ASSESSMENT OF THE TURKISH DISCOURSE BANK AND A CASCADED MODEL TO AUTOMATICALLY IDENTIFY DISCURSIVE PHRASAL EXPRESSIONS IN TURKISH

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

AYIŞIĞI BAŞAK SEVDİK ÇALLI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

FEBRUARY 2015

ASSESSMENT OF THE TURKISH DISCOURSE BANK AND A CASCADED MODEL TO AUTOMATICALLY IDENTIFY DISCURSIVE PHRASAL EXPRESSIONS IN TURKISH

Submitted by **AYIŞIĞI BAŞAK SEVDİK ÇALLI** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Cognitive Science Department, Middle East Technical University** by,

Prof. Dr. Nazife Baykal
Director,**Informatics Institute**

Prof. Dr. Hüseyin Cem Bozşahin
Head of Department, **Cognitive Science**

Prof. Dr. Deniz Zeyrek Bozşahin
Supervisor, **Cognitive Science**

Examining Committee Members:

Prof. Dr. Hüseyin Cem Bozşahin
Cognitive Science Department, METU

Prof. Dr. Deniz Zeyrek Bozşahin
Cognitive Science Department, METU

Assist. Prof. Dr. Cengiz Acartürk
Cognitive Science Department, METU

Assist. Prof. Dr. Murat Perit Çakır
Cognitive Science Department, METU

Prof. Dr. Ümit Deniz Turan
English Language Teaching, Anadolu University

Date: **27 February 2015**

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, Last Name: AYIŞIĞI BAŞAK SEVDİK ÇALLI

Signature :

ABSTRACT

ASSESSMENT OF THE TURKISH DISCOURSE BANK AND A CASCADED MODEL TO AUTOMATICALLY IDENTIFY DISCURSIVE PHRASAL EXPRESSIONS IN TURKISH

Sevdik Çallı, Ayışığı Başak
Ph.D., Department of Cognitive Science
Supervisor : Prof. Dr. Deniz Zeyrek Bozşahin

February 2015, 188 pages

This thesis presents a methodology for an overall assessment of the Turkish Discourse Bank (TDB), a linguistic resource where discourse relations overtly expressed by discourse connectives have been identified and annotated with the two arguments they relate. We provide a quantitative and qualitative assessment of the TDB in order to establish the reliability of this discourse resource for Turkish and suggest that our methodology can be utilized for reliability evaluations of other annotated corpora. Our quantitative evaluation consists of calculating in depth statistical measures using the Kappa statistic and extra evaluators originally used in evaluating information retrieval systems. A two-way methodology for calculating the agreement statistics is proposed: a Common Arguments approach and an Overall approach. Although the Overall approach is effective on its own, we propose a comparison of these two approaches, which enables to pin point sources of disagreements more accurately. As part of our qualitative evaluation we present a novel effort to automatically identify discursive uses of phrasal expressions that have been annotated systematically alongside explicit discourse connectives in the TDB, given any Turkish text. Our cascaded model, achieves full recall, provides 99.95% accuracy, and can be utilized to effortlessly enlarge the coverage of the TDB.

Keywords: Turkish Discourse Bank, discourse annotation, agreement measures, phrasal expressions, automatic identification

ÖZ

ODTÜ METİN DÜZEYİNDE İŞARETLENMİŞ DERLEM'İN DEĞERLENDİRMESİ VE TÜRKÇEDE DEYİMSEL İFADELERİN OTOMATİK BELİRLENMESİ İÇİN KADEMELİ BİR MODEL

Sevdik Çallı, Ayışığı Başak
Doktora, Bilişsel Bilimler Bölümü
Tez Yöneticisi : Prof. Dr. Deniz Zeyrek Bozşahin

Şubat 2015 , 188 sayfa

Bu doktora tezi söylem bağıntılarını açıkça ifade eden söylem bağlaçlarının belirlendiği ve birbiri ile ilişkilendirdiği iki üyesi ile beraber işaretlendiği bir dilbilimsel kaynak olan ODTÜ Metin Düzeyinde İşaretlenmiş Derlem'in (ODTÜ-MEDİD) kapsamlı değerlendirmesi için bir yöntem sunmaktadır. Bu çalışmada söz konusu Türkçe söylem kaynağının güvenilirliğini ortaya koymak amacıyla ODTÜ-MEDİD'in niceliksel ve niteliksel bir değerlendirmesi sunulmakta ve burada kullandığımız yöntemin diğer işaretlenmiş derlemlerin güvenilirlik değerlendirmeleri için kullanılabileceği önerilmektedir. Niceliksel değerlendirmemiz Kappa uyum istatistiği kullanılarak detaylı istatistiksel ölçütlerin ve daha önce bilgi erişim sistemlerinin değerlendirmesinde kullanılan bir takım ek değerlendiricilerin hesaplanmasını içermektedir. Uyum istatistiklerinin hesaplanmasında iki yönlü bir yöntem önerilmektedir: bir Ortak Üye yaklaşımı ve bir Kapsamlı yaklaşım. Kapsamlı yaklaşım tek başına etkili olsa da, uyumsuzluk kaynaklarının daha etkin bir biçimde saptanmasını sağlamak amacıyla bu iki yaklaşımın karşılaştırılması önerilmektedir. Niteliksel değerlendirmemiz kapsamında ise ODTÜ-MEDİD'te sistemli olarak söylem bağlaçları ile birlikte işaretlenen deyimsel ifadelerin metin düzeyinde kullanımlarının herhangi bir Türkçe metin üzerinde otomatik olarak tanımlanmasını sağlayan özgün bir girişim sunulmaktadır. Kademeli modelimiz tam geri çağırma ve %99.95 doğruluk sağlamaktadır. Bu modelin ODTÜ-MEDİD'in kapsama alanını geliştirmek için rahatlıkla kullanılabileceği öngörülmektedir.

Anahtar Kelimeler: ODTÜ Metin Düzeyinde İşaretlenmiş Derlem, metin düzeyinde işaretleme, uyum istatistikleri, deyimsel ifadeler, otomatik tanımlama

To my Gökyüzü and my Nebi.

ACKNOWLEDGMENTS

I would like to thank my supervisor Deniz Zeyrek Bozşahin for her guidance, constructive critism, support, kindness, patience and motivation throughout the development of this thesis. I would also like to thank the members of my thesis jury, Cem Bozşahin, Ümit Deniz Turan, Cengiz Acartürk and Murat Perit Çakır for their valuable comments and evaluations. I also would like to express my gratitude to Varol Akman and Fazlı Can at Bilkent University, who supported me in my decision to continue my research in Cognitive Science at METU, without whom I might not have envisioned a doctoral degree at all. Another sincere thank you is for Bilge Say, who has made my transition into Cogs easier with her support and advice.

A special thanks goes to Işın Demirşahin as a true friend, an office mate, an exceptional colleague and a fellow PhD candidate, who has been an immense support in every aspect of my METU Cogs days. I would also like to thank my colleagues in the METU TDB Project: Ruken Çakıcı, who guided me with her NLP knowledge at a time of being lost; Hale Ögel Balaban, who has been a good friend and an exemplary colleague; İhsan Yalçınkaya, who answered my endless questions despite having left the academic scene for a while; Berfin Aktaş and Utku Şirin for developing the TDB tools and always being helpful; and others who have participated in the TDB project.

I would like to acknowledge TÜBİTAK (The Scientific and Technological Research Council of Turkey) for the scholarship I received in the framework of the 1001 Research Project, No. 107E156.

I would also like to present my thanks to my friends at METU Cogs: Elif Selvi Gök for being a lovely office mate and a friend; Hatice Bayındır Karaaslan, Selgün Yüceil and Hande Kaynak for helping me get through the comprehensive exams; Özkan Kılıç for being an ever-encouraging friend; Zeynep Başgöze and Gözde Bahadır Mercan for making life of a PhD student bearable.

I would like to express my gratitude to my dear friend from my undergraduate studies at Başkent University and my graduate studies at Bilkent University, Rabia Nuray Turan, with

whom I started this academic journey together.

Very special thanks goes out to my true friends, who I consider sisters: Ayça Yetere Kurşun and Sibel Alumur Alev, thank you for always being there.

I would like to express deepest gratitude to my grandfather Recep Sevdik for always being an example in my head reminding me to be as diligent as I can; my grandfather Kamil Doğan for engraving his trio in my head: mind, body and heart, reminding me what's important in life; to my grandmother Ayten Doğan and my aunt Pınar Doğan for their love; without you life is missing a piece, I miss you all deeply.

I am indebted to my family, without their love and support I could not have done this. My parents, Bahar and Ömer Sevdik, I cannot find the words to express my gratitude, thank you for being such good role models. Thank you for making me who I am today. Thank you for taking care of my little one and for taking care of me while I worked on this thesis. Without your support none of this would have been possible. My brother, Alper Sevdik, though you are always far away, know that you are always close to my heart. To the rest of my family, thank you for your understanding, support and encouragement throughout this process.

My dear other half, Atılay Nebi Çallı, thank you for finding me, without your love and encouragement I could not walk through this life. You have been with me every step of the way, encouraged and motivated me on the darkest of days, stayed up beside me for support many many endless nights. This thesis could not have been completed without you. For that and much more, I am forever grateful.

My dear little one, Gökyüzü, you have changed my world in ways I could not have imagined. Thank you for allowing me to study and finish my thesis. You have provided me with the life energy to go on in times of despair. You are my joy, my ray of sunshine.

TABLE OF CONTENTS

CHAPTERS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AO	Abstract Object
CDT	Copenhagen Discourse Treebank
CDTB	Chinese Discourse TreeBank
D-LTAG	Lexicalized Tree-Adjoining Grammar for Discourse
DP	Discourse Purpose
DRS	Discourse Representation Structure
DRT	Discourse Representation Theory
DSP	Discourse Segment Purpose
DTC	Dutch Text Corpus
EDU	Elementary Discourse Unit
HDRB	Hindi Discourse Relational Bank
kNN	k-Nearest Neighbor
LADTB	Leeds Arabic Discourse TreeBank
L-TAG	Lexicalized Tree-Adjoining Grammar
MP	Minimality Principle
MST	METU-Sabancı Turkish Treebank
MTC	Metu Turkish Corpus
PCC	Potsdam Commentray Corpus
PDT	Prague Dependency Treebank
PDiT	Prague Discourse Treebank
PDTB	Penn Discourse TreeBank
PhD	Doctor of Philosophy (Latin: Philosophiae Doctor)
POS	Part Of Speech
RST	Rhetorical Structure Theory
SDRT	Segmented Discourse Representation Theory
TAM	Tense-Aspect-Mood
TDB	Turkish Discourse Bank
TDK	Türk Dil Kurumu ’Turkish Language Association’
WSJ	Wall Street Journal

CHAPTER 1

INTRODUCTION

This thesis is about discourse relations. It is about using language technology. It is about identifying phrasal expressions. It is about Turkish. It is about the mind. It is about corpora. It is about statistics.

No, these are not the ramblings of a PhD candidate. These sentences can all be tied up together. They can in fact form a coherent and cohesive whole, which forms this thesis. How? Let me try again.

This thesis presents an overall assessment of the Turkish Discourse Bank (TDB) (Zeyrek, Demirşahin, Sevdik Çallı, & Çakıcı, 2013), a linguistic resource where discourse relations overtly expressed by discourse connectives have been identified and annotated with the two arguments they relate. In order to unearth the value of such an annotated corpus for Turkish and linguistics as a universal concept, first we provide a statistical evaluation of this resource and make sure the annotations have been carefully carried out so as to produce a reliable, gold standard data. Then, after establishing the dependability of this resource, we utilize it for building language technology applications that can guide us in understanding the workings of Turkish discourse relations mainly by developing a model to automatically extract discursive uses of phrasal expressions in Turkish.

Observe the underlined words or word groups in the previous paragraph. All of these expressions signal a coherence relation between the clauses or sentences they relate to each other. These are what are known as *discourse connectives* (although they can be referred to with other names such as *discourse markers*, *cue phrases*, even *conjunctives*). They are used in many languages for the similar purpose of achieving coherence in discourse and they are what has been annotated in the TDB, along with the clauses or sentences providing the abstract objects they link for the Turkish language. Hence, our starting point in this dissertation is a discourse resource: the Turkish Discourse Bank.

The Turkish Discourse Bank (TDB) is the first publicly available corpus in Turkish annotated at the discourse-level (Zeyrek et al., 2013). This ~400.000-word language resource is a subcorpus of the 2 million-word METU Turkish Corpus (Say, Zeyrek, Oflazer, & Özge, 2004) consisting of texts of post-1990 written Turkish encompassing various genres. TDB contains Penn Discourse TreeBank (PDTB) (Prasad et al., 2008) style annotations of explicit discourse connectives, the two abstract object (Asher, 1993) arguments related by the connectives, as well as modifiers and supplementary materials. An example annotation of the explicit connective ama 'but' is provided in (1), where the argument syntactically hosting the connective is called the *second argument* and the other argument is the *first argument*. In the thesis, in all the examples provided from the TDB, the connective is shown underlined, the first argument is shown in italics and the second argument is rendered in bold face. The examples provided are from the TDB, unless otherwise stated.

(1) *Topu topu bir metrekarecik bir yer, belki daha da küçük.* Ama **barındırdığı binbir koku, binbir titreşim ve sesle, Artur için koca bir dünya**...

All in all just a square-meter of a space, maybe even smaller. But **with the thousands of scents, thousands of vibrations and sounds it houses, it is a big world for Artur**.

(00054123.txt)

Along with explicit connectives, phrasal expressions which contain a deictic demonstrative counterpart combined with a subordinating conjunction, e.g. *buna karşılık* 'despite this', have also been annotated in the TDB (2).

(2) *Bu konuda yapılan çalışmalar, ev kedilerinin gözlerinin, karanlıkta parlamaları gerekenden çok daha az parladıklarını gösteriyor. Yine başka bilimsel çalışmalar, insanlara bağımlılığımız arttıkça, koku alma ve görme yeteneğimizin yanında, duymamızda ve dişlerimizin sivriliğinde de gerilemeler olduğunu ortaya koyuyor.* Buna karşılık **kedi zekâmızda gelişme var mı**?

Studies conducted on this subject show that the eyes of house cats glow less than they should in the dark. Still other scientific studies put forward that as our dependence on humans increase, besides our smell and sight abilities, there is deterioration in our hearing and the sharpness of our teeth too. Despite this **is there increase in our cat intelligence**?

(00054223.txt)

However, implicit connectives defined in the PDTB framework to express discourse relations that are not explicitly signalled but inferred from adjacency as in (3), have not been annotated in TDB 1.0 (although there is ongoing work to annotate them in future versions).

(3) In July, the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos. (implicit = as a result) **By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed**. [wsj 0003]

(Prasad, Webber, & Joshi, 2014, p. 922, ex.2)

Although TDB 1.0 is not a resource covering all aspects of coherence, annotating explicit connectives is a first but important step towards a better coverage of discourse relations. This motivates the need to assess the validity and reliability of this resource.

In this thesis we are interested in establishing the TDB as a reliable discourse resource by providing a quantitative and qualitative assessment to the extent explicit connectives are concerned. Our quantitative evaluation includes providing basic descriptive statistics, as well as in depth statistical measures to ensure reliability of the annotations in the corpus. The qualitative evaluation involves putting this resource to use: using it to perform linguistic analyses, to add new annotations to it, to develop natural language applications and to test the abilities of these applications. In this way, we aim to confirm that the TDB can reliably be exploited in the future.

1.1 The Thesis

The main goal of the present work is to show that the Turkish Discourse Bank can successfully aid in the development of language technology applications such as automatically identifying discursive uses of phrasal expressions in Turkish. We show that discursive uses of Turkish phrasal expressions can be effectively detected from free text automatically, exploiting syntactic (e.g. part-of-speech) and morphological (e.g. case, person agreement) features, where the underlying hypothesis is that the discourse uses can be identified using lexical form, syntax and morphology. We propose a cascaded approach, where a lexical form filter is applied as a first step and the classifier is applied in a second step. We show that in the cascaded method, a perfect recall and a decrease in the total number of false predictions is achieved.

1.2 Motivation and Challenges

There is an increasing amount of effort to build discourse corpora in languages other than English in the linguistic community at this time. In the recent years, discourse corpora have been or is being built for Arabic, Chinese, Czech, Danish, Dutch, German, Hindi, Turkish and other languages. The Turkish Discourse Bank is the first resource for Turkish to annotate discourse relations.

In addition to providing an environment for scientists and researchers to make linguistic investigations, these corpora offer an important test bed for applications of Natural Language Engineering, as well as language technology. In this age of informatics, the growing interest for algorithms recognizing discourse structures in such applications is more pronounced for languages other than English as resources are scarce. Hence, the reliability of such linguistically annotated corpora gains prominence.

It is important to adopt an annotation approach which aims to be both linguistically sound and reliable at the same time. Such an approach will produce annotated corpora that can be reused by both NLP and Corpus Linguistics communities with confidence (Hovy & Lavid, 2010). The Computational Linguistics and NLP community can use these reliable resources to build linguistic applications with success, as well as enhance the resources by automatical means. On the other hand, the Corpus Linguistics community can solidify the theoretical soundness by ensuring the annotation reliability of these corpora. Hence, it is important to treat annotation as a ’science’ in order to achieve better annotated resources and better linguistic applications.

There are two main criteria in the way to ’science of annotation’ described by Hovy and Lavid (2010): 1.) a quantitative evaluation 2.) a qualitative evaluation. The quantitative assessment involves calculating the measures to prove reliability by mathematical or statistical means. The qualitative assessment further requires to make sure that this resource can in fact be used as a linguistic resource to analyze discourse structure, to model natural language algorithms, as well as to evaluate and test them. Hence, the challenges for the first criterion is calculating related statistical measures, evaluating their adequacy and efficiency and determining additional measures if necessary. The challenges for the second condition involves determining and addressing any deficiencies put forward in the first condition, making a linguistic analysis, building a model for a natural language algorithm, testing and evaluating this model.

The linguistic analysis we will make in this thesis will focus on the phrasal expressions systematically annotated in the TDB and their discursive use. Then in order to satisfy the rest of the second criterion, we will build a model to automatically extract discursive uses of phrasal expressions given any Turkish text and we will test and evaluate this model using the TDB as our gold standard data.

The automatic detection of discursive uses of phrasal expressions brings about more issues to tackle. The challenge in this quest is that, just doing a simple token search to match connective forms is not enough. Some phrasal expressions are not always used as discourse connectives, sometimes they are used non-discursively. Hence, additional methodology is required to effectively disambiguate discursive uses of phrasal expressions. Fulfilling this challenge will also provide a means towards extending the coverage of the TDB, as it will enable us to retrieve and annotate phrasal expressions in the rest of the corpus.

1.3 Contribution

The contributions this dissertation makes are the following:

- Assessment of the Turkish Discourse Bank as a reliable source for Turkish discourse coherence to the extent explicit connectives are concerned. All-encompassing reliability statistics for the first release of the TDB are calculated including descriptive statistics, inter-annotator agreement, gold standard agreement and intra-annotator agreement using Fleiss' Kappa measure, as well as additional measures of precision, recall and f-score. A computer software code to discretize all the annotations including discontinuous spans for overlapping and non-overlapping relations to calculate Fleiss' Kappa measure has been developed. The novel feature of this software is to include discontinuous spans, as well as include non-overlapping relations in the discretization. A discussion of the statistics results is provided and sources of disagreements between annotators is investigated. Re-annotations on a part of the corpus is done by the author and intra-annotator agreements for these annotations with previous annotations of the author are provided. The implications for the results of the proposed additional evaluation metrics are discussed, along with provided benefits of calculating these measures.
- A preliminary model to resolve Turkish demonstrative pronouns is developed. We apply known techniques building a feature set with Treebank information as a first effort in Turkish to resolve demonstrative pronoun reference. Within the context of this work, a 20K subpart of the TDB have been annotated for demonstrative pronoun reference (including explicit forms of the third person singular pronoun o due to its homonymy with the demonstrative *o* 'that'), resolving bare demonstrative and demonstrative +NP uses, as well as identifying the antecedents as abstract, concrete, or exophoric objects. The annotations have been double checked by two other annotators and exact match inter-annotator agreement has been provided. Our analysis on the distribution of demonstrative anaphora in the TDB showed that demonstrative + NP uses favor reference to concrete objects. This category also included phrasal expressions, as these expressions harbor demonstrative pronouns as deictic elements. Some phrasal expressions can be ambiguous in that they can also refer to concrete objects behaving non-discursively. However, our annotations of phrasal expressions in the TDB show that these forms more frequently make abstract object references. It is observed that there are other

forms in the demonstrative + NP category referencing abstract objects, which require further study. Nevertheless, a conclusion that can be drawn is that phrasal expressions do not act like the general category of demonstrative + NP uses. However, the know-how obtained from our experiment to resolve demonstrative pronouns in Turkish were successfully applied to the automatic identification of discursive uses of phrasal expressions. These include the importance of positive-negative instance balance, possibility of using filters to narrow down the search space, as well as aid in the instance balancing, and ideas in interpreting the results of the system.

- A cascaded model to automatically disambiguate discursive uses of phrasal expressions in Turkish from free text is developed. The model uses the features extracted from the morphologically, syntactically and dependency parsed TDB for the disambiguation process. Morphological, syntactic and dependency parsing of the TDB is provided as a by-product. Our cascaded model achieves full recall, high precision and f-score providing high accuracy. This model can be utilized to effortlessly enhance the coverage of the TDB by applying it to the other subcorpora of the MTC and manually adjudicating the predictions of the system.
- We have also shown that lexical word form is the main requirement for identifying the discourse uses of phrasal expressions in Turkish, as the best features in our models are found to be the word form and head word form. Hence, our observations of words enable us to reach discourse-level structures.
- An attempt to develop a preliminary model to extract discursive uses of explicit Turkish discourse connectives is made.

1.4 Outline

This thesis is organized as follows: In chapter 2, an introduction to discourse structure, cohesion, coherence and discourse relations will be given. The means to achieve coherence through intra- and inter-sentential connectives will be briefly described, followed by a brief overview of some prevalent discourse structure theories. Special attention will be given to discourse connectives, as a central focus item of this thesis.

In chapter 3, an overview of corpus studies on discourse will be presented and the Turkish Discourse Bank will be introduced.

In chapter 4, a general introduction to the importance of reliability studies on linguistic resources will be presented, followed by an assessment of the TDB using several evaluation methods.

In chapters 5-6, a set of additional annotations and some language technology applications built on the TDB will be presented. In chapter 5, we describe a small-scale annotation effort to annotate demonstrative pronominal reference in Turkish, followed by an investigation of techniques to automatically resolve demonstrative pronouns in Turkish, where an overview of previous work on this topic will also be presented.

In chapter 6, we will look into the phrasal expressions annotated as part of the TDB. The decision to annotate phrasal expressions as part of explicit discourse connectives in the TDB will be discussed in comparison to the approaches in other comparable corpora. Then, before developing a cascaded model to automatically identify these phrasal expressions for Turkish, other studies which develop similar applications will be briefly overviewed. Finally, benefits and future enhancements of the model will be discussed. In the last section of chapter 6, a brief look at a possible future application and an initial attempt will be presented to automatically discover discourse connectives from unannotated text. Possible future enhancements and benefits will be discussed.

In chapter 7, we conclude with an overview of the thesis, where the contributions of the thesis will be outlined, followed by some of the limitations of the current work and implications for future work.

