expressions (because of this, as a result of this), Predicate-argument structure annotations	Multi-genre corpus of modern written Turkish	8439	TCL
TDB 1.1	40000	Explicit relations, Implicit relations, EntRels, AltLexes, NoRels, Predicate-argument structure annotations and senses	Multi-genre subcorpus of TDB 1.0	1441 (updated) 912 (original)	TCL, disambiguator
T-TED-MDB	5215	Explicit relations, Implicit relations, EntRels, AltLexes, NoRels, Predicate-argument structure annotations and senses	6 TED talk transcripts in English and translations into Turkish	309 (updated) 276 (original)	TCL, disambiguator

MDB and TDB 1.1 datasets to perform this analysis. Firstly, we calculated the total usage count of connectives in the datasets using the TCD program. Then we counted the number of connectives marked as discourse by human annotators in the annotation files. Thus, we obtained the DC and NDC connective counts in the T-TED-MDB and TDB 1.1 datasets.

The result of this analysis showed that the number of non-discourse usages is approximately 1.5 times the number of discourse usages showing the challenge in usage disambiguation of Turkish connectives. Table 3.4 below shows these counts [25].

Table 3.4: DC and NDC connective counts in T-TED-MDB and TDB 1.1

Dataset	# of DC usage	# of NDC usage	Total
T-TED-MDB	309	323	632
TDB 1.1	1441	2170	3611
Total	1750	2493	4243

3.5.2 Sense Distribution of Turkish Connectives

Another main purpose of the studies carried out in this thesis is to identify the sense(s) of the discourse connectives. So we conducted an analysis for observing the sense distribution of discourse connectives in Turkish.

In order to carry out this analysis, we benefited from the infrastructures we established in the TCL program and the output of the program. We listed the connectives with their possible senses and hence we had the chance to observe the connectives with a single sense and with more than one sense.

In Appendix D, we provide the distribution of explicit discourse connectives with their sense(s). When we examine this table, it is seen that most of the connectives in Turkish are ambiguous in terms of sense. 5 of 15 suffixal connectives, 28 of 75 single connectives and 19 of 90 phrasal connectives are used in a single sense, while other connectives are used in more than one sense as shown in Table 3.5. According to these numbers, we can state that nearly 70 % of the discourse connectives in Turkish are ambiguous in terms of sense.

Table 3.5: Unambiguous and ambiguous connective counts in T-TED-MDB and TDB 1.1

Connective Type	# Unambiguous DC	# of Ambiguous DC	Total
Suffixal	5	10	15
Single	28	47	75
Phrasal	19	71	90

3.5.3 Usage Frequencies of Turkish Connectives

The usage frequencies of words in natural languages play a crucial role in natural language processing tasks such as language acquisition [70]. Interestingly, the usage frequency of words in natural languages shows a Zipfian distribution i.e. some words are used very frequently while some words are used very rarely.

To observe whether a similar situation is valid for discourse connectives in Turkish, we conducted various analyzes on TDB 1.1 and T-TED-MDB datasets. In these analyzes, we examined three types of distributions i.e. the distribution of all types, the distribution of single/phrasal type and the distribution of suffixal type discourse connectives. As a result of these analyses, we observed that discourse connectives in Turkish show a Zipfian distribution, where the single/phrasal type has a rather skewed distribution while the suffixal type has a less skewed distribution. Graphs 3.2, 3.3 and 3.4 below show these distributions [25].

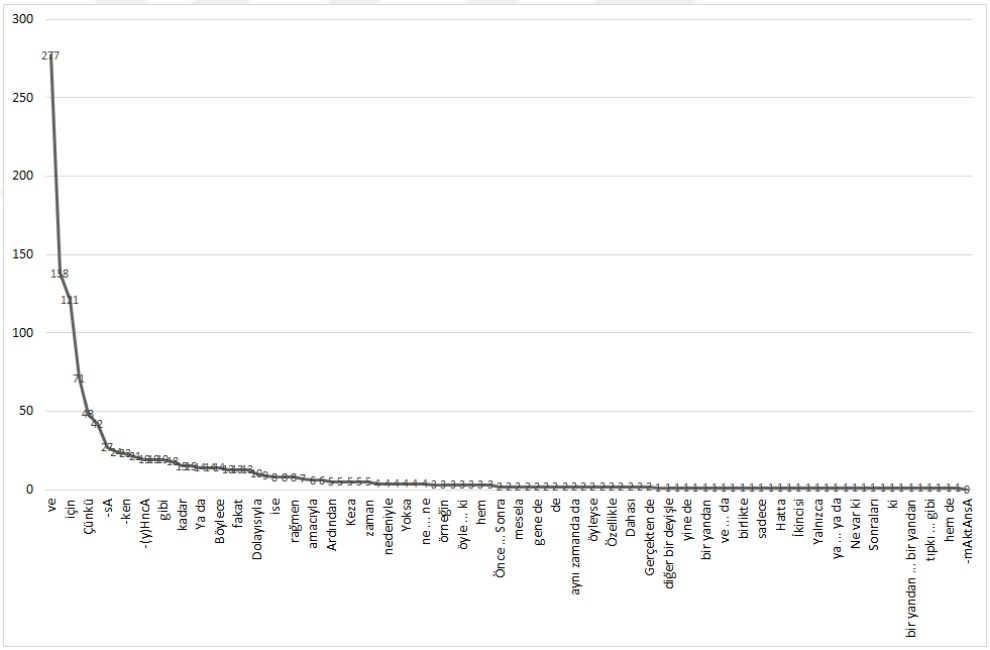

Figure 3.2: Usage frequencies of all types of connectives (single, phrasal and suffixal) in T-TED-MDB and TDB 1.1

In datasets where the data has a skewed distribution, machine learning evaluations should be done carefully. Shi and Demberg [71] recommend using the cross-validation approach in such datasets to avoid getting biased results. We followed this suggestion

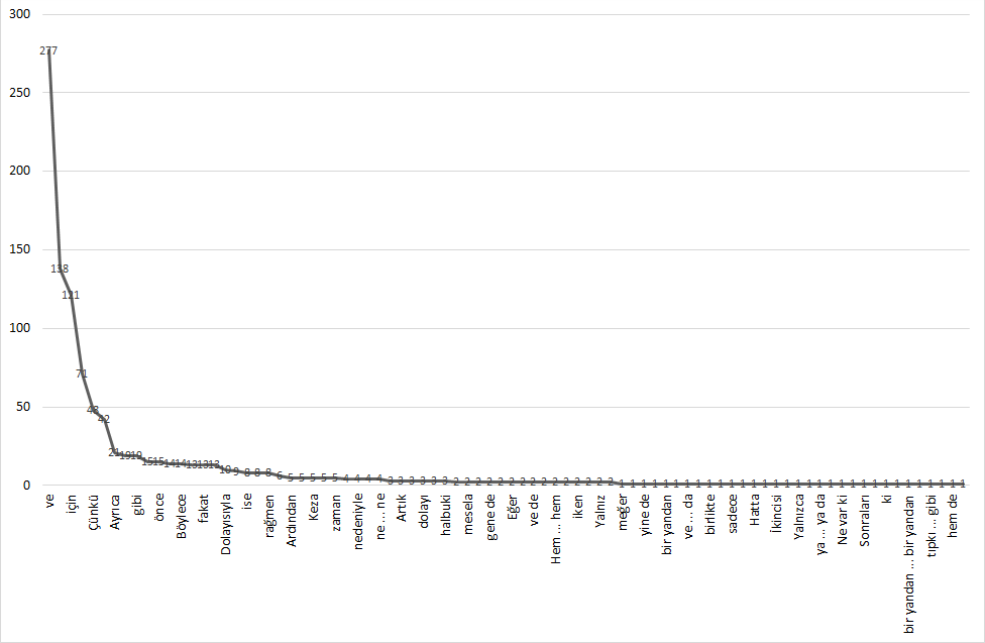

Figure 3.3: Usage frequencies of single and phrasal connectives in T-TED-MDB and TDB 1.1

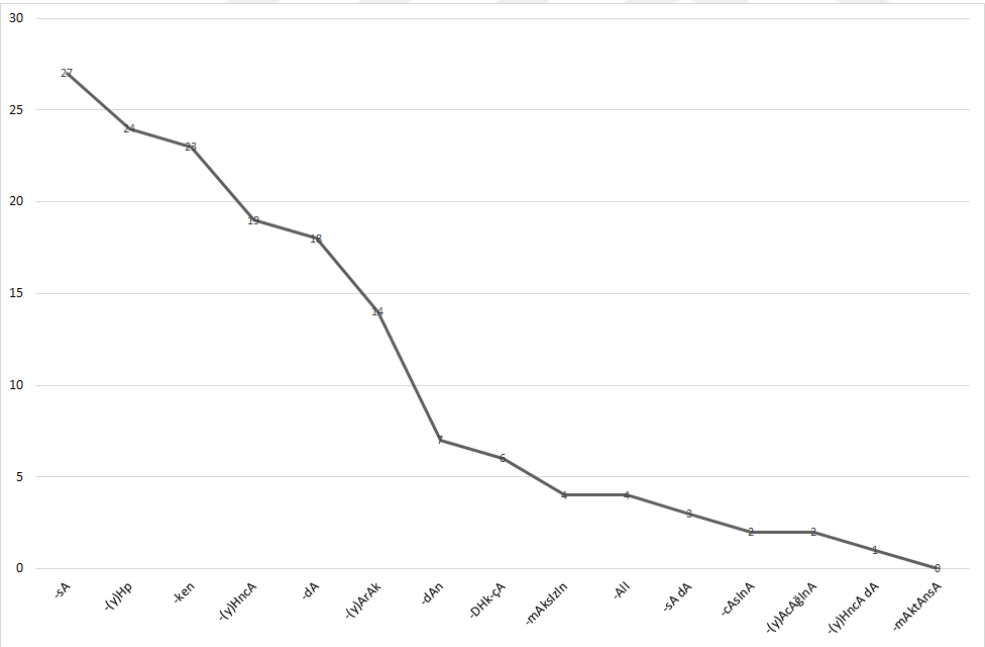

Figure 3.4: Usage frequencies of suffixal connectives in T-TED-MDB and TDB 1.1

in the thesis and applied cross-validation evaluation methodology in the machine learning evaluations.

To sum up, in this chapter, we provided detailed information about the data sources used in the thesis. In the next chapter, we will continue with the explanation of the TCL program.

CHAPTER 4

TURKISH CONNECTIVE LEXICON

A lexicon of discourse connectives can be defined as a dictionary that contains the discourse connectives in a language and provides several syntactic and semantic information about them. State-of-the-art discourse parsers generally use a lexicon of discourse connectives in annotating the discourse relations in a text [72], [73]. There is not such a discourse connective lexicon in Turkish, to the best of our knowledge. We wanted to fill this gap by building a comprehensive discourse connective lexicon for Turkish.

We developed the TCL, which includes the discourse connectives in Turkish with their syntactic and semantic information following the DiMLex XML format. In this chapter, we present the content, development steps and the interface of the TCL.

4.1 Data Sources

In the development of TCL, three PDTB-styled discourse-annotated corpora were used, namely, TDB 1.0 [22], TDB 1.1 [23] and T-TED-MDB [24]. Detailed information about these data sources was provided in Chapter 3. Here, we provide brief information about these data sources and explain how TCL makes use of them in order to compile the lexicon entries [18].

- TDB 1.0 is a multi-genre corpus of modern written Turkish containing about 400.000 words. In this corpus, explicit discourse connectives and phrasal expressions such as *bu sebeple* 'for this reason' (called as "secondary connectives" in the paper of Danlos [40]) are annotated together with their two arguments.

The semantic categories of the relations are not annotated in this corpus. In order to compile the list of connectives in TCL, we used the discourse relations established by explicit discourse connectives or phrasal expressions in this data source.

- TDB 1.1 is a subcorpus of TDB 1.0 containing about 40.000 words. This corpus is annotated with five relation types i.e., Explicit relations, Implicit relations, EntRels, AltLexes and NoRels, as well as the arguments and the sense(s). We used the explicit relations in this data source.

- TED-MDB is a multilingual discourse bank including the annotations of TED talks in English and translations into six languages (German, Polish, European Portuguese, Russian, Turkish and Lithuanian). We used the explicit discourse relations from the Turkish section of this data source.

We obtained several advantages by using multiple corpora in populating the TCL.

- First of all, some connectives that are not used in one corpus is used in another corpus. By creating the connective list from three corpora, we have gathered a more comprehensive list. We obtained 142 connectives from TDB 1.0, 78 connectives from TDB 1.1 and 55 connectives from T-TED-MDB and hence 180 connectives in total. The overlaps and differences between the data sources are shown in figure 4.1.

Figure 4.1: Distribution of connectives in data sources

- In some cases, a connective is used with a certain sense in a corpus, while it is used with another sense in another corpus. By collecting the senses of connectives from three corpora, we had the chance of listing as many senses as possible. Table 4.1 shows the different senses of various connectives obtained from different corpora.

Table 4.1: Different senses of various discourse connectives obtained from different corpora

DC	TDB 1.1	T-TED-MDB
ya da 'or'	EXPANSION: Substitution: Arg2-as-subst EXPANSION: Conjunction EXPANSION: Disjunction	Expansion.Disjunction
ancak 'however'	COMPARISON: Concession: Arg2-as-denier EXPANSION: Exception COMPARISON: Contrast	Comparison.Contrast Comparison.Concession.Arg2-as-denier Expansion.Conjunction Expansion.Exception.Arg2-as-excpt
yani 'that is'	EXPANSION: Equivalence EXPANSION: Level-of-detail: Arg2-as-detail CONTINGENCY: Cause+Belief: Result +Belief(Arg1-gives-evidence) EXPANSION: Substitution CONTINGENCY: Cause+Belief	Expansion.Equivalence Contingency.Cause+SpeechAct.Result+SpeechAct Expansion.Substitution.Arg2-as-subst
böylece 'thus'	CONTINGENCY: Cause: Result(Arg2-as-result)	Contingency.Cause.Result Contingency.Cause+Belief.Result+Belief (Arg1-gives-evidence)
keza 'also'	EXPANSION: Instantiation	Expansion.Conjunction
ise 'if'	COMPARISON: Contrast	Expansion.Conjunction Comparison.Concession.Arg2-as-denier Comparison.Contrast | Expansion.Level-of-detail.Arg2-as-detail
işte 'here'	CONTINGENCY: Cause+Belief: Reason+Belief (Arg2-gives-evidence)	Contingency.Cause.Result Expansion.Equivalence

- Phrasal expressions (a.k.a. secondary connectives) are only annotated in TDB1.0. Using this resource, we have added secondary connectives to the connective list in TCL.

The annotation files of TDB 1.1 are in XML file format while the annotation files of TDB 1.0 and T-TED-MDB are in pipe-delimited file format. These annotation file types are described in section 4.5.

4.2 Connective Categories in TCL

Discourse connectives in TCL can be listed under 5 main syntactic categories [18]:

- Coordinating conjunctions
  - Simple coordinating conjunctions
    *(*ama 'but', *çünkü* 'because/for', *ve* 'and' etc.)
  - Paired coordinating conjunctions
    *(*ne ... ne 'neither ... nor', *önce ... sonra* 'first ... then', *hem ...hem* 'both ... and' etc.)
- Subordinators
  - Simplex subordinators
    (-(y)ArAk 'by means of/and', *-sA* 'if', *-(y)HncA* 'when' etc.)
  - Complex subordinators
    (*için* 'since', *rağmen* 'despite', *gibi* 'as/like/such as' etc.)
- Adverbs
  (*yalnızca* 'only', *mesela* 'for example', *aslında* 'in fact' etc.)

Since detailed information about the connective categories has already been provided in Chapter 3.2, here we present some explanations about the *simplex subordinators* or in other words *suffixal connectives* in Turkish.

Representation of Suffixal Connectives in TCL

Simplex subordinators, i.e. converbs, can have discourse connective role since they establish a relation between a non-finite adverbial clause and the main clause. They are considered suffixal connectives in Turkish, and, just like other connective types, their discourse use is subject to certain linguistic constraints, as will be described later in the thesis. In the rest of this thesis, the term converb will be used interchangeably with the term suffixal connective.

Converbs have a rich allomorphy due to the vowel harmony and consonant harmony of Turkish. Their form varies depending on the words they are affixed. For example, the converb *-(y)Hp* can be used in the forms of *ıp, ip, up, üp, yıp, yip, yup, yüp*. In suffixal connective representation:

- **-** means that the suffixal connective is added to the end of a word

- **(y)** means that y becomes a part of the connective if the verb ends with a vowel
- **A** indicates the letter alternation as a, e
- **I** indicates the letter alternation as ı, i
- **D** indicates the letter alternation as t, d
- **H** indicates the letter alternation as ı, i, u, ü

Examples 28 and 29 show two different allomorphs of the suffixal connective *-(y)Hp*.

28. Ben Sierra Leone'da **doğ**-up *büyüdüm*
    I was **born** and *raised* in Sierra Leone
    (T-TED-MDB file 1971)

29. Yapraklar **açıl**-ıp *genişliyor.*
    The leaves are **opening** and *expanding.*
    (T-TED-MDB file 1976)

A discourse parsing pipeline requires the marking of connectives in a text in order to identify their discourse vs. non-discourse usage. In order to be able to mark the suffixal connectives in a text, the allomorphs of the connectives should be known. To meet this need, we provided all possible allomorphs of each converb in TCL as represented in Table 4.2.

4.3 The Structure of the Lexicon

The structure of the TCL is based on the DiMLex XML format. A sample discourse connective *gibi* 'like' in TCL is shown in figure 4.2. For each entry, the following information is supplied [18].

- **Entry** contains the id assigned to the connective and the connective string. (<entry id="32" word="gibi">)

Table 4.2: Converbs and their allomorphs in TCL

Converb	**Allomorphs**
-(y)ArAk	arak, erek
-sA	sam, san, sa, sak, sanız, salar, larsa, sem, sen, se, sek, seniz, seler, lerse
-(y)HncA	ınca, ince, unca, ünce, yınca, yince, yunca, yünce
-dAn	ımdan, imden, umdan, ümden, ından, inden, undan, ünden, ımızdan, imizden, umuzdan, ümüzden, ınızdan, inizden, unuzdan, ünüzden, larından, lerinden
-dA	ımda, imde, umda, ümde, ında, inde, unda, ünde, ımızda, imizde, umuzda, ümüzde, ınızda, inizde, unuzda, ünüzde, larında, lerinde
-(y)Hp	ıp, ip, up, üp, yıp, yip, yup, yüp
-DHk-çA	tıkça, tikçe, tukça, tükçe, dıkça, dikçe, dukça, dükçe
-AlI	alı, eli
-ken	ken
-cAsInA	casına, cesine
-mAksIzIn	maksızın, meksizin
-(y)AcAğInA	acağına, eceğine, yacağına, yeceğine
-mAktAnsA	maktansa, mektense
-sA dA	sam da, san da, sa da, sak da, sanız da, salar da, sem de, sen de, se de, sek de, seniz de, seler de
-(y)HncA dA	ınca da, ince de, unca da, ünce de, yınca da, yince de, yunca da, yünce de

```
<entry id="83" word="gibi">
  <orths>
    <orth canonical="0" orth_id="83o1" type="cont">
      <part type="single">gibi</part>
    </orth>
  </orths>
  <syn>
    <cat>ADP</cat>
    <sem>
      <pdtb2_relation anno_N="19" freq="3" sense="EXPANSION: Conjunction" />
      <example>kahve değirmeninin nerede olduğunu bilmediği gibi bulacağını da sanmıyordu.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="19" freq="15" sense="EXPANSION: Manner: Arg2-as-manner" />
      <example>beni büyülenmiş gibi dinliyorlardı.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="19" freq="1" sense="Comparison.Similarity" />
      <example>adeta çok küçük bir kasabada yaşıyormuşuz gibi gözüküyor.</example>
    </sem>
  </syn>
</entry>
```

Figure 4.2: A sample discourse connective *gibi* 'like' in TCL

- **Orthography** specifies whether the connective is made up of a single token (<part type="single">) or multiple tokens (<part type="phrasal">). In addition to these types, we added the type suffixal (<part type="suffixal">) in order to be able to define the suffixal connectives in Turkish. It, also, defines whether the connective is continuous (<orth type="cont">) or discontinuous (<orth type="discont">).

- **Variants** of a connective show different forms of the connective depending on its usage in the sentence. Allomorphs of a suffixal connective are also shown as variants in TCL. Figure 4.3 shows the variants of a single discourse connective *ve* 'and' while figure 4.4 shows the variants (allomorphs) of a suffixal discourse connective *-(y)ArAk* 'by means of'.

```
<entry id="162" word="ve">
  <orths>
    <orth canonical="0" orth_id="162o1" type="cont">
      <part type="single">ve</part>
    </orth>
    <orth canonical="1" orth_id="162o1" type="cont">
      <part type="single">Ve</part>
    </orth>
    <orth canonical="2" orth_id="162o1" type="cont">
      <part type="single">VE</part>
    </orth>
  </orths>
```

Figure 4.3: Variants of the single discourse connective *ve* 'and'

```
<entry id="2" word="-(y)ArAk">
  <orths>
    <orth canonical="0" orth_id="2o1" type="cont">
      <part type="suffixal">arak</part>
    </orth>
    <orth canonical="1" orth_id="2o1" type="cont">
      <part type="suffixal">erek</part>
    </orth>
  </orths>
```

Figure 4.4: Variants of the suffixal discourse connective *-(y)ArAk* 'by means of'

- **Canonical form** is the most commonly used form of the connective. In TCL, the canonical form of a connective is given as the first variant (<orth canonical="0">) of the connective. This property is set only for single and phrasal type connectives.

- **Syntactic category** gives the part of speech tag of the connective. In order to assign the syntactic categories of connectives in Turkish, we benefitted from various sources i.e. the Turkish section of UDPipe [1], TDK [2] and TRmorph [3].
- **Semantic category** presents the senses of the connective used in the corpora. A sample sentence is also provided for each sense of the connective in TCL.
- **Frequency** specifies how many times the connective is used in the corpora and how many times it is used in each sense.

4.4 TCL Search Tool

We have developed a connective search tool in the Visual Studio environment by using the C# programming language. The purpose of this tool is to create a lexicon containing the explicit discourse connectives in Turkish with their syntactic and semantic features, using the discourse relations in a given corpus. In this section, we explain the interface and the capabilities of the TCL search tool given in figure 4.5. We mentioned the TCL search tool in the paper [18], then we made some minor updates for this tool.

We explain the instructions for using the TCL search tool below:

- Select the path of the annotation files and raw text files of each corpus by using *Select* buttons.
- Check the *Use* checkbox next to the corpus from which you want to obtain discourse connectives
- Check the *Suffixal* radio button if you want to list only suffixal connectives
- Check the *Single* radio button if you want to list only single connectives
- Check the *Phrasal* radio button if you want to list only phrasal connectives
- Check the *Min Frequency* radio button and enter a number if you want to list the connectives used above a certain frequency in the corpora

[1] http://lindat.mff.cuni.cz/services/udpipe/
[2] http://www.tdk.gov.tr
[3] http://coltekin.net/cagri/trmorph/

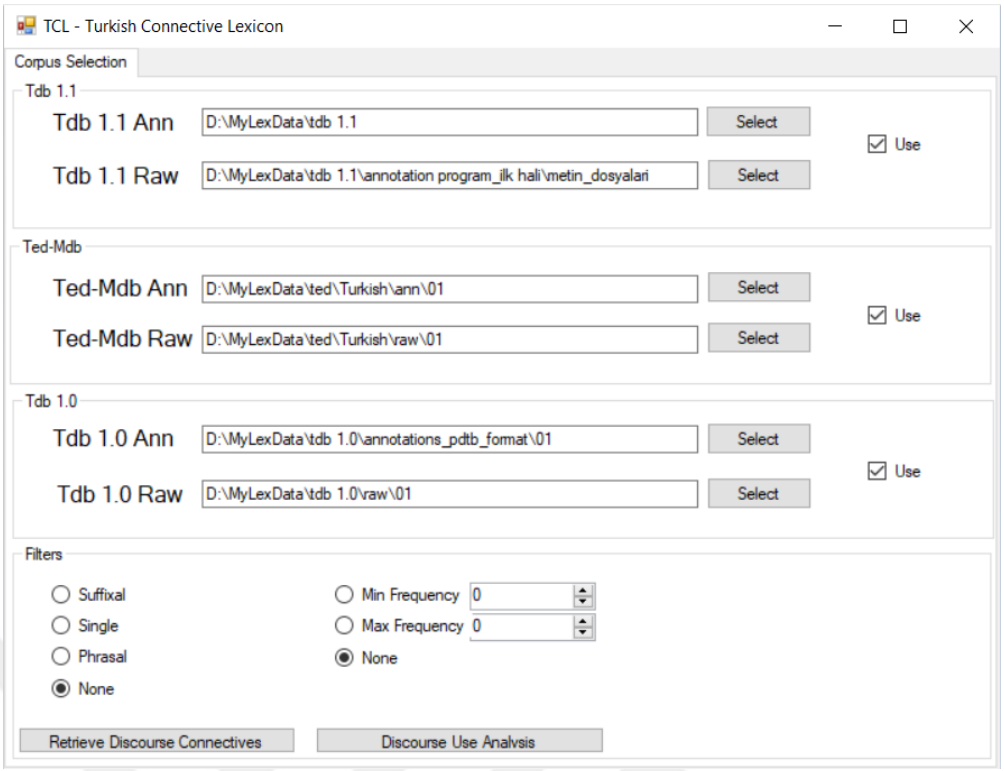

Figure 4.5: Interface of the TCL search tool

- Check the *Max Frequency* radio button and enter a number if you want to list the connectives used under a certain frequency in the corpora

- Do not use any of the filters if you want to list all the connectives used in the corpora

- Click on the *Retrieve Discourse Connectives* button

When the *Retrieve Discourse Connectives* button is pressed, the TCL search tool lists the connectives in the selected corpora filtered by the filters. The output screens of the TCL search tool are presented below. The first tab of the output screen presents discourse connectives and their syntactic categories (4.6). The second tab lists the syntactic and semantic information about the connectives (4.7). In the third tab, the PDTB 3.0 sense distributions of each connective are listed (4.8).

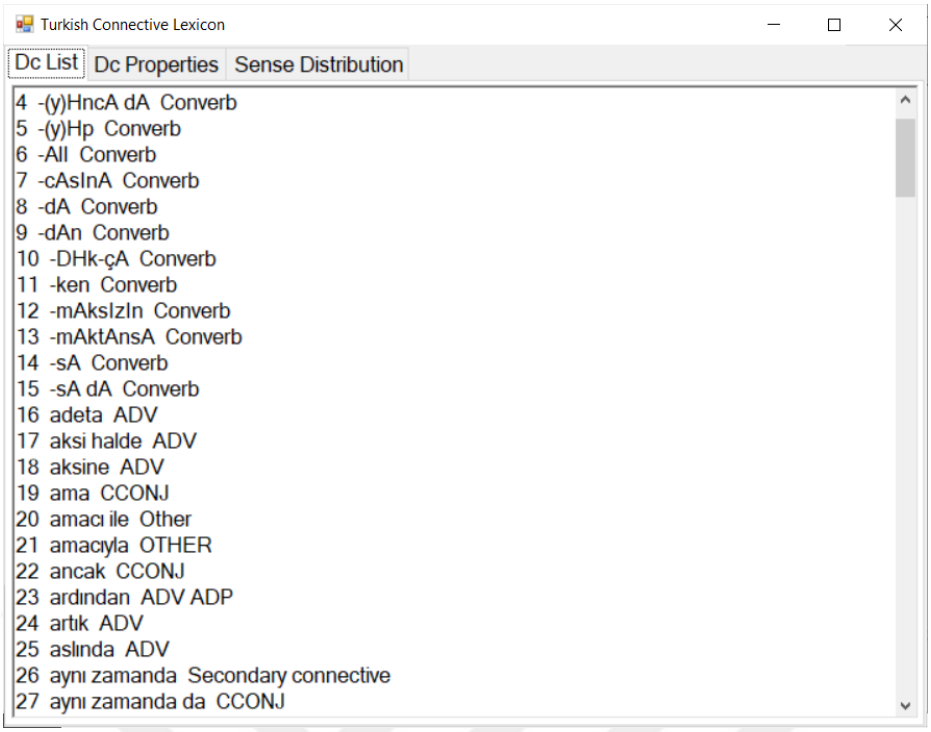

Figure 4.6: First tab of the TCL search tool result screen

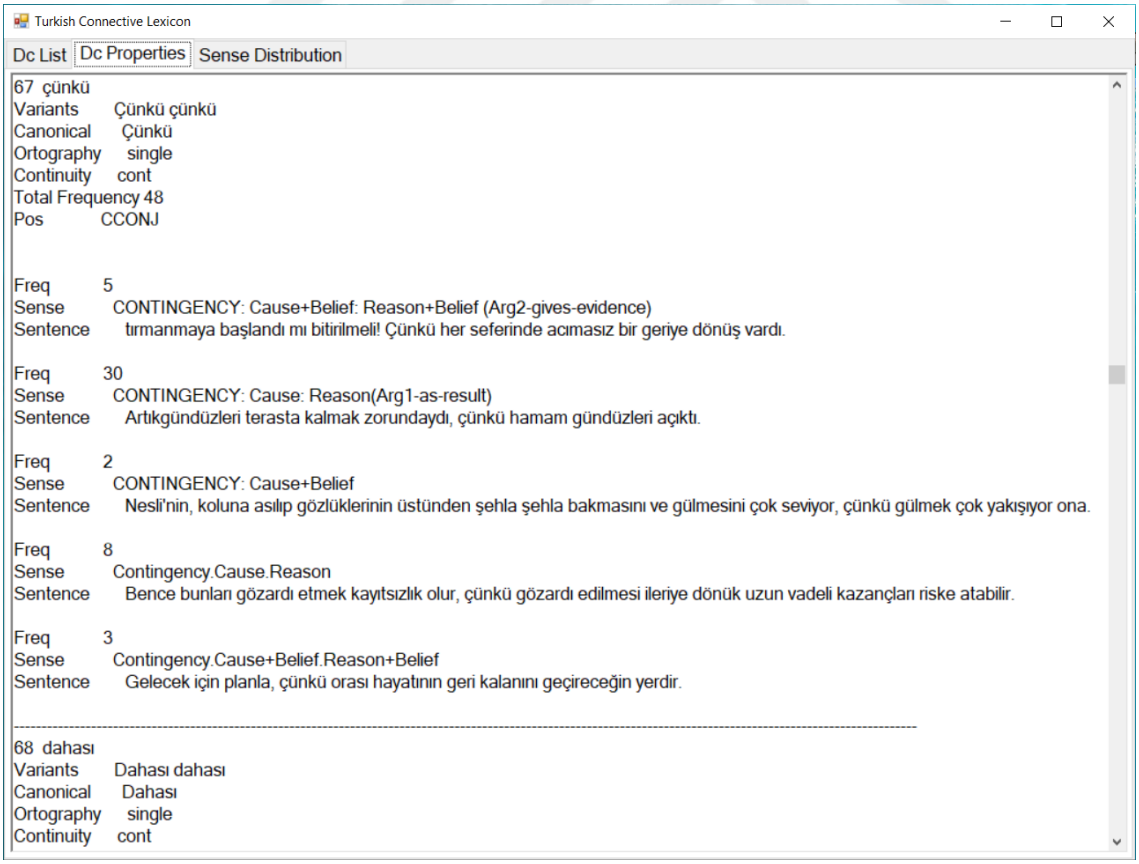

Figure 4.7: Second tab of the TCL search tool result screen

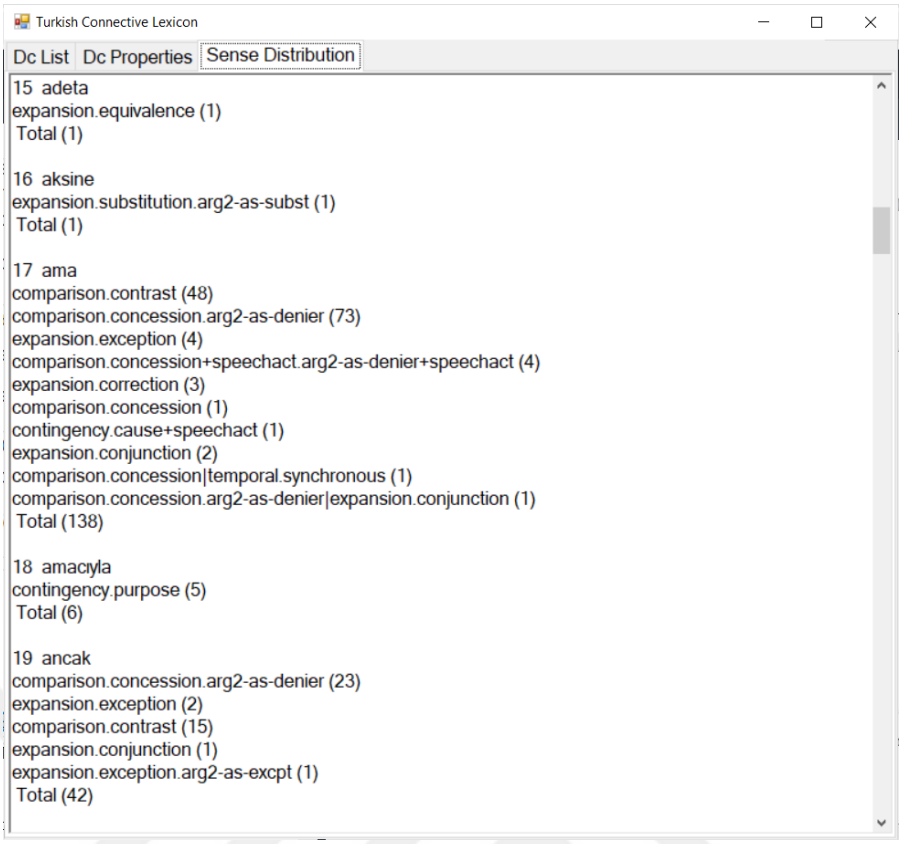

Figure 4.8: Third tab of the TCL search tool result screen

Many connectives in Turkish can be used in discourse or non-discourse roles depending on the context. In order to discover the discourse usage rates of the connectives, we added an analysis that counts the discourse and non-discourse usages of connectives and calculated the discourse usage ratio. This analysis is performed for the single and phrasal continuous type connectives on TDB 1.1 and T-TED-MDB datasets. When a user presses the *Discourse Use Analysis* button, the discourse use analysis results screen is shown (4.9). In this screen, each line shows the discourse use analysis results of a connective with the following information:

- Connective string
- Total usage count of the connective
- Discourse usage count of the connective
- Non-discourse usage count of the connective

- Discourse usage percentage of the connective

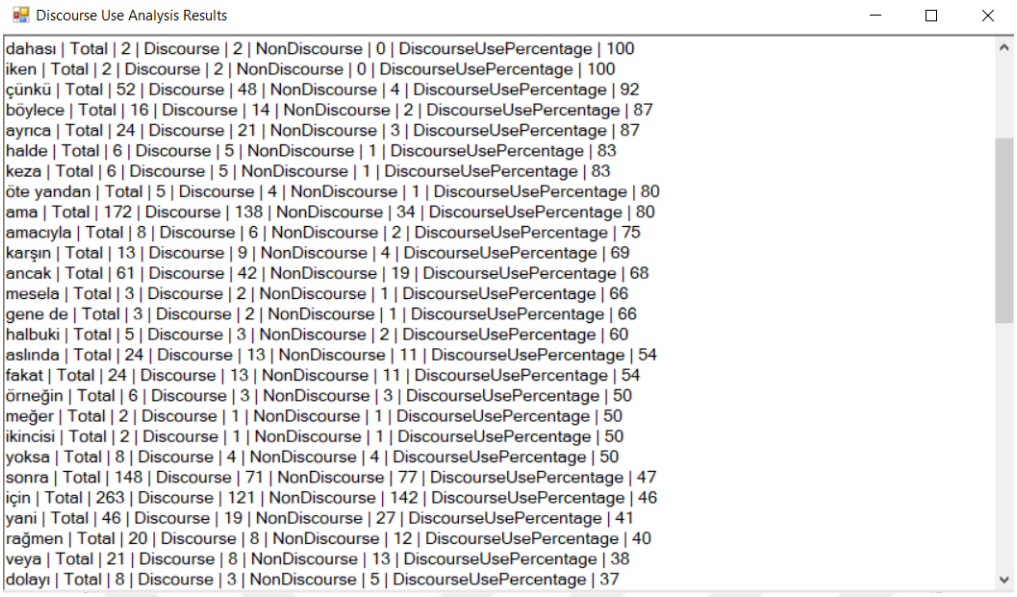

Figure 4.9: Discourse use analysis results screen

4.5 TCL Development Steps

The development of the TCL starts with reading the annotations from TDB 1.0, TDB 1.1 and T-TED-MDB annotation files and the relevant sentences from the raw text files. Then the process continues with the computation of syntactic and semantic features of connectives. TCL is finalized by mapping the connectives and their features into DiMLex XML format. The 4.12 shows the TCL flowchart. Details of the modules in the figure are explained below [18].

XML File Format: In XML type annotation files, each *Relation* node under *Document* node shows a discourse relation. Information about discourse relation such as the connective, its type, sense, arguments and supplementary material are given in the attributes and sub-nodes of the relation node. A sample annotation file in XML format from TDB1.1 is given in figure 4.10.

Pipe-delimited File Format: In this type of annotation file, each line shows a discourse relation. The parts separated by a pipe in a row gives information about the fields of the relation. The first field is *Relation Type* such as Explicit, Implicit, AltLex,

```
<Relation note="conjunction" sense="EXPANSION: Conjunction"
    sense2="" type="EXPLICIT" type2="intra_conj">
    <Conn>
        <Span>
            <Text>ve</Text>
            <BeginOffset>388</BeginOffset>
            <EndOffset>390</EndOffset>
        </Span>
    </Conn>
    <Mod/>
    <Arg1>
        <Span>
            <Text>İşbölümü ayrıntılanmış</Text>
            <BeginOffset>365</BeginOffset>
            <EndOffset>387</EndOffset>
        </Span>
    </Arg1>
    <Arg2>
        <Span>
            <Text>yönetici sınıf oluşmuş</Text>
            <BeginOffset>391</BeginOffset>
            <EndOffset>413</EndOffset>
        </Span>
    </Arg2>
    <Supp1/>
    <Supp2/>
    <Shared/>
    <Supp_Shared/>
</Relation>
```

Figure 4.10: A sample annotation file in XML file format from TDB1.1

EntRel, NoRel. The second field is *Conn SpanList* which shows the text spans of the connective. In the following fields, semantic classes, text spans of the arguments, text spans of supplementary materials and some other information are given. Detailed information about these fields can be found on the PDTB Annotator website [4]. The 4.11 shows an example annotation file from T-TED-MDB dataset.