CHAPTER 2

DISCOURSE STRUCTURE, DISCOURSE RELATIONS AND DISCOURSE CONNECTIVES

> *"Since that which is compounded out of something so that the whole is one, not like a heap but like a syllable - now the syllable is not its elements, ba is not the same as b and a, nor is flesh fire and earth (for when these are separated the wholes, i.e. the flesh and the syllable, no longer exist, but the elements of the syllable exist, and so do fire and earth); the syllable, then, is something - not only its elements (the vowel and the consonant) but also something else, and the flesh is not only fire and earth or the hot and the cold, but also something else: -if, then, that something must itself be either an element or composed of elements. . . "*
>
> *– Aristotle, Metaphysics, 350 BC.*
> *Book 7, Chapter 17*
> *(translated by W.D. Ross)*

Just as Aristotle's quote states above, our subject matter, discourse can be viewed in one respect as more than the sum of its parts.[1] It is more than merely a sequence of sentences. Thus, it conveys a wholesome understanding of the text by adding something to this mere sentence group. What is this something? The thing that makes a sentence group a text is the relationships between them providing them to function as a single meaningful unit, hence providing *coherence*, which distinguishes the text from some incomprehensible 'non-text' (Stede, 2012). If the linguistic means of achieving this coherence are verbally explicit (e.g. reference, substitution, ellipsis, conjunction, lexical cohesion), then there is *cohesion* (Halliday & Hasan, 1976; Stede, 2012). Hence, *coherence* involves a 'deep understanding' and interpretation of text and *cohesion* is 'identified at the text surface' as stated by Stede (2012). Thus, in a broad sense the relations between sentences (or any eventualities in clauses, or above sentence structures) that link them in a way to make them a whole, are *coherence relations* (Stede, 2012), also called *discourse relations* (Webber, Egg, & Kordoni, 2011), or *rhetorical relations* (as used to refer to these relations in the framework of Rhetorical Structure Theory of Mann & Thompson, 1988). Discourse relations, as we will be referring to them in this

[1] Here, it should be noted that we do not quote Aristotle as a statement against modularity (Fodor, 1983), but to put emphasis on that quality which differentiates a text from non-text, which is explained in what follows. Although it may not always be explicitly signalled, this quality may as well form a part in a modular discourse structure system. However, a discussion on modularity of discourse is out of the scope of this thesis. Graesser et al. (1997) provides a brief account of modularity theory and levels of discourse.

thesis, can be made explicit by discourse connectives as in (4-6)[2], where the connectives have been underlined.

(4) During daylight, the sky appears to be blue because air scatters blue sunlight more than it scatters red.

(5) During the day, the Sun can be seen in the sky unless obscured by clouds, whereas in the night sky, the moon, planets and stars are visible in the sky.

(6) Some of the natural phenomena seen in the sky are clouds, rainbows, and aurorae. In addition, lightning and precipitation can also be seen in the sky during storms.

It can be seen that discourse connectives can link both clauses within a sentence, and separate sentences to each other. In fact, they can even relate nominalizations, or larger discourse units consisting of several sentences. These discourse connectives are the main focus of the Turkish Discourse Bank (TDB) and thus this thesis. However, before going into the details of discourse connectives, let us understand the general concepts in discourse structure and the means they are made explicit in text to provide cohesion.

In what follows an overview of discourse structure and its building blocks will be presented. Then the means of achieving cohesion and coherence in text will be explained within the framework of Halliday and Hasan (1976), followed by a brief overview of some prevalent discourse structure theories. Finally, special attention will be put on discourse connectives and their types.

2.1 Discourse Structure

When we read a given text, we can find certain patterns in the way it is composed, such as a central idea of focus in parts of the text, or the whole, known as a *topic*. A group of sentences or a paragraph may be centered on one topic, while other paragraph groups may be centered on another topic, yielding perhaps a unifying topic for the text. *Topics* are discourse structures consisting of a set of entities and a restricted variety of things being said about them (Webber et al., 2011). Within a topic, the set of entities form *entity chains* referring to the same entity.

The class of texts that the given text belongs to (i.e. provides the same common communicative purpose (Stede, 2012)) is called the *genre* of the text, e.g. news, novel, academic writing, recipe, etc.

Elements of discourse have a *function* in terms of the part they play in the communication. Hence, discourse structures can be grouped by the functions they serve, for example segments in research paper abstracts providing the results of research – *Results* section, or providing information about methodology – *Methods* section, etc. (Webber et al., 2011). These segments of a given text with a particular function are sometimes called *content zones* (Stede, 2012).

Eventualities (events and states) and their spatiotemporal relations can also be used to provide structure to a given text (Webber et al., 2011). A typical example is found in narratives, which are by definition 'report of connected events presented in a sequence'[3], where the structure

[2] The content of the examples are taken from http://en.wikipedia.org/wiki/Sky.

[3] Definition from http://en.wikipedia.org/wiki /Narrative

can be divided into *exposition* (i.e. background information), *complication* (i.e. development of the story, the events) and *resolution* (i.e. ending of the story).

All of these patterns (i.e. topic, function, and eventualities) provide large discourse units (Stede, 2012). However, discourse can also be examined in small discourse units, which can be achieved by *discourse relations*. These are the relations that exist between the semantic content of two discourse units as stated by Webber, et al. (2011). Here, the semantic content can be a proposition, a fact, an event or state, i.e. an *Abstract Object* (AO) as classified by Asher (1993)[4] The discourse units can be clauses, groups of clauses, sentences, groups of sentences, or nominalizations, which express the semantic content. In RST (Mann & Thompson, 1988) the smallest discourse unit is called the *elementary discourse unit* (EDU) (Carlson, Marcu, & Okurowski, 2001; Stede, 2012). Hence, discourse relations can link EDUs, or larger units formed by these linked EDU pairs.

Discourse relations can be expressed explicitly (as in examples (4)-(6)) through the use of *discourse connectives*, where the connective can be viewed as the predicate of the two arguments (i.e. AOs put across by the discourse units) as in the framework of the PDTB (Prasad et al., 2008). They may also be left *implicit*, without an overt expression. Consider (7)[5] as a continuation of (4) above, repeated here for convenience. A contrastive relation can be found between these two sentences, but it is left implicit, whereas it could have been made explicit easily with the use of a discourse connective such as *in contrast* as portrayed in (8). Thus, such unstated discourse connectives are referred to as *implicit discourse connectives* by Prasad et. al (2008).

(4) During daylight, the sky appears to be blue because air scatters blue sunlight more than it scatters red.

(7) At night, the sky appears to be a mostly dark surface or region scattered with stars.

(8) During daylight, the sky appears to be blue because air scatters blue sunlight more than it scatters red. In contrast, at night, the sky appears to be a mostly dark surface or region scattered with stars.

A more detailed account of discourse connectives will be provided in Section 2.3. But, now having covered some basic notions of discourse structure, we first describe the means to achieve cohesion in text in the next section.

2.2 Cohesion

At the beginning of this chapter, we mentioned that a text is more than the sum of its sentences. There is something enabling a segment of speech or writing to become a text. We call it *coherence*, as previously stated. *Cohesion* exist when that something providing coherence in text is made verbally explicit (Halliday and Hasan, 1976), i.e. on the 'text surface' as expressed by Stede (2012).

Hence, cohesion can be tangibly observed in text through several linguistic means such as repetitions, omissions, patterns of occurrence between certain words. Halliday and Hasan

[4] A brief explanation of Asher's AO classification is provided in Chapter 5, Section 1 of this thesis.

[5] Also from http://en.wikipedia.org /wiki/Sky

(1976) classifies the concept of cohesion into five categories: 1.) reference, 2.) substitution, 3.) ellipsis, 4.) conjunction, 5.) lexical cohesion. We will briefly explain all.

2.2.1 Reference

Items in the language which refer to something else either within or outside the text for their interpretation are known as *reference* items. These are personals, demonstratives and comparatives in English. The information to be retrieved that is signaled by the reference items is the referential meaning, which provides a continuity of reference by enabling the item referenced to enter the discourse one more time. This in return provides cohesion.

(9)

(a) I like *nonsense*, it wakes up the brain cells.

(b) *Fantasy* is a necessary ingredient in living, it's a way of *looking at life through the wrong end of a telescope*.

(c) . . . Which is what I do, and *that* enables you to laugh at life's realities."

-Dr. Seuss

The quote above from Dr. Seuss makes use of reference as a cohesive agent, where in (a) *it* refers to *nonsense*, in (b) *it* refers to *fantasy* and in (c) *that* refers to *what I do*, which then refers to *looking at life through the wrong end of a telescope*.

There are two forms of reference: *exophoric*, which is referring to something situational and *endophoric*, which is referring to something textual. In example (10) below, it is possible that *that* is referring to some book mentioned previously in the text. However, it is also as possible that the speaker is simply pointing to the book in question in the environment where the utterance takes place. In the former case (e.g. if (11) preceded (10)), if the referent of that can be identified within the text somewhere, then there is *endophora* or *endophoric reference*, otherwise, if the referent is a particular book in the environment of the utterance that is being pointed to (i.e. it is situational), then there is *exophora* or *exophoric reference*.

(10) Did you read that book?

(11) I'm leaving the book by Dr. Seuss on the table.

Endophora is further divided according to the position of the referent. If the referent is in the preceding text, it is known as *anaphora* (as in the case of (11) preceding (10)). Otherwise, if the referent is found in the text that follows, it is known as *cataphora*, as in (12), where *she* refers to Alice.

(12) She was happy to have found a friend. Alice held her new friend’s hand and started playing.

Aside from the exophoric-endophoric distinction, reference is categorized into three types with respect to the reference items, as previously mentioned. *Personal reference* involves personal pronouns (e.g. he, she, it, we, etc.), possessive determiners (e.g. my, your, etc.) and possessive pronouns (e.g. mine, yours, etc.), where the referents are persons or objects. *Demonstrative reference* involves ’identifying the referent by locating it on a scale of proximity’ (Halliday & Hasan, 1976) as near (e.g. this) or far (e.g. that). A special case of the demonstrative reference is *extended reference* or reference to fact.

(13) Alice gave the book to her friend. That was a present from her father.

(14) Alice gave the book to her friend. That was nice.

In (13), the demonstrative *that* refers to *the book*, whereas in (14) it refers to *Alice giving the book to her friend*, Note that extended reference is only possible using the singular forms (i.e. *this* and *that*) without a following noun. Additionally, while *this* may be either anaphoric or cataphoric, *that* can only be anaphoric in extended reference.

Finally, *comparative reference* involves, as its name implies, reference by comparison (e.g. *identical* in *identical twins*, or *greater* in *greater responsibility*).

2.2.2 Substitution

Substitution is a cohesive relation on the grammatical level, i.e. it is relating words or phrases; whereas reference is on the semantic level relating meanings. Hence, instead of repeating a particular item in text, a substitute with the same structural function can be used as in (15) below.

(15) This plate is not clean, could I get a new one?

(16) She did not think of him as much as she used to do.

(17) Should I sign it? – It says so in the guidelines.

In this example, *one* is substituted for *plate*, where both have the grammatical function of Head in the nominal group. Substitution is thus divided into three types: *nominal* (uses *one, ones* or *same*) as in (15), *verbal* (uses *do*) as in (16), and *clausal* (uses *so* and *not*) as in (17).

2.2.3 Ellipsis

Ellipsis can be considered as substitution by zero, where something is left out of the text, i.e. it is not said but understood. The action of Alice in (18) is left unsaid, however it is identified easily as *drank* via ellipsis. Similarly in (19), although *what he had two in a row of* is left out in the second clause, it is understood to be *two cups of coffee*. Similar to substitution, ellipsis also has the three types *nominal*, *verbal* and *clausal*.

(18) The cat drank the milk and Alice the lemonade.

(19) I had one cup of coffee, he had two in a row.

2.2.4 Conjunction

Conjunctive elements are different from the other types of cohesive relations in that they are not cohesive themselves, but they express meanings which presuppose other components in text (Halliday & Hasan, 1976). Conjunction thus connects subsequent elements in the text to each other.

(20) She finished her writing. Subsequently, the sky cleared up.

In (20) the relation of time sequence is the only explicit form of connection between the event of her finishing her writing and the sky clearing up. This semantic relation of time sequence is the conjunction providing cohesion in this example. Here, the adverbial adjunct *subsequently* is making this conjunctive relation explicit. Hence, it is called a *conjunctive adjunct* or *discourse adjunct*.

(21) She finished her writing. Following that, the sky cleared up.

In (21), the reference item *that* relates the second sentence to the first one, thus, providing cohesion. The adverb *following* becomes cohesive through its structural relationship with *that* in this example. However, since such adjunct can also act cohesively on their own, Halliday and Hasan (1976) consider such prepositions + reference item constructions as a subtype of conjunctives. Furthermore, some current adverbs (e.g. therefore) have their linguistic origins in this kind of constructions, suggest that the whole phrase acts as a cohesive agent, rather than just the demonstrative. What's more, many adverbs/prepositional phrases which act as a conjunctive adjunct on their own (e.g. as a result), also act as a conjunctive as a prepositional phrase combined with a preposition (i.e. *of*) and a demonstrative (i.e. *this/that*) (e.g. as a result of this).

Therefore, in Halliday and Hasan's (1976) classification, conjunctive adjuncts consist of: 1.) adverbs (simple adverbs, e.g. *but, so, then*; compound adverbs in *–ly*, e.g. *subsequently, actually*; compound adverbs in *there-* and *where-*, e.g. *therefore*, *whereat*) 2.) other compound adverbs, e.g. *anyway*, *instead*, *furthermore*; prepositional phrases, e.g. *as a result*, *in addition* 3.) prepositional expressions with *that* or other deictic reference item (either optional or obligatory), e.g. *as a result of that, because of that*.

2.2.5 Lexical Cohesion

Lexical cohesion is the cohesive force introduced into the text by the relationships between the lexical items, either by *reiteration* of the same item or use of *collocations*. The reiteration may be achieved by repetition (22a), a synonym (or near-synonym) (22b), a super-ordinate (22c) or a general word (22d), where usually it is used together with a reference item such as *the*.

(22) There's a girl walking to the playground.

> **(a)** The girl is going to ride on a swing.
>
> **(b)** The gal is going to ride on a swing.
>
> **(c)** The kid is going to ride on a swing.
>
> **(d)** The cutie is going to ride on a swing.

Collocation, on the other hand, achieves cohesion through the relationship between lexical items that regularly co-occur. Moreover, long cohesive chains can be formed using collocations, e.g. *sky... sunshine... cloud... rain* (as in 23).

(23)

> **(a)** The sky was clear.
>
> **(b)** There was bright sunshine.
>
> **(c)** Suddenly a dark cloud appeared.
>
> **(d)** Rain fell down.

2.3 Prevalent Theories of Discourse Structure

In order to gain a basic understanding of where discourse relations and discourse connectives stand in the theoretical scene, we would like to give a brief account of some well known theories of discourse structure before going into the details of discourse connectives in the next section, These include the theory of Hobbs (1985), the Rhetorical Structure Theory of Mann and Thompson (1988), Grosz and Sidner's (1986) theory, Segmented Discourse Representation Theory of Lascarides and Asher (2007), and finally, Lexicalized Tree-Adjoining Grammar for Discourse of Webber (2004).

2.3.1 Hobb's Theory

Hobbs (1985) embeds a theory of coherence relations within a knowledge-based theory of discourse interpretation. Hobbs defines the minimal unit of discourse in written text as clauses or sentences and in spoken discourse as phrasal or smaller elements. Then, he identifies the source of discourse structure as the adjacency of two phrases, clauses, sentences, or larger stretches of discourse, which can be explained with coherence relations found between them. This theory integrates syntax, semantics and pragmatics by describing interpretation using abduction and parsing using deduction. Abduction is a method of reasoning where from an observable Q and a general principle explaining that observable such as P⊃Q, the underlying reason for the truth of Q is assumed to be P, whereas deduction is the process of concluding Q from P and P⊃Q. Discourse interpretation is explained in six subtheories: 1.) logical notation or knowledge representation, 2.) syntax and semantic translation of text to the logical notation, 3.) knowledge encoding (i.e. of world and language knowledge required to understand texts as what is called a *knowledge base*), 4.) deductive mechanism (i.e. to manipulate the knowledge stored as axioms in a logical notation), 5.) discourse operations or specification of possible interpretations (to constrain the deductive mechanism), and 6.) specification of the best interpretation.

In Hobbs' theory, coherence relations are divided into four classes: 1.) *occasion relation* denoting a change of state between the two discourse segments (e.g. cause and enablement relations), 2.) *evaluation relation* where it can be inferred from one of the discourse segments that the other segment is a plan towards a discourse goal, 3.) relations directed towards relating a segment of discourse to the listener's prior knowledge (e.g. background and explanation relations), and 4.) *expansion relations* which expand the discourse in place, rather than carrying it forward or filling in the background (e.g. parallel, generalization, exemplification, contrast and violated expectation relations).

According to Hobbs, the smallest unit of discourse is a clause. Discourse structure is built up recursively starting from a clause as a segment of discourse and linking two segments by a coherence relation to constitute a larger segment of discourse. Hobbs (1993) axiomatizes a tree-like structure of discourse, where a sentence describing an eventuality is a coherent discourse segment describing this eventuality and if two segments describe two eventualities which are related by some coherence relation, then the concatenation of these two segments is a coherent discourse segment.

2.3.2 Rhetorical Structure Theory (RST)

The Rhetorical Structure Theory of Mann and Thompson (1988) is a decriptive framework for discourse structure, identifying a functional hierarchical structure in text, describing relations between discourse segments in terms of their communicative role. The unit of discourse in RST is the non-overlapping text spans called *elementary discourse units* (edus). A rhetorical *relation* holds between two (or more) edus; where either one is the more salient *nucleus*, the other is the *satellite*, or all are equally important nuclei. A relation defines the constraints on the nucleus, the satellite and the combination of nucleus and the satellite, as well as the effect. The structure of a text are determined in this theory by an analyst, distinct from the writer and the readers of the text, through plausibility judgments.

The theory defines a small number of abstract patterns called *schemas*, which specify the relations between the constituent text spans and the identification of the nuclei as a discourse tree. Five basic schema types are defined as exemplified in Figure 2.1 below taken from Mann and Thompson (1988, p. 247, Fig.1), where curved lines define relations and straight lines identify nuclei. All text can be analyzed using these five schemas, with some variations defined by *schema applications*. According to these, the order of spans can change, individual relations in multi-relation schemas are optional where at least one must hold, and relations in a schema can be repeated.

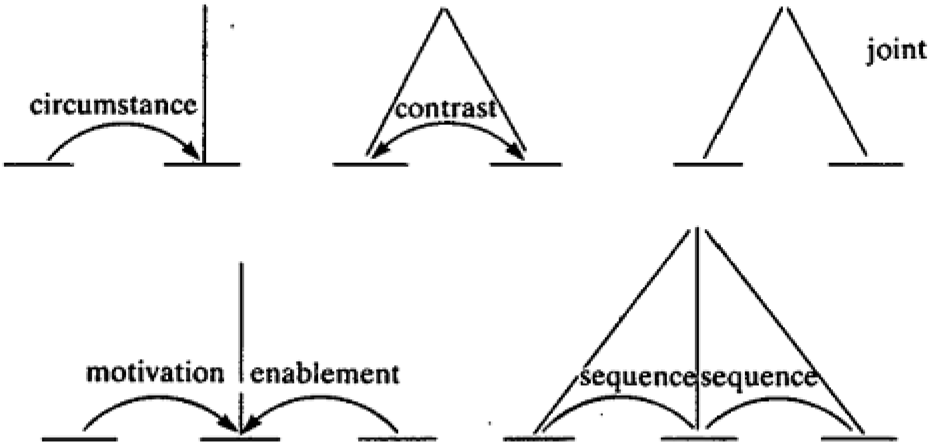


Figure 2.1: RST schema type examples (Mann and Thompson, 1988, p. 247)

RST defines four constraints on the schema applications: completedness, connectedness, uniqueness and adjacency. Completedness ensures that one schema application is applied to text spans constituting the whole text. Connectedness enables recursive application, where each text span is either a minimal unit or a constituent of another schema application. Uniqueness states that the text spans of schema applications do not overlap. Adjacency constrains the text spans of schema applications to be adjacent, providing one text span as a constituent of the schema application.

There are a predefined set of relations (but open to modification and extension) in the RST, which are defined independent of morphology or syntax, but are defined functionally and semantically. Hence, in RST relations are not lexically signalled, implying that there is no concept of cue phrases or discource connectives.

RST has been employed in the development of the RST Treebank described in Section 3.2.

2.3.3 Grosz and Sidner's Theory

Grosz and Sidner (1986) integrate the nonlinguistic concepts of intention and attention into their theory of discourse structure. Hence, they describe discourse structure as composed of three interrelated components: linguistic structure, intentional structure and attentional state. The linguistic structure is defined as the structure of the sequence of utterances, the intentional structure defines a structure of discourse-relevant purposes expressed in the linguistic utterances and their relationships, and the attentional state is the state of focus of attention of the participants. These components are used to explain the differentiations between discourse phenomena such as cue phrases, referring expressions and interruptions.

The basic elements of linguistic structure are utterances in this theory and they are said to be naturally aggregated into discourse segments. Utterances need not be consecutive to be in the same discourse segment. The linguistic structure consists of discourse segments and the embedding relation between them, which are surface reflections of relations between intentional structure elements. There is a two-way interaction between the linguistic structure elements and the discourse structure, where utterances give information about the discourse structure and the discourse structure constrains their interpretation. Discourse segment boundaries are mainly indicated by linguistic expressions such as certain words or phrases, as well as intonation, tense and aspect changes. Grosz and Sidner refer to these special words or phrases as *cue phrases* and classify these markers according to the chages they explicitly indicate, either in intentional structure or attetional state. Cue phrases are at the discourse-level and do not contribute to the sentence-level properties such as truth conditions of the sentences.

In the intentional structure, a *discourse purpose* (DP) is the underlying intention of engaging in the particular discourse, which provides the reason for the linguistic act and the reason of the chosen content of the discourse. The intention of each discourse segment is called *discourse segment purpose* (DSP) specifying how the segment contributes to the overall discourse purpose. Two structural relations that relate the DSPs or DPs are identified as dominance and satisfaction-precedence. Furthermore, the intentions providing discourse purpose are from an open-ended range.

The attentional state is defined as an abstraction of the participant's focus of attention as a property of the discourse, which dynamically records salient objects, properties and relations in discourse. Grosz and Sidner model this attentional state using a set of *focus spaces* containing the salient entities including the DSP, and changes in attentional state using a set of transition rules for adding/deleting these spaces. The process of manipulating spaces is called *focusing* and the collection of focus spaces at any one time is called *focus structure*, where the focus process relates each focus space with a discourse segment. This focus structure is modeled as a stack acting as a central repository for contextual information needed to process the utterances. The pushes and the pops of this stack are determined by the relationships between DSPs. Grosz and Sidner also distinguish the attentional state from the cognitive state, where the former is merely a component of the latter.

Grosz and Sidner's theory define the function of cue phrases as providing abbreviated, indirect means of indicating changes in intentional structure and attentional state. The information that needs to be provided is determined to be the approaching of change of attention, the return to a previous focus space or the creation of a new focus space, the relationship of the intention to other intentions, the relevant precedence relationships and the intention entering focus. Cue phrases are said to provide all this information but the newly focused intention.

2.3.4 Segmented Discourse Representation Theory (SDRT)

Segmented Discourse Representation Theory (Lascarides & Asher, 2007) is a dynamic semantic theory of discourse interpretation. It provides a logic of information content for interpreting logical forms of discourse and a *glue logic* for constructing these logical forms. SDRT of Lascarides and Asher (2007) extends Discourse Representation Theory (DRT) of Kamp and Reyle (1993) by employing rhetorical relations.

In dynamic semantics, a sentence S is interpreted as a relation between an input and output contexts, consisting of variable assignment functions. In DRT, logical forms are called discourse representation structures (DRSs), which are pairs of a set of discourse referents and a set of DRS-conditions. SDRT adds *speech act discourse referents*, which label content and keep track of the token utterances in discourse, and *rhetorical relations*, which relate speech act discourse referents, to DRT, calling the resulting structures segmented DRSs.

2.3.5 Lexicalized Tree Adjoining Grammar for Discourse (D-LTAG)

In a Lexicalized Tree-Adjoining Grammar (LTAG) each word is associated with a tree set, where the tree structures represent one of each minimal syntactic constructions that the word can appear in. These tree structures can be either *initial trees* or *auxiliary trees* allowing trees to be modified by *substitution* or *adjunction* operations and hence introducing recursion.