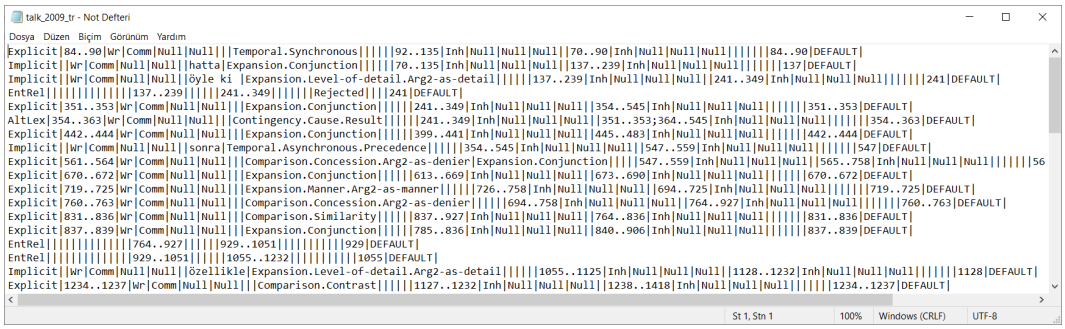

Figure 4.11: A sample annotation file in pipe-delimited file format from T-TED-MDB dataset

XML File Parser and Pipe-delimited File Parser: TDB 1.1 annotation files are in XML file format. So, an XML File Parser has been developed to read the information in these annotation files. TDB 1.0 and T-TED-MDB annotation files are in pipe-delimited file format. In order to read these annotation files, a Pipe-delimited File Parser has

[4] https://www.cis.upenn.edu/ pdtb/annotator.html

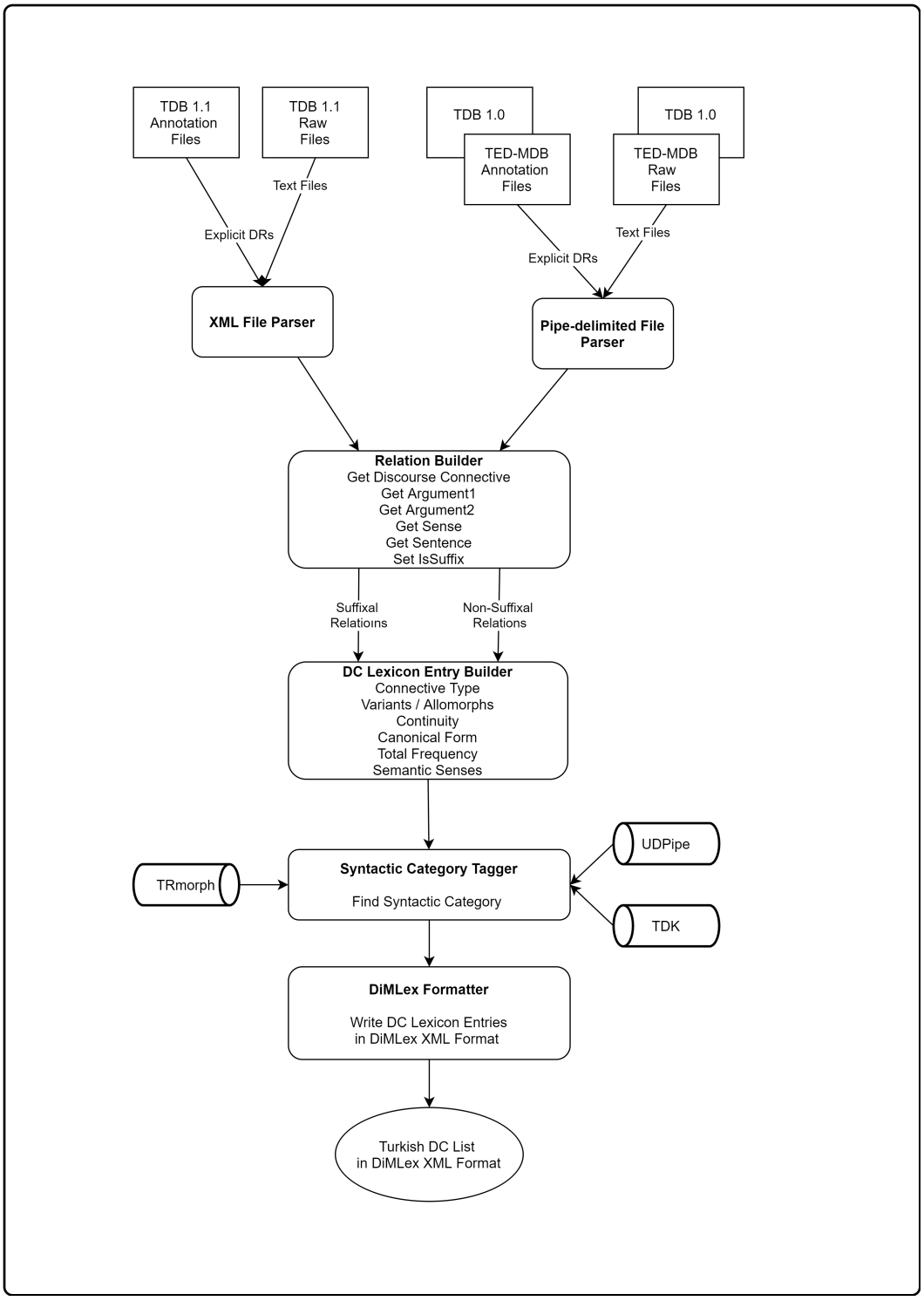

Figure 4.12: TCL Flowchart

been developed. The parsers read the annotation files and pull the EXPLICIT type relations in them.

Relation Builder: The Relation Builder populates the following information about each relation drawn from annotation files. Annotation files either directly give this information or give the text spans. In case where text spans are given in the annotation files, the relevant text is pulled from the raw text files.

- Connective
- Modifier of the connective
- Arguments of the connective
- Sense of the connective

Using the connective, modifier, argument1 and argument2 text spans in the annotation file, the sentence in which the connective occurs is drawn from the raw text file. Also, whether the connective is suffixal or not is calculated using the raw text file. The Relation Builder checks the previous character of the DC token. If the previous character is a white space, a newline or one of these separators (. , : ; ! ? ( ) ‘ “ -), the DC is not suffixal, otherwise, it is suffixal. Hence, the Relation Builder creates two relation lists, suffixal and non-suffixal.

DC Lexicon Entry Builder: This module takes the suffixal and non-suffixal relation lists from the Relation Builder module and calculates the following properties for the connectives in these lists.

- **Connective Type:** The type of the connectives in the suffixal relation list are *suffixal* while the type of the connectives in the non-suffixal relation list can be *single* or *phrasal*. The type of non-suffixal connectives consisting of only one word is set as *single* and the type of those consisting of more than one word is set as *phrasal*.
- **Orthographical Variants:** Connectives may have several orthographical variants depending on their usage in text. By searching all usages of connectives through the corpora, the variants have been found. For suffixal connectives, the allomorphs have been listed in TCL instead of orthographical variants.

- **Allomorphs:** Suffixal connectives in Turkish have complex allomorphy based on vowel harmony and consonant harmony. Suffixal relation list entries were used to obtain the allomorphs of connectives and the missing allomorphs not included in these lists were added manually.

- **Continuity:** Suffixal and single type connectives are continuous in all cases. However, phrasal connectives may be continuous or discontinuous. To set the continuity property of the phrasal type connectives, it is checked whether another word is inserted between the parts of the connective in the sentence. If no word is used between the parts of the connective, it is marked as continuous, otherwise, it is marked as discontinuous.

- **Canonical Form:** Canonical form is the most commonly used form of a connective. In order to find the canonical form of a connective, the frequency of all variants of the connective is calculated. The most frequent variant is labelled as canonical. This property is not calculated for suffixal connectives.

- **Total Frequency:** By counting the number of usages of a connective in the entire corpora, the frequency of the connective is calculated.

- **Semantic Category:** To find the semantic category of a connective, all entries of the connective in the relation list are checked. The semantic category in these entries are added to the TCL as the senses of the connective together with a sample sentence and the sense frequency.

Syntactic Category Tagger: The Syntactic Category Tagger assigns the syntactic categories of the connectives. We assigned the syntactic category of the *suffixal* connectives as *converb*, while we used various sources to find the syntactic category of the *single* and *phrasal* type connectives. These sources are listed below.

- **UDPipe:** UDPipe [5] presents a pipeline for tokenization, tagging, lemmatization and dependency parsing of text files in many languages. The Turkish section of UDPipe has been used in finding the PoS tags of discourse connectives in TCL.

[5] http://lindat.mff.cuni.cz/services/udpipe/

- **TRmorp:** TRmorph [6] is a morphological analyzer for Turkish. A given word is split into its root and suffixes and then PoS tags of the split parts are assigned.
- **TDK:** TDK [7] offers a comprehensive Turkish dictionary that gives the definition of words and their role in sentences.

Some discourse connectives in Turkish are lexicalized forms of nominal roots. For such connectives, PoS taggers may give incomplete information. While assigning the PoS tags of such connectives, we tried to reach a common result by comparing the PoS tag assignments of UDPipe, TRmorph and TDK. In case we cannot reach a common PoS tag, we used the *Other* tag.

Table 4.3: Different parses for noun based DCs and assigned syntactic categories

Connective	UDPipe	TRmorph	Result
amacıyla	Noun	Noun	Other
ardından	Adv	Adv, Noun	Adv, Adp
keza	Noun	Cnj:adv	Adv
zaman	Noun	Noun, Postp:adv	Other
halde	Noun	Noun, Postp:adv	Other
nedeniyle	Noun	Noun, Adv:qst	Other
sonuçta	Noun	Noun, Adv	Other
yahut	Noun	Cnj:adv	Cconj
sayesinde	Noun	Noun	Other
yüzünden	Noun	Noun, Num, Adv:partial	Other
karşılık	Noun	Noun, Adj, Postp:adv, Adv	Other
sonucunda	Noun	Noun	Other

It is possible for a connective to have more than one syntactic category. In such cases, more than one syntactic category has been assigned to connectives in TCL. In example 30 below, the connective *önce* 'before/first' is used as ADP while in example 31 it is used as ADV.

[6] http://coltekin.net/cagri/trmorph/
[7] http://www.tdk.gov.tr

30. **genel başkanlığa adaylığını koymadan** önce *destek turu için İstanbul İl Merkezi'ni ziyaret etti.*
He visited Istanbul City Center for the support tour before **nominating his candidacy for the chairmanship.**
(ADP)
(TDB 1.1 file 20580000)

31. **Gökkafes'in 23'üncü katında yer alan süper lüks daire,** önce **1 trilyon 710 milyar liradan peşin olarak satışa çıkarıldı.** *Peşin alıcı çıkmayınca, aynı daire 24 ay vadeli olmak üzere 2 trilyon 120 milyar 400 milyon liradan satışa sunuldu.*
The super luxury flat located on the 23rd floor of Gökkafes was first **put up for sale for 1 trillion 710 billion liras.** *When there were no buyers in advance, the same flat was offered for sale for 2 trillion 120 billion 400 million liras with a maturity of 24 months.*
(ADV)
(TDB 1.1 file 20570000)

Similarly, examples 32 and 33 illustrate the sample usages of the connective *sonra* 'after/then' as ADP and ADV.

32. **Bir süre bakıştıktan** sonra,*"Tanıyamadın değil mi?" dedi.*
After **staring for a while**, *"You couldn't recognize, could you?" he said.*
(ADP)
(TDB 1.1 file 00065111)

33. *Keban'la başladık,* sonra **yine aynı çerçevede Aşağı Fırat'a indik.**
We started with Keban, then **went down to Lower Euphrates within the same frame.**
(ADV)
(TDB 1.1 file 00013112)

The following examples 34 and 35 show the different syntactic categories of the connective *ardından* 'after/then'.

34. *Kalanların* **prosedürlerin tamamlanmasının** ardından *ihale yöntemiyle satılacağı ...*
The remainder to be sold by the tender method after **the completion of the procedures ...**
(ADP)
(TDB 1.1 file 20590000)

35. *Araçlarını terk edip yürüyerek Pile'ye geliyorlar.* Ardından **kolayca Rum tarafına geçiyorlar.**
They walk to Pile by leaving their vehicles. Then **they easily pass to the Greek side.**
(ADV)
(TDB 1.1 file 20570000)

The following syntactic categories are assigned to the discourse connectives in TCL: CCONJ, Converb (Simplex Subordinator), ADP (Postposition, Complex Subordinator), ADV, Secondary Connective (Phrasal Expressions) and Other categories. These syntactic categories are mapped onto the syntactic categories of the connective-lex website as follows.

Table 4.4: Syntactic category mapping between Connective-Lex and TCL

Connective-Lex	TCL
cco	CCONJ
csu	Converb, ADP
adv	ADV
prep	-
other	Secondary Connective, Other

Dimlex Formatter: After all the information about connectives are obtained through data mining methods and listed in the TCL, the Dimlex Formatter converts all the connectives and their information into the DiMLex XML format. For this purpose, an XML file with root <dimlex> is created and an <entry> node is added for each connective. The 4.13 shows a sample connective *mesela* ''for example' in TCL while

figure 4.14 shows how the same connective is exported to DiMLex format.

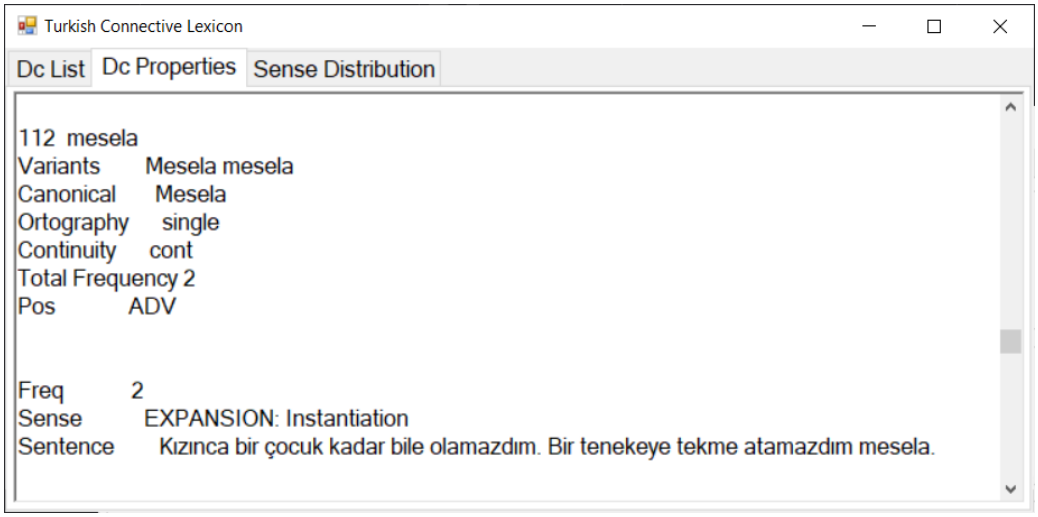

Figure 4.13: Connective *mesela* 'for example' in TCL

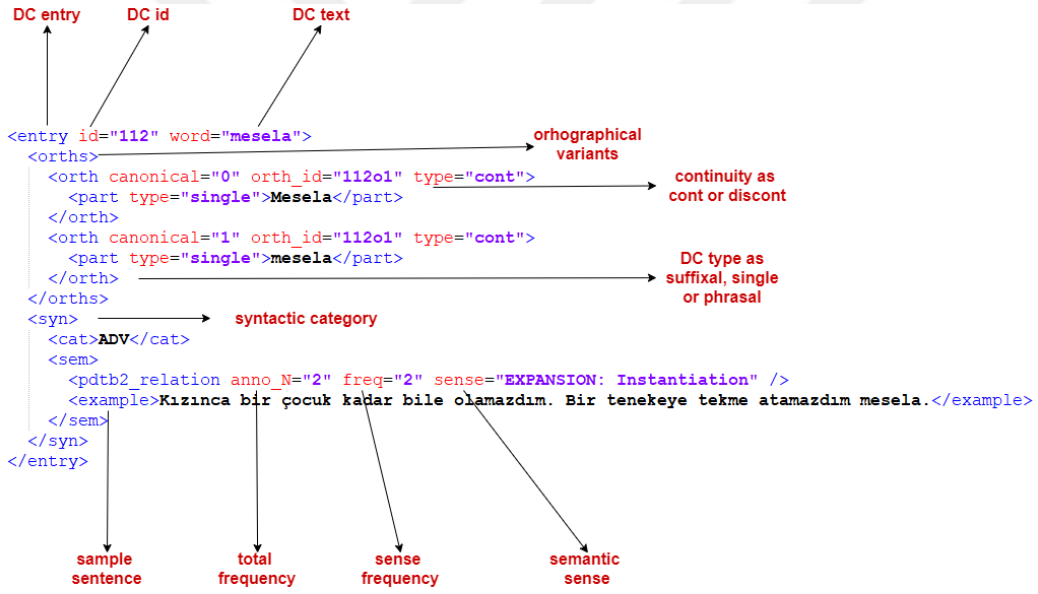

Figure 4.14: Connective *mesela* 'for example' in DiMLex format

Turkish DC List in DiMLex XML Format: As a result, TCL was populated with a total of 180 connectives. Of these, 15 are suffixal, 75 are single and 90 are phrasal. These connectives with their syntactic and semantic information are written to an XML file in DiMLex file format. We give examples of suffixal, single and phrasal type connectives in DiMLex format in the appendices section C.

In summary, we have created a list of discourse connectives in Turkish with their

syntactic and semantic properties both through a program interface and as a list in DiMLex format with this study. We used this list of connectives in our discourse disambiguation studies in the thesis. We hope that the researchers in the NLP literature would benefit from the TCL in their discourse parsing or similar discourse-related studies.

By using the infrastructure and the data structures we established, TCL can be used to conduct several analyses related to Turkish connectives and it can be enriched in various ways in the future. For example, while it currently lists only the Explicit type connectives, it may be expanded to include other types of connectives such as AltLexes. Additionally, TCL now contains samples of discourse usage of connectives. Samples of non-discourse usage may be added in the future.

Having described the TCL program in the present chapter, we continue with the TCD program in the next chapter.

CHAPTER 5

TURKISH CONNECTIVE DISAMBIGUATOR

The aim of the TCD program is to find all the connectives in a given Turkish text and disambiguate their usage as discourse or non-discourse. For this purpose, TCD uses the Turkish connective list from TCL and two rule-based decision trees, one developed for solving the usage ambiguity of single/phrasal connectives and the other developed for solving the usage ambiguity of suffixal connectives.

As preparatory work for TCD, an analysis has been carried out about the discourse usage rates of Turkish connectives. In this section, we first explain how we performed the analysis, and we use the analysis results in the disambiguator also detailing the specific characteristics of the disambiguator.

5.1 Discourse Usage Analysis

Some of the connectives in Turkish are mostly used as discourse while some of them are rarely used as discourse. The purpose of the *Discourse Usage Analysis* study is to discover the discourse usage rates of the Turkish connectives. The analysis results are used in the discourse usage disambiguation process of single/phrasal type connectives in TCD. Thus, the TCD program classifies some of the connectives using a probabilistic approach before using the classification rules.

The discourse usage analysis was carried out by using the TDB 1.1 and T-TED-MDB datasets. A total of approximately 4000 sentences were used from these two datasets. The analysis steps can be summarized as follows. For each connective (single or phrasal cont.) in the TCL connective list:

- The total occurrence of the connective in the raw text files of datasets has been counted and this is called as *usage count*.
- The total occurrence of the connective annotated as discourse in the annotation files of datasets has been counted and this is called as *annotation count*.
- Discourse usage percentage is calculated with the ratio of the *annotation count* to the *usage count* times a hundred.

This analysis capability has been developed through the TCL search tool as mentioned in section 4.4. A sample part of the analysis results is listed in Table 5.1 and the all results are given in Appendix A.

Table 5.1: A sample part of discourse usage analysis results

Connective	PoS	Connective Type	Gloss	Total (#)	DC Use (#)	NDC Use (#)	DC Use (%)
ne var ki	ADV	phrasal	nevertheless	1	1	0	100
ve de	CCONJ	phrasal	also	2	2	0	100
oysa	CCONJ	single	hovewer	13	13	0	100
dolayısıyla	ADV	single	therefore	10	10	0	100
dahası	ADP	single	furthermore	2	2	0	100
böylece	ADV	single	thus	16	14	2	87
ama	CCONJ	single	yet	172	138	34	80
halbuki	CCONJ	single	whereas	5	3	2	60
fakat	CCONJ	single	but	24	13	11	54
için	ADP	single	since	263	121	142	46
tersine	ADV	single	in contrast	3	1	2	33
ve	CCONJ	single	and	905	277	628	30
önce	ADV ADP	single	before	74	15	59	20
kadar	ADP	single	until	144	15	129	10
eğer	CCONJ	single	if	30	2	28	6
zaten	ADV	single	after all	27	1	26	3
diye	ADP	single	so that	97	1	96	1

As a result of this analysis, connectives are grouped under three categories.

- The connectives that are always used as discourse are grouped under the *Mostly*

DC category. TCD directly assigns the *discourse* label to the connectives under this category.

- As a result of this analysis, we observed that there are no connectives that always serve as non-discourse. Therefore, connectives used as discourse under a certain percentage (20 percent), in other words, rarely used as discourse, are grouped under the *Mostly NDC* category. TCD directly assigns the *non-discourse* label to the connectives under this category.

- Connectives that can be used as discourse or non-discourse are grouped under a third category, *DC or NDC*. For the connectives in this group, TCD decides between discourse vs. non-discourse roles using classification rules developed for single/phrasal type connectives.

These categories and the list of connectives under the categories are given in Appendix B [25].

5.2 Turkish Connective Disambiguator Workflow

TCD is a program developed in Visual Studio .Net environment using C# programming language. The general workflow of this program is given in figure 5.1 and can be summarized as follows:

- Single, phrasal and suffixal connective lists produced by TCL are used as input.

- C# library offered by UDPipe is integrated into the Visual Studio environment and the library's methods are used for text processing purposes.

- To disambiguate the single/phrasal type connectives, discourse usage analysis results and the single/phrasal decision tree are used .

- To disambiguate the suffixal type connectives, the suffixal decision tree is used.