D-LTAG of Webber (2004) basically defines a new LTAG for discourse. In D-LTAG predicates on clausal arguments define the domain of locality and hence, are associated with initial trees. These predicates on clausal arguments are discourse connectives, i.e. subordinate conjunctions and other subordinators, anchors of parallel constructions, coordinate conjunctions, and also some specific verb forms. For example, subordinate conjunctions anchor initial trees, where clauses substitute as arguments. D-LTAG has taken two auxiliary trees, one anchored by an explicit coordinating conjunctions or null connective, and the other anchored by a discourse adverbial. The null connective is used to express implicit discourse relations. The arguments of discourse connectives are taken as text spans denoting abstract objects of Asher (1993).

The PDTB, which will be described in Section 3.1, stems from D-LTAG's main idea of a lexical basis for discourse.

2.4 Discourse Connectives

Words, expressions, or sometimes certain morphological suffixes that signal discourse relations in a given text are called *discourse connectives*. If there is an explicit signal denoting that two discourse units are related to each other, then this overt signal is referred to as an *explicit discourse connective*. Contrastively, if the signal is implicit and there is no overt expression present, then the relation is an *implicit relation* and the connective possibly associated with this relation has been called an *implicit discourse connective* by Prasad et al. (2008). Examples of explicit and implicit uses of connectives have been provided above, at the beginning of this chapter.

Explicit discourse connectives for English have been identified to be from 4 different well-defined syntactic classes within the PDTB framework (Prasad et al. 2014): 1.) Subordinating Conjunctions, 2.) Coordinating Conjunctions, 3.) Prepositional Phrases, and 4.) Adverbs. The first class includes connectives such as *because, although, when, if*, etc., while coordinating conjunctions include *and, but, nor, either... or, neither... nor* and *so*. Prepositional phrases used as explicit discourse connectives include *as a result, on the one hand... on the other hand*, etc. Finally, discursive adverbs consist of *then, however, instead*, etc. In the LADTB (Al-Saif & Markert, 2010, 2011), clitics and prepositions called *Al-Masdar* have also been annotated as explicit discourse connectives for Arabic. The *Al-Masdar* consists of forms that span several grammatical and morphological categories in English (i.e. gerund, nominalization, noun which is not a nominalization).

Implicit connectives were identified in the PDTB between paragraph internal adjacent sentences, if there was no explicit connective present, however, there was a coherence relation between the sentences where one or more connectives could be placed without causing redundancy (Prasad et al., 2014). In the HDRB (Kolachina, Prasad, Sharma, & Joshi, 2012; Oza, Prasad, Kolachina, Sharma, & Joshi, 2009; Sharma, Dakwale, Sharma, Prasad, & Joshi, 2013), implicit connectives have also been identified across paragraph boundaries for Hindi.[6]

Expressions outside of these predefined terms were considered as *alternative lexicalizations*, if they caused redundancy in an attempt to add an implicit connective, hence signaling a discourse relation (Prasad, Joshi, & Webber, 2010). In the CDTB (Y. Zhou & Xue, 2012, 2014) a small number of verbs were identified as alternative lexicalizations for Chinese. Phrasal expressions containing a deictic item which signal discourse relations were treated in the PDTB as alternative lexicalization, although they were not systematically annotated due to limited resources (Prasad et al., 2014).

(24) John decided to live in Istanbul, but he nonetheless refused to learn Turkish.

(Stede, 2012: p.98, ex. 4.22)

Stede (2012), on the other hand, brings forward a slightly different classification of connectives, where he identifies connectives as belonging to the closed-class lexical items, which are not easily invented and are not inflected. In this definition, a connective syntactically can be a subordinating or coordinating conjunction, an adverbial, as in the PDTB classification; however, it can arguably be a preposition, but not a prepositional phrase. He identifies phrases lexically signaling coherence relations that are more open to lexical modification or extension as *cue phrases*, such as *for this reason, for all these reasons* (Stede, 2012). In this classification, connectives consisting of multiple tokens are called *complex connectives*, as in (24) where the two non-adjacent connectives but and *nonetheless* signal the coherence relation.

In Rysová and Rysová (2014), discourse connectives are classified into two main categories: *primary connectives* and *secondary connectives*. Primary connectives are defined as expressions from selected parts of speech (i.e. conjunctions and some types of particles), which primarily function to connect two units of text. Secondary connectives, on the other hand, are multiword expressions such as *that is the reason why, from these reasons, because of this*, which do not belong to the generally accepted POS for connectives (i.e. conjunctions and particles), but clearly signal discourse relations (Rysová & Rysová, 2015). They are not lexically frozen, not grammaticalized, but they can be inflected (e.g. *from this reason – from these reasons*) and have certain different forms (e.g. *due to this, due to this fact, due to this situation*). However, only non-context dependent forms of these secondary expressions are referred to as *universal* secondary connectives (e.g. *because of this*), are included in the discourse connective classification. The other expressions which can only be used in limited contexts (e.g. *because of this increase*) are not included in the discourse connectives and are referred to as *non-universal connecting phrases*.

[6] Note that the decision to limit the implicit connective annotation to paragraph internal sentences does not assert that there are no implicit connectives across paragraph boundaries, just that they have not been annotated for English in the PDTB.

2.5 Turkish Discourse Connectives

Discourse connectives in Turkish can be spotted by analyzing three main syntactic classes (Zeyrek & Webber, 2008; Zeyrek et al., 2009, 2010, 2013):

1.) Coordinating conjunctions:

a. simple coordinators, e.g. *ve* ’and’, *ama* ’but’, *ya da* ’or,’,

b. paired conjunctions, e.g. *hem... hem* ’both..and’, *ne... ne* ’neither... nor’;

2.) Subordinators:

a. simplex subordinators (also called *converbs*), e.g. *-(y)ArAk* ’by means of’, *-IncA* ’when’, *-(y)ken* ’while/now that’, *-DIkçA* ’as’, *-mAksIzIn* ’without’[7],

b. complex subordinators, e.g. *için* ’for’, *rağmen* ’despite’, *kadar* temporal ’until’, *beri* causal ’since’;

3.) Anaphoric connectives:

a. discourse adverbials, e.g. *oysa* ’however’, *öte yandan* ’on the other hand’, *ayrıca* ’in addition/separately’, *aksine* ’on the contrary’, *ne var ki* ’however’,

b. phrasal expressions, e.g. *buna rağmen* ’despite this’, *onun için* ’for this/that’, *bu nedenlerle* ’for these reasons’,

Simple coordinators are single lexical items coordinating two clauses, or sentences, whereas paired conjunctions link two clauses, where one element of the pair is associated with each clause in the discourse relation. Simplex subordinators are suffixes attached to the verb roots forming non-finite adverbial clauses. For example in (25) *-(y)ArAk* ’by means of’ is attached to the nominalized adverbial clause *onayla* ’approve’. Complex subordinators involve two parts: a postposition and a suffix on the subordinate verb, assigning case, e.g. *-nA rağmen* ’despite’ (Zeyrek, Çakıcı, Sevdik Çallı, & Demirşahin, 2015) (see example (26)).

(25) Hükümet ... uyum paketini onaylayarak ... Erdoğan’ın önündeki engellerden birini kaldırdı.

By approving the harmonization package ..., the government alleviated one of the obstacles for Erdoğan ....

(Zeyrek et al., 2009: p. 45, ex.1)

(26) Gerçeği bilmesi-ne rağmen sustu.

Despite knowing the truth, she kept quiet.

(Zeyrek et al., 2015)

[7] The capital letters denote allomorphy, where A: a/e, I: i/ı/u/ü, D: d/t.

Discourse adverbials are considered to be anaphoric connectives which can access the inferences in prior discourse (Webber, Stone, Joshi, & Knott, 2003) as in (27) below, where the discourse adverbial *yoksa* ’or else, otherwise’ accesses the inference that the organizations have not united and hence did not introduce political strategies unique to Turkey.

(27) Bu örgütlerin birleşerek Türkiye’yi etkilemesi ve Türkiye’ye özgü politikaları gündeme getirmesi lazım. Yoksa Tony Blair söyle yaptı simdi biz de simdi böyle yapacağızla olmaz.

These organizations must have an impact on Turkey by uniting and introduce political strategies unique to Turkey. Or else saying Tony Blair did this and now let’s do that is outright wrong.

(Zeyrek and Webber, 2008: p. 69, ex. 14)

Similarly, in this classification we include phrasal expressions that have a deictic demonstrative counterpart combined with a subordinating conjunction, e.g. *buna rağmen* ’despite this’ as in (28).

(28) Ahmet nezle olmuş. Buna rağmen yüzmek istiyor.

Ahmet’s got a cold. Despite this, he wants to go swimming.

(Göksel & Kerslake, 2005: p.520, ex. 58)

Major Turkish grammars such as Banguoğlu (1990) and Göksel and Kerslake (2005) also categorize such phrasal expressions as a sub-type of discourse connectives in Turkish. Moreover, Halliday and Hasan (1976, p. 231) include such expressions referring to them as "prepositional expressions with *that* or other reference item", where the reference item should be functioning as deictic for the expression to be conjunctive.

In line with this classification, phrasal expressions with deictic elements such as *bu nedenle* (for this reason), *onun için* (for that), or *bunların sonucunda* (as a consequence of these) have been systematically annotated in the TDB. However, for example, in the PDTB they have either been annotated as alternative lexicalizations (Prasad et al., 2010), or the deictic item has been selected as one of the arguments related by an explicit connective (Miltsakaki et al., 2004a). In the Potsdam Commentary Corpus, phrasal expressions were annotated as part of complex connectives (Stede & Heintze, 2004). In the extended annotations of the PDiT 1.0, these expressions were annotated as universal secondary connectives (Rysová & Rysová, 2014). Phrasal expressions will be further discussed in more detail in Section 6.1.

In this chapter we presented an overview of discourse structure and discourse relations, introducing some basic concepts. We also provided a brief overview of well known theories of discourse structure. The concept of cohesion was classified in five categories: reference, substitution, ellipsis, conjunction and lexical cohesion as described by Halliday and Hasan (1976). Our main concern has been focused on *conjunction* in this classification, as a means of achieving coherence and hence cohesion through the use of explicit discourse connectives. Then, we explained how different studies define discourse connectives for languages other than Turkish and how discourse connectives are identified in Turkish. Finally, we introduced phrasal expressions as a kind of discourse relational devices, which will be subject of Chapter 6.

CHAPTER 3

CORPUS STUDIES ON DISCOURSE AND THE TURKISH DISCOURSE BANK

Since the 1990's an increasing amount of work has been done on building linguistic corpora (e.g. the Penn Treebank: Marcus, Santorini, & Marcinkiewicz, 1993; the British National Corpus: Leech, Garside, & Bryant, 1994) and later from the 2000's onwards there have been many efforts to add annotation to these resources with some form of discourse structure. Some of the earlier work for discourse annotated corpora include the RST Discourse Treebank (Carlson et al., 2001) and the Penn Discourse Treebank (Prasad et al., 2008), which is perhaps one of the most popular. Recent additions to the discourse annotated corpora are mainly studies done in languages other than English. These include the Potsdam Commentary Corpus (PCC) (Stede & Neumann, 2014; Stede, 2004) for German, the Hindi Discourse Relational Bank (HDRB) (Kolachina et al., 2012; Oza et al., 2009; Sharma et al., 2013) for Hindi, Leeds Arabic Discourse Treebank (LADTB) (Al-Saif & Markert, 2010, 2011) for Modern Standard Arabic, the Dutch Text Corpus (DTC) (Redeker, Berzlánovich, & Vliet, 2012; Vliet, Berzlánovich, Bouma, Egg, & Redeker, 2011) for Dutch, the Prague Dependency Treebank (PDT 3.0) (Bejček et al., 2013) and the Chinese Discourse Treebank (CDTB) (Y. Zhou & Xue, 2014) for Chinese, as well as parallel corpora such as the Copenhagen Dependency Treebank (CDT) (Buch-Kromann, Korzen, & Müller, 2009) for Danish, English, Italian and Spanish. Table 3.1 below presents an overview of discourse annotated corpora.

In this chapter, we first overview discourse annotation efforts on other comparably annotated corpora and explain their method of assessing reliability. This will provide a basis for the assessment of the TDB and in applications built on this resource. Then, we will describe the main resource used in this thesis, namely the Turkish Discourse Bank in detail.

Table3.1: Overview of Discourse Annotated Corpora

Name	**Coverage**	**Count**	**Mods**	**Supp**	**Phr Exp**	**Alt Lex**	**Sense**	**Impl**	**Agrmt Spans**	**Agrmt Sense**
TDB 1.0[1]	novels, news, article, story, research/survey, travel, interview, memoir	8,483	Y	Y	Y	-	N	N	Y (Fleiss' Kappa, K)	N
PDTB 2.0	WSJ news, essays	40,600	Y	Y	Y[2]	Y	Y	Y	Y (exact match)	Y (%)
RST-DT	WSJ news	21K	N	N	N	N	Y	Y	Y (K)	Y (K)
PCC 2.0	German newspaper commentaries	32K	N	N	Y[3]	N	Y	Y	Y (%)	N
HDRB	Hindi news	~5K	Y	N	Y[2]	Y	Y	Y	Y (exact match)	Y (K)
LADTB	Arabic news	6,328	Y	N	N	N	Y	N	Y (agr[4], exact match)	Y(K)
PDT 3.0 (PDiT 1.0)	news	20,542	Y	N	Y[5]	Y	Y	N		
CDTB	Xinhua news	3,951	Y	N	N	Y	Y	Y	Y (exact match, P,R,F)	Y (exact match, P,R,F)
CDT	Excerpts from general purpose texts	~12K	N	N	N	N	Y	Y	N	N
DTC	Encyclopedia entries, science news, fundraising letters, commercial advertisements	NA[6] 80 texts btw.190-400 words)	N	N	N	N	Y	Y	Y (K)	Y (K)

[1] Count: number of annotated relations, Mods: modifiers, Supp: supplementary material, PhrExp: phrasal expressions, Impl: implicit relations, Agrmt Spans: agreement on span selection, Agrmt Sense: agreement on sense identification, Y: yes, annotated, N: no, not annotated, %: percent agreement, P: precision, R: recall, F: f-score.

[2] Some phrasal expressions were annotated as alternative lexicalization, but not systematically.

[3] Phrasal connectives were annotated as part of complex connectives.

[4] *agr* is a directional measure described in Al-Saif and Markert (2010).

[5] Phrasal expressions are annotated as *secondary* connectives.

[6] The relation count is not available, only the corpus size has been provided.

3.1 The Penn Discourse Treebank (PDTB)

The Penn Discourse Treebank (henceforth PDTB) (Prasad et al., 2008) is a lexically grounded annotation effort to manually annotate the 1 million word Wall Street Journal (WSJ) Corpus with discourse relations holding between two abstract object (AO) arguments. The AOs are taken to be eventualities (events or states), facts or propositions as in the classification of Asher (1993). Both explicit discourse relations signaled by *discourse connectives* and implicit relations have been annotated with their two arguments. Since the WSJ Corpus has been previously annotated for sentence-level syntax in the Penn Treebank (Marcus et al., 1993), temporal relations TimeBank 1.2 (Pustejovsky et al., 2003) and predicate-argument structure in the Propbank (Palmer, Gildea, & Kingsbury, 2005), these other annotation levels are also available apart from the discourse annotations. Hence many applications have been built on this resource which provides syntactic, semantic and discourse annotation together (cf. Miltsakaki, Dinesh, Prasad, Joshi, & Webber, 2005; Pitler & Nenkova, 2009). The latest release of the PDTB is version 2.0, which is available from the Linguistics Data Consortium.[7]

An explicit discourse connective of the PDTB is annotated as in (29), where the argument syntactically bound to the connective (underlined in the example) is marked as Arg2 (shown in bold face), and the other argument is marked as Arg1 (shown in italics). The explicit connectives were determined from four well-defined syntactic classes: subordinating conjunctions, coordinating conjunctions, prepositional phrases and adverbs. It is noted in Prasad et al. (2014) that a resource-limited subset of these discourse relations were annotated.

(29) *U.S. Trust, a 136-year-old institution that is one of the earliest high-net worth banks in the U.S., has faced intensifying competition from other firms that have established, and heavily promoted, private-banking businesses of their own.* As a result, **U.S. Trust's earnings have been hurt**.[8]

(Prasad et al., 2008: p.2961, ex.1)

However, not all discourse relations are signalled explicitly. Some may be inferred from argument adjacency (Webber et al., 2011). An implicit relation is annotated, when there is no overt connective present in text, but the annotators inferred a discourse relation connecting the two consecutive sentences (and inserted the connective that best expressed the relation) as in (30). These types of relations were annotated for all sentence pairs, as well as complete clauses separated by a colon or a semicolon within a sentence.

(30) But a few funds have taken other defensive steps. *Some have raised their cash positions to record levels.* Implicit = BECAUSE **High cash positions help buffer a fund when the market falls**.

(Prasad et al., 2008: p.2963, ex.6)

[7] http://www.ldc.upenn.edu, Catalog No = LDC2008T05

[8] All examples in this section about the PDTB are taken from either Prasad et al., 2008 or Prasad et al., 2010. The exact reference of each example is given below its text, along with its page number and corresponding enumeration in the original document.

One of the claims of the PDTB is that it is a theory-neutral approach to discourse; it does not commit itself to any theory specifying the kinds of high-level structures that can be obtained from the annotated discourse relations. The PDTB annotates local discourse by doing sentence by sentence annotation, without keeping track of the global structure of discourse. This approach allows to investigate how sentence structure relates to local discourse structure at the sentence level. It also allows researchers to test existing theories of discourse structure (Prasad et al., 2008).

3.1.1 Attribution

Apart from the annotation of explicit and implicit discourse relations, the PDTB annotates attribution of these relations and their arguments as in (31) (shown between square brackets).

(31) "*The public is buying the market* when **in reality there is plenty of grain to be shipped**," [said Bill Biedermann, Allendale Inc. director].

(Prasad et al., 2008: p.2966, ex.15)

Although attribution is not a discourse relation, but a relation between agents and abstract objects, it was annotated in the PDTB mainly due to its interaction between sentence-level structure and discourse structure. The text span for the attribution phrase was annotated with its *source* (either *Wr*: writer, *Ot*: other, *Arb*: arbitrary), *type* of the relation with AOs (either *Comm*: verbs of communication, *PAtt*: propositional attitude verbs, *Ftv*: factive/semi-factive verbs, *Ctrl*: control verbs), *scopal polarity* (*Neg*: surface negation, *Null*: default) and *determinacy* (*Indet*: indeterminate, *Null*: determinate).

3.1.2 Sense

Senses of the discourse relations were also annotated in the PDTB, according to a hierarchical classification system. The sense labels describe the discourse relation between the arguments semantically. The sense hierarchy was divided on its top-level to four categories: temporal, comparison, contingency and expansion with each category further divided up to two more levels, where multiple senses could be assigned. Sense annotation can be used to disambiguate polisemous connectives, where a given connective displays different senses (e.g. the connective *and* can have the senses conjunction, list, result, juxtaposition, as well as some others).

3.1.3 Modifiers

Modified forms of connectives have also been annotated in the PDTB. Adverbs modifying explicit discourse connectives, including productive ones such as *apparently, at least, partly, in large part, even only* were also annotated in the PDTB as modifiers (Prasad et al., 2014). This enabled the modified forms of the connectives to be annotated as the same types of a given head (i.e. the bare form of the connective).

3.1.4 Entity Relations, Alternative Lexicalizations and No Relations

In the cases where an implicit connective could not be provided, annotators were asked to label the relations as either EntRel, AltLex or NoRel. If the annotators could not find an inferred implicit relation, but they identified entity-based relations between sentences, these relations were marked as *EntRel*. The EntRel annotations were based on adjacency, where the same entity was realized in both sentences.

In cases where an implicit connective was redundant because there was a non-connective expression (i.e. outside the set defined for explicit connectives) present in the text which signaled the discourse relation, these expressions were marked as *Alternative Lexicalizations (AltLex)* with their two arguments as in (32) (where the AltLex is shown underlined and the sense of the relation is given inside parenthesis in small capitals).

(32) Now, GM appears to be stepping up the pace of its factory consolidation to get in shape for the 1990s. (CONTINGENCY.CAUSE.REASON) **One reason is mounting competition from new Japanese car plants in the U.S. that are pouring out more than one million vehicles a year at costs lower than GM can match**. [wsj 2338]

(Prasad et al., 2014: p. 926, ex.8)

Alternatively lexicalized expressions may have different types: (1) two part expressions where one refers to the relation and the other anaphorically to Arg1; (2) one part expressions referring anaphorically to Arg1; (3) one part expressions referring to the relation. However, the PDTB annotation does not distinguish between these different types of AltLex expressions.

When the annotators could not find any explicit, implicit, entity-based, or alternatively lexicalized relation between the sentences, they marked the relation as *NoRel* denoting that there was no relation.

Overall, about 18K explicit, 16K implicit, 600 AltLex relations, 5K EntRel and 250 NoRel, for a total of around 40K relations were annotated in the PDTB 2.0. (Prasad et al., 2014).

3.1.5 Minimality Principle

One important convention used in the PDTB is the *Minimality Principle* (Prasad et al., 2008), which restricts the annotators to select only the span that is necessary for the complete interpretation of the discourse relation as an argument. Hence, parenthetical clauses, non-restrictive clauses and such were excluded from the argument spans. An example is given in (33), where the arguments do not include the attribution (shown in square brackets) or the non-restrictive relative clause, since they are not required to interpret the relation conveyed by the connective *But*.

(33) '*I'm sympathetic with workers who feel under the gun*', [says Richard Barton of the Direct Marketing Association of America], which is lobbying strenuously against the Edwards beeper bill. 'But **the only way you can find out how your people are doing is by listening**'. [wsj 1058]

(Webber et al., 2011, p. 447, ex.13)

3.1.6 Inter-Annotator Agreement

The inter-annotator agreement for explicit and implicit connectives have been reported using the *exact match* criterion (Miltsakaki, Prasad, Joshi, & Webber, 2004a), where agreement was recorded as 1 for identical span selection and 0 otherwise. This is provided for two separate calculations, where one calculation involves taking each argument of a connective as distinct tokens, and the other involves taking the two arguments of a connective as a combined token. An agreement of ~86% for Arg1 and ~94% for Arg2 is reported for explicit connectives for an overall agreement of ~90%. When the agreement over both spans were calculated the overall agreement drops to ~83%. It is reported that the combined method presents lower agreement for subordinating conjunctions and adverbials.

3.1.7 PDTB as a Resource in Language Technology Applications

As previously mentioned, the PDTB has become a frequently utilized resource in language technology applications. Many studies have used this corpus for a variety of applications including automatic identification of discourse relations and/or their arguments (Elwell & Baldridge, 2008; Ghosh, Johansson, Riccardi, & Tonelli, 2011; Lin, Ng, & Kan, 2011; Louis, Joshi, Prasad, & Nenkova, 2010; Pitler & Nenkova, 2009; Pitler et al., 2008; Polepalli Ramesh, Prasad, Miller, Harrington, & Yu, 2012; Prasad et al., 2010; Torabi Asr & Demberg, 2013; Wellner, Pustejovsky, Havasi, Rumshisky, & Sauri, 2006; Wellner & Pustejovsky, 2007; L. Zhou, Li, Gao, Wei, & Wong, 2011) statistical machine translation (Meyer & Webber, 2013), content selection in summarization (Louis, Joshi, & Nenkova, 2010), among others.

3.1.8 Nominal Phrases and Discourse Deictic Expressions

Miltsakaki et al. (2004a) states that nominal phrases and discourse deictic expressions denoting events or states were annotated in the PDTB. A nominal phrase such as "fainting spells" denotes an event and was selected as an argument in (34) below.

(34) Its symptoms include a cold sweat at the sound of debate, clammy hands in the face of congressional criticism, and *fainting spells* <u>when</u> **someone writes the word "controversy"**.

(Miltsakaki et al., 2004a, p. 4, ex.12)

Forms such as *this* and *that* which denote clausal textual spans from preceding discourse may also denote events or states. These are called discourse deictic expressions and were annotated in the PDTB as in (35) below. In cases where there was an anaphoric or deictic expression, the annotators were asked to annotate as if these expressions were resolved, hence in (35) *that's* was selected as Arg1 for the relation linked by *because*.

(35) Airline stocks typically sell at a discount of about one third to the stock market's price-earnings ratio – which is currently about 13 times earnings. *That's* <u>because</u> **airline earnings, like those of auto makers, have been subject to the cyclical ups- and-downs of the economy**.

(Miltsakaki et al., 2004a, p. 4, ex.13)

Hence, the PDTB considers groups of sentences, single sentences or clauses, NPs specifying events or states, and discourse deictic expressions as legal arguments of discourse relational devices as long as they have an abstract object interpretation.

The annotation of alternative lexicalizations, as well as discourse deictic expressions in the PDTB require special attention for this thesis. Our phrasal expressions correspond to PDTB's alternative lexicalizations where a discourse deictic expression is present. In the TDB, phrasal expressions corresponding to forms such as *that's because* in example (35) above are annotated as a kind of discourse relational device including the deictic item. Unlike the PDTB convention to assume that the deictic expressions have been resolved, in the TDB the referents of the deictic expressions are resolved as one of the arguments of these discourse relational devices. Although they have been briefly described here, we will provide a wider explanation and discussion in Chapter 6.

3.2 The Rhetorical Structure Theory (RST) Discourse Treebank

RST Discourse Treebank is a resource of a 385-article (176K-word) portion of the WSJ corpus of the Penn Treebank. The RST-based discourse annotation includes the identification of elementary discourse units (EDUs), considered as the minimal units of an RST discourse tree. The EDUs were chosen to be clauses, with the exception of subject or object clauses and clauses complementing a main verb. Relative clauses and nominal postmodifiers were considered embedded discourse units, and a small number of phrasal EDUs beginning with strong discourse markers (i.e. when a discourse marker only cues a single relation and is not ambiguous) such as *because, in spite of, as a result of, according to* were also permitted. Relations between the adjacent spans of the EDUs were identified as mononuclear (i.e. the salient span is the *nucleus*, the other span is the *satellite*) or multinuclear (i.e. all spans are of equal weight). A total of 78 relations (53 mononuclear, 25 multinuclear) were used in the annotation of the corpus, which resulted in 21K EDUs annotated for rhetorical relations. The RST Discourse Treebank is available from the Linguistics Data Consortium.[9]

The inter-annotator agreement was tracked during separate phases of the project using the Kappa statistic of Siegel and Castellan (1988), which was adapted for application to hierarchical structures (Marcu, Amorrortu, & Romera, 1999). In this method, hierarchical structures were mapped to labeled sets of units. The agreements were calculated for the three annotators on selected documents for each phase on the EDU, hierarchical spans, hierarchical nuclearity and hierarchical relation assignment judgments.

3.3 Potsdam Commentary Corpus (PCC)

The PCC is a collection of 175 German newspaper (~32K tokens) commentaries manually annotated with syntax trees, nominal coreference, explicit discourse connectives and their arguments, and RST discourse structure trees. The sentence syntax annotation was done semi-automatically, where a parse tree suggested by the annotation tool is inspected by a manual

[9] http://www.ldc.upenn.edu, Catalog No=LDC2002T07

annotator. Only nominal coreference is annotated in the PCC, excluding event anaphora and indirect coreference. The rhetorical discourse structure was annotated manually, where annotation guidelines were devised in the revision of the corpus to form PCC 2.0 (Stede & Neumann, 2014). Additions in the second version include the annotation of PDTB-style explicit discourse connectives and their arguments as *external* (corresponding to Arg1 in the PDTB-style) and *internal* (corresponding to Arg2 in the PDTB-style), where only a closed-class set of German connectives were considered. The PCC 2.0 corpus is available from the corpus web site.[10]

Inter-annotator agreements for 20 of the texts for two annotators have been reported using a percent match measure. On the distinction of discourse vs. non-discourse relations annotators are reported to agree 83.3% of the cases, and within the commonly annotated connectives, agreement of ~91% is reported for argument spans (Stede & Neumann, 2014).

3.4 Hindi Discourse Relational Treebank (HDRB)

HDRB is an effort to manually annotate 200K-word portion of a Hindi text corpus of newspaper articles. It follows the PDTB-style annotation with some exceptions and additions. Explicit, implicit connectives and alternative lexicalizations, entity-based coherence relations (EntRels) and no relation cases (NoRel) were annotated, along with the senses of the relations. As part of the explicit connectives, the HDRB annotated *sentential relatives* (i.e. relative pronouns conjoining a relative clause with a matrix clause), *subordinators* (including postpositions, verbal participles and suffixes introducing non-finite clauses of AOs), and *particles* which act as discourse connectives, in addition to the subordinating conjunctions, coordinating conjunctions and adverbials annotated in the PDTB. Also different from the PDTB-style was the labeling of the two arguments of a relation, which was done semantically. Hence, the sense of the relation was used as a guide to determine which argument was Arg1 and which was Arg2, rather than assigning the Arg2 label to the syntactically hosting argument. Another difference of HDRB is the annotation of implicit connectives across paragraph boundaries, in addition to adjacent sentences in a paragraph. The sense annotations were also done with some modifications to the PDTB scheme, where argument-specific labels were eliminated, pragmatic relations were treated uniformly and a *goal* sense was added. The HDRB annotations are ongoing, as about 75K part of the HDRB annotation has been completed at this time (Sharma et al., 2013).

An inter-annotator agreement study on a 5K-word part of the HDRB corpus is reported in Kolachina et al. (2012), which was conducted using the exact match criterion of Miltsakaki et al. (2004a) separately for connective identification and argument identification. Sharma et al. (2013) also reports agreement over sense identification calculated using the Fleiss' Kappa measure, along with exact-match measures for connective and argument identification calculated for a 12K part of the corpus.

[10] http://www.ling.uni-potsdam.de/acl-lab/Forsch/pcc/pcc.html

3.5 Leeds Arabic Discourse Treebank

This Treebank is an effort to manually annotate Modern Standard Arabic news articles from the Arabic Penn Treebank for explicit discourse connectives with their two arguments (Al-Saif & Markert, 2010, 2011), as well as the senses for the relations. LADTB uses a PDTB-style annotation scheme, with a few exceptions. First of all, it uses a more coarse-grained sense classification, as well as two additional senses (i.e. BACKGROUND and SIMILARITY). Another shift from the PDTB-style is the annotation of clitics as connectives and prepositions used as connectives. The special form of the Arg2 of these prepositions is called *Al-Masdar*, which consists of forms that span several grammatical and morphological categories in English (i.e. gerund, nominalization, noun which is not a nominalization), similar to infinitive forms in Turkish such as *gelmek*, which are called *mastar* forms. The final gold standard corpus has 125K tokens annotated with about 6K discourse connectives.

Inter-annotator agreement for two annotators is reported for discourse vs. non-discourse distinction, sense labeling and argument selection over 537 news texts with 107 discourse connectives (Al-Saif & Markert, 2011). The agreement for the first two were calculated using Siegel and Catellan's Kappa statistic. Argument selection agreement was calculated with a directional measure agr, which measured the word overlap between text spans of two judges in relevance to each other, taken from Wiebe, Wilson and Cardie (2005), and an exact-match measure.

Work on automatically identifying discourse connectives, sense relations and relation arguments have been done on the LADTB (Al-Saif & Markert, 2011) using supervised algorithms. Surface features such as the position of the potential connective in the sentence (i.e. sente-initial, -medial, -final), type of the sentence (i.e. *Simple*: unattached single token, *Pot-Clitic*: attached token, *MoreThanToken*: formed of more than one token), lexical features of surrounding three words, POS features (where multi-token POS were combined, and clitics untagged by the tagger were labeled with the POS *None*), syntactic category of parent, left sibling and right sibling, as well as an *Al-Masdar* feature were used in several different models. The best model achieved 92% accuracy using syntax with no lexical patterns.

3.6 The Prague Dependency Treebank (PDT 3.0)

The discourse annotated portion of the PDT 3.0 (Bejček et al., 2013) is a revised version of the Prague Discourse Treebank (PDiT 1.0) (Poláková et al., 2013). This revised discourse Treebank includes discourse annotations for explicit connectives, pronominal textual coreference annotation for first and second person. The PDiT consists of about 49K manually annotated sentences from Czech newspapers, where the discourse relations were directly annotated on the syntax trees different from the PDTB. The explicit discourse relations between nominalizations or deictic expressions were not annotated. All relations between sentences (i.e. inter-sentential) and a part of the relations within sentences (i.e. intra-sentential) were annotated manually, and the rest of the intra-sentential connectives were automatically annotated. In the extended discourse annotation, multiword connectives in the form of certain collocations (e.g. *that is the reason why, the only condition was, from this reason) were also annotated as what they refer to as secondary connectives* (Rysová & Rysová, 2014). These correspond to the phrasal expressions annotated in the TDB and will be explained in greater detail in chapter 4, where we discuss types of discourse connectives. The discourse relations were also labeled for sense. The Prague Dependency Treebank 3.0 can be obtained from the LINDAT-Clarin repository.[11]

Inter-annotator agreement results are reported on a 4% portion of the corpus where the agreements were calculated using a connective-based F1-measure (Mírovský, Mladová, & Zikánová, 2010) for relation existence agreement, simple ratio and Cohen's Kappa for relation type agreement.

3.7 The Chinese Discourse Treebank (CDTB)

CDTB 0.5 annotates about 70K-words of Chinese newswire text of the Chinese Treebank with discourse relations with their two arguments in the PDTB-style. About 5K explicit and implicit discourse relations were manually annotated, including sense labeling. Different from the PDTB, a semantically-based argument selection approach was applied and a flat sense classification was used. It is reported that discourse relations are mostly implicit in Chinese as they are dropped (i.e. there are about four times more implicit relations identified than explicit relations) (Y. Zhou & Xue, 2012, 2014) and hence, the implicit and explicit connective annotation was done simultaneously. Alternative lexicalizations are also annotated in the CDTB, including the annotation of cases where a small number of verbs indicating discourse relations. Another deviation from the PDTB-style was that the arguments of a discourse relation were identified semantically, instead of syntactically. Entity coherence relations (EntRel) and cases where there were no discourse relations identified (i.e. NoRel) were annotated as in the PDTB. The CDTB is available from the Linguistic Data Consortium.[12]

All the annotations in the CDTB were done by two annotators and controlled by a third annotator. Inter-annotator agreement results have been reported for relation identification (discourse vs. non-discourse), relation type (i.e. explicit, implicit, AltLex, etc.), sense, argument span exact match, overall span match and boundary match using precision, recall, f-score and accuracy measures.

3.8 The Copenhagen Dependency Treebanks (CDT)

The CDT includes five parallel open-source treebanks (for Danish, English, German, Italian and Spanish) manually annotated with syntax, discourse, anaphora and morphology, where the Danish Treebank is word-aligned with the other corpora. All relations in the corpora are represented as directed labelled relations between words and morphemes (Buch-Kromann & Korzen, 2010). The annotations are based on the Discontinuous Grammar dependency theory of Buch-Kromann (2006) as described in Buch-Kromann et al. (2009). The corpus is annotated using a multi-level approach which treats discourse structures as dependency trees linking discourse parts of sentence and sentence fragments separated by full stops. A relation has a *head* (i.e. nucleus) and a *dependent* (i.e. satellite), or in the case of no head, it has a multi-nuclear structure as in RST. Dependency is represented as a linear relation linking a second segment dependent on a first segment. Explicit and implicit connectives, attribution

[11] http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3

[12] http://www.ldc.upenn.edu, Catalog No = LDC2014T21

and senses of the discourse relations have been identified. The CDT is freely available from the CDT website.[13]

The CDT only reports inter-annotator agreement for the annotation of coreferential and associative anaphoric relations as percent agreement (Korzen & Buch-Kromann, 2011).

3.9 The Dutch Text Corpus

The Dutch Text Corpus (Redeker et al., 2012; Vliet et al., 2011) is a compilation of 80 Dutch texts manually annotated for rhetorical and genre-specific discourse structure, as well as lexical cohesion (*repetition, systematic semantic relation,* or *collocation*). An RST analysis is done for the discourse-structure annotation. In order to analyze relational cohesion (i.e. lexical or phrasal elements that signal coherence relations), all lexical and phrasal elements (discourse markers) in the text that signal cohesion at local and global levels were considered. Coordinated elliptical clauses and clauses that share a noun phrase as subject were also treated as EDUs. Non-restrictive relative clauses and embedded clauses within parentheses were taken as embedded discourse units, whereas restrictive relative clauses, subject and object clauses and complement clauses were not taken as EDUs as in RST. In the genre-specific discourse structure annotation, genre-specific moves (i.e. functional components; e.g. move types for the encyclopedia genre were *name, define* and *describe*) were identified and overlaid with the RST-tree with a segmentation into a sequence of moves.

All annotations were done by two separate annotators and inter-annotator agreements for 16 texts have been reported using Kappa measure on the identification of segment boundaries, discourse spans, nucleus assignment and rhetorical relation assignment. Separate agreements were calculated for the lexical cohesion relation identification and relation assignment (Redeker et al., 2012).

3.10 The Turkish Discourse Bank (TDB)

The Turkish Discourse Bank (TDB) is the first publicly available corpus in Turkish annotated at the discourse-level (Zeyrek et al., 2013). This ~400.000-word language resource is a subcorpus of the 2 million-word METU Turkish Corpus (MTC) (Say et al., 2004) consisting of texts of post-1990 written Turkish encompassing various genres. TDB contains PDTB-style annotations of explicit discourse connectives, the two arguments related by the connectives, as well as modifiers and supplementary materials. The arguments of the connectives are taken as text spans denoting abstract objects (Asher, 1993).

The first release of the TDB (February, 2011) includes annotations of 8483 relations for 144 discourse connective forms (for 77 search items[14]). It is freely distributed online to researchers upon request at the project website[15] along with a browser (Şirin, Çakıcı, & Zeyrek, 2012). In the following subsections, the data, the annotation scheme, annotation cycle and annotation procedures are described in more detail.

[13] https://code.google.com/p/copenhagen-dependency-treebank/wiki/CDT

[14] A list of search items and corresponding connective forms are provided in Appendix B. For example, the connective forms provided by the search item dolayı are dolayı, bundan dolayı and bu sebepten dolayı.

[15] The TDB Project Website can be found at http://medid.ii.metu.edu.tr

3.10.1 Data

TDB annotations are done on 197 text files that constitute one of MTC's subcorpora. Initially, the MTC was divided into four equal-size subcorpora obtained by converting the MTC to raw text files with UTF-8 encoding and discarding tags such as <p> (paragraph), <list> (list), <hi> (highlight), but retaining information such as genre, author, publication date. Two genres, namely essays (76 files) and columns (83 files) were excluded from the four subcorpora obtained in this way, as they did not have conventional paragraph structure, which may have interfered with the annotation of the discourse connectives. The rest of the genres (787 files) were kept by maintaining the same genre distribution of the MTC for each subcorpus. The TDB is annotated on the first subcorpus. (See Table 3.2.[16])

Table3.2: The Distribution of the genres in the MTC and the TDB

Genre	**MTC**		**TDB**	
	#	%	#	%
Novel	123	15.63	31	15.74
Story	114	14.49	28	14.21
Research/Survey	49	6.23	13	6.60
Article	38	4.83	9	4.57
Travel	19	2.41	5	2.54
Interview	7	0.89	1	1.02
Memoir	18	2.29	4	2.03
News	419	53.24	105	53.30
TOTAL	787	100	197	100

The annotations of the TDB are done using an annotation tool (Discourse Annotation Tool for Turkish-DATT) especially developed for this purpose (Aktaş, Bozşahin, & Zeyrek, 2010). DATT performs stand-off annotation in the form of XML files, where beginning and end offsets of the annotated text spans, as well as their content text are kept. A sample xml relation is provided in Appendix A.

3.10.2 Annotation Scheme

The TDB 1.0 annotates explicit discourse connectives that relate two abstract objects (Asher, 1993), which primarily correspond to verbal clauses and certain nominal clauses in Turkish. The connectives that relate arguments that are not abstract objects are not annotated. Each annotated connective is tagged with the *connective* tag (*Conn*). For the two arguments that the connective relates, following the PDTB, the argument syntactically hosting the connective is called the *second argument* and the other argument is called the *first argument*. The first argument is tagged as *Arg1* and the second argument is tagged as *Arg2*. An example annota-

[16] This table has been taken from Demirşahin, Sevdik Çallı, Ögel Balaban, & Zeyrek (2012), p.1.

tion is presented in (36)[17]. In all examples Arg1 is presented in italics, Arg2 is represented in bold and the connective is underlined, unless stated otherwise.

(36) *Dışa karşı güçlüydü*, ama **içe, kendi yüreğine yıkılmak üzereydi**.

He was strong against the outside, but **inside, he was about to collapse on his heart**.

(00001131.txt)

A further example is provided in (37), where both arguments *bekleme* (waiting) and *arama* (looking) for the connective *ve* (and) are nominal clauses denoting abstract objects (AOs). Whereas in (38) below both arguments are verbal clauses. If there is a modifier for the connective as in this case, where *hemen* (right/just) modifies *önce* (before), this is tagged with the *modifier* tag (*Mod*). In the examples the modifier is shown underlined and marked with Mod in subscripts.

(37) ||Sabah uyanır uyanmaz Beril'i bulması gerektiğini düşündü||$_{\text{Supp\_Shared}}$. .. Artık {onu}$_{\text{Shared}}$ *beklemenin* **ve** **aramanın** boşuna olduğunu anlamıştı.

||He thought he needed to find Beril as soon as he woke up in the morning||$_{\text{Supp\_Shared}}$. .. Now he had realized it was pointless *waiting* **and** **looking** {for her}$_{\text{Shared}}$.

(00001231.txt)

(38) Sabah çok erken saatte *bir önceki akşam gün batmadan* hemen$_{\text{Mod}}$ **önce** **astığı** çamaşırları toplamaya çıkıyordu...

At a very early time in the morning she went to pick up the laundry **she had put up** right$_{\text{Mod}}$ **before** *sun came down in the evening the day before.*

(00001131.txt)

The material supplementing or supporting the arguments are tagged as *supplementary material1* (*Supp1*) and *supplementary material2* (*Supp2*) numbered according to the argument they are related to (See examples (39) and (40) respectively). Supp1 and Supp2 are given inside square brackets marked with the related tag in subscripts in the examples.

(39) [Duvarlardan ürkütücü sesler geliyordu]$_{\text{Supp1}}$. İçi titredi. Paniğe kapıldı. Aslında *böyle şeyler onu asla korkutmazdı*, **ama** **bu sefer ne yapacağını şaşırmıştı**.

[Scary noises came from the walls]$_{\text{Supp1}}$. He shivered inside. He panicked. Actually *these things would never scare him*, **but** **this time he was confused about what to do**.

(00001231.txt)

(40) [Duvarlardan ürkütücü sesler geliyordu]$_{\text{Supp2}}$. İçi titredi. *Paniğe kapıldı.* **Aslında** **böyle şeyler onu asla korkutmazdı**, ama bu sefer ne yapacağını şaşırmıştı.

[Scary noises came from the walls]$_{\text{Supp2}}$. He shivered inside. *He panicked.* **Actually** **these things would never scare him**, but this time he was confused about what to do.

[17] This example is from the file 00001131_agreed_ama.xml of the TDB.

(00001231.txt)

The text spans that are common for both arguments, such as a common subject, object, or an adjunct, are annotated as *shared material* with the *Shared* tag (41). If any supplementary material for the shared material is needed, this text span is annotated with the *Supp_Shared* tag (See example (37) above). The Shared tag is represented in the examples in curly brackets marked with Shared in subscripts, and the Supp_Shared tag is shown within double vertical bars marked with Supp_Shared in subscripts.

(41) {Öğretmenimiz}$_{\text{Shared}}$ *bize bu hastane gezisini*, **iyilik yapmak, başkalarını düşünmek ve yurda vefa borcunu ödemek gibi birtakım kavramları öğretmek amacıyla** *düzenlemişti.*

{Our teacher}$_{\text{Shared}}$ *for us this hospital trip*, **teach some concepts such as doing good, thinking of others and honor one's duty of loyalty to the country for the purpose of** *organized.*

(00008113.txt)

Finally, each relation for a given discourse connective may have a *note* attribute, where the annotators add their notes if needed when they are annotating. There is also a *type* attribute, which in the current version is set to the value "*EXPLICIT*", as only explicit connectives are annotated (See Appendix A).

The TDB XML schema keeps offsets of the text spans and this allows for discontinuous text spans to be selected where needed. Thus, different conditions such as crossing arguments can easily be captured in the annotations (Aktaş et al., 2010). In (41) above notice how the first argument is selected discontinuously, whereas example (42) has both of its arguments selected discontinuously.

(42) ... *gerçekte evli iki insan gibi değil de* (**evlilikler sıradanlaşıyordu çünkü, tekdüze ve sıkıcıydı**; biz farklı olacaktık), *aynı evi paylaşan iki öğrenci gibi yaşayacaktık.*

... *in reality we wouldn't be like two married people* (**marriages were becoming ordinary because, they were monotonic and boring**; we would be different), *we would live like two students sharing the same house.*

(00003121.txt)

A total of 833 discontinuous arguments (i.e. 574 discontinuous Arg1s, 252 discontinuous Arg2s and 7 relations where both arguments are discontinuous) were selected in the TDB, as opposed 7650 non-discontinuous arguments (See Appendix F, Table F.1).

In the following part we describe the annotation cycle and the procedures used in the annotations of the TDB.

3.10.3 Annotation cycle and annotation procedures

The annotations included in TDB 1.0 are agreed upon gold standard annotations, which are established by the agreement of three independent annotators, or one independent annotator and a pair of annotators (called *pair annotation*) (Demirşahin, Yalçınkaya, & Zeyrek, 2012). In the case of no initial agreement among annotators, a decision is reached in agreement meetings. This procedure is called as *group annotation*, which is also used for annotating connectives that are few in number, where inter-annotator agreement is difficult to calculate. The annotation cycle of the TDB was as follows:

First, a preliminary set of guidelines was prepared taking the PDTB style as a baseline and an initial set of connectives were determined for annotation. Three independent annotators annotated all the relations in the corpus, relating two abstract object arguments, along with their supplementary materials, modifiers and shared materials. Non-discourse usages of the connectives were determined and they were not annotated. The annotators added their notes when necessary, in order to ease the resolution of any disagreement later on. Then, the annotations were examined by a member of the research group; the fully agreed annotations were set apart as agreed annotations and the cases where the annotators were not in full agreement were extracted to be resolved. These cases were discussed in agreement meetings, where all annotators and at least one researcher attended. Each annotator explained the reasons behind their annotation and the researchers gave feedback. A unanimous decision was reached at the end of the discussions to be set as an agreed annotation for that relation. The interactive discussions usually resulted in some new guidelines to be determined, or old ones to be updated and the set of annotation guidelines were updated as needed. If the updated guidelines required any changes to the previously agreed annotations, these were rechecked to ensure their conformity to the new guidelines. This process was called *proofing*. The proofed annotations were set as the gold standards of the TDB. This procedure was repeated for another set of connectives and the annotation cycle continued in this way.

Prior to the annotation of the initial connective set, the annotators were trained to form their own ideas about annotating discourse connectives by reading articles on other discourse and annotation studies, and by discussions on topics such as what constitutes a discourse connective, how to distinguish non-discourse usages, etc. Then the preliminary annotation guidelines were prepared, and the annotators started annotating the initial set of connectives.

The annotation guidelines do not explicitly specify how to annotate discourse relations or what discourse segments to choose. They merely provide some principles to abide by to achieve uniformity in the annotations, leaving the annotators room to use their native-speaker intuitions. For example, in determining the span of the connective's arguments, the guidelines specify to adhere to the "minimality principle" (MP) (Prasad et al., 2008).

The interpretation of the MP frequently caused the annotators to disagree on the exact boundaries of the argument spans as in (43) (Zeyrek et al., 2010). In this example, a disagreement on the boundary of the Arg2 span was observe, where one annotator selected (43b), while the other annotators selected a longer span (8b-c), including an additional clause where the cataphor is resolved. Similar inconsistencies were observed when the annotators felt that the MP conflicted with their basic insights of discourse interpretation. Cases like (43) were resolved with a methodological decision biased towards the MP, to use the *Supp* label for anaphor /coreference chain resolution.