Preprocessing

In this stage, firstly the C# library offered by UDPipe has been integrated into the

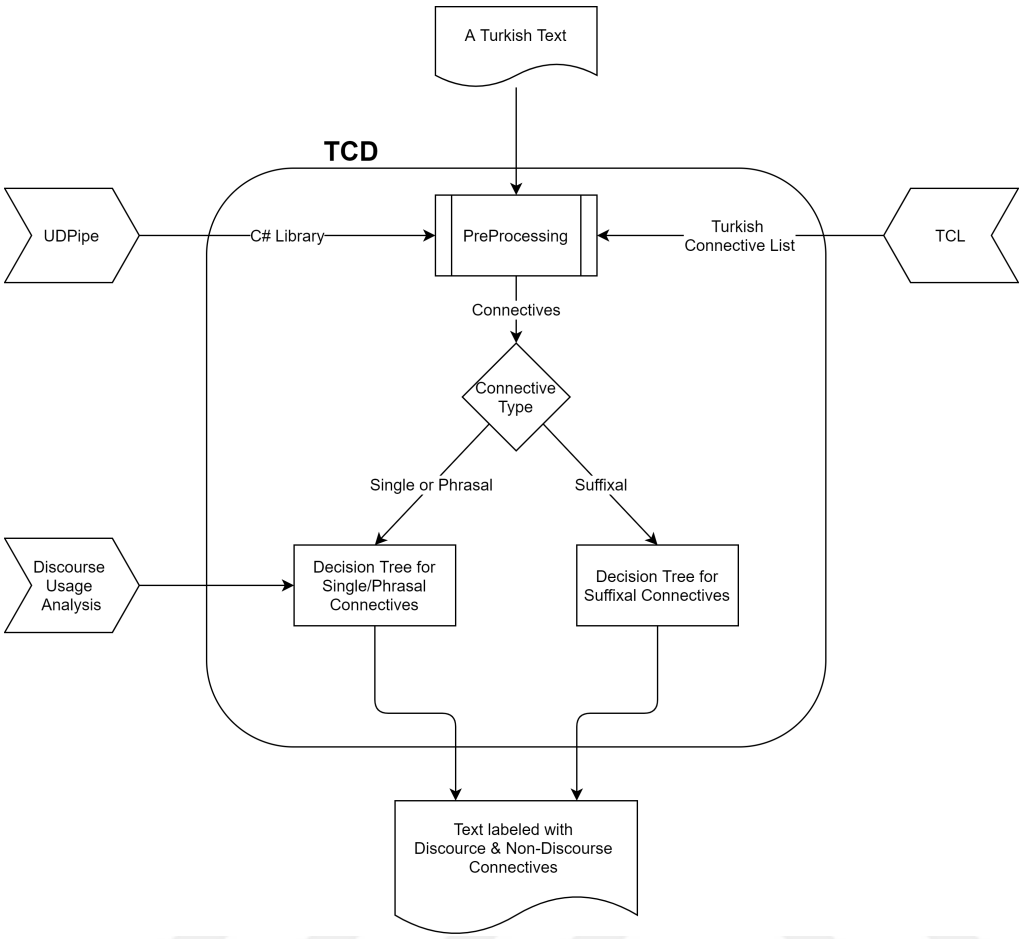

Figure 5.1: Turkish Connective Disambiguator workflow

TCD program. UDPipe presents a pipeline for tokenization, tagging, lemmatization and dependency parsing of files [74]. Using these methods offered by UDPipe, the following operations are carried out on the given text in order.

- The given text is divided into sentences. Then the Id of each sentence (sentence number), the starting position of the sentence, the text of the sentence and the list of words in the sentence are kept in a data structure named as *MySentence*.
- Each word's Id, PoS tag, text, start and end positions are kept in a data structure named as *MyWord*.
- To determine single/phrasal connectives, single and phrasal connective lists of TCL are searched within the sentences by using the string match method.
- To determine suffixal connectives, the allomorph lists of suffixal connectives in

TCL are used to find the words ending with one of the suffixes.

After the connectives in a given text are found, the discourse usage analysis results and the classification rules for single/phrasal type connectives are used for the usage disambiguation of single/phrasal type connectives. These classification rules are given in section 5.3 below. Since the grammatical structure of suffixal connectives and their usage in the sentences are quite different from those of the single/phrasal type, completely different classification rules have been developed for their usage disambiguation. These classification rules are given in section 5.4 below.

5.3 Classification Rules For Single and Phrasal Connectives

To understand the situations where the single and phrasal type connectives are used as discourse, many samples from Turkish discourse-annotated corpora have been examined. Additionally, the literature about the discourse-hood of single and phrasal type connectives are reviewed. As a result of these studies, we developed several classification rules for solving the usage ambiguity of single and phrasal type connectives and presented these rules by a decision tree in figure 5.2. These classification rules are described below [25].

RULE1 Position of the connective in the sentence

The connective usage at the beginning, at the end or in the middle of a sentence gives an idea about whether the connective is discourse or non-discourse. In many sentence samples examined through this thesis, it has been observed that a connective used at the beginning or at the end of the sentence often acts as discourse. A connective used in the middle of the sentence can function either as discourse or non-discourse, so other rules should be followed for such cases.

Sentence-Initial: In the sentences below, the connectives used at the beginning of the sentences play a discourse role.

36. *Sebep sonuç değil*. Ancak **bu, çevresel liderliğin iyi kazançla bağdaştığını gösteriyor**.

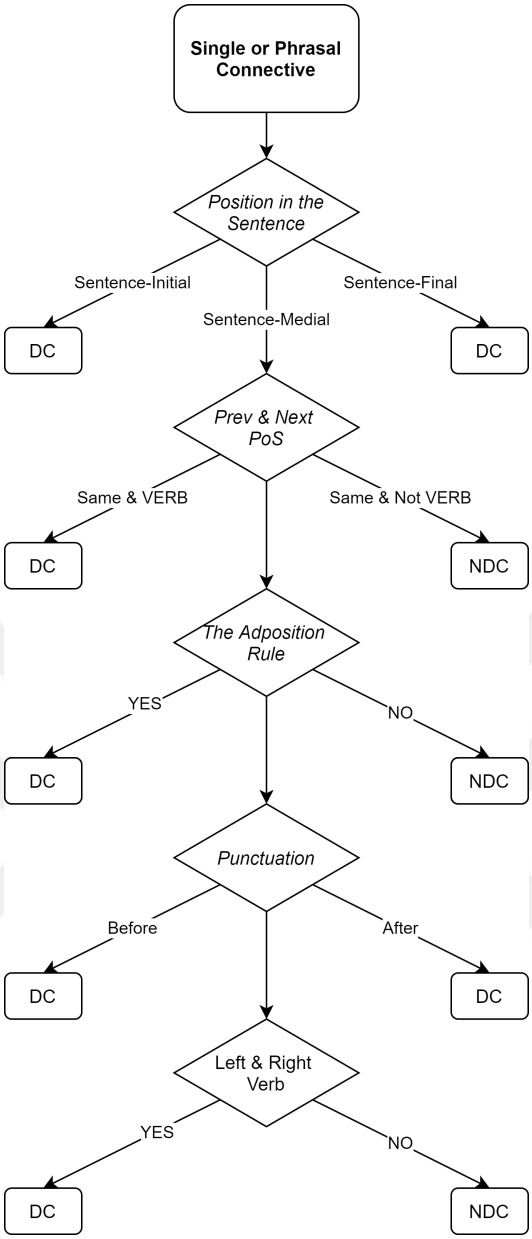

Figure 5.2: Decision tree for the usage disambiguation of single and phrasal connectives

'*Cause is not effect.* However, **this shows that environmental leadership is correlated with good earnings**.'
(DC)
(T-TED-MDB file 1927)

37. *Hem de çok.* Yani, **gerçekten çok fazla**.

'*A lot indeed*. In a word, **it's really too much**.'
(DC)
(T-TED-MDB file 1927)

38. *Carl Sagan'ın "Soluk Mavi Nokta"sı işte bu*. Ama **hadi 10 kat daha uzağa gidelim**.
'*This is Carl Sagan's "Pale Blue Dot"*. But **let's go 10 times further**.'
(DC)
(T-TED-MDB file 1976)

39. *Sanırım, bazıları coğrafi bölge ya da sokak ve bina topluluğu diyebilir*.
Fakat **inanıyorum ki bir şehir orada yaşayan insanlar arasındaki ilişkiler bütünüdür**.
'*I guess, some might call it geographic territory or street and building community*.
But **I believe that a city is a collection of relationships between people living there**.'
(DC)
(T-TED-MDB file 2150)

In some cases, a connective at the beginning of the sentence can act as non-discourse as shown in the following sample.

40. Artık $_{\text{NDC}}$ bir şey çizemez olmuştu.
'Anymore $_{\text{NDC}}$ he couldn't draw anything.'
(NDC)
(TDB 1.1 file 00001131)

Sentence-Final: In the sentences below, the connectives used at the end of the sentences play a discourse role.

41. *onlar iyi yaşamak için çalışacaklar sadece hayatta kalmak için değil*, **daima bir anlamda kaçmak** için,
'*They will work for good, not only to survive*, **always in a sense escape** to,'

(DC)
(T-TED-MDB file 2009)

42. *TEDx Rio, sağ aşağıda, blogcuların ve yazarların tam yanındadır*. Fakat, **değişik müzik tarzlarıyla ilgilenen büyük miktarda insan çeşitliliğine de sahipsiniz** ayrıca.
'*TEDx Rio is right below, right next to bloggers and writers*. But **you have a large variety of people who are interested in different styles of music** in addition.'
(DC)
(T-TED-MDB file 2150)

In some cases, a connective at the end of the sentence can act as non-discourse as shown in the following sample.

43. Sen bana anlatamıyordun. Anlatmanı istemiyordum zaten $_{\text{NDC}}$.
'You couldn't tell me. I didn't want you to tell anyway $_{\text{NDC}}$'
(NDC)
(TDB 1.1 file 00001131)

Sentence-Medial: The connective *için* '(in order) to' is used in the middle of the sentences below. It is used as a discourse connective in sentence 44 while it is used as a non-discourse connective in sentence 45. Similarly, the connective *ancak* 'however/but/just' is used sentence-medially in sentences 46 and 47. While the first usage is discourse, the second usage is non-discourse.

44. *onlar iyi yaşamak için çalışacaklar* **sadece hayatta kalmak** için **değil**,
'*They will work to live well* **not only** to **survive**
(DC)
(T-TED-MDB file 2009)

45. Sürdürülebilirliği sadece önemli olarak değil, şirketin başarısı için $_{\text{NDC}}$ de kritik olarak görmeye başladılar.

'They began to see sustainability not only as important but critical to NDC the company's success.'
(NDC)
(T-TED-MDB file 1927)

46. *Başarının bir an olduğunu*, ancak **her zaman göklere çıkardığımız şeyin yaratıcılık ve ustalık olduğunu** anladım.
'I realized that *success is a moment*, but **what we always praise is creativity and ingenuity**.'
(DC)
(T-TED-MDB file 1978)

47. Bu küçük ölçekli bir yıldız gölgeleyici. Az önce gördüğünüzün ancak NDC yarısı kadar.
'This is a small scale star shader. Just NDC half of what you have just seen.'
(NDC)
(T-TED-MDB file 1976)

RULE2 PoS tags of the previous word and the next word

By checking the PoS tags of the previous and next words of a connective, we can judge the discourse-hood of that connective. When the PoS tags of the previous and next words are the same and *Verb*, the connective mostly acts as discourse. On the other hand, when the PoS tags of the previous and next words are the same and *Not Verb such as Noun, Adj, Adv etc.*, the connective mostly acts as non-discourse.

Same & VERB: In the following sample sentences 48, 49 and 50, connectives connect two *Verbs* and so they are used as discourse.

48. **onlar iyi yaşamak** VERB için *çalışacaklar* VERB
'*they will work* (in order) to **live well**'
(DC)
(T-TED-MDB file 2009)

49. *yük trenleri öldürüyor* $_{\text{VERB}}$ ve **sakatlıyor** $_{\text{VERB}}$
    '*wagon trains kill* and **maim**'
    (DC)
    (T-TED-MDB file 2009)

50. Ancak, şehirlerin haritalarını oluşturmayı *düşünmenin* $_{\text{VERB}}$ ve **yapmanın** $_{\text{VERB}}$ başka yolları da var.
    'However, there are other ways of *thinking* and **making** maps of cities.'
    (DC)
    (T-TED-MDB file 2150)

Same & Not VERB: The connective *ve* 'and' in sentence 51 connects two *Adjectives* while the connective *ve* 'and' in sentence 52, the connective *veya* 'or' in sentence 53 connect two *Nouns*. In all three sentences, connectives are used as non-discourse.

51. Bunlar belli ki çevresel $_{\text{ADJ}}$ ve $_{\text{NDC}}$ sosyal $_{\text{ADJ}}$ sorunlar, ama hepsi bununla kalmıyor.
    'These are obviously environmental and $_{\text{NDC}}$ social issues, but that's not all.'
    (NDC)
    (T-TED-MDB file 1927)

52. Bunlar ekonomik sorunlar ve bu da bu sorunları risk $_{\text{NOUN}}$ ve $_{\text{NDC}}$ kazanç $_{\text{NOUN}}$ ile ilgili hâle getiriyor.
    'These are economic issues, and this makes these issues related to risk and $_{\text{NDC}}$ gain.'
    (NDC)
    (T-TED-MDB file 1927)

53. Bu gölgeye bir teleskop $_{\text{NOUN}}$ veya $_{\text{NDC}}$ kamera $_{\text{NOUN}}$ koyuyoruz,
    'We put a telescope or $_{\text{NDC}}$ a camera in this shadow'
    (NDC)
    (T-TED-MDB file 1976)

RULE3 The adposition rule

If the PoS of the connective is ADPOSITION and the PoS of the word previous to the connective is VERB or AUXILIARY, the connective plays a discourse role otherwise it plays a non-discourse role.

For example, in the following sentences the PoS of the connective *için* 'to/for' is ADPOSITION. In sentences 54 and 55, the PoS tag of the previous word is VERB, so the connectives play a discourse role. On the other hand, in sentences 56 and 57, the PoS tag of the previous word is NOUN, so the connectives play a non-discourse role.

54. **bu sorunları çözmek** $_{\text{VERB}}$ için $_{\text{ADP}}$ *en fazla çalışan*
    '*working hardest* in order to **solve these problems**'
    (DC)
    (T-TED-MDB file 1927)

55. **Bunu anlamak** $_{\text{VERB}}$ için $_{\text{ADP}}$, *Hubble Uzay Teleskobu'nu tutup döndürdüğümüzü ve onu Mars'ın yörüngesine doğru gönderdiğimizi varsayalım.*
    'In order to **understand this**, *suppose we grab and spin the Hubble Space Telescope and send it into Mars' orbit*.'
    (DC)
    (T-TED-MDB file 1976)

56. Sürdürülebilirliği sadece önemli olarak değil, şirketin başarısı $_{\text{NOUN}}$ için $_{\text{NDC ADP}}$ de kritik olarak görmeye başladılar.
    'They began to see sustainability not only as important but critical to $_{\text{NDC}}$ the company's success.'
    (NDC)
    (T-TED-MDB file 1927)

57. Daha fazla uzun dönem kazanç $_{\text{NOUN}}$ için $_{\text{NDC ADP}}$ bunun kritik olduğuna inanıyorlar, nokta.
    'They believe it's critical for $_{\text{NDC}}$ more long-term gains, period.'
    (NDC)
    (T-TED-MDB file 1927)

RULE4 Punctuation before and after the connective

If there is a PUNCTUATION mark previous to the connective or next to the connective, the connective most likely plays a discourse role. The following sentences 58, 59 and 60 show samples of discourse usage of connectives when there is a PUNCTUATION mark previous to the connective.

58. *Işığın çoğunu engelliyor*, böylece **etrafındaki soluk koronayı görebiliyoruz**.
    '*It blocks most of the light*, so **we can see the faint corona around it**.'
    (DC)
    (T-TED-MDB file 1976)

59. *Bir şeyler anlatacaktın*, ama **öyküyü bitirmemi bekliyordun**.
    '*You were going to tell something*, but **you were waiting for me to finish the story**.'
    (DC)
    (TDB 1.1 file 00001131)

60. *Avcı bir toplum, tarım henüz başlamış*; ama **son derece gelişmiş bir toplumsal örgütlenme var**.
    '*A hunter society, agriculture has just begun*; but **there is a highly developed social organization**.'
    (DC)
    (TDB 1.1 file 00013112)

Similarly, sentence 61 below shows the discourse usage of the connective *ama* 'but' when there is a PUNCTUATION mark after the connective.

61. *Yazdıklarımı anlamıştı* ama, **beni hiçbir zaman anlayamadı ne yazık ki**.
    '*She understood what I wrote* but, **unfortunately she could never understand me**.'
    (DC)
    (TDB 1.1 file 00065111)

Although there is a punctuation mark before or after the connective, there are cases where the connectives are used as non-discourse. The following sentences shows the samples of such situations.

62. Antrenmanın sonunda, okçulardan biri öylesine zorlanmıştı ki $_{\text{NDC}}$, yere pestil gibi serildi,
'At the end of the practice, one of the archers was so strained that $_{\text{NDC}}$, she fell on the ground like a pulp,'
(NDC)
(T-TED-MDB file 1978)

63. 1988'de heptatlonda altını kazandı ve 7.921 puanla rekor kırdı, ki $_{\text{NDC}}$ bu o zamandan beri hiçbir sporcunun yaklaşmadığı bir skor.
'He won the gold in the heptathlon in 1988 and set a record 7,921 points, which $_{\text{NDC}}$ a score no athlete has come close to since.'
(NDC)
(T-TED-MDB file 1978)

RULE5 Verb usage before and after the connective

If one of the three words before the connective has PoS tag VERB/AUXILIARY and one of the words until the end of the sentence after the connective has the PoS tag VERB/AUXILIARY, the connective most probably plays a discourse role, otherwise it most probably plays a non-discourse role. In this rule, the window size *three* is obtained experimentally.

Samples 64 and 65 below show the connective usages that comply with this rule and classified as discourse, while samples 66 and 67 show the connective usages that do not comply with this rule and hence classified as non-discourse.

64. *Başvurdum* $_{\text{VERB}}$ ve **yüzlerce adayla birlikte sınava girdim** $_{\text{VERB}}$.
'*I applied* and **took the exam with hundreds of candidates**.'
(DC)
(TDB 1.1 file 00065111)

65. *Bunun ona daha çok kararlılık verdiğini gördüm* VERB ve **tekrar tekrar denedi** VERB.
'*I saw that this gave him more determination* and **he tried over and over again**.'
(DC)
(T-TED-MDB file 1978)

66. Ama Spitzer aslında NDC yanıtı biliyordu.
'But Spitzer actually NDC knew the answer.'
(NDC)
(T-TED-MDB file 1976)

67. Rio, Baltimore ve San Francisco'nun tersine NDC tam bir heterojen şehir
'Rio is a heterogeneous city unlike NDC Baltimore and San Francisco'
(NDC)
(T-TED-MDB file 2150)

5.4 Classification Rules For Suffixal Connectives

In this section, we present the classification rules which we developed for solving the usage ambiguity of suffixal connectives in Turkish. In the literature, we could not see any study about the usage disambiguation of Turkish suffixal connectives. So, we have developed the classification rules which will be described in this section, original to this thesis.

Converbs (suffixal connectives) are added to the end of a *Verb* and they can be in different forms depending on the vowel harmony and consonant harmony with the *Verb* it is affixed. We list the possible forms of the converbs (allomorphs of the converbs) in Table 4.2. TCD program marks the usages of the allomorphs of the converbs in a text by using this table. However, some marked usages may actually be a set of phonemes or a syllable of the word, and these usages are non-discourse usages.

In example 68 below, the suffix *se* (an allomorph of the converb -sA 'if') is added to the end of a *Verb* and acts as a suffixal discourse connective. But, in example 69, the syllable *se* is at the end of a word but it is actually a part of the word *elbise* 'dress', it

is not a suffix. So, it should be disambiguated as non-discourse by the classification rules.

68. **Erken gelir** - se *dışarı çıkarız.*
    'If **he comes early**, *we'll go out*'
    (DC)

69. Bu ne güzel bir elbi • se $_{\text{NDC}}$.
    'what a beautiful dress'
    (NDC)

Another case where suffixes do not serve as discourse connectives is when they are appended to a word that has a PoS tag other than *Verb*. In sentence 70, the suffix *ında* (an allomorph of the converb -dA 'when') is appended to a word with PoS tag *Verb* and it plays a discourse role. However, in sentence 71, the suffix *de* is appended to a word with PoS tag *ADJ* and it does not play a discourse role. Such cases should be disambiguated by the classification rules of the TCD program.

70. **Kitap okumayı bıraktığ** - ında *akşam olmuştu.*
    '*It was evening* when **she stopped reading**.'
    (DC)

71. Sözlük masanın üstün - de $_{\text{NDC}}$.
    'The dictionary is on the table.'
    (NDC)

We show the classification rules used for the usage disambiguation of suffixal connectives by a decision tree in figure 5.3 and describe the details of the rules in the following paragraphs [25].

RULE1 Part of Speech

It is known that converbs are suffixed to *Verbs*. So, if the PoS tag of the word ending with a suffix is other than *Verb*, the suffix is used as non-discourse. Or else, the TCD program continues with other rules.

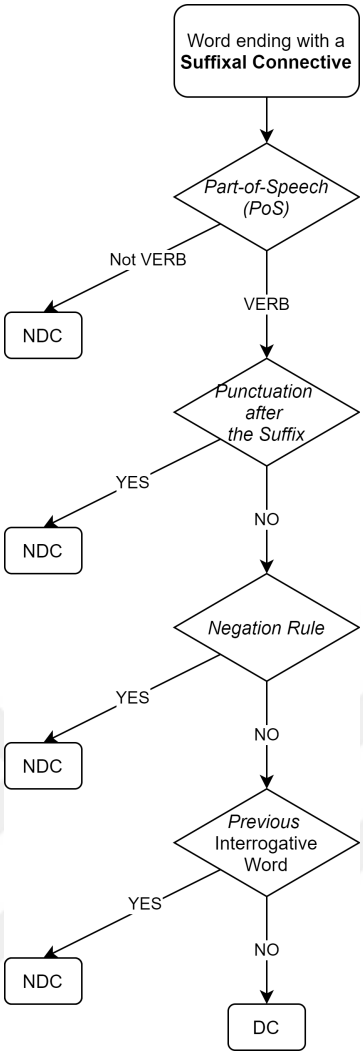

Figure 5.3: Decision tree for the usage disambiguation of suffixal connectives

In the following sentences, 72 and 73 are samples of cases where suffixes are added to words with the PoS tag *Verb* and act as discourse, while 74 and 75 are samples of cases where suffixes are added to words with the PoS tags *Noun* and *ADJ* and serve as non-discourse.

72. **okçuluk takımını seyretmeye gittiğ** $_{\text{VERB}}$ - imde *bunu anlamaya başladım*.
    '*I started to understand this* when **I went to watch the archery team**.'
    (DC)
    (T-TED-MDB file 1978)

73. **Diğerleri çalışmalarını yalnızca övgüye değer bulur** $_{\text{VERB}}$ - ken, *Franz Kafka*

bitmemiş olarak gördü,
'While **others found his work only commendable**, *Franz Kafka saw it as unfinished*,'
(DC)
(T-TED-MDB file 1978)

74. Voyager uzay aracı tarafın $_{\text{NOUN}}$ - dan $_{\text{NDC}}$ 1990'da çekildi.
'It was filmed by $_{\text{NDC}}$ the Voyager spacecraft in 1990.'
(NDC)
(T-TED-MDB file 1976)

75. Teleskobun için $_{\text{ADJ}}$ - de $_{\text{NDC}}$ saçılarak, gezegeni görülemeyecek hâle getiren bu aşırı parlak görüntüyü yaratıyor.
'It scatters through $_{\text{NDC}}$ the telescope, creating this extremely bright image that makes the planet invisible.'
(NDC)
(T-TED-MDB file 1976)

RULE2 Punctuation after the Suffix

Another feature we discovered about the usage disambiguation of converbs is related to punctuation after the word ending with a converb. When a punctuation mark is used after a converb, the punctuation mark mostly indicates a sentence boundary, and these usages are considered as non-discourse (see examples 76 and 77).

76. Gelişmiş ülkelerde bile bir hastanın rahat bir soket alabilmesi 3 haftadan yıllara kadar çıkabiliyor, tabi alabilir - se $_{\text{NDC}}$.
'Even in developed countries, a patient can get a comfortable socket from 3 weeks to several years, if $_{\text{NDC}}$ he can.'
(NDC)
(T-TED-MDB file 1971)

77. Peki ya çalışıp, biriktirip, yatırım yapsak ve emekli olduğumuzda dünyanın şimdikinden daha stresli ve daha az güvenli olduğunu gör - sek $_{\text{NDC}}$?

'What if $_{NDC}$ we work, save, invest, and find that the world is more stressful and less secure than it is now when we retire?'
(NDC)
(T-TED-MDB file 1927)

RULE3 Negation Rule

When a *Verb* ending with a converb is followed by the negation of this *Verb* such as *olup olmadığı* 'whether or not', *gidip gitmediği* 'whether she goes or not', the converb is most probably non-discourse. When a *Verb* and the negation of the *Verb* are used consecutively, the converb suffixed to the *Verb* is usually not used in the discourse sense, but it is used in the non-discourse sense with the meaning of *whether*.

78. Ama bu beni bugünkü yatırım kurallarının yarınki amaca uygun ol $_{VERB}$ - up $_{NDC}$ ol - ma $_{NEG}$ - dığı konusunda meraklandırıyor.
'But this makes me wonder whether $_{NDC}$ today's investment rules are suitable for tomorrow's purpose.'
(NDC)
(T-TED-MDB file 1927)

RULE4 Previous Interrogative Word

If the previous word of the *word ending with a converb* is an interrogative such as *kim* 'who,' *ne* 'what/whatever', the converb is most probably non-discourse. In such cases, interrogative words plays a clause-level syntactic role rather than a connective role.

79. kamyonculardan futbolcu annelerine kim $_{Q}$ gelir $_{VERB}$ - se $_{NDC}$ onlarla yolculuk ediyorlar.
'they travel with whoever $_{NDC}$ comes from truckers to soccer mothers.'
(NDC)
(T-TED-MDB file 2009)

The classification rules we mentioned above describe the situations where converbs are used as non-discourse. The converbs that do not comply with the cases described by these rules are classified as discourse as shown in samples 80 and 81 below.

80. **Kafamızı kuma göm** - üp, *bunun hakkında düşünmemek*.
'By **burying our heads in the sand**, *not thinking about it*.'
(DC)
(T-TED-MDB file 1927)

81. Bir kez daha **10 kat uzaklaş** - arak *Alpha Centauri'ye varıyoruz.*
'Once again, by **10 times zooming out**, *we arrive at Alpha Centauri*.'
(DC)
(T-TED-MDB file 1976)

5.5 TCD Interface

We developed a user interface for the TCD program in the Visual Studio .Net environment using the C# programming language. When a user selects a Turkish text using this interface, the TCD program finds all occurrences of connectives in the text and distinguishes discourse usages from non-discourse usages. The program marks discourse connectives with blue colour and non-discourse ones with red colour on the text. This section describes the interface of the TCD program and its capabilities. Figure 5.4 shows the interface of the program.

In the TCD program, there are three panels from left to right named as *Text Viewer*, *Predicted Connectives* and *Actual Connectives*. The usage flow of the TCD program can be summarized as follows:

- User selects a Turkish text file using the *Select Raw File* button and presses the *Parse* button.
- The content of the selected text file is loaded into the *Text Viewer* panel. TCD program finds all the connectives in the text by using the connectives listed by the TCL program. Then TCD disambiguates the discourse usages from the non-discourse ones by enabling discourse usage analysis and rule-based decision trees in the background. According to the disambiguation results, discourse connectives are marked with blue colour while non-discourse connectives are marked with red colour on the text.

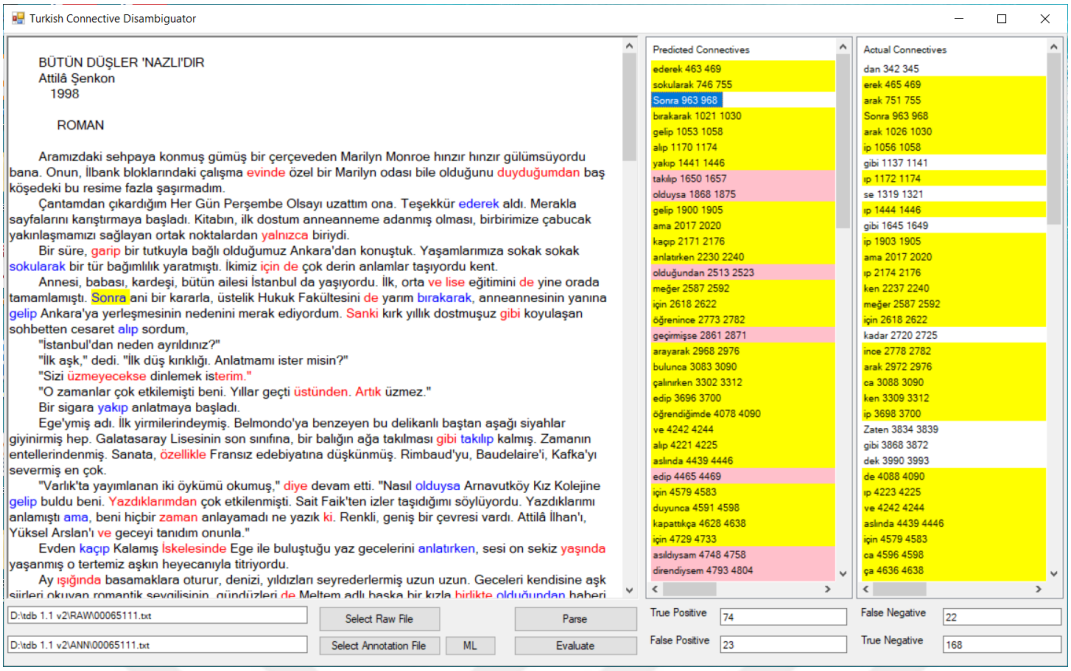

Figure 5.4: Turkish Connective Disambiguator interface

- If the selected text file has a corresponding annotation file prepared by humans, the user selects this annotation file using the *Select Annotation File* button and presses the *Evaluate* button. Discourse connectives predicted by the TCD program are shown in the *Predicted Connectives* panel with their start and end offsets in the text. Discourse connectives annotated by humans are shown in the *Actual Connectives* panel with their start and end offsets in the text. The annotation file should be in pipe-delimited file format.

- The TCD program compares the *Predicted Connectives* list with the *Actual Connectives* list and the following counts are computed as an evaluation.

  – Connectives included in both lists are counted as *True Positive* (yellow background).

  – Connectives predicted as discourse by the program but not annotated as discourse by humans are counted as *False Positive* (pink background).

  – Connectives not predicted as discourse by the program but annotated as discourse by humans are counted as *False Negative* (white background).

  – Connectives predicted as non-discourse by the program and not exists in the annotation file are counted as *True Negative*.

For the evaluation of the TCD program, these numbers are calculated for each file in the TDB 1.1 and T-TED-MDB datasets.

- When a user double-clicks on a connective in the *Predicted Connectives* or *Actual Connectives* panels, the connective is selected in the text with the yellow background. In this way, the user can easily observe the position of the connective in the text. This capability has greatly contributed to the development of classification rules used in the rule-based decision trees of the TCD program. We had the chance of updating the classification rules by examining incorrectly disambiguated connectives and their usage in the sentence.

- We also used the TCD program to collect data to use in the machine learning experiments. After we decided on the features to be used in the machine learning experiments, we collected these features for each connective in the TDB 1.1 and T-TED-MDB files. When the user presses the ML button, data is collected in the background and written to the text files to be used in the machine learning experiments.

We designed the infrastructure of the TCD program to be suitable for adding new capabilities later. Currently, the discourse and non-discourse usages are identified and marked in the *Text Viewer* panel. If the TCD program can compute the sense(s) and the arguments of the connectives in future work, they can also be marked in the *Text Viewer* panel.

In summary, in this chapter, we explained the details of the TCD program that we developed for solving the usage ambiguity of Turkish connectives. In the next chapter, we describe the machine learning models that we designed for solving the usage and sense ambiguity of Turkish connectives.

CHAPTER 6

MACHINE LEARNING MODELS

In this thesis, we firstly conducted rule-based classification for the usage disambiguation of the Turkish explicit connectives. Then, we developed machine learning models for the connectives' usage and sense disambiguation by designing disambiguating features. In designing these features, we inspired by the classification rules of the TCD and the features used in the machine learning models of other languages. In this chapter, we mention about how we developed the machine learning models and the features for the usage and sense disambiguation of Turkish explicit connectives.

6.1 Machine Learning Methodology

In this thesis, we used the **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM) process model in our machine learning studies. This model consists of 6 phases and describes the data science life cycle as shown in figure 6.1 [1]. We briefly mention about what we conducted at these phases in our machine learning studies below.

- **Business Understanding:** For business understanding, we analyzed the concept of usage ambiguity and sense ambiguity of connectives in Turkish.
- **Data Understanding:** For data understanding, we analyzed the Turkish discourse annotated datasets i.e. TDB 1.1 and T-TED-MDB.
- **Data Preparation:** In order to prepare data, we processed the raw files of the datasets by splitting the sentences and assigning the PoS tags of each word in the sentences.

[1] https://www.datascience-pm.com/crisp-dm-2/

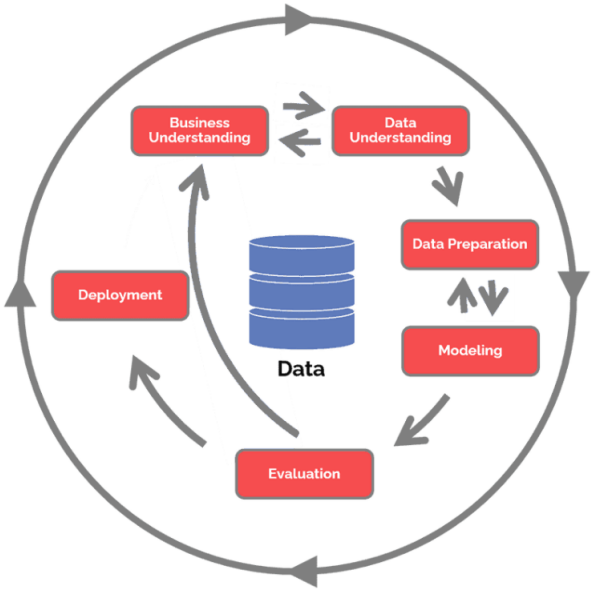

Figure 6.1: CRISP-DM Process Model

- **Modeling:** For developing ML models, we firstly explored disambiguating features for the classification algorithms and experimented with several ML models using different classification algorithms.
- **Evaluation:** We evaluated the results of the ML models using the five fold cross validation approach.
- **Deployment:** We repeated the procedure several times to get better results and deployed the models with best results.

For developing machine learning models, we used the ML.NET [2] framework which is an open-source framework and integrated this framework to the Visual Studio environment. For usage disambiguation, we conducted binary classification by classifying connective usages as DC or NDC. On the other hand, for sense disambiguation, we conducted multi-class classification by classifying the senses of discourse connectives according to the PDTB 3.0 sense hierarchy as detailed below.

- **For Sense Level 1,** discourse connectives are classified into 4 sense classes which are *Temporal*, *Contingency*, *Comparison*, *Expansion*.

[2] https://dotnet.microsoft.com/apps/machinelearning-ai/ml-dotnet

- **For Sense Level 2,** discourse connectives are classified into 16 sense classes such as *Temporal.Synchronous*, *Contingency.Cause*, *Comparison.Concession*, *Expansion.Conjunction* etc.
- **For Sense Level 3,** discourse connectives are classified into 30 sense classes such as *Temporal.Asynchronous.Precedence*, *Contingency.Cause.Reason* etc.

In both binary classification and multi-class classification, we tried several classification algorithms and deployed the algorithm giving the best result. For the usage disambiguation of single and phrasal type connectives, we selected the *Fast Forest* algorithm which is an implementation of Random Forest algorithm. On the other hand, for the usage disambiguation of suffixal type connectives and sense disambiguation of all type of connectives, we selected the *Averaged Perceptron* algorithm which calculates a function by multiplying the features by certain coefficients. Hence the algorithm finds a separating hyper-plane.

6.2 Features for the Usage Disambiguation of Single/Phrasal Connectives

To design machine learning features for solving the usage ambiguity of single and phrasal type connectives in Turkish, we first searched the literature to observe how other languages solved this problem. The most featured studies on this subject are briefly summarized below.

Pitler and Nenkova [50] used the *connective itself* with syntactic features extracted from constituent parse trees which are *parent category*, *left sibling category*, *right sibling category*, *whether right sibling contains a VP* and *whether right sibling contains a trace*.

Lin et al. [19] state that the context and the PoS of a connective are powerful features in determining the discourse role of the connective and they model the connective's context and POS with the features; *the connective, connective PoS, previous word, previous PoS, next word, next PoS*.

In a recent study, Gopalan and Devi [52] made use of the following features for usage disambiguation; *previous word, current word, next word, PoS of previous word, PoS of*

current word, PoS of next word and chunk of current word.

Laali et al. [51] used ten features in total which are *connective*, *the case of connective (all lowercase or initial uppercase)*, *the highest node in the parse tree (SelfCat)*, *the parent of the SelfCat*, *the left sibling of the SelfCat*, *the right sibling of the SelfCat*, *the left word*, *PoS of the left word*, *the right word* and *PoS of the right word* in their CLaC discourse parser for usage disambiguation.

By analysing the machine learning features designed for other languages and the classification rules described in chapter 5 for the usage disambiguation of single and phrasal type connectives in Turkish, we designed the machine learning features. Table 6.1 shows the classification rules and the machine learning features inspired by the classification rules. The details of the rules can be found in subsection 5.3.

Table 6.1: Classification rules of the TCD program for solving the usage ambiguity of single/phrasal connectives and the machine learning features inspired by the rules

Classification Rule	**Machine Learning Feature**
Rule1	Sentence-Initial, Sentence-Final
Rule2	PoS of Previous Word, PoS of Next Word
Rule3	PoS of Connective, PoS of Previous Word
Rule4	PoS of Previous Word, PoS of Next Word
Rule5	LeftVerb, RightVerb

Another source we employed for determining machine learning features is the result of *discourse usage analysis* mentioned in subsection 5.1. The result of this analysis tells us whether a connective plays mostly a discourse role *(MostlyDC)* or plays mostly a non-discourse role *(MostlyNDC)*. We added these features obtained as a result of *discourse usage analysis* to the machine learning feature set.

As a result, a set of machine learning features were designed for the usage disambiguation of single/phrasal connectives and are listed in Table 6.2 below [25].

Table 6.2: Features used for the usage disambiguation of single/phrasal connectives

Feature	Explanation
Gopalan and Devi (2016) Features	
Previous word	The word before the connective
Current word	The connective
Next word	The word after the connective
PoS of previous word	The PoS tag of the word before the connective
PoS of current word	The PoS tag of the connective
PoS of next word	The PoS tag of the word after the connective
Chunk of current word	The feature to identify the boundary of the connective
Features added by the present thesis	
MostlyDC	Whether the connective mostly fulfills a discourse use
MostlyNDC	Whether the connective mostly fulfills a non-discourse use
Sentence-Initial	Whether the connective is sentence-initial
Sentence-Final	Whether the connective is sentence-final
LeftVerb	Whether one of the three words previous to the connective contains a word with the PoS tag VERB
RightVerb	Within a sentence, whether the words to the right of the connective contain a word with the PoS tag VERB

6.3 Features for Usage Disambiguation of Suffixal Connectives

We developed machine learning features for the usage disambiguation of suffixal connectives in Turkish inspired by the classification rules we developed for suffixal connectives. Table 6.3 shows the classification rules and the machine learning features inspired by the classification rules. The details of the rules can be found in subsection 5.4.

As a result, the machine learning features we designed for the usage disambiguation of suffixal connectives in Turkish are given in Table 6.4 below [25].

Table 6.3: Classification rules of the TCD program for solving the usage ambiguity of suffixal connectives and the machine learning features inspired by the rules

Classification Rule	**Machine Learning Feature**
Rule1	Connective, PoS
Rule2	HasPunct
Rule3	Negation
Rule4	PrevQuestion

Table 6.4: Features for the usage disambiguation of suffixal connectives

Feature	**Explanation**
Connective	The suffixal connective corresponding to an actual allomorph of a converb
PoS	The PoS tag of the word carrying the suffix
HasPunct	Whether there is a punctuation mark (? ! .) after the word that has the suffix
Negation	Whether the following verb is the negation of the current verb that has the suffix
PrevQuestion	Whether the previous word is an interrogative word

6.4 Features for Sense Disambiguation

In the next step, we aimed to carry out a sense disambiguation task of Turkish discourse connectives for each level of the PDTB 3.0 sense hierarchy, i.e. Level-1, Level-2 and Level-3 (see table 2.1). As far as we know, this is the first attempt to disambiguate the sense of Turkish explicit discourse connectives. Therefore, we referenced the study of Lin et al. [19] for English sense disambiguation. Lin et al. [19] emphasize that *the connective itself* is an important feature since many connectives are unambiguous and the ambiguous ones tend to convey a specific sense in most cases. They used three features in their sense classifier; *connective string*, *PoS of connective* and *previous word*.

We made use of the same features in the sense disambiguation of Turkish discourse connectives. The resulting machine learning features we designed for the sense disambiguation of discourse connectives are given in Tables 6.5 and 6.6 below [25].

Table 6.5: Features used for the sense disambiguation of single/phrasal connectives

Features	**Explanation**
Current word	The connective
PoS of the current word	The PoS tag of the connective
Previous word	The word before the connective

Table 6.6: Features for the sense disambiguation of suffixal connectives

Features	**Explanation**
Current word	The suffixal connective corresponding to an actual allomorph of a converb
PoS of the current word	The PoS tag of the word carrying the suffix
Previous word	The word before the word that includes the suffix

In the next chapter, we present the results and evaluation of the TCD program and the machine learning models developed by using the features mentioned in this chapter.