(43)

(a) . . . *ikincisindeki ayrıntı bolluğu Recaizade Ekrem'in gerçekçiliğine atfedilmiştir.*

... the richness of details in the second (novel) was attributed to Recaizade Ekrem's realism.

(b) Oysa **asıl dikkat çekmesi gereken şudur**:

However, **this is what should be noted**:

(c) Araba Sevdasının Çamlıca'sı yitik bir Çamlıca'dır.

The Çamlıca described in Araba Sevdası is a lost place.

(Zeyrek et al., 2010, p. 286, ex.4)

In the later stages of the project when inter-annotator agreement stabilized, an annotation procedure called *pair annotation* (Demirşahin, Yalçınkaya, et al., 2012) was adopted. In this method, one annotator annotated a connective for a set of files independently.Two annotators annotated the same files together, where one sat by the computer and did the actual annotation, the other provided an opinion and they formed a common annotation at the end. The fully agreed annotations of the independent annotator and the group were set as agreed annotations, and the disagreements were again resolved in interactive agreement meetings as usual. This procedure both shortened the annotation time and lowered the number of disagreements to be resolved in the annotation meetings. Pair annotation procedure was used for 3985 annotations (about 47%), whereas 3804 relations (about 45%) were annotated by three independent annotators and the remaining 694 relations (about 8%) were annotated by group annotation (Zeyrek et al., 2013). Group annotation was either done on the early stages of the project, or when the relation count of a given connective was very low and all the annotators went through the occurrences together and decided on the annotations through common discussion. The agreement for the pair annotation is calculated by taking the pair of annotators as a single annotator and comparing their agreement with the independent annotator.

In this chapter, we tried to provide an overview of discourse annotated corpora and their reliability assessment methods. Finally, we described the TDB which is our main resource in this thesis. We explained the data used and the basic principles applied in the annotation of the TDB including the annotation schema. Then we described the procedures involved in the annotations. A quantitative assessment of the TDB will be provided in Chapter 4.

CHAPTER 4

ASSESSMENT OF THE TURKISH DISCOURSE BANK

It is important in any annotation study to provide an assessment of the annotations so as to determine the reliability of the final product. Only if the final annotations are reliable, then they can be used by the research community to investigate the properties of the language at hand and for further studies of that language. It is even more imperative to evaluate and assure the integrity of the annotations in a corpus because a large amount of data to be built on is involved. If the annotated corpus is not reliable, then any additional annotations on the corpus or any study utilizing the corpus data would be compromised. In order to prevent this, the reliability of any such annotation effort needs to be evaluated and the reliability of the manual annotations needs to be corroborated.

In the field of computational linguistics, it is seen that discussions on reliability of manual annotations began in the 1990's, when researchers have started to require evidence for reliability of a given work involving subjective judgments (Carletta, 1996). About ten years later, Spooren (2004) argues that providing reliability measures are needed especially in discourse studies, where coherence relations are analyzed. Arstein and Poesio (2008) explicitly state the necessity to show the reliability of hand-coded data. However, although there is an increasing number of corpus annotation efforts in the field of discourse analysis (Al-Saif & Markert, 2010; Oza et al., 2009; Prasad et al., 2008; Stede, 2004), it is not yet standard practice to report reliability measures (Spooren & Degand, 2010). In fact, Spooren and Degand (2010) plead the researchers for explicitness and to report their agreement results.

In this chapter we present our methodology for an overall assessment of the Turkish Discourse Bank (TDB) Version 1.0 (Zeyrek et al., 2013). In order to unearth the value of a discourse annotated corpus for Turkish linguistics, we provide a statistical evaluation of this resource and make sure the annotations have been carefully carried out so as to produce a reliable, gold standard data. Then, after establishing the dependability of this resource, we can utilize it in future studies for building language technology applications, adding new annotations or making linguistic analyzes. As a consequence, it will help us gain an understanding of the workings of Turkish discourse relations.

Hence, the aims of this chapter are to (a) present the measures taken to ensure TDB's reliability and validity (agreement measurements) (b) suggest a methodology to follow when building corpora in order to have a reliable resource as an end product. For this purpose, annotations in the TDB are evaluated using the Kappa agreement measure. Three kinds of agreement are measured: (a) inter-annotator agreement, (b) intra-annotator agreement, and (c) gold-standard agreements. The inter-annotator agreements check the reliability of the annotations of the annotators with respect to each other, whereas the intra-annotator agreement is aimed at assessing the integrity of the independent annotations over time. It is calculated for the re-annotation of a certain portion of the independent annotations of one of the annotators. Furthermore, agreements with the gold standard data, which we call *gold standard agreements*, display content reliability. Finally, we propose calculating some extra evaluators that are originally used in evaluating information retrieval systems, in order to assess an annotator's reliability when comparing independent annotations with the gold standards.

4.1 Evaluating the TDB

Annotated corpora are important resources that can be used in many applications in natural language processing (NLP) and computational linguistics (CL). Especially for languages where such resources are scarce such as Turkish, and for research areas such as discourse, which are less frequently tackled, the value of such corpora increase. As Leech (2005, Section 1)[1] states: "adding annotation to a corpus is giving 'added value', which can be used for research by the individual or team that carried out the annotation, but which can also be passed on to others who may find it useful for their own purposes". Hence, one of the most important functions of annotating corpora is reusability by other researchers. In order to provide this, the quality of the annotations need to be supplied via accuracy and consistency, where accuracy is usually used for automatic annotation to express how much of the annotation is correct, and consistency refers to the agreement of the annotators with each other (Leech, 2005). The researchers should be able to provide evidence that different people can agree on the results, hence the results can be replicated (Carletta, 1996). Otherwise, the annotations cannot be reused.

The TDB is the first annotated discourse resource for the Turkish language, hence it is important to validate its annotation integrity in order to use it in future research as a gold standard data. For this purpose, throughout the development of this annotated corpus, statistical analyses were done (cf. Demirşahin, Yalçınkaya, et al., 2012; Zeyrek et al., 2009, 2010, 2013). In Zeyrek et al. (2009) initial inter-annotator and gold standard agreements using Cochran's Q statistic for three subordinators (i.e. *rağmen* 'despite', *karşın* 'on the contrary', *halde* 'despite') were presented. On the other hand, in Zeyrek et al. (2010) eight connectives (i.e. *yandan* 'on the other hand', *ayrıca* 'in addition, separately', *rağmen* 'despite, despite this', *fakat* 'but', *tersine* 'on the contrary', *dolayısıyla* 'as a result', *oysa* 'however', *amaçla* 'for this purpose') displaying low Kappa values for inter-annotator agreement and two discontinuous connectives (i.e. *ne* .. *ne* 'neither ..nor', *hem* .. *hem* 'both ..and') were presented and the most common inconsistencies for these connectives were discussed. Demirşahin, Yalçınkaya, et al. (2012) compares Kappa values for gold standard agreements of independent annotators with gold standard agreement of pair annotations for six connectives (i.e. *aslında* 'actually', *halde* 'despite', *nedeniyle* 'because of', *nedenle* 'for this reason', *ötürü* 'due to', *yüzden* 'so, because of this'). The study also presents a comparison of inter-annotator agreements of three independent annotators with inter-annotator agreements of an individual annotator and a pair of annototars for four connectives (i.e. *ama* 'but', *sonra* 'after', *ve* 'and', *ya da* 'or'). Finally, Zeyrek et al. (2013) present inter-annotator Kappa agreements for search items that retrieve subordinators and related phrasal expressions and inter-annotator Kappa agreements for discourse adverbials. One of the contributions of this thesis is to provide an all encom-

[1] Page numbers for his resource are not available as it is an online version of the book. The quote is from Section 1, paragraph 2.

passing, overall assessment of the TDB, where all agreement statistics (i.e. not just inter-annotator agreement but also intra-annotator agreements, gold standard agreements and extra evaluators) for all the annotated connectives, along with evaluations and discussions will be presented.

In order to quantitatively evaluate the first release of TDB, one can first examine the descriptive statistics of the annotations. In TDB 1.0 there are a total of 8483 relations for 77 search tokens annotated. The search tokens cover 144 distinct discourse connective forms and 121 modifications of these forms (See Table B.1 in Appendix B). There were a total of 21,710 occurrences of the search tokens in the TDB corpus, where 8483 of these were found to be discourse usages amounting to about 39 percent of the overall usage. The relation numbers, discourse and non-discourse usages and annotation procedure(s) used for each connective are given in Appendix C (Table C.1). A comparison of some descriptive statistics of the TDB with other discourse annotated corpora was presented in Chapter 3 (see Table 3.1).

The next step in evaluating the TDB is determining the level of confidence in the annotations (or consistency). The most extensively used confidence measure in manual annotation studies is the agreement between the annotators annotating the same material, known as *inter-annotator agreement* or *inter-coder agreement* or *inter-rater agreement* (cf. Artstein & Poesio, 2008; Bayerl, Gut, Lüngen, & Karsten I., 2003; Leech, 2005; Sporleder & Lascarides, 2008; among others). Field (2009) defines *content validity* as the degree to which the construct being measured is represented. Hence, for manual annotation we can calculate content validity as the agreement of the annotators with the gold standards. We will call these *gold standard agreements* and they will provide a second means of assessing the annotations. A third measure is the agreement of a single annotator over the re-annotation of the same material, called *intra-annotator (intra-coder)* agreement, which ensures that the annotations are reproducible. This is the test-retest reliability measure used in general statistics, which asserts that at different points in time, an instrument should produce similar scores in order to be accepted as reliable (Field, 2009).

A simple statistic of inter-annotator agreement is *percent agreement*, where the percentage of agreed annotations by the annotators is presented. Many early studies in computational linguistics (Chamberlain, Kruschwitz, & Poesio, 2009; Marcus et al., 1993; Passonneau & Litman, 1993; Prasad et al., 2008) have utilized this measure to present the validity of the judgments used. However, this measure can be biased to yield higher agreement since it is not corrected for chance agreement and thus, it is better to use chance-corrected measures to calculate agreement instead (cf. Artstein & Poesio, 2008; Carletta, 1996).

For this purpose, most of the recent studies in this field (Mírovský et al., 2010; Palmer et al., 2005; Poesio, Chamberlain, Kruschwitz, Robaldo, & Ducceschi, 2013) use the κ, the K or α measures since these measures remove annotator bias, where κ refers to Cohen's Kappa statistic (Cohen, 1960), K is the kappa statistic discussed by Siegel and Castellan (1988) referring to Fleiss' Kappa[2] (also argued for by Carletta, 1996) and α is Krippendorff's (Krippendorff, 1995, 2004) alpha measure (Artstein & Poesio, 2008).

A different statistic, the *exact match* criterion is used in studies such as the PDTB (Miltsakaki et al., 2004a; Miltsakaki, Prasad, Joshi, & Webber, 2004b; Prasad, Miltsakaki, Joshi, & Web-

[2] Although Fleiss' terminology calls this measure Kappa, Artstein and Poesio (2008) express that the coefficient of agreement proposed by Fleiss (1971) for multiple coders is actually a generalization of Scott's π. We will use Fleiss' terminology and refer to this coefficient as Kappa but use the capital letter K for shorthand.

ber, 2004) and HDRB (Kolachina et al., 2012) for their annotation span agreement assessments. They argue that the kappa-like coefficients require classification in discrete categories and cannot be applied to spans of text of indeterminate length.

The agreement statistic used to calculate the agreement measurements for the TDB over the spans of the two arguments of each relation annotated is Fleiss' Kappa (Fleiss, 1971) as suggested to be one of the valid and sound methods for the calculation of agreement of the TDB annotations in Yalçınkaya (2010). The other measure suggested by Yalçınkaya (2010) was Krippendorff's alpha, which will be used to compare and contrast the results of the Fleiss' Kappa measure as in Section 4.3.2. We use the methodology described in Zeyrek et al. (2013) to discretize the argument text spans, so that these kappa-like statistics can be applied to our data (see below).[3]

The formula used to calculate Fleiss' Kappa is given in Equation 4.1, where A_0 denotes the observed agreement and A_e denotes expected agreement.[4]

Equation 4.1. Fleiss' Kappa Measure

$$K = \frac{A_0 - A_e}{1 - A_e} = \frac{The\ degree\ of\ agreement\ actually\ attained\ above\ chance}{The\ degree\ of\ agreement\ attainable\ above\ chance}$$

The agreement over the text spans of the two arguments of the relations annotated was calculated using agreement tables similar to Fleiss' (1971, p. 379) for each argument, where the **items** (markables) are the words of a text span which are decided to be in one of two **categories** {k | k ∈ {0,1}} (i.e. exclude/select) and the number of judgments by at least two **coders** {c | c ∈ {1,..n},n≥2} are recorded for each category. The agreement over the first and last words (denoting the boundaries) of each text span selected as an argument by separate annotators was measured.

Table4.1: Degree of agreement for Kappa measures

K–value < 0.0	poor agreement
0.0< *K*–value < 0.2	slight agreement
0.2< *K*–value < 0.4	fair agreement
0.4< *K*–value < 0.6	moderate agreement
0.6< *K*–value < 0.8	substantial agreement
0.8< *K*–value	perfect agreement

The results of Kappa agreement measures are considered to indicate the degree of agreement as given in Table 4.1 (Artstein & Poesio, 2008; Landis & Koch, 1977). Arstein and Poesio (2008) consider a *K*–value < 0.80 to be unacceptable for inter-rater agreements. However, Spooren and Degand (2010) note the difficulty of achieveing high agreements in studies involving coherence relations and argue that low agreements do not always indicate inadequacy

[3] For a detailed critical evaluation of Miltsakaki et al. (2004a) see Yalçınkaya (2010, sec. 3.1.5 pp. 31-34).

[4] We adopt Artstein and Poesio (2008)'s notation A0 and Ae instead of P(A) and P(E), respectively, in order to eliminate any confusion arisen by the use of the letter P.

of annotation guidelines but may reflect the underdeterminacy of language due to the different mental interpretations of text by the annotators. Hence they relax this standard and suggest that K–value > 0.70 can be accepted for coherence relations and if this level cannot be reached, researchers should account for the reasons behind the low agreement. In this thesis we use the 0.80 value as a clear indicator of agreement, however we do not dismiss lower results discussing linguistic reasons behind them.

The agreement scores were determined by preparing span-wise agreement of word boundaries as suggested in Yalçınkaya (2010) and using this data to calculate the Fleiss' Kappa statistic. An initial set of agreement scores for common annotations (i.e. annotations of relations that have been agreed to be discursive by all the annotators) were determined using the Rater Agreement Tool (RAT) (Yalçınkaya, 2010), which is a computer software to calculate inter-annotator agreement. A second set of overall agreement scores were established, where the annotated relations having discontinuous argument spans and set-theoretic differences[5] of the annotated relations were also taken into consideration. This was a methodology chosen to set apart disagreements on deciding the 'discoursehood' of the relations (i.e. discursive or non-discursive) and disagreements caused by discontinuous span selections, from span disagreements of relations whose discursiveness has been agreed by all. As a novel contribution of this thesis a computer program was written to include discontinuous argument spans into the agreement calculations, as well as to compare uncommon relations (i.e. relations that have not been agreed to be discursive by all annotators) annotated by at least one of the annotators. Furthermore, other original contributions of this work include the calculation of all agreement statistics (i.e. inter-annotator, intra-annotator, gold standard agreements) and extra evaluators (i.e. precision, recall and f-measure) for all the annotations of TDB 1.0.

In the following subsections, first, the two means of agreement score calculation will be explained. Then, the inter-annotator agreements and agreements of the independent annotators with the gold standards will be presented, followed by the explanation of the procedure for re-annotating the TDB in order to calculate intra-annotator agreements and the presentation of the intra-annotator agreement scores along with the gold standard agreements for the re-annotation. The calculation of some extra measures to evaluate the TDB will be discussed in Section 4.5. Then, in Section 4.6, a comparison of the two methods used to calculate the Kappa agreement score will be presented. Finally in Section 4.7, we conclude with a methodology proposal for the assessment of annotated corpora.

4.2 Two-Way Methodology to Calculate the Kappa Agreement Scores

The methodology we propose involves two separate means of calculating the agreement scores. The first method, which we will call the *Common Arguments* approach (henceforth the *Common* approach), to calculating the Fleiss' Kappa statistic for the two arguments of a discourse relation annotated in the TDB involves the identification of argument spans of a discourse relation, which has already been agreed to be discursive by all the annotators. This was the method used in Yalçınkaya (2010) to calculate inter-annotator agreement. Using

[5] Set-theoretic difference A-B (Partee, Wall, & Meulen, 1990, pp. 14–16) is defined in this case, as the set of all relations that are annotated by annotator A but not by annotator B, where a set is the group of relations annotated by an individual annotator.

RAT[6], the agreements are calculated on word boundary data of these argument spans. This Common method considers an intersection (A$\bigcap$B) of the relations annotated by the annotators, i.e. the relations marked to be discursive by all the annotators involved in the agreement were compared.

A secondary calculation incorporates the agreement evaluation of arguments with discontinuous spans, as well as the agreement evaluation of the set difference of the relations annotated by separate annotators, i.e. relations that are marked to be discursive by at least one annotator. This we will refer to as the *Overall* approach. Since the agreement scores were calculated using argument spans, incorporating all relations that are annotated by at least one annotator into the agreements, is important in terms of the final assessment of the TDB as it also involves the agreement for relation vs. non-relation distinction[7]. Hence, the Overall method compared all the relations annotated, even if it was found to be discursive by only one of the annotators.

In the TDB, 833 relations have discontinuous argument spans (574 relations have discontinuous arg1, 252 relations have discontinuous arg2 and 7 relations have both). The Overall approach captures the agreement over these relations as well as the remaining 7650 relations with non-discontinuous spans. In terms of a final assessment of the annotated corpus, it is essential to validate and account for all the relations annotated, and this is the intended result by the Overall method.

In order to calculate the annotator agreements in the Overall approach; first, the argument spans of the annotated discourse relations were tokenized (unitized/discretized) by selecting words as our markables (i.e. items) and categorizing each markable as belonging to category 1 (i.e. included in the argument boundary) or category 0 (i.e. excluded from the argument boundary). Here, the word boundaries of the selected arguments are considered for agreement. This procedure is done for all the annotations of each annotator for Arg1 and Arg2 separately, taking the extended common text span of the arguments annotated[8]. For the relations with discontinuous spans each word boundary of the separate parts are taken. A software program was written, where the annotations of different annotators over the same relation are compared to determine the extended common text span and are discretized. If one or more annotators have not annotated a particular relation, then their discretized annotations show that no span was selected for that relation (i.e. marked with 0s). After the discretization process, the software prepares an output file for each connective and argument pair (e.g. ama_arg1, ama_arg2), which contains the discretized annotations of each annotator listed in separate columns. Both character boundary and word boundary comparisons are prepared in discretized format as output. However, in this thesis we present only the word boundary agreement results. Finally, the Fleiss' Kappa statistic is calculated on this data with a readily available software package for statistical analysis.[9]

[6] Only the Fleiss Kappa for word boundary agreement feature of the RAT was used in the initial approach, as it was one of the suggested approaches for the TDB (Yalçınkaya, 2010, pp. 96, 107), however, there are other statistical methods and coding and context units available in the tool.

[7] The agreement of relation vs. non-relation decision may also be calculated separately as in Palmer et. al (2005), where they distinguish the role identification (role vs. non-role) and role classification (Arg0 vs. Arg1 vs. ...) and calculate Kappa for each decision separately.

[8] Detailed explanation of the discretization can be found in Yalçınkaya (2010).

[9] The IBM SPSS Statistics (Version 21) software's STATS FLEISS KAPPA extension bundle is used for this purpose.

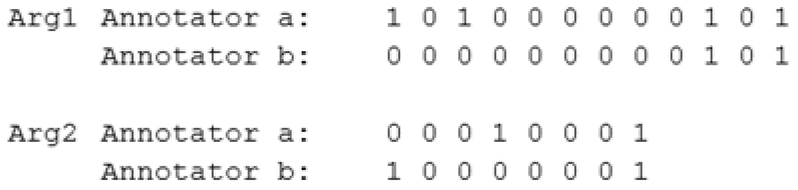
(44)

a. *Aşıklı Höyük'deki yerleşmenin,* **o zamanki bitki örtüsünü belirlemek amacıyla** *polen analizi yapılmıştır.*

Of the settlement in Aşıklı Mound, **in order to determine the vegetation at that time***, pollen analysis was done.*

b. Aşıklı Höyük'deki yerleşmenin, o zamanki bitki örtüsünü belirlemek amacıyla *polen analizi yapılmıştır.*

(00013212.txt)



```
Arg1 Annotator a:    1 0 1 0 0 0 0 0 0 1 0 1
     Annotator b:    0 0 0 0 0 0 0 0 0 1 0 1

Arg2 Annotator a:    0 0 0 1 0 0 0 1
     Annotator b:    1 0 0 0 0 0 0 1
```

Figure 4.1: Discretization of Arg1 and Arg2 for Example 9

An example discretization for the two annotations of (44) is provided in Figure 4.1, where *Annotator a* selected a discontinuous span for Arg1, and *Annotator b* selected a continuous span. The extended common text span for the two annotators for Arg1 starts at the beginning of the sentence at "*Aşıklı*" and ends at the end of the sentence at "*yapılmıştır*", since *Annotator a* has included "*Aşıklı Höyük'deki yerleşmenin*" into the selection of Arg1. Hence there are 12 items in the extended text span of Arg1 in this case. The first three items are included in *Annotator a*'s discontinuous selection as part1 and the first and third items are marked as boundaries (i.e. 1). The next 6 items are intervening text, since only Arg1 is considered. Then the final three items are again included in *Annotator a*'s selection as part2 and the 10th and 12th items are marked as boundaries (i.e. 1). Since *Annotator b* only selected the last three words as Arg1, all but the 10th and 12th items are marked as 0 (i.e. unmarked). Similarly for Arg2, the extended text span is selected as "*Aşıklı. . . belirlemek*" having 8 items, where all but the boundaries of the selections are marked with 0, and the boundaries are marked with 1 for each annotator separately as in Figure 4.1.

4.3 Inter-annotator and Gold Standard Agreements

An overview of TDB annotations reveal that 45% of the overall annotations in the TDB (3804 relations) were annotated by three independent annotators, 47% of the annotations were by pair annotation (3985 relations) and 8% (694 relations) were by group annotation (Zeyrek et al., 2013). In order to assess the confidence in these manual annotations, the agreement between the annotators for the annotations of the same material has been calculated as previously explained. For this purpose, the Kappa measure was used for the two argument text spans of the relations annotated in the TDB. The inter-annotator Kappa agreement values are presented in Appendix C (Table C.1), where the number of annotators specifies if the annotations were done by 3 independent annotators (represented with 3), by pair annotation procedure (represented with 2), or by group annotation (represented with 1). The agreement for the pair annotation is calculated by taking the pair of annotators as a single annotator and comparing their agreement with the independent annotator. No agreement could be calculated for the annotations done by group work, since there were only the agreed annotations available. Some of the connectives had too few discourse-relations annotated; hence agreement for these could not be calculated either. 15 of the search tokens were annotated as a group, 14 of the search tokens revealed too few annotations for agreement to be calculated; for 22 of the search tokens part of the relations were annotated by pair annotation and for 31 of the search tokens there were 3 independent annotations available.