CHAPTER 7

RESULTS AND EVALUATION

In this section, we first provide the results of the TCD program and evaluate the program's success in solving the usage ambiguity of the connectives. Then, we give the results of the machine learning models and evaluate their success in solving the usage and sense ambiguity of the connectives.

7.1 Results and Evaluation of the Turkish Connective Disambiguator

For the evaluation of the TCD program, we used the raw text files and annotation files of the TDB 1.1 and T-TED-MDB datasets. The metrics Accuracy (7.1), Precision (7.2), Recall (7.3) and F1-score (7.4) were calculated by using the TP, FP, FN and TN counts. These counts were computed by comparing the results of the TCD program with the human annotations in the corresponding annotation files. The definitions of these metrics are given below:

True Positive (TP): predicted as DC and annotated as DC
False Positive (FP): predicted as DC but not annotated as DC
False Negative (FN): not predicted as DC but annotated as DC
True Negative (TN): not predicted as DC and not annotated as DC

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{7.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{7.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{7.3}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7.4}$$

We tested the TCD program on the TDB 1.1 and T-TED-MDB datasets for the usage disambiguation of single/phrasal and suffixal connectives and we obtained the results shown in Table 7.1 [25].

Table 7.1: Usage disambiguation results of TCD

Dataset	**Target**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
T-TED-MDB	Single and Phrasal Connectives	85.17	85.22	78.28	**81.60**
TDB 1.1	Single and Phrasal Connectives	86.17	75.34	79,56	77.39
T-TED-MDB	Suffixal Connectives	88.52	71.15	86.04	**77.89**
TDB 1.1	Suffixal Connectives	87.01	64.19	90.84	75.22

The results show that the TCD program performed better in the usage disambiguation of both single/phrasal and suffixal type connectives in the T-TED-MDB dataset. After this evaluation, we performed an analysis about the errors of the TCD program in both datasets. We give the details of the error analysis results below.

7.1.1 Error Analysis

After obtaining the TCD results, we analyzed the sentences where the connectives were disambiguated incorrectly to understand where the TCD program fails. As a result of this analysis, we realized that some of the errors were caused by the incorrect PoS tagging process. The PoS tagging success of UDPipe directly affects the TCD disambiguation success because the PoS feature is used in both the classification rules developed for single/phrasal connectives and the classification rules developed for suffixal connectives. On the other hand, in some erroneous cases, it was seen that some discourse usage scenarios could not be covered by the classification rules. We present the details of some erroneous cases below.

Result 1 There are both verbal and nominal predicates in Turkish which are typically used at the end of the sentence. In addition to their position in the clause, these verbal or nominal predicates may also have a suffix showing the tense/aspect/mood attributes. This suffix is marked as AUXILIARY by the PoS tagger (e.g. -dir). However, this tense/aspect/mood suffix may be non-existent but the human reader still infers the predicate from its clause-final position. In cases when the tense/aspect/mood suffix is not used, the AUXILIARY PoS tag cannot be assigned and the PoS tagger only gives the PoS tag of the stem of the nominal predicate. This yields an error in TCD, as exemplified below.

In sentence 7.2 below, the connective *ve* 'and' is correctly marked as *discourse* according to RULE5 stating that *if a connective has VERB/AUX in the preceding three words and in following words to the end of the sentence, then it is discourse*.

Table 7.2: A nominal predicate with tense/aspect/mood suffix -dir

Correct disambiguation of the connective *ve* 'and' as discourse							
Kuşlar	özgürlüğün	göstergesi	-dir	ve	şair	bunu	vurgulamaktadır
NOUN	NOUN	NOUN	AUX	CCONJ	NOUN	PRON	VERB
Birds are the sign of freedom and the poet emphasizes this							

However, in sentence 7.3 below, the suffix -dir to be marked as AUX is dropped; the remaining word *gösterge* is marked as NOUN. Not being able to apply Rule5, TCD marks *ve* 'and' incorrectly as *non-discourse*.

Table 7.3: A nominal predicate with a dropped tense/aspect/mood suffix

Incorrect disambiguation of the connective *ve* 'and' as non-discourse						
Kuşlar	özgürlüğün	göstergesi	ve	şair	bunu	vurgulamaktadır
NOUN	NOUN	NOUN	CCONJ	NOUN	PRON	VERB
Birds are the sign of freedom and the poet emphasizes this						

Result 2 The existential stems *var* 'there is / there are' and *yok* 'there is not / there are not' in Turkish are often used as predicates, typically at the end of clauses. However, these words can be used with or without inflection. When an existential stem is used without inflection, the PoS tagger does not assign the AUX PoS, which results in

incorrect disambiguation.

In sentence 7.4 below, the connective *ve* 'and' is correctly marked as *discourse* according to RULE5 stating that *if a connective has VERB/AUX in the preceding three words and in following words to the end of the sentence, then it is discourse*.

Table 7.4: An existential stem with inflection

Correct disambiguation of the connective *ve* 'and' as discourse								
Bahçede	rengarenk	çicekler	var	-dır	ve	çok	hoş	görünüyorlar
NOUN	ADJ	NOUN	ADJ	AUX	CCONJ	ADV	ADJ	VERB
There are colorful flowers in the garden and they look very nice								

However, in sentence 7.5 below, the suffix *-dır* to be PoS tagged as AUX is dropped and the connective *ve* 'and' is incorrectly marked as *non-discourse*.

Table 7.5: An existential stem without inflection

Incorrect disambiguation of the connective *ve* 'and' as non-discourse							
Bahçede	rengarenk	çicekler	var	ve	çok	hoş	görünüyorlar
NOUN	ADJ	NOUN	ADJ	CCONJ	ADV	ADJ	VERB
There are colorful flowers in the garden and they look very nice							

Result 3 Another error encountered in usage disambiguation of discourse connectives is related to suffixal connective disambiguation. In case the syllable at the end of a word is identical to a converb, this syllable is marked as a candidate suffixal connective by the TCD. If this candidate connective complies with the linguistic rules we designed for the usage disambiguation of suffixal connectives, it is marked as a suffixal connective incorrectly, hence increasing the number of False Positives.

For example, in sentence 7.6, *-ken* 'while' is a suffix appended to the end of the verb *izle-* 'watch' and acts as a suffixal discourse connective. TCD correctly marks this suffix as *discourse*.

However, in sentence 7.7 below, *ken* is a syllable of the word *giriş.ken* 'sociable' and so it does not act as a discourse connective. However, TCD marks this syllable as a suffixal discourse connective incorrectly.

Table 7.6: A converb as a suffixal discourse connective

Correct disambiguation of the connective *-ken* 'while' as discourse				
Film	izler -ken	mısır	keyfi	yaptı
NOUN	VERB	NOUN	NOUN	VERB
She enjoyed corn while watching movies				

Table 7.7: A syllable mistakenly identified as a converb

Incorrect disambiguation of the connective *-ken* 'while' as discourse				
Çalışkan	olduğu	kadar	giriş.ken	birisi
ADJ	VERB	ADP	VERB	PRON
He is as sociable as he is hardworking				

7.2 Results and Evaluation of Machine Learning Models

In this thesis, we used the ML.NET [1] framework to conduct the machine learning experiments. ML.NET is an open-source framework that can be integrated into the Visual Studio environment to train and use machine learning models for several machine learning tasks such as *text classification*, *value prediction*, *image classification*, *recommendation* and *object detection*.

When a dataset is provided for training purposes, ML.NET tries several models with different algorithms and presents the models with the best evaluation scores. For usage disambiguation, a binary classification (discourse vs. non-discourse) task is performed and *Accuracy* and *F1-Score* metrics are provided. On the other hand, for sense disambiguation, a multi-class classification (senses in the PDTB 3.0 sense hierarchy) task is performed and *Micro-Accuracy* and *Macro-Accuracy* metrics are calculated. For both kinds of classification, a 5-fold cross-validation methodology is used in the evaluation process. In the tables showing evaluation results, the highest F1-Score values are written in bold font.

[1] https://dotnet.microsoft.com/apps/machinelearning-ai/ml-dotnet

7.2.1 Results of the Usage Disambiguation of Single and Phrasal Connectives

For the usage disambiguation of single/phrasal type connectives, we first developed machine learning models with the feature set suggested by Gopalan and Devi [52]. Then, we added the features explored in this thesis and we developed machine learning models with the total set of features.

In this task, the machine learning models developed by using the *Fast Forest* algorithm gave the best results. The *Fast Forest* algorithm is an implementation of Random Forest algorithm and consists of a group of decision trees. Random forest algorithm generally performs well on skewed data [75] and fits our dataset in this respect. The distribution of single/phrasal type connectives in the TDB 1.1 and T-TED-MDB datasets has a Zipfian distribution and is fairly skewed. As a result, the *Fast Forest* algorithm gave the highest success among many algorithms tried by the ML.NET framework.

Table 7.8 below gives the results of two feature sets on the TDB 1.1 and T-TED-MDB datasets [25].

Table 7.8: Usage disambiguation results of the machine learning models for single/phrasal connectives

Dataset	**Method**	**Features**	**Accuracy**	**F1-Score**
T-TED-MDB	FastForestBinary	Gopalan and Devi's (2016) Features	78.38	70.37
TDB 1.1	FastForestBinary	Gopalan and Devi's (2016) Features	87.10	77.78
T-TED-MDB	FastForestBinary	Features added by the present thesis	86.54	**84.44**
TDB 1.1	FastForestBinary	Features added by the present thesis	88.73	**79.49**

The results show that the classification success of the usage disambiguation of single/phrasal connectives has increased with the new features added by this thesis. Additionally, the classification results obtained by adding new features is better than the results we get in TCD using the classification rules.

7.2.2 Results of the Usage Disambiguation of Suffixal Connectives

For the usage disambiguation of suffixal connectives, we developed machine learning models using the features we mentioned in section 6.2. We wanted to compare machine learning models using two feature sets. In the first experiment, we developed models with *Connective* and *PoS* features, which are frequently used features in the literature, while in the second experiment we developed a model with the features we explored in this thesis. Classification results of these models on the TDB 1.1 and T-TED-MDB datasets are given in Table 7.9 below [25].

Table 7.9: Usage disambiguation results of the machine learning models for suffixal connectives

Dataset	**Method**	**Features**	**Accuracy**	**F1-Score**
T-TED-MDB	AveragedPerceptronBinary	Connective, PoS	85.29	76.19
TDB 1.1	AveragedPerceptronBinary	Connective, PoS	87.32	75.68
T-TED-MDB	AveragedPerceptronBinary	Connective, PoS, HasPunc Negation, PrevQues	88.24	**80.00**
TDB 1.1	AveragedPerceptronBinary	Connective, PoS, HasPunc Negation, PrevQues	88.45	**76.80**

Machine learning models developed with the *Averaged Perceptron* algorithm gave the best results in the usage disambiguation task of suffixal connectives. The *Averaged Perceptron* classification algorithm calculates a function by multiplying the features by certain coefficients. Hence the algorithm finds a separating hyperplane.

The results show that the models developed by using Turkish-specific features yielded better results than the models developed by using only *Connective* and *PoS* features. Additionally, its results are better than the results we get in the TCD using the classification rules.

7.2.3 Results of the Sense Disambiguation of Discourse Connectives

In this experiment, we aimed to implement a task for the sense disambiguation of Turkish discourse connectives for all three levels in the PDTB 3.0 sense hierarchy. As far as we know, this is the first attempt in the literature which carries out a sense

disambiguation of explicit discourse connectives in Turkish. Table 7.10 shows the results obtained at all three levels of PDTB for single/phrasal connectives while Table 7.11 gives the results for suffixal connectives [25].

Table 7.10: Sense disambiguation results of the machine learning models for single/phrasal connectives

Dataset	**Method**	**Features**	**Accuracy Level-1**	**Accuracy Level-2**	**Accuracy Level-3**
T-TED-MDB	AveragedPerceptronOva	Current word, PoS of the current word, Previous word	**87.50**	75.86	78.26
TDB 1.1	AveragedPerceptronOva	Current word, PoS of the current word, Previous word	**91.18**	72.37	69.01

Table 7.11: Sense disambiguation results of the machine learning models for suffixal connectives

Dataset	**Method**	**Features**	**Accuracy Level-1**	**Accuracy Level-2**	**Accuracy Level-3**
T-TED-MDB	AveragedPerceptronOva	Current word, PoS of the current word, Previous word	**80.00**	75.00	70.00
TDB 1.1	AveragedPerceptronOva	Current word, PoS of the current word, Previous word	**66.67**	58.75	59.26

The best results in sense disambiguation were obtained with the *Averaged Perceptron* algorithm. The results show that the best results were obtained in both single/phrasal type and suffixal type at Level-1 of PDTB. The success at Level-2 and Level-3 changes according to the dataset and the connective type.

After the sense classification task for single/phrasal and suffixal type discourse connectives, we analyzed the confusion matrices to observe the errors made in the sense classification process. Since the evaluation results were obtained with a 5-fold cross-validation method, the confusion matrices were calculated by summing the confusion matrices of the 5 evaluation results. Tables E.11 and E.22 in the Appendices shows the

confusion matrices of the sense classification results at the sense level 1.

The confusion matrices show that there are very few errors in the sense classification of single/phrasal discourse connectives. The most common error in the single/phrasal type is confusing the Comparison sense with the Contingency or Expansion senses.

The sense classification of suffixal discourse connectives is more error-prone. The Contingency sense is often predicted as the Expansion sense or sometimes as the Comparison sense. Errors in the classification of the Temporal sense are also noticeable. It is seen that the Temporal sense is sometimes mapped to the Expansion sense or sometimes to the Comparison sense.

To summarize, in this chapter, we presented the results and the evaluation of the TCD program and the machine learning experiments we conducted. In the next chapter, we conclude the thesis.

CHAPTER 8

CONCLUSION

In this thesis, we explored the possibilities to classify discourse relations in Turkish by discourse processing techniques. Therefore in the present thesis, linguistic concepts are used for the problem definition and the solutions, while Computer Engineering capabilities are used to develop the solutions, especially using the techniques of Natural Language Processing and Machine Learning branches.

Various ambiguity problems occur in Turkish and other languages in automatic discourse parsing. In this thesis, we aimed to solve the usage ambiguity and sense ambiguity of Turkish connectives. Additionally, we have addressed the PoS ambiguity problem by using the PoS feature as a distinguishing feature within the scope of the solution of the sense ambiguity.

The first step in solving the ambiguity problems of connectives in a text is to determine the connective occurrences in the text mostly using a discourse connective lexicon. We realized that there is not such a comprehensive lexicon for Turkish discourse connectives in the literature. Therefore, we started our studies by creating a Turkish Connective Lexicon both to use it in the present thesis and to fill this gap for the Turkish language. This lexicon is automatically built by using the annotations of several Turkish discourse annotated corpora i.e. TDB 1.0, TDB 1.1 and T-TED-MDB, and it presents syntactic and semantic information about the connectives in Turkish. The connective list produced by TCL is displayed in a human-readable format through the user interface of the TCL and also saved to a file in DiMLex XML format. The TCL interface provides various extra capabilities such as selecting the discourse annotated corpus to be used in the creation of lexicon, filtering the connective lists according to the connective type or connective frequency, listing the sense distributions

of connectives and calculating the discourse usage percentages of connectives.

The main purpose of this thesis has been to solve the usage and sense ambiguity of Turkish discourse connectives. For this purpose, we first developed the TCD program, which solves the usage ambiguity of connectives by enabling two sets of classification rules one for single/phrasal type connectives and one for suffixal type connectives. We developed the classification rules by analysing how connectives are used in the Turkish sentences, considering the Turkish grammar rules and surveying the related studies of other languages.

We developed a user interface for the TCD program. Through this interface, a user can load any Turkish text into the program and observe the discourse and non-discourse usages of connectives on the text. Additionally, if the file has a corresponding annotation file in the pipe-delimited file format, the TCD program presents an evaluation by comparing the program results with the human annotations.

The TCD program interface has contributed a lot to the development of the classification rules. Through this interface, we easily observed where the rules failed and improved the rules accordingly. Also, the future users can benefit from the TCD program interface to observe the connective occurrences in the text and the disambiguation results before the human annotation process.

In the next step of the thesis, we developed machine learning models by designing various features that can be used to solve the usage and sense ambiguity of connectives. We evaluated the classification power of both the TCD program and the machine learning models on TDB 1.1 and T-TED-MDB datasets and obtained quite satisfactory results.

Although discourse connective disambiguation studies are carried out in several languages, each language needs to reveal its own approach by exploring language-specific features. For example, Gopalan and Devi [52] suggested a set of features for discourse usage disambiguation and achieved an f-score of 86.42%. When we tried this feature set for the usage disambiguation of Turkish connectives, we obtained a much lower success as presented in subsection 7.2.1. Hence, we developed several language-specific features original to this thesis in addition to the features inspired by other languages in

the related literature.

There are a few studies in the literature on solving the usage ambiguity of Turkish single/phrasal connectives. In this thesis, we made considerable progress towards solving this problem. Additionally, we attempted to solve the usage ambiguity of Turkish suffixal connectives and the sense ambiguity of all types of connectives for the first time in the literature, to the best of our knowledge.

The research done and the infrastructures established in this thesis have been designed in a way that can be extended by future research. The capabilities that can be added are listed below.

- TCL currently includes only *Explicit* discourse connectives. *Alternative Lexicalizations* can also be added to TCL.
- We provide sample sentences for the discourse usage of connectives in TCL. Sample sentences for the non-discourse usage of connectives can also be added.
- Various analyses can be conducted about connective usage frequencies and connective sense frequencies by using discourse annotated corpora developed for Turkish.
- Some new classification rules can be explored to cover the scenarios of false-positives and false-negatives. Hence the discourse usage disambiguation success of the TCD program can be increased.
- The TCD program can be transformed to a shallow discourse parser which finds the discourse connectives, their arguments and their sense(s).
- Some new machine learning features can be discovered for the machine learning models in order to increase their performance.

Finally, when the state-of-the art discourse parsing techniques for several languages are considered, the results obtained in this thesis will form a baseline in similar discourse parsing researches and studies for Turkish language.

REFERENCES

[1] B. Webber and A. Joshi, "Discourse structure and computation: past, present and future," in *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pp. 42–54, 2012.

[2] B. Webber, M. Egg, and V. Kordoni, "Discourse structure and language technology," *Natural Language Engineering*, vol. 18, no. 4, pp. 437–490, 2012.

[3] N. Asher, *Reference to abstract objects in discourse*, vol. 50. Springer Science & Business Media, 2012.

[4] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text*, vol. 8, no. 3, pp. 243–281, 1988.

[5] R. Prasad, B. Webber, and A. Joshi, "Reflections on the penn discourse treebank, comparable corpora, and complementary annotation," *Computational Linguistics*, vol. 40, no. 4, pp. 921–950, 2014.

[6] A. Al-Saif and K. Markert, "The leeds arabic discourse treebank: Annotating discourse connectives for arabic.," in *LREC*, pp. 2046–2053, 2010.

[7] N. Xue, "Annotating discourse connectives in the chinese treebank," in *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pp. 84–91, 2005.

[8] N. Xue and Y. Zhou, "Applying syntactic, semantic and discourse constraints in chinese temporal annotation," in *Coling 2010: Posters*, pp. 1363–1372, 2010.

[9] L. Mladová, S. Zikanova, and E. Hajicová, "From sentence to discourse: Building an annotation scheme for discourse based on prague dependency treebank.," in *LREC*, 2008.

[10] M. Buch-Kromann and I. Korzen, "The unified annotation of syntax and discourse in the copenhagen dependency treebanks," in *Proceedings of the Fourth Linguistic Annotation Workshop*, pp. 127–131, 2010.

[11] N. Van Der Vliet, I. Berzlánovich, G. Bouma, M. Egg, and G. Redeker, "Building a discourse-annotated dutch text corpus," *Bochumer Linguistische Arbeitsberichte*, vol. 3, pp. 157–171, 2011.

[12] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber, "The penn discourse treebank 2.0.," in *LREC*, Citeseer, 2008.

[13] M. Stede, "The potsdam commentary corpus," in *Proceedings of the Workshop on Discourse Annotation*, pp. 96–102, 2004.

[14] M. Stede, "Disambiguating rhetorical structure," *Research on Language and Computation*, vol. 6, no. 3, pp. 311–332, 2008.

[15] U. Oza, R. Prasad, S. Kolachina, D. M. Sharma, and A. Joshi, "The hindi discourse relation bank," in *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pp. 158–161, 2009.

[16] D. Zeyrek, U. D. Turan, C. Bozsahin, R. Cakıcı, A. B. Sevdik-Calli, I. Demirsahin, B. Aktaş, I. Yalcinkaya, and H. Ö. Balaban, "Annotating subordinators in the turkish discourse bank," in *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pp. 44–47, 2009.

[17] K. Oflazer, "Turkish and its challenges for language processing," *Language resources and evaluation*, vol. 48, no. 4, pp. 639–653, 2014.

[18] D. Zeyrek and K. Başıbüyük, "Tcl-a lexicon of turkish discourse connectives," in *Proceedings of the First International Workshop on Designing Meaning Representations*, pp. 73–81, 2019.

[19] Z. Lin, H. T. Ng, and M.-Y. Kan, "A pdtb-styled end-to-end discourse parser," *Natural Language Engineering*, vol. 20, no. 2, pp. 151–184, 2014.

[20] B. Webber, R. Prasad, and A. Lee, "Ambiguity in explicit discourse connectives," in *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pp. 134–141, 2019.

[21] D. Zeyrek and M. Kurfalı, "An assessment of explicit inter-and intra-sentential discourse connectives in turkish discourse bank," in *Proceedings of the Eleventh*

International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[22] D. Zeyrek, I. Demirsahin, A. Sevdik-Çallı, and R. Cakıcı, "Turkish discourse bank: Porting a discourse annotation style to a morphologically rich language.," *Dialogue Discourse*, vol. 4, no. 2, pp. 174–184, 2013.

[23] D. Zeyrek and M. Kurfalı, "Tdb 1.1: Extensions on turkish discourse bank," in *Proceedings of the 11th Linguistic Annotation Workshop*, pp. 76–81, 2017.

[24] D. Zeyrek, A. Mendes, and M. Kurfalı, "Multilingual extension of pdtb-style annotation: The case of ted multilingual discourse bank," in *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[25] K. Başıbüyük and D. Zeyrek (submitted), "Turkish discourse connective disambiguation: Identification of explicit discourse connectives and their senses," *Language Resources and Evaluation*, 2021.

[26] M. Stede and C. Umbach, "Dimlex: A lexicon of discourse markers for text generation and understanding," in *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pp. 1238–1242, Association for Computational Linguistics, 1998.

[27] T. Scheffler and M. Stede, "Adding semantic relations to a large-coverage connective lexicon of german," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1008–1013, 2016.

[28] B. Webber, R. Prasad, A. Lee, and A. Joshi, "A discourse-annotated corpus of conjoined vps," in *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pp. 22–31, 2016.

[29] C. Roze, L. Danlos, and P. Muller, "Lexconn: a french lexicon of discourse connectives," *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, no. 10, 2012.

[30] N. Asher, N. M. Asher, and A. Lascarides, *Logics of conversation*. Cambridge University Press, 2003.

[31] I. Keskes, F. B. Zitoune, and L. H. Belguith, "Learning explicit and implicit arabic discourse relations," *Journal of King Saud University-Computer and Information Sciences*, vol. 26, no. 4, pp. 398–416, 2014.

[32] A. Feltracco, E. Jezek, B. Magnini, and M. Stede, "Lico: A lexicon of italian connectives," *CLiC it*, p. 141, 2016.

[33] A. Feltracco, E. Jezek, and B. Magnini, "Enriching a lexicon of discourse connectives with corpus-based data," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[34] J. Mírovskỳ, P. Synková, M. Rysová, and L. Poláková, "Czedlex–a lexicon of czech discourse connectives," *The Prague Bulletin of Mathematical Linguistics*, vol. 109, no. 1, pp. 61–91, 2017.

[35] A. Mendes, I. D. R. Gayo, M. Stede, and F. Dombek, "A lexicon of discourse markers for portuguese–ldm-pt," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[36] D. Das, T. Scheffler, P. Bourgonje, and M. Stede, "Constructing a lexicon of english discourse connectives," in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 360–365, 2018.

[37] P. Bourgonje, J. Hoek, J. Evers-Vermeul, G. Redeker, T. Sanders, and M. Stede, "Constructing a lexicon of dutch discourse connectives," *Computational Linguistics in the Netherlands Journal*, vol. 8, pp. 163–175, 2018.

[38] M. Stede, T. Scheffler, and A. Mendes, "Connective-lex: A web-based multilingual lexical resource for connectives," *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, no. 24, 2019.

[39] M. Stede, "Dimlex: A lexical approach to discourse markers," *A. Lenci and*, vol. 501, pp. 1–15, 2002.

[40] L. Danlos, K. Rysova, M. Rysova, and M. Stede, "Primary and secondary discourse connectives: definitions and lexicons," *Dialogue & Discourse*, vol. 9, no. 1, pp. 50–78, 2018.

[41] P. R. Group *et al.*, "The penn discourse treebank 2.0 annotation manual," *December*, vol. 17, pp. 26–37, 2007.

[42] M. Stede and S. Heintze, "Machine-assisted rhetorical structure annotation," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pp. 425–431, 2004.

[43] A. Mendes and I. del Río, "Using a discourse bank and a lexicon for the automatic identification of discourse connectives," in *International Conference on Computational Processing of the Portuguese Language*, pp. 211–221, Springer, 2018.

[44] M. Stede, "Resolving connective ambiguity: A prerequisite for discourse parsing," in *The Pragmatics of Discourse Coherence*, pp. 121–141, John Benjamins, 2014.

[45] T. Meyer, A. Popescu-Belis, S. Zufferey, and B. Cartoni, "Multilingual annotation and disambiguation of discourse connectives for machine translation," in *SIGDIAL 2011 Conference*, p. 194, 2011.

[46] T. Meyer and L. Poláková, "Machine translation with many manually labeled discourse connectives," in *Proceedings of the Workshop on Discourse in Machine Translation*, pp. 43–50, 2013.

[47] D. Meurers and M. Dickinson, "Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics," *Language Learning*, vol. 67, no. S1, pp. 66–95, 2017.

[48] A. Kehler, L. Kertz, H. Rohde, and J. L. Elman, "Coherence and coreference revisited," *Journal of semantics*, vol. 25, no. 1, pp. 1–44, 2008.

[49] B. Webber, R. Prasad, A. Lee, and A. Joshi, "The penn discourse treebank 3.0 annotation manual," *Philadelphia, University of Pennsylvania*, 2019.

[50] E. Pitler and A. Nenkova, "Using syntax to disambiguate explicit discourse connectives in text," *ACL-IJCNLP 2009*, p. 13, 2009.

[51] M. Laali, A. Cianflone, and L. Kosseim, "The clac discourse parser at conll-2016," *arXiv preprint arXiv:1708.05798*, 2017.

[52] S. Gopalan and S. L. Devi, "Biodca identifier: A system for automatic identification of discourse connective and arguments from biomedical text," in *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pp. 89–98, 2016.

[53] B. Polepalli Ramesh, R. Prasad, T. Miller, B. Harrington, and H. Yu, "Automatic discourse connective detection in biomedical text," *Journal of the American Medical Informatics Association*, vol. 19, no. 5, pp. 800–808, 2012.

[54] N. Xue, H. T. Ng, S. Pradhan, R. Prasad, C. Bryant, and A. Rutherford, "The conll-2015 shared task on shallow discourse parsing," in *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pp. 1–16, 2015.

[55] N. Xue, H. T. Ng, S. Pradhan, A. Rutherford, B. Webber, C. Wang, and H. Wang, "Conll 2016 shared task on multilingual shallow discourse parsing," in *Proceedings of the CoNLL-16 shared task*, pp. 1–19, 2016.

[56] A. Zeldes, D. Das, E. G. Maziero, J. Antonio, and M. Iruskieta, "The disrpt 2019 shared task on elementary discourse unit segmentation and connective detection," in *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pp. 97–104, 2019.

[57] A. Zeldes, D. Das, E. G. Maziero, J. Antonio, and M. Iruskieta, "Introduction to discourse relation parsing and treebanking (disrpt): 7th workshop on rhetorical structure theory and related formalisms," in *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pp. 1–6, 2019.

[58] Y. Yu, Y. Zhu, Y. Liu, Y. Liu, S. Peng, M. Gong, and A. Zeldes, "Gumdrop at the disrpt2019 shared task: A model stacking approach to discourse unit segmentation and connective detection," *arXiv preprint arXiv:1904.10419*, 2019.

[59] P. Bourgonje and R. Schäfer, "Multi-lingual and cross-genre discourse unit segmentation," in *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pp. 105–114, 2019.

[60] M. Iruskieta, K. Bengoetxea, A. A. Salazar, and A. D. de Ilarraza, "Multilingual segmentation based on neural networks and pre-trained word embeddings," in

Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019, pp. 125–132, 2019.

[61] P. Muller, C. Braud, and M. Morey, "Tony: Contextual embeddings for accurate multilingual discourse segmentation of full documents," in *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pp. 115–124, Association for Computational Linguistics, 2019.

[62] K. Oflazer and M. Saraçlar, *Turkish natural language processing*. Springer, 2018.

[63] Ö. Yüksel and C. Bozsahin, "Contextually appropriate reference generation," *Natural Language Engineering*, vol. 8, no. 1, pp. 69–89, 2002.

[64] S. Yıldırım, Y. Kılıçaslan, and R. E. Aykaç, "A computational model for anaphora resolution in turkish via centering theory: an initial approach," in *International Conference on Computational Intelligence*, pp. 124–128, 2004.

[65] D. Küçük and M. T. Yöndem, "A knowledge-poor pronoun resolution system for turkish," *arXiv preprint arXiv:1504.04751*, 2015.

[66] M. Kurfalı, D. Zeyrek, and T. Gonçalvesb, "automatic prediction of implicit discourse relations in turkish," in *TextLink–Structuring Discourse in Multilingual Europe Second Action Conference Károli Gáspár University of the Reformed Church in Hungary Budapest, 11–14 April, 2016*, p. 65, 2016.

[67] B. Aktaş, C. Bozsahin, and D. Zeyrek, "Discourse relation configurations in turkish and an annotation environment," in *Proceedings of the fourth linguistic annotation workshop*, pp. 202–206, 2010.

[68] D. Zeyrek and B. Webber, "A discourse resource for turkish: Annotating discourse connectives in the metu corpus," in *Proceedings of the 6th workshop on Asian language resources*, 2008.

[69] D. Zeyrek, A. Mendes, Y. Grishina, M. Kurfalı, S. Gibbon, and M. Ogrodniczuk, "Ted multilingual discourse bank (ted-mdb): A parallel corpus annotated in the pdtb style," *Language Resources and Evaluation*, pp. 1–27, 2019.

[70] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical

review and future directions," *Psychonomic bulletin & review*, vol. 21, no. 5, pp. 1112–1130, 2014.

[71] W. Shi and V. Demberg, "On the need of cross validation for discourse relation classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 150–156, 2017.

[72] M. Laali and L. Kosseim, "Improving discourse relation projection to build discourse annotated corpora," *arXiv preprint arXiv:1707.06357*, 2017.

[73] P. Bourgonje and M. Stede, "Identifying explicit discourse connectives in german," in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 327–331, 2018.

[74] M. Straka and J. Straková, "Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe," in *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88–99, 2017.

[75] C. Chen and L. Breiman, "Using random forest to learn imbalanced data; 666," *University of California, Berkeley: Berkeley, CA, USA*, 2004.

Appendix A

DISCOURSE USAGE ANALYSIS RESULTS

Connective	PoS	Conn. Type	Total (#)	DC Use (#)	NDC Use (#)	DC Use (%)
ne var ki	ADV	phrasal	1	1	0	100
ve de	CCONJ	phrasal	2	2	0	100
olsa bile	ADP	phrasal	1	1	0	100
aksine	ADV	single	1	1	0	100
oysa	CCONJ	single	13	13	0	100
diğer bir deyişle	ADV	phrasal	1	1	0	100
öyleyse	ADV	single	2	2	0	100
ne ki	CCONJ	phrasal	1	1	0	100
kısacası	ADV	single	1	1	0	100
yine de	ADV	phrasal	1	1	0	100
dolayısıyla	ADV	single	10	10	0	100
aynı zamanda da	CCONJ	phrasal	2	2	0	100
dahası	ADP	single	2	2	0	100
iken	ADP	single	2	2	0	100
çünkü	CCONJ	single	52	48	4	92
böylece	ADV	single	16	14	2	87
ayrıca	ADV	single	24	21	3	87
halde	OTHER	single	6	5	1	83
keza	ADV	single	6	5	1	83
öte yandan	ADV	phrasal	5	4	1	80
ama	CCONJ	single	172	138	34	80
amacıyla	OTHER	single	8	6	2	75
karşın	ADP	single	13	9	4	69
ancak	CCONJ	single	61	42	19	68
mesela	ADV	single	3	2	1	66
gene de	ADV	phrasal	3	2	1	66

Connective	PoS	Conn. Type	Total (#)	DC Use (#)	NDC Use (#)	DC Use (%)
halbuki	CCONJ	single	5	3	2	60
aslında	ADV	single	24	13	11	54
fakat	CCONJ	single	24	13	11	54
örneğin	OTHER	single	6	3	3	50
meğer	CCONJ	single	2	1	1	50
ikincisi	OTHER	single	2	1	1	50
yoksa	CCONJ	single	8	4	4	50
sonra	ADV ADP	single	148	71	77	47
için	ADP	single	263	121	142	46
yani	CCONJ	single	46	19	27	41
rağmen	ADP	single	20	8	12	40
veya	CCONJ	single	21	8	13	38
dolayı	ADP	single	8	3	5	37
tersine	ADV	single	3	1	2	33
sonraları	ADV	single	3	1	2	33
gerçekten de	ADV	phrasal	6	2	4	33
bir yandan	ADV	phrasal	3	1	2	33
ve	CCONJ	single	905	277	628	30
ya da	CCONJ	phrasal	52	14	38	26
ardından	ADV ADP	single	21	5	16	23
nedeniyle	OTHER	single	18	4	14	22
önce	ADV ADP	single	74	15	59	20
işte	ADV	single	38	5	33	13
ise	ADP	single	63	8	55	12
hem	CCONJ	single	27	3	24	11
yalnız	ADV	single	18	2	16	11
gibi	ADP	single	169	19	150	11
o zaman	ADV	phrasal	19	2	17	10
kadar	ADP	single	144	15	129	10
hem de	CCONJ	phrasal	11	1	10	9
beraber	OTHER	single	11	1	10	9
özellikle	ADV	single	24	2	22	8
zaman	OTHER	single	67	5	62	7
adeta	ADV	single	13	1	12	7
sonra da	ADV	phrasal	13	1	12	7

Connective	PoS	Conn. Type	Total (#)	DC Use (#)	NDC Use (#)	DC Use (%)
eğer	CCONJ	single	30	2	28	6
peki	ADV	single	16	1	15	6
artık	ADV	single	61	3	58	4
hatta	CCONJ	single	25	1	24	4
üzere	ADP	single	25	1	24	4
yalnızca	ADV	single	21	1	20	4
sanki	ADV	single	22	1	21	4
zaten	ADV	single	27	1	26	3
sadece	ADV	single	29	1	28	3
birlikte	ADP	single	45	1	44	2
diye	ADP	single	97	1	96	1
de	OTHER	single	432	2	430	0
ki	CCONJ	single	117	1	116	0

Table A.1: Discourse Usage Analysis Results

Appendix B

DISCOURSE USAGE ANALYSIS CATEGORIES

Category	**Connective list**
Mostly DC	ne var ki, ve de, olsa bile, aksine, oysa, diğer bir deyişle, öyleyse, ne ki, kısacası, yine de, dolayısıyla, aynı zamanda da, dahası, iken
Mostly NDC	önce, işte, ise, hem, yalnız, gibi, o zaman, kadar, hem de, beraber, özellikle, zaman, adeta, sonra da, eğer, peki, artık, hatta, üzere, yalnızca, sanki, zaten, sadece, birlikte, diye, de, ki
Connectives used in a DC or a NDC sense	çünkü, böylece, ayrıca, halde, keza, öte yandan, ama, amacıyla, karşın, ancak, mesela, gene de, halbuki, aslında, fakat, örneğin, meğer, ikincisi, yoksa, sonra, için, yani, rağmen, veya, dolayı, tersine, sonraları, gerçekten de, bir yandan, ve, ya da, ardından, nedeniyle

Table B.1: The categories formed as a result of the discourse usage analysis of single and phrasal connectives and the list of connectives in each category

Appendix C

CONNECTIVE SAMPLES FROM TCL IN THE DIMLEX FORMAT

```
<entry id="2" word="-(y)ArAk">
  <orths>
    <orth canonical="0" orth_id="2o1" type="cont">
      <part type="suffixal">arak</part>
    </orth>
    <orth canonical="1" orth_id="2o1" type="cont">
      <part type="suffixal">erek</part>
    </orth>
  </orths>
  <syn>
    <cat>Converb</cat>
    <sem>
      <pdtb2_relation anno_N="14" freq="6" sense="EXPANSION: Manner: Arg2-as-manner" />
      <example>Sandalyemin tekerleklerini çevirerek koltuğunun önüne gelmiştim.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="14" freq="1" sense="TEMPORAL: Asynchronous: Succession" />
      <example>Peştamalı ve havluyu ıslatarak banyodan çıktı.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="14" freq="1" sense="CONTINGENCY: Cause: Reason(Arg1-as-result)" />
      <example>onu günde dört-beş kez mahzene inen Abid Efendi sanarak istifini bozmadı.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="14" freq="2" sense="EXPANSION: Conjunction" />
      <example>paketten çıkarılarak Bakanlar Kurulu'na sevk edildi.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="14" freq="1" sense="Contingency.Cause.Result" />
      <example>Bir kez daha 10 kat uzaklaşarak Alpha Centauri'ye varıyoruz.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="14" freq="1" sense="Expansion.Manner.Arg2-as-manner | Contingency.Cause.Result" />
      <example>Teleskobun içinde saçılarak, gezegeni görülemeyecek hâle getiren
      bu aşırı parlak görüntüyü yaratıyor.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="14" freq="1" sense="Temporal.Synchronous" />
      <example>Küçük bir kız olarak, bir gün kaçan giden biri olmayı hayal ettim.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="14" freq="1" sense="Expansion.Manner.Arg2-as-manner | Temporal.Synchronous" />
      <example>Hiç kimse yollara düşerek kendi içindeki şeytanından kurtulamıyor.</example>
    </sem>
  </syn>
</entry>
```

Figure C.1: A sample suffixal connective -(y)ArAk ‘by means of/ and’ in the DiMLex format

```
<entry id="6" word="-AlI">
  <orths>
    <orth canonical="0" orth_id="6o1" type="cont">
      <part type="suffixal">alı</part>
    </orth>
    <orth canonical="1" orth_id="6o1" type="cont">
      <part type="suffixal">eli</part>
    </orth>
  </orths>
  <syn>
    <cat>Converb</cat>
    <sem>
      <pdtb2_relation anno_N="4" freq="4" sense="TEMPORAL: Asynchronous: Succession" />
      <example>üç yıl oldu siz evleneli.</example>
    </sem>
  </syn>
</entry>
```

Figure C.2: A sample suffixal connective -AlI ‘since’ in the DiMLex format

```
<entry id="67" word="Çünkü">
  <orths>
    <orth canonical="0" orth_id="67o1" type="cont">
      <part type="single">Çünkü</part>
    </orth>
    <orth canonical="1" orth_id="67o1" type="cont">
      <part type="single">çünkü</part>
    </orth>
  </orths>
  <syn>
    <cat>CCONJ</cat>
    <sem>
      <pdtb2_relation anno_N="48" freq="5" sense="CONTINGENCY: Cause+Belief: Reason+Belief (Arg2-gives-evidence)" />
      <example>tırmanmaya başlandı mı bitirilmeli! Çünkü her seferinde acımasız bir geriye dönüş vardı.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="48" freq="30" sense="CONTINGENCY: Cause: Reason(Arg1-as-result)" />
      <example>Artıkgündüzleri terasta kalmak zorundaydı, çünkü hamam gündüzleri açıktı.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="48" freq="2" sense="CONTINGENCY: Cause+Belief" />
      <example>Nesli'nin, koluna asılıp gözlüklerinin üstünden şehla şehla bakmasını ve gülmesini çok seviyor,
      çünkü gülmek çok yakışıyor ona.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="48" freq="8" sense="Contingency.Cause.Reason" />
      <example>Bence bunları gözardı etmek kayıtsızlık olur, çünkü gözardı edilmesi
      ileriye dönük uzun vadeli kazançları riske atabilir.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="48" freq="3" sense="Contingency.Cause+Belief.Reason+Belief" />
      <example>Gelecek için planla, çünkü orası hayatının geri kalanını geçireceğin yerdir.</example>
    </sem>
  </syn>
</entry>
```