The inter-annotator agreement results showed that 20 of the search tokens (*ama* ’but/yet’, *ardından* ’afterwards’, *böylece* ’thus’, *bu yana* ’since this time’, *çünkü* ’because’, *ister* ’either..or’, *ne..ne* ’neither..nor’, *nedeniyle* ’due to the reason’, *nedenle* ’for (this/that) reason’, *önce* ’prior to’, *örneğin* ’for example’, *ötürü* ’due to’, *sayede* ’thanks to (this/that)’, *sonra* ’after’, *ve* ’and’, *veya* ’or’, *ya da* ’or’, *yüzden* ’due to’, *yüzünden* ’since’, and *zaman* ’when’) revealed perfect agreement (above 0.80 K-value) for both argument spans (3 Individual comparisons and 15 Pair Annotation comparisons), 14 search items (*ama**[10]’but/yet’, *amaçla* ’with this aim of’, *amacıyla* ’with the aim of’, *dahası* ’furthermore’, *dolayısıyla* ’in consequence of”, *fakat* ’but’, *halde* ’inspite of’, *hem*$^{+}$[11]’at the same time/both..and’, *için* ’for’, *kadar* ’as well as’, *oysa* ’however’, *sonuç olarak* ’as a result’, *ve** ’and’ and *ya*$^{+}$ ’or/either..or’) displayed perfect agreement for only arg2 and the connective *zamanda* ’at the same time’ presented perfect agreement for solely its first argument. The other arguments of the connectives having perfect agreement for one of their arguments (either Arg1 or Arg2), also presented substantial agreement (i.e. 0.60< K-value<0.80). Twenty comparisons for 17 search tokens displayed less than perfect inter-annotator agreement for both of their arguments, however, four of these presented perfect agreement in other inter-annotator comparisons (*ama* ’but/yet’, *önce* ’prior to’, *sonra* ’after’, *ya da* ’or’) and twelve displayed not perfect but substantial agreement for both arguments. The remaining connectives with less than even a substantial agreement are *ama* ’but/yet’*, *gibi* ’as’, *önce** ’prior to’, *sonucunda* ’result of’, *sonuçta* ’finally/in the end’ and *yandan* ’on the one hand’.

Before discussing the reasons behind less than substantial inter-annotator agreements, it would be beneficial to first investigate the individual annotator agreements with the gold standards in order to see if the individual annotator’s annotations are in agreement with the final gold standard annotations for a given connective. This will help in identifying the reasons for disagreements. Agreement for each annotator with the gold standard annotations is presented in Appendix C (Table C.2). There were 46 search tokens, for which independent annotators’ agreement with the gold standards could be calculated and 22 search tokens, for which pair annotators’ agreement with the gold standards was calculated. 15 search tokens were annotated by group work and 17 tokens had too few annotations, hence agreement could not be calculated for them. According to the agreements with the gold data, some of the independent annotations for the connectives *gibi* ’as’, *sonuçta* ’finally/in the end’, *ya* ’or’ and *yandan* ’on the one hand’ display only fair agreement, whereas there are also moderate agreements for 12

[10] There are several inter-annotator comparisons available for the connectives marked with an *, where at least one of them shows perfect agreement for both arguments.

[11] There is only a single agreement score available for the connectives marked with a + and their use as a parallel connective as *hem..hem,* or *ya..ya.*

connectives, namely *ama* 'but/yet', *aslında* 'in fact', *ayrıca* 'in addition', *gibi* 'as', *önce* 'prior to/first', *sonucunda* 'result of', *sonuçta* 'finally/in the end', *tersine* 'in contrast', *ve* 'and', *ya* 'or', *yandan* 'on the one hand' and *yüzünden* 'since'. Eleven connectives (*ancak* 'however', *ardından* 'afterwards', *bu yana* 'since this time', *çünkü* 'because', *gene de* 'still', *ister* 'either..or', *karşılık* 'despite', *ne ki* 'howbeit', *nedenle* 'for (this/that) reason', *yüzden* 'due to' and *zaman* 'when') show perfect agreement with the gold data for all their annotators. Four connectives (*hem* 'at the same time/both..and', *nedeniyle* 'due to the reason', *nedenle* 'for (this/that) reason', and *sayede* 'thanks to (this/that)') display substantial agreement for all annotators, and the remaining 22 connectives display perfect or at least substantial agreement with the gold standards.

In the following section, we investigate the reasons behind low Kappa values observed for both inter-annotator and gold standard agreements. The disagreements for connectives with less than substantial Kappa scores will be examined to determine the implications of these reliability measures. As Reidsma and Carletta (2008) suggests, we will try to determine patterns in the disagreement of the annotators.

4.4 Investigating the Reasons Behind Low Kappa Values

4.4.1 Ama ('but/yet')

The lowest inter-annotator scores are observed for the connective *ama* 'but/yet', where one comparison for the coordinator *ama* 'but/yet' (0.01, -0.03 agreement scores for Arg1 and Arg2 respectively) displayed poor agreement (i.e. K-value <0.0). The reason behind this low agreement can be attributed to the merely moderate agreement (i.e. 0.4< K-value<0.6) of one of the annotators (Ann3) with the gold standards (see Appendix C, Table C.1), whereas the other two annotators compared (Ann5 and Ann7) present perfect agreement. In fact, an agreement score calculation between the two other annotators reveal perfect agreement (0.95 for both arguments) between each other. The poor agreement score of the three annotators may be due to the lack of adherence by Ann3 to the guidelines. The fact that other comparisons for this connective provide either perfect or substantial agreement, also corroborates our hypothesis of annotator error in this case.

The rest of the connectives with less than substantial agreement are *gibi* 'as' (0.57, 0.43), *önce** 'piror to/first' (0.58, 0.58), *sonucunda* 'result of' (0.56, 0.48), *sonuçta* 'finally/in the end' (0.51, 0.62), *ya* 'or' (0.55, 1.00), and *yandan* 'on the one hand' (0.46, 0.56), where the number in parenthesis show agreement scores for Arg1 and Arg2, respectively. In what follows, we will deal with the reasons of low Kappa scores obtained for these connectives.

4.4.2 Gibi ('as')

The subordinator *gibi* 'as' displayed one of the lowest inter-annotator agreement Kappa scores, where both of its arguments present only moderate agreement (0.4< K-value <0.6). Again, looking at the gold standard agreements reveal that the main reason behind less than substantial agreement is the merely fair agreement of the individual annotator with the gold standards (Arg1: 0.31, Arg2: 0.20). This may be due to the progressive development of the annotation guidelines especially for this connective, as some final decisions were made during the annotation process. The procedure applied for this connective was that, initially the independent / individual annotator (IA) annotated according to previous guidelines and her native speaker intuitions. Then, the pair of annotators (PA) annotated the same material for this connective separately from the IA. They made note of specific points they thought caused difficulties in the annotation of *gibi* 'as' or any annotations they thought should be discussed in an agreement meeting. The breakdown of commonly annotated relations (i.e. relations annotated by both parties) for the connective *gibi* 'as' is given in Table 4.2[12].

Table4.2: Commonly Annotated Relation Breakdown for the connective *gibi* 'as'

Annotated by	**Individual**	**PA**	**Gold**	**All**	**TOTAL**
Individual	35	92	0	103	230
PA	92	68	120	103	383
Gold	0	120	5	103	228

Another reason for the low agreement scores appears to be lack of adherence to annotation guidelines. In particular, the annotation guidelines exclude annotations of *gibi* 'as' used together with "verbs of saying/reporting" such as *söylediği gibi* 'as he/she says', or used together with "verbs of perception/cognition" such as *bilindiği gibi* 'as it is known', or "non-discursive uses" such as *-Ir/mIş gibi* 'as if' (See Table 4.3). Finally, sentence-final *gibi* were not included in the gold standards as they are predicative.

Table4.3: Syntactically frozen uses of *gibi* 'as' excluded from the gold standards

Syntactically Frozen Uses	**Individual**	**PA**	**Both**
olduğu gibi 'as it is'	-	10	39
verbs of saying/reporting + gibi (e.g. söylediği 'as he/she says')	9	-	36
verbs of perception/cognition + gibi (e.g. bilindiği gibi 'as it is known')	26	1	15
others (e.g. -mIş gibi 'as if')	-	57	3
TOTAL	**35**	**68**	**93**

It is seen that the individual annotator did not leave out uses with verbs of saying/reporting and uses with verbs of perception/cognition, whereas the PA and the final gold annotations did, and this is one of the main disagreement reasons (see example 45). All 35 relations annotated only by the individual annotator that are excluded from the golden data are of this kind.

[12] The number given in the crossing of vertical and horizontal tabs denote the number of relations annotated by both the annotator specified in the vertical tab and the horizontal tab, and by no other.

(45) Bildiğimiz gibi, efsane katına yükselmiş kişiler ortalıkta pek görünmezler, günlük kargaşanın örseleyemeyeceği bir zırhın arkasına gizlenirler.

As we know, people who have become legends are not seen around much, they hide behind an armor which cannot be crumpled by the daily havoc.

(00048220.txt)

However, there are also 92 relations annotated both by the individual annotator and the PA, which are left out of the gold standards. Such relations were another cause of disagreement as seen in (46) and (47). In (46) there are dropped elements making it difficult to decide whether it is a discursive or a non-discursive use of *gibi*. A syntactically frozen use of "istediğim gibi" 'as I want' where *gibi* takes a nominalized cognition verb as its Arg2 is observed in example (47).

(46) ..seslere de, karanlığa olduğu gibi, zamanla alışılıyordu.

.. one got used to the sounds, as one (did) to darkness.

(00001231.txt)

(47) istediğim gibi davranamıyorum.

I cannot act as I want.

(00002213.txt)

4.4.3 Sonucunda ('result of')

Another connective with moderate inter-rater agreement results for both of its arguments was *sonucunda* 'result of'. Out of the 21 separate relations annotated, only 12 are accepted to be discursive in the gold standards.

(48) Türkiye, neler pahasına yaratmış olduğu birikimini, sermayenin kısa süreli çıkarları uğruna seçtiği plânsız gelişme sonucunda Körfez'e gömmüştür.

Turkey, as a result of the structureless development chosen for the stock's short term interests, has buried the savings it has created at the expense of the world, in the Gulf.

(00018112.txt)

There are 3 relations annotated by Ann1 and Ann5, but not by Ann2. In these cases *sonucunda* 'result of' binds an abstract object (AO) and an NP. For example, in (48), the two annotators misinterpreted the nominalizing –me suffix of "gelişme" 'development' as a verb forming suffix, in violation of the annotation guidelines. Five other relations are annotated by only one of the annotators and present similar misinterpretations of the nominalizations, where "bu çarpma sonucunda" 'result of this collision', "inceleme sonucunda" 'result of this examination', "birleşme sonucunda" 'result of the merger, "çalışması sonucunda" 'result of his study' are selected as one of the arguments. This may be due to the fact that the guideline for the nominalizations had not been finalized at the time of the annotation of this connective, and only after the agreement meetings the group set the final standard for distinguishing the nominalizations from the AOs.

4.4.4 Sonuçta ('finally/in the end')

The connective *sonuçta* 'finally/in the end' has also displayed only moderately agreed results for its Arg1 (i.e. 0.51 K-value) and barely substantial agreement for Arg2 (i.e. 0.62 K-value). All the annotators showed moderate agreement with the gold standards for this connective. There are 7 relations annotated by all the independent annotators and accepted to be gold standards, where only 5 of these have been annotated by all (complete agreement for 2 relations and partial overlap for the other 3). Five relations have been annotated by the independent annotators which are excluded from the gold data (49-53), and yet another 3 added to the gold standards after the agreement meeting.

Three relations were annotated by all the annotators (49, 50, 51), where all three annotators annotated (50) differently but with overlapping annotations, and one of the annotators made a span selection error in selecting the connective (51). Another relation was annotated by only two of the annotators (52) and one by only a single annotator (53).

(49) O dönem çok da ciddi olmadığım bir erkek arkadaşım vardı. Ama sevgilimdi sonuçta...

I had a not so serious boyfriend at the time. But he was my lover in the end. . .

(20240000.txt)

(50) Genetik suç, henüz ne kanıtlanmış ne de belirlenmiş bir olgu değil. Bunlar sadece hipotez ama sonuçta insanlar hiçbir şekilde hiçbir konuda hiç bir durumlarıyla sınıflandırılmamalı diye düşünüyorum.

Genetical crime, is not a fact that has neither been proved, nor identified. These are just hypotheses but in the end I think people should not be categorized in any way about anything with any of their conditions.

(20440000.txt)

(51) Yaklaşık 1 ay boyunca böyle konuştuktan sonra yüz yüze görüşmeye karar verdik. Görüştük ve birbirimizi beğendik. Bir hafta boyunca birbirimizden ayrılmadık. Sanal cinselliği gerçeğe dönüştürdük yani. Sonuçta, birinci haftanın sonunda ben onun evli olduğunu öğrendim.

After talking like this about for 1 month we decided to meet face to face. We met and liked each other. We did not leave each other's sight for a week. We turned virtual sex to reality hence. Finally, at the end of the first week I found out he was married.

(20360000.txt)

(52) Irak'a 1991'de yapılan operasyonun faturası 100 milyar dolar oldu. Bu coğrafyada Iraklılar'la sonuçta beraber olacağız.

The cost of the operation on Iraq in 1991 was 100 billion dollars. In this geography we will be together with the Iraqi in the end.

(20510000.txt)

(53) İdeal çiftimizin dağlarda balayı şeklindeki, filmin baştan beri süregelen ’arızalı’ yapısına karşıt, alışılmış bir Hollywood romantik komedisinin alışılmış mutlu sonuna bağlanıyor ”Sekreter." Çiftimiz, fantezilerini birbirlerine yaşatarak onca kontrol edilemezliğine karşın aşklarını memnun mesut biteviye sürdüreceklerdir diğer çiftlere benzeyene kadar. Sabun köpüğü finaliyle sonuçta biraz irtifa kaybetse de bu düzeyli ve eğlendirici ”yolları kesişen yalnız egoların uyumu" çeşitlemesi, yine de meraklısını hoşnut bırakıyor.

Irrespective of the continuous ’defected’ structure since the beginning of the movie of our ideal couple’s honeymoon in the mountains, "The Secretary" ties up to a customary happy ending of a customary Hollywood romantic comedy. Our couple, will continue their love happy as ever day after day making each other live their fantasies regardless of all the uncontrollability, until they turn into other couples. With its soap opera ending in the end although it loses some altitude this refined and entertaining "lone egos that cross paths" variation, still leaves its fans content.

(10650000.txt)

As far as can be seen, the main cause for disagreement for *sonuçta* ’finally/in the end’ is the decision of discursive vs. non-discursive relation (49-53). Another factor that can be speculated to affect the agreement results would be the partial overlaps of the argument spans. Observation of the partial overlaps, especially for the selection of Arg1, suggests that the annotators had difficulty abiding by the minimality principle because they did not want to leave out certain information which they thought was relevant. Example (54) shows such a difference in Arg1 selection by the three annotators, where one annotator selected (a-c) as Arg1, another annotator selected (b-c), whereas the third annotator selected only (c), which was accepted to be the annotation for the gold standards. There was no disagreement for the span of Arg2.

(54)

(a) Ahlak kurallarının ve toplumsal değerlerin sürekli değiştiği bir ülkede yaşamaktan bıkkın ve umarsız olduk çoğumuz. Değişimin, gelişmenin bir etmeni olduğunu yadsımamakla birlikte bunca değer değişiminin gelişmeye öncü olduğu düşüncesine pek katılamıyorum.

Most of us have become hopeless and tired of living in a country where ethical rules and societal values constantly change. Acknowledging that change is a factor of development, I do not quite agree with the idea that all this change of values is a precursor of development.

(b) İnsanların kişisel çıkarları öylesine önde olmaya başladı ki, herkese, her kesime göre ayrı bir doğru kavramı oluştu. Ana babanın doğrusu çocuğun, işçinin doğrusu işverenin, varsılın doğrusu yoksulun, seçenin doğrusu seçileninkine benzemez oldu. Kişisel çıkarlarımız adına karşımızda ne varsa ezip geçmek, toplumsal çıkarlar için komşu bir ülkede olacak savaştan bile umar beklemek doğal oldu. Gençlerimiz için yalnızca varsıllığa ulaşmak değerlerin en önemlisi. Varsılın varlığını sürdürebilmesi için, yoksulun daha da yoksul olması, siyasetçinin koltuğunu yitirmeme uğruna olmadık ödünleri vermesi de yadsınamaz oldu.

People’s personal interest have become so prominent that there is a concept of truth for everyone, every fraction. The truth of the parents to the child’s, truth of the employee’s to the employer, the truth of the rich to the poor, the truth of the elector to the electee do not correspond anymore. It has become natural to run over anything infront of us in the name of our personal interests, to expect hope from war in a neighboring country for societal interests. Getting rich is the most important value for our youth. It has become undeniable for poor to get poorer in order to make rich to keep his wealth, for the politician to make inappropriate compromises to hold his position.

(c) ”Doğru insan olmak” kavramı giderek genel anlamından sıyrılıp, kişiye özel olmaya başladı.

The concept of "being an honest person" has started to be tailor-made by gradually losing its general meaning.

(d) Sonuçta toplumsal bir açmaza girdik.

In the end we have entered a societal conundrum.

(10560000.txt)

4.4.5 Yandan (’on the one hand’)

The discourse adverbial yandan ’on the one hand’ also displayed moderate inter-annotator agreement results for both its arguments (i.e. 0.46 for Arg1, 0.56 for Arg2). The gold standard agreement results for the connective *yandan* ’on the one hand’ suggest that although two annotators have at least substantial agreement with the gold data, one of the annotators presents fair and moderate agreement for Arg1 and Arg2, respectively. Inspection of the disagreements revealed that the main reason behind this is a technical issue, where the particular annotator annotated on the wrong set of text files causing a difference in the offset values of the annotations. As the annotations of the TDB are kept in a stand-off fashion, with their character offset values and the disagreement comparisons are done based on these offset values, the annotator presented low agreement results with the gold data, causing the inter-annotator agreement to also be lower than expected. Since, the stand-off annotations also keep the text data of the annotations, a brief visual inspection shows that in fact the annotators mostly agree on the discursive relations. However, the calculation of the correct results are left for future work, where the offsets of the annotator will be corrected for 54 files.

4.4.6 Önce (’prior to/first’)

The inter-annotator agreement results for the subordinator *önce* ’prior to/first’ display a 0.58 K-value for both arguments. These show that one of the comparisons present moderate agreement between the individual annotator and the PA. Observation of the gold standard agreement of this PA, reveals perfect agreement, while the individual annotator displays moderate agreement. Hence, at first glance, the disagreements can be attributed to the annotations of the individual annotator. There are 58 relations not annotated by the individual annotator, but included in the gold standards, 21 of which were annotated by the first PA (PA1) and 9 of which were annotated by the second PA (PA2). Sixteen relations annotated by the individual annotator were excluded from the gold standard annotations, and 28 relations were added to the gold data where none of the annotators annotated them.

There are hard cases where the –sE (if) suffix linking the two arguments confuses the annotators to annotate the *önce* 'prior to/before' as in (55). There are also annotations the PA or the individual annotator annotated but marked with a note indicating their hesitation. An example of this is (56), which the individual annotator simply noted her uncertainty and PA1 added a note saying the temporal relation holds for the order of thoughts not between the selected arguments, which becomes clear in the English translation.

(55) ..onu bu halde gördüğün zaman çörekotuyla karıştırabilirsin. Tabii daha önce çörek otunu öğrenmişsen.

..when you see it in this manner you might confuse it with nigella. Surely if you have learned nigella before.

(00035220.txt)

(56) Mide bulantısından nasıl kurtulacağından önce, o günün bir iş günü olup olmadığını düşünmüş.

Prior to thinking about how he would get rid of the stomach sickness, (he thought about) if that day was a work day or not.

(00060111.txt)

Another example shows an error in following the guidelines, where the individual annotator annotated a relation as in (57) which actually links two NPs *İngiltere'de* (in England) and *Paris'te* (in Paris).

(57) Amerikalı gazeteci Kressmann Taylor'ın İkinci Dünya Savaşı'ndan hemen önce, 1938'de yazdığı 'Bu Adreste Bulunamadı' adlı kitabı, önsözde de belirtildiği gibi, Neonazizmin doğup yükseldiği 1990'lı yıllarda yeniden basıldı. Önce İngiltere'de ve geçen sezon da Paris'te tiyatro sahnesine taşındı.

The American journalist Kressmann Taylor's book titled 'Not Found in This Address' which he wrote in 1938, right before the Second World War, was reprinted in the 1990's when Neonazizm was born and risen, as also mentioned in the foreword. First in England and last season in Paris it was brought to the theatre.

(10150000.txt)

The addition of independently unannotated relations to the gold standard annotations suggest a change in the guidelines which caused a systematic addition of a certain use. In fact, the modified uses of *önce* 'prior to ' such as *daha önce* 'previously' (58), *ilk önce* 'first of all' and *uzun yıllar önce* 'many years ago' of which there are 17 uses, were added to the gold data after a decision in the agreement meetings. The procedure of annotation for this connective involved first an annotation of all the relations by the individual annotator. Then, PA1 annotated half of the relations and an agreement meeting was held, after which PA2 annotated rest of the relations. Variations of the parallel uses of *önce* 'first' with *sonra* 'after', such as *ardından* 'afterwards' (59), *arkasından* 'following', *sonradan* 'later', *şimdi de* 'and now' (60) were added to the final set of annotations, adding up to 7 new relations. Note that PA2's annotations included modified uses such as *daha önce* 'previously' and parallel connective use variations, yielding better agreement results with the gold standards.

(58) *Sezer'in* {daha}$_{\text{Mod}}$ önce **"kişiye özel düzenleme" olduğu gerekçesiyle veto ettiği** *değişiklikleri onaylaması*, ..

Sezer's approving the changes that he previously **vetoed with the reason that they are "tailor-made adjustments"**.

(20220000.txt)

(59) ..{ilk}$_{\text{Mod}}$ önce, *gerçekçi yöntemden kurtulma isteğinin saydığımız roman öğelerini ne yönde etkilediğini kısaca araştırmakta yarar var.* Ardından **Latife Tekin'in bu "yeni bir biçim geliştirme" çabasının nasıl sonuçlandığına ve bize ne tür bir roman kazandırdığına bakmak istiyorum**.

..first of all, *it is beneficial to briefly investigate in what way the desire to get rid of the realist method affects the novel constituents.* Afterwards **I would like to look at how Latife Tekin's effort to "develop a new format" resulted and what type of novel it brought us**.

(00026131.txt)

(60) Önce **sıfır zam aldık**, şimdi {de}$_{\text{Mod}}$ *lojmanları verdik.*

First **we got no raise**, and now *we gave up the housing.*

(20620000.txt)

Apart from these corrections, there are also 4 sentence-medial *önce* 'first' uses added to the gold standards as in (61) as a result of agreement meetings.

(61) **Adını** önce **İngiltere ve Amerika'da duyuran** *Çağlayan, Paris'te yükselen yıldızını, başkentin gözde konser salonlarından 'Salla Gaveau'daki çılgın defilesiyle 'en iyiler' arasına yazdırdı.*

Çağlayan, **who** first **became known in England and United States**, *had his star risen in Paris, among the 'best' with his mad fashion show in one of the top concert halls of the capital 'Salla Gaveau'.*

(20210000.txt)

A final cause of lower agreement regarding *önce* 'first' was the annotations interpreted as non-discursive by the individual annotator, but as discursive by the PA, which were included in the gold standards. An example is given in (62), where the issue that may have confused the individual annotator could be the seemingly parallel construction of *önce..sonra* 'before.. after' with *sonra* 'after' taking another argument before being linked to *önce* 'first'. Hence, in the final annotations the whole *sonra* 'after' relation is taken as the second argument of *önce* 'first'.

(62) *Halimi görünce* önce *korkan*, **anlattıklarımı dinledikten sonra ise üzülen** anneannem, ..

My grandmother who first *got scared seeing my situation*, **after listening to what I told was sad**..

Overall, the majority of the disagreements for *önce* 'prior to/first' appear to stem from guidelines being updated along the annotation process. The difficulties posed by this particular connective had not been foreseen and hence the need to update the annotation guidelines was deemed necessary.

4.5 Inter-Annotator Results of Krippendorff's Alpha Measure

In order to compare and contrast the Fleiss' Kappa calculated for inter-annotator agreements of the TDB, Krippendorff's alpha was also calculated. This measure was the other suggested measure to be used in agreement calculations of the TDB in Yalçınkaya (2010) and it is a chance-corrected measure used to calculate inter-annotator reliability in many recent studies in this field.

Krippendorff's alpha (α) is developed in the field of content analysis to measure agreement between coders (Krippendorff, 2011) as briefly mentioned in Section 4.1 above. The formula of Krippendorff's alpha used in our calculations is given in Equation 4.2 below (as given in Krippendorff, 2011). Krippendorff's alpha uses the observed disagreement and the expected disagreement (i.e. "disagreement one would expect when the coding of units is attributable to chance rather than to the properties of these units" as stated in Krippendorff, 2011, p. 1), whereas Fleiss' Kappa uses the observed agreement and expected agreement.[13]

Equation 4.2. Formula for Krippendorff's alpha

$$\alpha = 1 - \frac{D_0}{D_e} = 1 - \frac{Observed\ disagreement}{Expected\ disagreement}$$

An α value of 1 indicates perfect reliability and α=0 indicates no reliability. For reliability considerations Krippendorff (2011) defines α's range as:

$$1 \geq \alpha \geq 0 \begin{cases} -Systematic\ disagreement \\ \pm Sampling\ errors \end{cases}$$

The observed and expected disagreements for each argument was calculated where the **units** are the words of a text span assigned one of two **categories** {k | k ∈ {0,1}} (i.e. exclude/select) by at least two **coders** (i.e. annotators). As in our Kappa calculations, word boundary agreement over the first and last words of each argument text span selected by separate annotators was measured. The same discretized output of the software program decribed in Section 3.2 for word boundary comparisons of the Overall method are used in the Krippendorff's alpha calculations. This output data is input into the aforementioned software package for statistical analysis, this time using the KALPHA extension of Hayes and Krippendorff (2007) to calculate Krippendorff's alpha. The acceptable threshold value for Krippendorff's alpha is taken as above 0.80, similar to the threshold for Fleiss' Kappa (Artstein & Poesio, 2008).

[13] For a detailed comparison of Fleiss' Kappa and Krippendorff's alpha measures for nominal data see Artstein and Poesio (2008).

The results of Krippendorff's alpha measure calculations for inter-annotator agreements of TDB 1.0 are provided in Appendix G (Table G.1). These results show that there are no systematic disagreements among the annotators and except for one comparison of the connective *ama* 'but/yet', they are nearly the same values with the Fleiss' Kappa calculations (i.e. within 0.10 difference value, not affecting the reliability interpretation). The last comparison for the connective *ama* 'but/yet' displays a greater value for α (i.e. 0.01, -0.03 κ–value and 0.44, 0.42 α –value for arg1 and arg2 respectively), but still represents a similar poor agreement, for which the reasons were explained in Section 3.3.1. Hence, the Krippendorff's alpha results corroborate our previous findings using Fleiss' Kappa.

4.6 Re-Annotating the TDB to calculate Intra - annotator Agreements

Intra-annotator agreements are calculated to ensure that the annotations are reproducible, indicating the test-retest reliability in general statistics terms. This measure has not been used in the reliability evaluations of annotated corpora, as far as we know, except for the Sporleder and Lascarides (2008) study, where they used it along with inter-annotator agreements to test the reliability of the manual labelling of relations. In order to determine the integrity of the independent annotations of the TDB over time, a certain portion of the independent annotations was re-annotated after more than two years from the initial annotations by one of the annotators following the guidelines. The author was the independent annotator chosen in this case. The portion to re-annotate was determined as 20 percent of the original independent annotations done by the same annotator, as was done in the study by Sporleder and Lascarides (2008).[14] The annotator/author is an experienced annotator as she had previously independently annotated 2860 relations in the TDB and participated in the development of TDB at all stages. The number of files and relations for each discourse connective of the original annotation of the particular annotator and the intended number of relations to be re-annotated (calculated as a percentage of the initial annotation counts) is given in Table 4.4. A total of 569 relations for 30 connectives were re-annotated.

Table4.4: Original and re-annotation relation counts

Connective	Gloss	Original		Re-Annotations	
		# of Files	# of Relations	# of Relations (Goal)	Actual # of Relations
ama	but, yet	35	308	62[15]	94
amacı ile	with the aim of	1	1	0	0
amacıyla	with the aim of	50	67	13	13
amaçla	with this aim of	11	11	2	2
çünkü	because	128	304	61	68
dahası	furthermore	9	12	2	2
dolayı	owing to	17	24	5	5

[14] In Sporleder and Lascarides (2008) 200 re-annotations were done to assess an initial 1051 annotations. This was taken to be roughly 20 percent of the initial annotations.

[15] Twenty percent of the number of primary relations annotated was rounded to the closes integer to determine the necessary re-annotation counts. Since their initial relation counts were too low, some of the connectives were not re-annotated. 8 connective forms were discarded in the re-annotation in this way.

Table 4.4 (continued)

Connective	**Gloss**	**Original**		**Re-Annotations**	
		# of Files	**# of Relations**	**# of Relations (Goal)**	**Actual # of Relations**
dolayısı ile	in consequence of consequently	1	1	0	0
dolayısıyla	in consequence of	46	67	13	13
ek olarak	in addition to (this)	1	1	0	0
gerek	both..and	2	2	0	0
ha	either..or	2	2	0	0
hem/ hem..hem	at the same time / both..and	63	101	20	20
için	for, so as to, for (this/that), for..for	78	365	73	141
ister	either..or	5	6	1	1
karşın	despite	36	46	9	12
mesela	to exemplify	11	12	2	2
ne..ne	neither..nor	40	51	10	10
oysa	however	73	134	27	28
önce	prior to, first	63	97	19	33
örneğin	for example	42	64	13	13
örnek olarak	to illustrate	2	2	0	0
sonra	after	62	257	51	86
sonucunda	result of	11	15	3	5
sonuç olarak	as a result	6	6	1	2
sonuçta	finally, in the end	11	11	2	3
söz gelimi	for instance	1	1	0	0
sözgelimi	for instance	2	5	1	5
taraftan	on the other hand	4	4	1	1
tersine	in contrast	9	10	2	2
ve	and	68	664	133	234
veyahut	or	1	4	1	4
ya	or	8	9	2	2
ya da	or	36	59	12	13
yahut	or	2	3	1	2
yandan	on the one hand	51	68	14	15
yoksa	otherwise	45	66	13	13
TOTAL		1033	2860	569	844

4.6.1 The Re-Annotation Procedure

In order to randomize the re-annotation process, for each connective initially annotated, a list of the annotated files was created and a file was chosen randomly to re-annotate. The randomization was done by generating a random number between 1 and file count for that connective.[16] Then, the file enumerated by this random number was selected to re-annotate. If a certain file was already re-annotated, a new random number was generated. Since each file may contain a different number of relations, a new file was chosen until the desired relation count was achieved. After the anticipated counts for all connectives were reached, a total of 844 relations were re-annotated (Table 4.4[17]), amounting to 606 relations annotated independently twice by the annotator. We will refer to the original annotations of the annotator as the *primary annotations* and the re-annotations as the *final annotations*.

4.6.2 Intra-annotator and Gold Standard Agreements for the Re-Annotation

In order to assess the annotator's annotation stability over time, intra-annotator agreement for the primary and final annotations was calculated. The agreement statistic used for this purpose was chosen as word boundary Fleiss' Kappa as in the inter-annotator agreements. The Overall method was used to calculate these statistics using the discretization software developed by the author for the doubly annotated files. The primary and final annotations were fed into the software as annotations conducted by two separate annotators (i.e. annotator 1: primary, annotator 2: final annotations). The results are presented in Table 4.5, where Kappa values above 0.80 are given in bold face and the table is grouped with respect to arguments. For seven discourse connectives (*dahası* 'furthermore', *dolayı* 'owing to', *ister* 'either..or', *mesela* 'to exemplify', *sonuç olarak* 'as a result', *taraftan* 'on the other hand' and *ya* 'or'), complete agreement is observed for both arguments. An anticipated value of above 0.80 (Artstein & Poesio, 2008) was found for both arguments of six connectives (*ama* 'but/yet', *amacıyla* 'with the aim of', *çünkü* 'because', *dolayısıyla* 'in consequence of', *ne* 'neither..nor', and *ve* 'and'). The second arguments of five other connectives (*amaçla* 'with this aim of', *tersine* 'in contrast', *veyahut* 'or', *ya da* 'or' and *yahut* 'or') displayed perfect agreement, whereas their first arguments showed values between 0.60 and 0.80, except for *yahut* 'or', which showed no agreement for its first argument. Eight connectives (*hem* 'at the same time/both..and', *için* 'for', *karşın* 'despite', *oysa* 'however', *örneğin* 'for example', *sonuçta* 'finally/in the end', *yandan* 'on the one hand' and *yoksa* 'otherwise') displayed above 0.80 threshold values for their second arguments, whereas the first arguments showed above 0.65 values for all but *sonuçta* 'finally/in the end', which had only 0.47 agreement. (See Table 4.5). Amongst the remaining connectives, *sonra* 'after' had above 0.60 agreement for both arguments, *sonucunda* 'result of' and *sözgelimi* 'for instance' had above 0.80 agreement for its first argument and above 0.60 for its second argument. Finally, there was a 0.65 agreement for the first argument of *önce* 'prior to/first' and only a 0.55 agreement for the second argument.

[16] A generic random number generator was used for this purpose, which is available online at random.org.

[17] Due to an initial procedural error, some files that were not included in the initial independent annotations of the particular annotator were annotated. This resulted in 44 additional files containing 238 relations being re-annotated.

Table4.5: Intra-annotator agreement Kappa measures for the re-annotated relations

K > 0.80 for	**Connective**	**Gloss**	**Overall**		
			Arg1	**Arg2**	**# of Annotations Compared**
Arg1 and Arg2	ama	but/yet	**0.80**	**0.87**	73
	amacıyla	with the aim of	**0.82**	**0.95**	13
	çünkü	because	**0.82**	**0.89**	69
	dahası	furthermore	**1.00**	**1.00**	2
	dolayı	owing to	**1.00**	**1.00**	5
	dolayısıyla	consequently	**0.83**	**0.95**	14
	ister	either..or	**1.00**	**1.00**	2
	mesela	to exemplify	**1.00**	**1.00**	2
	ne	neither..nor	**0.85**	**1.00**	10
	sonuç olarak	as a result	**1.00**	**1.00**	2
	taraftan	on the other hand	**1.00**	**1.00**	2
	ve	and	**0.80**	**0.91**	147
	ya	or, either..or	**1.00**	**1.00**	2
Arg2	amaçla	with this aim of	0.64	**1.00**	2
	hem/ hem.. hem	at the same time/ both.. and	0.75	**0.91**	22
	için	for, so as to	0.76	**0.88**	78
	karşın	despite	0.78	**0.83**	9
	oysa	however	0.68	**0.90**	28
	örneğin	for example	0.78	**0.80**	13
	tersine	in contrast	0.64	**1.00**	2
	veyahut	or	0.73	**1.00**	4
	ya da	or	0.66	**1.00**	13
	yandan	on the one hand	0.77	**0.92**	15
	yoksa	otherwise	0.73	**0.90**	13
	sonuçta	finally	0.47	**0.80**	3
	yahut	or	-0.2	**1.00**	2
Arg1	sonucunda	result of	**0.83**	0.73	5
	sözgelimi	for instance	**0.82**	0.67	4
Neither	önce	prior to	0.65	0.55	20
	sonra	after	0.68	0.75	56
	TOTAL				632

Table4.6: Kappa measures for Agreement of the re-annotated relations with gold standard annotations

K > 0.80 for	**Connective**	**Gloss**	**Overall**		
			Arg1	**Arg2**	**# of Annotations Compared**
Arg1 and Arg2	ama	but/yet	**0.84**	**0.91**	71
	amacıyla	with the aim of	**0.89**	**1.00**	13
	çünkü	because	**0.88**	**0.91**	72
	dahası	furthermore	**1.00**	**1.00**	2
	dolayı	owing to	**1.00**	**1.00**	5
	hem/ hem..hem	at the same time/ both.. and	**0.92**	**0.94**	20
	için	for, so as to	**0.84**	**0.92**	76
	ister	either..or	**1.00**	**1.00**	2
	karşın	despite	**0.93**	**0.92**	9
	mesela	to exemplify	**1.00**	**1.00**	2
	ne	neither..nor	**0.85**	**1.00**	10
	örneğin	for example	**0.82**	**0.90**	13
	sözgelimi	for instance	**0.82**	**0.83**	4
	taraftan	on the other hand	**1.00**	**1.00**	2
	tersine	in contrast	**1.00**	**1.00**	2
	ve	and	**0.85**	**0.94**	145
	ya	or, either..or	**1.00**	**1.00**	2
	ya da	or	**0.81**	**0.96**	13
	yandan	on the one hand	**0.81**	**0.85**	15
	yoksa	otherwise	**0.86**	**0.81**	13
Arg2	amaçla	with this aim of	0.64	**1.00**	2
	dolayısıyla	consequently	0.74	**0.86**	13
	oysa	however	0.64	**0.93**	28
	sonuç olarak	as a result	0.74	**1.00**	2
	veyahut	or	0.21	**1.00**	4
	yahut	or	-0.2	**1.00**	2
Arg1	sonucunda	result of	**0.81**	0.62	6
Neither	önce	prior to	0.74	0.75	18
	sonra	after	0.68	0.78	54
	sonuçta	finally	0.18	0.43	3
	TOTAL				623

Agreement of the secondary independent annotations with the gold standards was also calculated with the intention of understanding the reliability of the annotator after more than two years from the initial annotations. The annotator, in fact, displayed better agreement with the gold standard annotations (see Table 4.6). Of the seven connectives for which there was complete agreement between original annotations and re-annotations, *dahası* 'furthermore', *dolayı* 'owing to', *ister* 'either..or', *mesela* 'to exemplify', *taraftan* 'on the other hand' and *ya* 'or' are in complete agreement with the gold standards as well. This means that there was complete agreement with the gold standards to begin with, and there still is after the re-annotation. However, the connective *sonuç olarak* 'as a result' displayed 0.74 agreement for its first argument with the gold standards, meaning the annotator kept loyal to her original annotations, but this was not the gold standard annotation. Similarly, *dolayısıyla* 'in consequence of' showed high agreement for both arguments with the original annotations, but slightly lower agreement values with the gold standards although still having substantial agreement (i.e. 0.74 and 0.86 for arg1 and arg2, respectively). Thirteen of the connectives, namely *ama* 'but/yet', *amacıyla* 'with the aim of', *çünkü* 'because', *hem* 'at the same time/both..and', *için* 'for', *karşın* 'despite', *ne* 'neither..nor', *örneğin* 'for example', *sözgelimi* 'for instance', *ve* 'and', *ya da* 'or', *yandan* 'on the one hand' and *yoksa* 'otherwise' present perfect agreement above the expected threshold value of 0.80 with the gold standards. Of these, *ama* 'but/yet', *amacıyla* 'with the aim of', *çünkü* 'because', *ne* 'neither..nor', and *ve* 'and' already had high agreement with the original annotations, meaning they were already close to gold standards. There was substantial agreement for the first arguments of *hem* 'at the same time/both..and', *için* 'for', *karşın* 'despite', *örneğin* 'for example', *yandan* 'on the one hand' and *yoksa* 'otherwise' with the original annotations, whereas there is perfect agreement with the gold standards. This may mean that working with guidelines that have been revised since the time of the original annotations enabled the annotator to have better agreement with the gold standard annotations. Although the connective *tersine* 'in contrast' presented only substantial agreement for its first argument with the original annotations, its re-annotation displays complete agreement with the gold standards.

The worst performance is observed for the first argument of *yahut* and the two arguments of *sonuçta* 'finally / in the end' in both the intra-annotator agreement and gold standard agreements. There were two relations annotated for *yahut* 'or', where one showed complete agreement with both the original annotations and the gold standards, and the other had partial overlap for Arg1. The original annotation of this relation had complete agreement with the gold standards for its two arguments, however, the re-annotation had very low agreement for its Arg1 span as seen in (63). Although there was partial overlap, since the number of relations compared were very low, the Kappa score was very low also. In the re-annotation the annotator was careless in identifying the shared part and was in contrast with the guidelines. This was simply an annotator error pertaining to lack of adherence to the guidelines.

(63) re-annotation: .. {Müslüman tarafından}$_{\text{Shared}}$ *hamr satma* $\underline{\text{yahut}}$ **sıkma**

original: .. {hamr}$_{\text{Shared}}$ *satma* $\underline{\text{yahut}}$ **sıkma**

(00023213.txt)

The other disagreed connective was *sonuçta* 'finally / in the end', for which 3 relations were re-annotated. This connective was found to display disagreement in the inter-annotator agreements as described above in Section 4.4.4. For one of the re-annotated relations, there was complete Arg2 agreement in all annotations, but partial overlap for Arg1. The disagreement involved the decision of length of the Arg1 span. The gold standards, in accordance with the minimality principle, selected a single sentence (c), whereas in the re-annotation 2 sentences (b-c) were selected and in the original 3 sentences (a-c) were selected all inclusive of the other (see 64).

(64)

(a) Antalya, Ankara, İzmir gibi Orta ve Batı Anadolu'daki en başarılı iller, Hakkâri, Şırnak, Ardahan gibi Doğu Anadolu'daki en başarısız iller.

The most successful cities in Middle and West Anatolia such as Antalya, Ankara, İzmir, the least successful cities in East Anatolia such as Hakkâri, Şırnak, Ardahan.

(b) Ortaöğretimde gençler değişik düzeylerde yetiştirilmektedir.

Youngsters are raised in different degrees in secondary education.

(c) *Fen liselerinde, Anadolu liselerinde, gelişmiş kent liselerinde gençler daha iyi olanaklarla eğitilirken doğudaki okullarda öğretmen, donanım eksikliği yüzünden gençlerin yeterince eğitilemediğini herkes biliyor.*

Youngsters are educated with better opportunities in science high schools, Anatolian high schools, developed city high schools, whereas in eastern schools everyone knows that youngsters cannot be provided the necessary education due to lack of teachers, hardware.

(d) Sonuçta **76 çeşit lisede değişik ortamlarda yetişmiş gençleri aynı sınavda aynı sorularla değerlendirmek eşitlik değildir**.

In the end **it is not equality to evaluate youngsters raised in different environments in 76 types of high schools in the same exam with the same questions.**

(10390000.txt)

In the second relation re-annotated, a similar difference in span length is observed for both arguments. Again, the original annotations have longer spans and the re-annotation has more minimal spans providing partial overlap, while in the third relation, the partial overlap is only in the first argument. In the re-annotation the extra sentence of the original annotation was marked as supplementary material and the same Arg2 is annotated for both original and re-annotated versions. However, both of these relations were excluded from the gold standards. This shows that the annotator had better grasp of the minimality principle, however there was still some confusion as to the discursive uses of *sonuçta* 'finally / in the end'.

Overall, the re-annotation process displayed that the annotator was more careful in abiding by the guidelines, especially in terms of the minimality principle. This resulted in better agreement scores with the gold standard data. However, in several cases, the annotator seems insistent on her original instincts and failed to match the gold standards, especially in the identification of discursive *sonuçta* 'finally / in the end' and in the span selection of the first argument of *sonuç olarak* 'as a result'. It should be noted that the small number of relations compared also had an affect in the agreement scores, as even one partial overlap lowered the score more noticeably. For example, in the case of *sonuç olarak* 'as a result' only one of the relations re-annotated had only a partial overlap of Arg1 while the rest were in complete agreement, resulting in less than perfect agreement for Arg1 of this connective.

4.7 Using Extra Evaluation Measures: Calculating Precision and Recall for Manual Annotations

In information retrieval and database systems, the measures of reliability mainly used are precision and recall values. These measures evaluate the text-retrieval performance of systems. Precision is defined as the ratio of retrieved relevant documents to the number of retrieved documents, whereas recall is the ratio of retrieved relevant documents to the total number of relevant documents in the database (Bird, Nagappan, Gall, Murphy, & Devanbu, 2009; Can, Nuray, & Sevdik, 2004). Precision and recall measures have also been used to evaluate automatic annotation models and tools, as well as other NLP applications (Sazedj & Pinto, 2005; Sporleder & Lascarides, 2008). However, they are deemed inapplicable to evaluate manual annotation tools (Sazedj & Pinto, 2005). Since they are used to evaluate the systems, they have not been used to evaluate annotators of manual annotations. There are two exceptions, namely studies by Burstein, Marcu and Knight (2003) and Mírovský et al., (2010). In Burstein et al. (2003), precision, recall and F-measure were used to observe the relative performance among human judges. The annotation in this case involved manually labeling all sentences of an essay as belonging to one of seven specified categories. In the Burstein et al. (2003) study, precision is defined as the number of cases in which J1 and J2 agree divided by the number of cases labeled by human J2, and recall is defined as the number of cases in which J1and J2 agree divided by the number of cases labeled by J1, where J1 = human judge 1 and J2 = human judge 2. The F-measure is calculated in the usual manner as 2 × precision × recall/(precision + recall). In Mírovský et al. (2010), the F-measure is used to evaluate the agreement on existence of discourse relations. It should be mentioned that this study also uses Cohen's kappa to evaluate the agreement on types of discourse relations.

In our case, we propose to utilize precision, recall and F-measure as extra evaluators to assess an annotator's reliability when comparing independent annotations with the gold standards. In this way, we hope to get an idea of how many of the gold standard relations the independent annotator captured, regardless of their arg1 and arg2 agreement. Since the annotation in our case does not merely involve labeling as in Burstein et al. (2003) study, but involves selecting text spans, looking for exact agreement is too strict, and the agreement of the text spans is already analysed using the Kappa statistic. The intended purpose here is to capture if the annotator annotated the same connective instances to be discourse connectives as the gold standard.

The *precision of an annotator* can thus be defined as the ratio of the relations annotated by the annotator that are also in the gold standard annotations to the total number of relations annotated by the annotator, in this case. Hence, the *recall of an annotator* would be defined as the ratio of the relations annotated by the annotator that are also in the gold standard annotations to the total number of relations in the gold standard (See Equation 4.3 and Equation 4.4). Thus, these measures will provide a sense of how much of the relations annotated by the annotator are relevant, therefore are in the gold standards.

Since precision and recall measures will produce the reliability of the annotator's independent annotations with respect to the gold standards, in a re-annotation task as discussed in Section 3.4, they will produce any changes in the stability of that annotator's annotation.

Equation 4.3. Precision

$$Precision_{annotator1} = \frac{|relations\ in\ the\ gold\ standard \bigcap relations\ annotated\ by\ annotator1|}{|relations\ annotated\ by\ annotator1|}$$

Equation 4.4. Recall

$$Recall_{annotator1} = \frac{|relations\ in\ the\ gold\ standard \bigcap relations\ annotated\ by\ annotator1|}{|relations\ in\ the\ gold\ standard|}$$

We calculated these extra evaluation measures, namely precision, recall and f-measure for each annotator with respect to the gold standard annotations. These are given in Appendix E (Table E.1). Lower recall values are due to an annotator only annotating some fraction of the total files for that connective. For example, for the connective *ama* Ann3 has annotated 176 relations and looked at only 40 files, where there are a total of 1024 relations annotated in 173 files in the gold standards. This situation where an annotator did not annotate all the files for a particular connective brings forth a potential drawback if this annotator's annotations are compared with the total annotations.

This drawback led to the idea of calculating a second kind of recall value, where the relations in the gold standard are not the total annotations but the number of gold standard annotations in the file span annotated by that annotator. Thus, this new measure will look at only the span annotated by the annotator and present if the annotator extracts all the gold data in that span. Since usually, annotators only annotated partial set of files for a given connective, we take this as a more appropriate way to calculate the recall (referred to as *Recall in Span*). Hence the equation for *Recall in Span* is given below in Equation 4.5.

Equation 4.5. Recall in Span

$$RecallinSpan_{annotator1} = \frac{|relations\ in\ the\ gold\ standard \bigcap relations\ annotated\ by\ annotator1|}{|relations\ in\ the\ gold\ standard\ for\ the\ files\ annotated\ by\ annotator1|}$$

Especially for connectives with a large number of relations annotated, for which many annotators participated in the annotations, the recall in span calculations greatly differ from the initial recall calculations with the grand total of gold annotations (e.g. this difference can be drastically observed for the connectives *ama, aslında, birlikte, için, karşın, önce, sonra, ve* and *ya da*). Hence we use the Recall in Span measure instead of the Recall defined in Equation 4.4 above.