Figure C.3: A sample single connective *çünkü* ‘because’ in the DiMLex format

```
<entry id="94" word="için">
  <orths>
    <orth canonical="0" orth_id="94o1" type="cont">
      <part type="single">için</part>
    </orth>
  </orths>
  <syn>
    <cat>ADP</cat>
    <sem>
      <pdtb2_relation anno_N="121" freq="72" sense="CONTINGENCY: Purpose" />
      <example>Sokakta birlikte olmak için sandalyemi itmen gerekiyordu.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="121" freq="17" sense="CONTINGENCY: Cause: Reason(Arg1-as-result)" />
      <example>Tüm gücünü kullandığı için ter içinde kalmıştı.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="121" freq="1" sense="CONTINGENCY: Cause+Belief: Result+Belief (Arg1-gives-evidence)" />
      <example>özgürlüğünü teslim ettiği için sanki rahatlamıştı.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="121" freq="1" sense="COMPARISON: Degree" />
      <example>Tamamen pilottan kaynaklanan hata olduğunu söylemek için erken"</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="121" freq="23" sense="Contingency.Purpose.Arg2-as-goal" />
      <example>Portfolyomuzu politika beyanı yapmak için kullanmak istemiyoruz.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="121" freq="5" sense="Contingency.Cause.Reason" />
      <example>bu kabul edilmiş bilimsel mutabakata dayalı olduğu için tamamen yanlış olmaması ihtimali evimizin
      yanması veya araba kazası geçirmemiz ihtimalinden daha fazla.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="121" freq="1" sense="Expansion.Level-of-detail.Arg2-as-detail" />
      <example>gezegeni görmek için tüm bu ışığa ilişkin bir şeyler yapmamız gerek.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="121" freq="1" sense="Contingency.Cause+Belief.Reason+Belief (Arg2-gives-evidence)" />
      <example>Ustalar bir konuyu kavramsal sona getirdikleri için uzman değillerdir.</example>
    </sem>
  </syn>
</entry>
```

Figure C.4: A sample single connective *için* 'in order to' in the DiMLex format

```
<entry id="80" word="gene de">
  <orths>
    <orth canonical="0" orth_id="80o1" type="cont">
      <part type="phrasal">gene de</part>
    </orth>
    <orth canonical="1" orth_id="80o1" type="cont">
      <part type="phrasal">Gene de</part>
    </orth>
  </orths>
  <syn>
    <cat>ADV</cat>
    <sem>
      <pdtb2_relation anno_N="2" freq="2" sense="COMPARISON: Concession: Arg2-as-denier" />
      <example>Henüz çok iyi öğrenememiştim New York metrosunu ama gene de her gece
      gideceğim yere varabiliyordum.</example>
    </sem>
  </syn>
</entry>
```

Figure C.5: A sample phrasal continuous connective *gene de* 'nevertheless' in the DiMLex format

```
<entry id="48" word="Bu sebeple">
  <orths>
    <orth canonical="0" orth_id="48o1" type="cont">
      <part type="phrasal">Bu sebeple</part>
    </orth>
  </orths>
  <syn>
    <cat>Secondary connective</cat>
    <sem>
      <pdtb2_relation anno_N="1" freq="1" sense="" />
      <example>İmparatorluğu meydana getiren azınlıkların milliyetçi kıpırdanmaları, devletin geleceği önünde en büyük
      tehlikeyi oluşturmaktaydı. Bu sebeple, İmparatorluk halklarını birarada tutacak, ayrılıkçı eğilimleri frenleyecek
      yeni bir dâyanışmanın yaratılması gerekmekteydi.</example>
    </sem>
  </syn>
</entry>
```

Figure C.6: A sample phrasal continuous connective *bu sebeple* ‘for this reason’ in the DiMLex format

```
<entry id="135" word="Önce ... Sonra">
  <orths>
    <orth canonical="0" orth_id="135o1" type="discont">
      <part type="phrasal">önce sonra</part>
    </orth>
    <orth canonical="1" orth_id="135o1" type="discont">
      <part type="phrasal">önce Sonra</part>
    </orth>
    <orth canonical="2" orth_id="135o1" type="discont">
      <part type="phrasal">Önce sonra</part>
    </orth>
    <orth canonical="3" orth_id="135o1" type="discont">
      <part type="phrasal">Önce Sonra</part>
    </orth>
  </orths>
  <syn>
    <cat>CCONJ</cat>
    <sem>
      <pdtb2_relation anno_N="2" freq="1" sense="TEMPORAL: Asynchronous: Precedence" />
      <example>Önce biraz oturdu. Sonra duvara sırtını verdi ve adımlayarak yukarıdan
      deliği açacağı yeri hesapladı.</example>
    </sem>
    <sem>
      <pdtb2_relation anno_N="2" freq="1" sense="TEMPORAL: Synchronous" />
      <example>önce hükümete brifing verdi, sonra da kamuoyuna açıklama yaptı.</example>
    </sem>
  </syn>
</entry>
```

Figure C.7: A sample phrasal discontinuous connective *önce* ... *sonra* ‘first ... then’ in the DiMLex format

```
<entry id="160" word="tıpkı ... gibi">
  <orths>
    <orth canonical="0" orth_id="160o1" type="discont">
      <part type="phrasal">tıpkı gibi</part>
    </orth>
  </orths>
  <syn>
    <cat>ADP</cat>
    <sem>
      <pdtb2_relation anno_N="1" freq="1" sense="Comparison.Similarity" />
      <example>Bunun nedeni, ışık dalgalarının tıpkı teleskobun içinde olduğu gibi,
      gölgelikten sapma yapmaları.</example>
    </sem>
  </syn>
</entry>
```

Figure C.8: A sample phrasal discontinuous connective *tıpkı* ... *gibi* ‘as if’ in the DiMLex format

Appendix D

EXPLICIT CONNECTIVES AND THEIR SENSES

Connective	Freq	Sense
-(y)AcAğInA	2	expansion.substitution.arg1-as-subst
-(y)ArAk	6	expansion.manner.arg2-as-manner
	1	temporal.asynchronous.succession
	1	contingency.cause.reason(arg1-as-result)
	2	expansion.conjunction
	1	contingency.cause.result
	1	expansion.manner.arg2-as-manner|contingency.cause.result
	1	temporal.synchronous
	1	expansion.manner.arg2-as-manner|temporal.synchronous
-(y)HncA	10	temporal.synchronous
	3	contingency.cause.reason(arg1-as-result)
	5	temporal.asynchronous.succession
	1	expansion.level-of-detail.arg2-as-detail
-(y)HncA dA	1	temporal.asynchronous
-(y)Hp	22	expansion.conjunction
	1	expansion.manner.arg2-as-manner
	1	temporal.synchronous|expansion.conjunction
-AlI	4	temporal.asynchronous.succession
-cAsInA	2	expansion.manner.arg2-as-manner
-dA	17	temporal.synchronous
	1	contingency.condition.arg2-as-cond
-dAn	6	contingency.cause.reason(arg1-as-result)
	1	contingency.cause.result(arg2-as-result)
-DHk-çA	6	temporal.synchronous
-ken	20	temporal.synchronous
	1	expansion.conjunction
	1	comparison.concession+speechact
	1	comparison.contrast
-mAksIzIn	3	expansion.manner.arg2-as-manner
	1	temporal.asynchronous.precedence
-sA	19	contingency.condition.arg2ascondition
	1	comparison.concession.arg1-as-denier
	2	contingency.negative-condition.arg2-as-negcond
	1	contingency.condition.arg2-as-cond|contingency.purpose.arg2-as-goal

Connective	Freq	Sense
	3	contingency.condition+speechact
	1	expansion.substitution.arg2-as-subst
-sA dA	2	comparison.concession.arg1-as-denier
	1	comparison.concession.arg2-as-denier
adeta	1	expansion.equivalence
aksine	1	expansion.substitution.arg2-as-subst
ama	48	comparison.contrast
	73	comparison.concession.arg2-as-denier
	4	expansion.exception
	4	comparison.concession+speechact.arg2-as-denier+speechact
	3	expansion.correction
	1	comparison.concession
	1	contingency.cause+speechact
	2	expansion.conjunction
	1	comparison.concession|temporal.synchronous
	1	comparison.concession.arg2-as-denier|expansion.conjunction
amacıyla	5	contingency.purpose
ancak	23	comparison.concession.arg2-as-denier
	2	expansion.exception
	15	comparison.contrast
	1	expansion.conjunction
	1	expansion.exception.arg2-as-excpt
ardından	3	temporal.asynchronous.precedence
	2	temporal.asynchronous.succession
artık	1	expansion.level-of-detail.arg2-as-detail
	1	temporal.synchronous
	1	contingency.cause.result
aslında	6	expansion.level-of-detail.arg2-as-detail
	1	expansion.exception
	1	comparison.concession.arg2-as-denier
	1	comparison.concession+speechact.arg2-as-denier+speechact
	1	expansion.conjunction
	1	expansion.correction
	1	expansion.level-of-detail.arg2-as-detail|expansion.substitution.arg2-as-subst
	1	expansion.instantiation
aynı zamanda da	1	temporal.synchronous
	1	expansion.conjunction
ayrıca	21	expansion.conjunction
beraber	1	expansion.conjunction
bir tarafta ... bir tarafta da	1	comparison.contrast
bir yandan	1	temporal.synchronous
bir yandan ... bir yandan	1	expansion.conjunction
birlikte	1	comparison.concession.arg1-as-denier
böylece	8	contingency.cause.result(arg2-as-result)
	5	contingency.cause.result

Connective	Freq	Sense
	1	contingency.cause+belief.result+belief(arg1-gives-evidence)
çünkü	5	contingency.cause+belief.reason+belief(arg2-gives-evidence)
	30	contingency.cause.reason(arg1-as-result)
	2	contingency.cause+belief
	8	contingency.cause.reason
	3	contingency.cause+belief.reason+belief
dahası	1	expansion.level-of-detail.arg2-as-detail
	1	expansion.conjunction
de	1	expansion.conjunction
	1	comparison.similarity
diğer bir deyişle	1	expansion.equivalence
diye	1	expansion.disjunction
dolayı	3	contingency.cause.reason(arg1-as-result)
dolayısıyla	9	contingency.cause.result(arg2-as-result)
	1	contingency.cause.result
eğer	2	contingency.condition.arg2-as-cond
fakat	10	comparison.contrast
	2	comparison.concession.arg2-as-denier
	1	comparison.concession
gene de	2	comparison.concession.arg2-as-denier
gerçekten de	1	contingency.cause+belief.reason+belief(arg2-gives-evidence)
	1	expansion.level-of-detail.arg1-as-detail
gibi	3	expansion.conjunction
	15	expansion.manner.arg2-as-manner
	1	comparison.similarity
halbuki	3	comparison.contrast
halde	1	contingency.condition.arg2ascondition
	2	expansion.manner.arg2-as-manner
	2	comparison.concession.arg1-as-denier
hatta	1	expansion.conjunction
hem	1	expansion.level-of-detail.arg2-as-detail
	2	expansion.conjunction
hem ... hem	2	expansion.conjunction
hem de	1	expansion.conjunction|expansion.equivalence
için	72	contingency.purpose
	17	contingency.cause.reason(arg1-as-result)
	1	contingency.cause+belief.result+belief(arg1-gives-evidence)
	1	comparison.degree
	23	contingency.purpose.arg2-as-goal
	5	contingency.cause.reason
	1	expansion.level-of-detail.arg2-as-detail
	1	contingency.cause+belief.reason+belief(arg2-gives-evidence)
iken	1	comparison.contrast
	1	contingency.condition.arg2ascondition
ikincisi	1	expansion.level-of-detail.arg2-as-detail

Connective	Freq	Sense
ise	3	comparison.contrast
	2	expansion.conjunction
	2	comparison.concession.arg2-as-denier
	1	comparison.contrast|expansion.level-of-detail.arg2-as-detail
işte	1	contingency.cause+belief.reason+belief(arg2-gives-evidence)
	3	contingency.cause.result
	1	expansion.equivalence
kadar	11	comparison.degree
	2	temporal.asynchronous.precedence
	1	comparison.similarity
	1	contingency.purpose.arg2-as-goal
karşın	1	comparison.concession.arg2-as-denier
	8	comparison.concession.arg1-as-denier
keza	4	expansion.instantiation
	1	expansion.conjunction
kısacası	1	expansion.level-of-detail.arg1-as-detail
ki	1	expansion.conjunction
meğer	1	contingency.cause+belief.result+belief(arg1-gives-evidence)
mesela	2	expansion.instantiation
ne ... ne	4	expansion.conjunction
ne ki	1	comparison.contrast
ne var ki	1	comparison.contrast
nedeniyle	4	contingency.cause.reason(arg1-as-result)
o zaman	1	contingency.cause+speechact.result+speechact
	1	contingency.cause.result
olsa bile	1	contingency.condition.arg2-as-cond
oysa	9	comparison.contrast
	1	comparison.concession
	3	comparison.concession.arg2-as-denier
önce	8	temporal.asynchronous.precedence
	7	temporal.asynchronous.succession
önce ... ardından	1	temporal.asynchronous.precedence
önce ... sonra	1	temporal.asynchronous.precedence
	1	temporal.synchronous
örneğin	3	expansion.instantiation
öte yandan	4	expansion.conjunction
öyle ... ki	1	contingency.cause.result(arg2-as-result)
	2	expansion.instantiation
öyleyse	1	contingency.cause.result(arg2-as-result)
	1	contingency.cause+speechact.result+speechact(arg1-as-reason)
özellikle	2	expansion.conjunction
peki	1	expansion.level-of-detail.arg2-as-detail
rağmen	8	comparison.concession.arg1-as-denier
sadece	1	expansion.substitution.arg1-as-subst

Connective	Freq	Sense
sanki	1	comparison.similarity
sonra	27	temporal.asynchronous.precedence
	44	temporal.asynchronous.succession
sonra da	1	temporal.asynchronous.precedence
sonraları	1	temporal.asynchronous
tersine	1	comparison.contrast
tıpkı ... gibi	1	comparison.similarity
üzere	1	contingency.purpose.arg2-as-goal
ve	229	expansion.conjunction
	19	contingency.cause.result(arg2-as-result)
	1	contingency.cause+belief.result+belief(arg1-gives-evidence)
	12	temporal.asynchronous.precedence
	4	temporal.synchronous
	2	expansion.level-of-detail.arg2-as-detail
	3	expansion.conjunction|expansion.level-of-detail.arg2-as-detail
	1	expansion.conjunction|comparison.contrast
	1	contingency.cause.result
	1	expansion.conjunction|temporal.synchronous
	3	expansion.conjunction|contingency.cause.result
	1	expansion.conjunction|expansion.instantiation
ve ... da	1	expansion.conjunction|temporal.asynchronous.precedence
ve de	2	expansion.conjunction
veya	8	expansion.disjunction
ya ... ya da	1	expansion.disjunction
ya da	1	expansion.substitution.arg2-as-subst
	7	expansion.conjunction
	6	expansion.disjunction
yalnız	1	comparison.contrast
	1	expansion.conjunction
yalnızca	1	expansion.exception.arg2-as-excpt
yani	9	expansion.equivalence
	2	expansion.level-of-detail.arg2-as-detail
	3	contingency.cause+belief.result+belief(arg1-gives-evidence)
	1	expansion.substitution
	2	contingency.cause+belief
	1	contingency.cause+speechact.result+speechact
	1	expansion.substitution.arg2-as-subst
yine de	1	comparison.concession.arg2-as-denier
yoksa	2	contingency.negative-condition.arg1-as-negcond
	2	expansion.disjunction
zaman	2	contingency.condition.arg2ascondition
	2	temporal.synchronous
	1	temporal.asynchronous
zaten	1	expansion.conjunction

Appendix E

CONFUSION MATRICES

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**8**	1	0	0
Expansion	0	**16**	0	0
Comparison	1	2	**5**	0
Temporal	0	1	0	**0**

Table E.1: The confusion matrix for the sense classification of single/phrasal connectives - T-TED-MDB fold1

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**11**	1	1	0
Expansion	4	**28**	1	1
Comparison	0	1	**16**	0
Temporal	0	0	0	**1**

Table E.2: The confusion matrix for the sense classification of single/phrasal connectives - T-TED-MDB fold2

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**24**	1	1	0
Expansion	0	**6**	0	0
Comparison	0	0	**6**	0
Temporal	1	0	0	**0**

Table E.3: The confusion matrix for the sense classification of single/phrasal connectives - T-TED-MDB fold3

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**16**	1	0	0
Expansion	2	**11**	0	0
Comparison	0	1	**5**	0
Temporal	0	0	0	**0**

Table E.4: The confusion matrix for the sense classification of single/phrasal connectives - T-TED-MDB fold4

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**17**	2	0	0
Expansion	0	**21**	3	0
Comparison	1	0	**4**	0
Temporal	1	1	0	**0**

Table E.5: The confusion matrix for the sense classification of single/phrasal connectives - T-TED-MDB fold5

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**30**	0	3	1
Expansion	0	**26**	4	1
Comparison	4	1	**56**	0
Temporal	0	0	2	**11**

Table E.6: The confusion matrix for the sense classification of single/phrasal connectives - TDB 1.1 fold1

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**48**	0	3	0
Expansion	1	**34**	6	0
Comparison	1	1	**54**	0
Temporal	0	0	5	**23**

Table E.7: The confusion matrix for the sense classification of single/phrasal connectives - TDB 1.1 fold2

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**28**	1	1	0
Expansion	1	**47**	0	0
Comparison	2	2	**48**	0
Temporal	0	0	1	**17**

Table E.8: The confusion matrix for the sense classification of single/phrasal connectives - TDB 1.1 fold3

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**34**	3	0	1
Expansion	5	**55**	0	1
Comparison	1	2	**15**	1
Temporal	0	2	0	**27**

Table E.9: The confusion matrix for the sense classification of single/phrasal connectives - TDB 1.1 fold4

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**38**	2	1	0
Expansion	0	**26**	1	0
Comparison	6	1	**49**	0
Temporal	2	2	4	**22**

Table E.10: The confusion matrix for the sense classification of single/phrasal connectives - TDB 1.1 fold5

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**254**	12	10	2
Expansion	13	**270**	15	3
Comparison	16	11	**258**	1
Temporal	4	6	12	**101**

Table E.11: The confusion matrix for the sense classification of single/phrasal connectives - total

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**4**	0	0	1
Expansion	0	**5**	1	0
Comparison	0	0	**1**	0
Temporal	0	1	0	**0**

Table E.12: The confusion matrix for the sense classification of suffixal connectives - T-TED-MDB fold1

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**11**	1	0	0
Expansion	0	**8**	0	0
Comparison	1	0	**1**	0
Temporal	1	0	0	**0**

Table E.13: The confusion matrix for the sense classification of suffixal connectives - T-TED-MDB fold2

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**4**	0	1	0
Expansion	1	**3**	0	0
Comparison	0	1	**7**	0
Temporal	2	1	0	**0**

Table E.14: The confusion matrix for the sense classification of suffixal connectives - T-TED-MDB fold3

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**7**	1	0	0
Expansion	0	**3**	0	0
Comparison	0	0	**5**	0
Temporal	0	0	1	**0**

Table E.15: The confusion matrix for the sense classification of suffixal connectives - T-TED-MDB fold4

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**6**	1	1	0
Expansion	1	**5**	0	1
Comparison	0	0	**0**	0
Temporal	0	0	0	**7**

Table E.16: The confusion matrix for the sense classification of suffixal connectives - T-TED-MDB fold5

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**0**	22	1	0
Expansion	1	**65**	1	0
Comparison	0	11	**11**	0
Temporal	0	2	1	**1**

Table E.17: The confusion matrix for the sense classification of suffixal connectives - TDB 1.1 fold1

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**1**	20	1	0
Expansion	0	**95**	2	0
Comparison	1	6	**8**	0
Temporal	0	7	1	**1**

Table E.18: The confusion matrix for the sense classification of suffixal connectives - TDB 1.1 fold2

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**0**	1	22	0
Expansion	0	**9**	8	0
Comparison	0	1	**65**	0
Temporal	0	1	7	**2**

Table E.19: The confusion matrix for the sense classification of suffixal connectives - TDB 1.1 fold3

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**0**	24	0	0
Expansion	1	**64**	2	0
Comparison	0	11	**7**	0
Temporal	0	1	0	**4**

Table E.20: The confusion matrix for the sense classification of suffixal connectives - TDB 1.1 fold4

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**0**	26	0	0
Expansion	0	**68**	0	0
Comparison	0	2	**4**	1
Temporal	0	15	0	**10**

Table E.21: The confusion matrix for the sense classification of suffixal connectives - TDB 1.1 fold5

Truth \ Predicted	Contingency	Expansion	Comparison	Temporal
Contingency	**33**	96	26	1
Expansion	4	**325**	14	1
Comparison	2	32	**109**	1
Temporal	3	28	10	**25**

Table E.22: The confusion matrix for the sense classification of suffixal connectives - total

CURRICULUM VITAE

PERSONAL INFORMATION