Equation 4.6. F-Measure

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

F-measures calculated using the precision and recall in span measures using Equation 4.6 show that for 4 connectives (*gibi, nedeniyle, nedenlerle, sonuçta, yüzünden)* and for one of the individual annotators in another 7 connectives (*ama, içindir, önce, örnek olarak, sonucunda, tersine* and *yandan)* there are lower agreements with the gold standards. These results coincide with some of the results obtained from the gold standard agreements of the Overall approach. Hence, this coincidence suggest that the P, R and F-measure can be used as complementary measures for the Common method, in order to obtain an approximation to the Overall method.

For example, the results for *gibi* with low precision and low recall, amounting to a lower F-measure point to the results obtained for Kappa agreements, i.e. 0.45 precision of the individual annotator shows that more than half of her annotations were not included in the gold standards, and 0.46 recall shows that there were almost as many annotations added to the gold data that the individual annotator did not annotate. Similarly for the PA 0.58 precision suggests that the annotations made by this pair were excluded from the gold data, but the high recall indicates that their annotations captured most of the gold data. This is inline with what has been discussed for the Kappa agreement results in Section 3.3.1. Similarly for *sonucunda*, there is a 0.75 F-Measure due to a 0.60 precision value for Ann5, although a perfect recall is observed; suggesting the annotator annotated other relations that were not added to the gold data. In fact, it was seen in prior investigation that there were 12 relations in the gold data, whereas 21 relations were annotated by the annotators and five of these were only annotated by Ann5, hence the lower precision.

To sum up, all these deviations in precision, recall and f-measure values can point us in the direction of the cause of disagreements in terms of the relation vs. non-relation distinction. However, these statistics do not tell us anything about the agreement in terms of the argument spans and they are sensitive to the total number of relations, presenting greater deviations for connectives with a small number of total relations, as discussed in other studies (Artstein & Poesio, 2008; Mírovský et al., 2010). The information that can be gained from these measures can be summarized as follows:

- Low precision: indicative of exclusion of annotator(s)'s annotations from the gold standards

- Low recall: indicative of annotator(s) failing to annotate relations that are in the gold standards

- Low f-measure: indicative of either low precision, low recall, or both.

Hence, precision, recall and f-measure can be used as extra measures to understand the agreement in terms of discursive-non-discursive relation distinction, but the sensitivity to the total amount of relations (i.e. low number of annotations, tend to skew the results in favor of disagreement) should be taken into consideration when making predictions. In terms of the Common method, since they provide agreement on the whole set of relations including the non-intersecting and discontinuous, these statistics can be beneficial in understanding the agreements on the relation vs. non-relation distinction over the whole set, aiding in the determination of the extent of uncommon relations. This leads us to consider them as complementary measures for the Common method. However, these extra measures cannot provide argument based agreement for the non-intersecting and discontinuous relations, which the Overall approach easily provides.

4.8 Evaluation of the Different Approaches to Reliability Calculations

Comparisons of inter-annotator agreement values and gold standard agreement values for the Common approach and the Overall approach are presented in Appendix D (Table D.1 and Table D.2). In order to get a better understanding of both approaches, the number of relations compared for each inter-annotator comparison is also presented in Table D.1. Examination of the number of relations compared for the two approaches reveal that for 26 connectives, the Overall method includes at least 10 more relations in its calculations. The most striking differences are observed for the connectives *ve* 'and' (1558), *sonra* 'after' (439), *gibi* 'as' (322), *ama* 'but/yet' (300) and *için* 'for' (114), with numbers in the parenthesis indicating the additional relations compared for inter-annotator agreement in the Overall method, where non-intersection relations and discontinuous spans are included. Thus, the difference in relations compared are due to those relations having discontinuous arguments and non-intersecting relations which have been annotated by at least one annotator but not all.

Appendix F (Table F.1) presents the information for discontinuity of the arguments for each connective in the TDB. According to this, there are a total of 833 relations which have at least one of its arguments selected discontinuously. Hence, part of the difference of relations compared for the two procedures stem from these relations with discontinuous spans. However, for the five connectives with the greatest difference in number of relations compared, discontinuous relation count is not as much; *ve* 'and' (38), *sonra* 'after' (102), *gibi* 'as' (69), *ama* 'but/yet' (27) and *için* 'for' (231). Only for *için* 'for' can the whole difference possibly be accounted for by the discontinuous relations. Hence, the rest of the difference must be due to those non-intersecting relations, which at least one annotator chose not to annotate.

The results of the Common method show us that the inter-annotator agreements of *yandan* has moderate agreement for its Arg1, whereas it has substantial agreement for Arg2. 13 other connectives (*ama**[18]'but/yet', *amaçla* 'with this aim of', *amacıyla* 'with the aim of', *ayrıca* 'in addition', *dahası* 'furthermore', *dolayısıyla* 'in consequence of', *oysa* 'however', *rağmen* 'despite', *sonucunda* 'result of', *sonuç olarak* 'as a result', *sonuçta* 'finally / in the end', *tersine* 'in contrast', *ve** 'and') display substantial agreement for either Arg1 or both arguments. Among all the connectives presenting disagreement in the Common method, 14 of them (*ama*** 'but/yet', *amaçla* 'with this aim of', *amacıyla* 'with the aim of', *ayrıca*$^{+}$ 'in addition', *dahası* 'furthermore', *dolayısıyla* 'in consequence of', *oysa* 'however', *rağmen* 'despite', *sonucunda*** 'result of', *sonuç olarak* 'as a result', *sonuçta*$^{+}$ 'finally / in the end', *tersine*$^{+}$ 'in contrast', *ve* 'and', and *yandan* 'on the one hand') were correspondingly found to exhibit disagreement in the Overall approach. On the other hand, 16 connectives (*aslında* 'in fact', *dolayı* 'owing to', *fakat* 'but', *gibi* 'as', *halde* 'inspite of', *hem* 'at the same time/both..and', *için* 'for', *kadar* 'as well as', *karşın* 'despite', *mesela* 'to exemplify', *önce* 'prior to/first', *sonra* 'after', *ya* 'or', *ya da* 'or', *yoksa* 'otherwise' and *zamanda* 'at the same time') that present perfect agreement in the Common approach inter-annotator results, displayed lower agreements in the Overall approach.

In the gold standard agreements Common method presents 15 connectives (*amaçla** 'with this aim of', *amacıyla** 'with the aim of', *aslında** 'in fact', *ayrıca* 'in addition', *dahası**

[18] * indicates that there are other annotators for the connective that present perfect agreements. ** indicates that much lower agreement was observed in the Overall approach. + indicates that the other argument was also found to have less than perfect disagreement in the Overall approach. - indicates that one of the arguments presented perfect agreement in the Overall approach.

’furthermore’, *fakat** ’but’, *kadar** ’as well as’, *karşın** ’despite’, *mesela** ’to exemplify’, *oysa** ’however’, *rağmen** ’despite’, *sonuç olarak** ’as a result’, *sonuçta** ’finally / in the end’, *ya** ’or’ and *yoksa** ’otherwise’) having substantial agreement for either Arg1 or both arguments. Two connectives (*gibi** ’as’, *yandan** ’on the one hand’) are shown to display moderate agreement for both of their arguments. The Common - gold standards agreement results show perfect agreement for 19 connectives (*dolayı* ’owing to’, *dolayısıyla* ’in consequence of’, *halde* ’inspite of’, *hem* ’at the same time / both..and’, *için* ’for’, *ne* ’neither..nor’, *nedeniyle* ’due to the reason’, *önce* ’prior to/first’, *örneğin* ’for example’, *ötürü* ’due to’, *sayede* ’thanks to (this/that), *sonra* ’after’, *sonucunda* ’result of’, *tersine* ’in contrast’, *ve* ’and’, *veya* ’or’, *ya da* ’or’, *yüzünden* ’since’ and *zamanda* ’at the same time’), whereas the Overall method presents lower agreement values for them. Hence, the disagreements for these connectives are missed by the Common procedure. Also for the connectives evaluated to have less than substantial agreement commonly by both approaches have lower values in the Overall approach. Looking at both the inter-annotator and gold-standard results of the Common method, the disagreements for *dolayı* ’owing to’, *halde* ’inspite of’, *hem* ’at the same time / both..and’, *için* ’for’, *önce* ’prior to/first’, *sonra* ’after’, *ya da* ’or’ and *zamanda* ’at the same time’ cannot be observed. It is expected that the Common method would miss some of the disagreements as the set of annotations used in the two approaches is different and the additional relations considered in the Overall method produce previously unmeasured disagreements. This suggests that for the aforementioned 8 connectives, there have either been many non-intersecting relations annotated, or many relations with discontinuous spans annotated resulting in disagreements either in between the annotators or between the annotators and the gold standards.

On the whole, our comparison of the two procedures show that inclusion of non-intersecting relations annotated by the annotators and inclusion of discontinuously annotated relations in the statistical evaluations is important in getting a clear comprehension of the disagreements of a given annotated resource. Although extra evaluation measures like precision, recall and f-measure can help in understanding disagreements due to non-intersection (i.e. relation-non-relation distinction), they cannot handle disagreements due to argument span selection. Hence, including the whole set of annotations and evaluating their agreement using the Kappa measure, as in our Overall approach, is a better method than calculating Kappa on only the common relations and trying to compensate with other evaluation metrics. However, calculating both approaches together provides a better understanding of the data and the disagreements. Hence, as a methodology to assess the reliability of corpus annotations, we suggest to use both methods to better comprehend the characteristics of the disagreements between the annotators. In this way, the sources of disagreement can be identified as either due to the discursive relation identification, or discontinuous span selection, or difference in argument span selection within common relations.

4.9 Conclusions

In this chapter, we presented our methodology for an overall assessment of the TDB 1.0. We first reviewed the descriptive statistics of the TDB and explained the agreement measures used to evaluate this resource. We then presented our two-way methodology to calculate reliability statistics: (1) calculate reliability statistics for the common set of annotations (i.e. relations identified as discursive by all annotators), (2) calculate an overall set of agreements for all the annotations including uncommon annotations (i.e. where some annotator(s) found the relation to be discursive, whereas other(s) found it nondiscursive), as well as including annotations having discontinuous spans. These two methods enabled us to set apart the sources of disagreement with more ease. The results of the agreement statistics were discussed and the reasons behind low Kappa values were identified. We also calculated inter-annotator agreement results using Krippendorff's alpha measure. These showed very similar results to our Kappa calculations. We explained the re-annotation procedure of the TDB to calculate intra-annotator agreement. Then we calculated extra evaluators, namely precision, recall and f-measure to assess an annotator's reliability when comparing independent annotations with the gold standards. Finally, we presented a discussion on calculating reliability for just common annotations with respect to calculating reliability including uncommon annotations and discontinuous spans, comparing our two methods. Our comparison of the two procedures showed the benefits of the overall method, where it was seen that the inclusion of non-intersecting relations annotated by the annotators and inclusion of discontinuously annotated relations in the statistical evaluations is important in getting a clear comprehension of the disagreements of a given annotated resource. However, as a methodology to assess the reliability of corpus annotations, we suggest to use both methods to better comprehend the characteristics of the disagreements between the annotators.

CHAPTER 5

DEMONSTRATIVE PRONOUNS IN TURKISH AND A STUDY ON AUTOMATIC IDENTIFICATION OF DEMONSTRATIVE PRONOUNS

In Chapter 4, we have put forward an assessment of the TDB. After the assessment of its reliability, a resource can be utilized in various ways: new markup can be added, the annotated data can be used to train models for automatic identification or other natural language applications. In order to establish the TDB as a gold standard resource for discourse studies to the extent explicit connectives are concerned, we must show that this resource can be exploited towards the development of efficient natural language applications and systems.

In Chapter 6, we will be utilizing this reliably annotated gold standard data to investigate methods for automatic identification of phrasal expressions in Turkish. Before presenting an effort to automatically identify phrasal expressions in the TDB, in this chapter we look at demonstrative pronouns, which are the deictic items present in phrasal expressions. We will present some background information about the concepts of anaphora and deixis, as well as examine the uses of demonstrative pronouns in phrasal expressions. The chapter introduces prior studies on extracting pronouns in other languages and in Turkish. Finally, a first effort to identify demonstrative pronouns in the Metu-Sabancı Turkish Treebank (Atalay, Oflazer, & Say, 2003; Oflazer, Say, Hakkani Tür, & Tür, 2003) by the author will be presented. The chapter ends with a discussion on the advantages and limitations of the resource used and what the outcome of this experiment means for an automatic identification of phrasal expressions.

5.1 Anaphora, Deixis and the Use of Demonstrative Pronouns in Phrasal Expressions

Demonstrative pronouns, demonstrative adjectives and demonstrative adverbs are deictic (Lyons, 1977). Lyons explains that a definiteness component exists for the demonstrative pronouns *this* meaning "the one here" and *that* meaning "the one there", as well as a distinction of proximity vs. non-proximity. The function of the demonstrative pronoun is described to be "to draw the attention of the addressee to a referent which satisfies the description implied by the use of the pronoun in terms of gender, number, status, etc." (Lyons, 1977). Demonstrative pronouns in English are not distinguished for gender, but are for number and proximity, and they have the same forms as demonstrative adjectives. When demonstratives are used as deictics, they can refer the addressee to a particular region of the environment to find the person/object that is being referred to. The deictic pronoun displays a presupposition of existence. Whether a pronoun has anaphoric or deictic reference is determined by the context-of-utterance. Linguistic entities in the co-text of the utterance may also be referred to by demonstrative pronouns. In this perspective, anaphora presupposes that the referent is already in the universe-of-discourse, whereas deixis is a means of placing entities into the universe-of-discourse so they can be referred to later.

5.1.1 Anaphora and Abstract Objects

The term *anaphora* refers to objects previously introduced into the discourse through the use of a pronoun or other linguistic entities. The previously introduced object is known as the *antecedent* or the *referent* of the anaphor. Some studies differentiate between the use of the term referent and antecedent, where the former denotes objects referenced by making inferences from textual entities present in the text (as in abstract object reference or discourse deixis), whereas the latter is reference to the linguistic expression itself, (such as the case of reference to noun phrases (NPs)) (cf. Hedberg, Gundel, & Zacharski, 2007).

Anaphoric expressions can be personal pronouns, possessive pronouns, demonstratives, named entities such as proper names or definite noun phrases. They may also refer to abstract entities such as events, facts or propositions. This last type of anaphora, where the antecedent is an abstract object is termed by Webber (1988a, 1988b) as *discourse deixis*. Asher (1993) provides an extensive examination of abstract object anaphora. He states that abstract objects (AOs) may be introduced into a discourse by constructions like verb phrases (VPs), or whole sentences, which may then be referenced by anaphoric pronouns. Asher (1993) classifies AOs as *eventualities*, which he further divides as events and states, and *purely abstract objects*, which he divides as fact-like objects and proposition-like objects. These are divided into more fine grained categories as given in Figure 5.1.

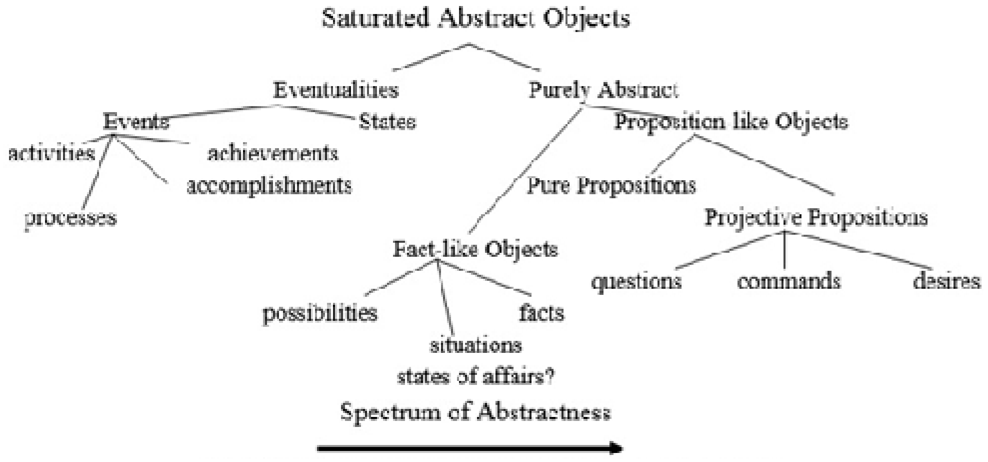


Figure 5.1: Classification of Abstract Objects in Asher (1993).

Asher also identifies the overt pronouns for abstract entity anaphora as *this*, *that*, and *it*, and specifies some differences in their anaphoric uses. For example, demonstrative pronouns are usually used to refer anaphorically to AOs that are not adjacent (i.e. more than one sentence away, or in another discourse segment), whereas *it* references AOs within the previous sentence or in the same discourse segment. He divides the referring expressions into six kinds: *that* clauses (65), infinitival phrases (66), gerund phrases (67), naked infinitive complements (68), noun phrases that denote proposition-like entities (69), clauses with parasitic gaps (70), chunks of text (71).[1]

(65) John believed [that Mary was sick]$_i$. The teacher believed *it*$_i$ too.

(reference to a that clause)

(66) Fred wanted [to go to the movies]$_i$. But his mother wouldn't allow *it*$_i$.

(reference to an infinitive)

(67) [John's hitting Fred]$_i$ got everyone in trouble, for *it*$_i$ led to a brawl. *It*$_i$ also indicated that they must have been pretty mad at each other.

(reference to a gerund)

(68) [The claim that Susan got a C on the test]$_i$ was surprising. John did not believe *it*$_i$.

(naked infinitive complement)

(69) Fred [hit a home run]$_i$, and then Sally did *it*$_i$.

(reference to a verb phrase)

(70) Fred believed that [Mary was not nice enough to try to please]$_i$. But Bill didn't believe *that*$_i$ (*this*$_i$, *it*$_i$).

(reference to a clause with a parasitic gap)

(71) The "liberation" of the village had been a disaster. [First on a sweep through the town some of the Marines had gone crazy and killed some innocent villagers. To cover up the "mistake," the rest of the squad had torched the village. To cap it off, the lieutenant called in an air strike.]$_i$ At first the battalion commander hadn't believed *it*$_i$.

(chunks of text)

In Turkish, references to noun clauses as in (65) and naked infinitive complements as in (68) are denoted with verbal complements marked with –DIK and nominal complements marked with –(s)I. The nominalizer for the infinitive exemplified in (66) by Asher is –mAk in Turkish (known as "mastar"). The gerund in (67) and the verb phrase in (69) denoting a concept are formed in Turkish using nominalized clauses with –mA and –Iş, whereas the clause with a parasitic gap in (70) is treated as a verbal complement in Turkish. The chunks of text references in (71) are also similarly represented as chunks of text in Turkish.

[1] Examples are from Asher (1993, p. 226, ex. 1.a-g).

5.1.2 Pronouns in Turkish

The independent/overt pronouns in Turkish are *ben* 'I', *sen* 'you-Singular', *o* 'he/she/it', *biz* 'we', *siz* 'you-plural', and *onlar* 'they'. Depending on their grammatical role in the sentence these pronouns get inflected for case (Erguvanlı Taylan, 1986). The subject agreement is marked on the verbal element using a person suffix as in (72). This makes it optional to use a subject pronoun in cases where the subject does not have an emphatic or contrastive function (73).

(72) *Ben* iş-e gecik-ti-m.

I work.DAT be late.PAST.1SG

I'm late for work.

(Erguvanlı Taylan, 1986, p. 210, ex.1.a.)

(73) ∅ İş-e gecik-ti-m.

work.DAT be late.PAST.1SG

I'm late for work.

(Erguvanlı Taylan, 1986, p. 210, ex.1.b.)

Examples such as (73) where the subject pronoun is left out are called *Pro drop* (or *NP drop*, or *Zero Pronoun*, or *Zero Anaphora*, or *Null Subject*) (Enç, 1986). In cases where the subject has emphatic or contrastive functions, then the pronominal form is required as in (74) where the subjects of the conjoined sentences are in contrast and independent subjects are required (Erguvanlı Taylan, 1986). Omitting the explicit pronouns in (74) makes the sentence ungrammatical. In cases where the subject of a sentence is providing new information, then the explicit pronominal subject is also obligatory.

(74) *Ben* işe geciktim ama *sen* henüz gecikmedin.

I work.DAT be late.PAST.1SG but you yet be late.NEG.PAST.2SG

I'm late for work but *you*'re not late for work yet.

(Erguvanlı Taylan, 1986, p. 210, ex.2.a.)

In Turkish, as long as the speaker/writer remains in the same topic (see below), coreference with subject NPs is expressed by zero anaphora (75). Coreference with a non-subject NP can be expressed using either zero or pronominal anaphora, but this requires that the antecedent precedes the anaphor. The presence of an overt pronoun indicates distinct reference as in (76).

(75) Erol$_i$ ∅$_i$ karısı için her şeyi, yapar.

Erol$_i$ does (will do) everything for his$_i$ wife.

(Erguvanlı Taylan, 1986, p. 213, ex.8.a.)

(76) Erol$_{i}$ *onun*$_{*i/j}$ karısı için her şeyi yapar.

Erol$_{i}$ does everything for *his*$_{*i/j}$ wife.

(Erguvanlı Taylan, 1986, p. 213, ex.8.c.)

In certain cases (e.g. when the topic changes) coreference with another NP requires an overt pronoun and dropping the pronoun makes the sentence ungrammatical (77). When an object is a required argument of the verb then an overt pronoun object is obligatory (Erguvanlı Taylan, 1986).

(77) Nazan'ın Erol'u$_{i}$ herkese şikayet etmesi *onu*$_{i}$ çok üzmüş.

That Nazan complained about Erol$_{i}$ to everyone has upset *him*$_{i}$ very much.

(Erguvanlı Taylan, 1986, p. 218, ex.16.a.)

As can be seen from the examples above Turkish is a null subject/ pro-drop language. However, we should note that in the scope of this thesis we are not interested in this fact since our main concern is the explicit uses of demonstrative pronouns in phrasal expressions.

However, we are interested in anaphoric relations (especially pronominal anaphora) extending beyond the boundaries of a sentence. In this case, the antecedent of an anaphor would depend on the discourse context. Within the sentence, we have seen that zero anaphora is resolved using the semantic and syntactic information associated with the null representation, which is recoverable/predictable (e.g. person agreement on the verb and the possessed NP aid in the resolution). Across the sentence boundary, zero anaphora is resolved through the discourse context supplying the recoverable information needed. (Erguvanlı Taylan, 1986)

Enç (1986) argues that an overt pronominal subject use in a sentence signals topic change in Turkish and a null subject sentence is just used for continuing with a previous topic. According to her analysis, subject pronouns have topic-switching function (78) and the use of Turkish pronouns generally indicate contrast (79). When topic-switching is viewed as a form of contrast, semantically redundant subject pronouns in Turkish have the sole function of contrast.

(78) *Sen* Ali'nin Ankara'ya gideceğini biliyordun.

You knew that Ali was going to Ankara.

(Enç, 1986, p. 204, ex.17)

(79) Arabayı Ahmet yıkamadı *ben* yıkadım.

Ahmet didn't wash the car, *I* did.

(Enç, 1986, p. 204, ex.18)

Enç (1986) points to the correspondence between the complexity of the surface form and the amount of information carried by the sentence, where "the more complex form with the semantically redundant pronoun provides additional pragmatic information" (*ibid.*, p.206).

Similar to Enç (1986) and Erguvanlı-Taylan (1986), Turan (1995) shows that there are discourse-level well-formedness rules for the use of null and overt pronouns and full NPs in subject position in Turkish within a Centering Theory framework. According to her analyses multiple overt pronouns with the same person and number features are ungrammatical in Turkish.

For this thesis, a more central kind of pronouns than personal pronouns are demonstrative pronouns. Demonstrative pronouns for Turkish have three main types. These are given in Lewis (1967) as *bu* 'this', *şu* 'this/that', and *o* 'that'. Göksel and Kerslake (2005) also provides the plural forms as *bunlar* 'these', *şunlar* 'these/those', and *onlar* 'those'. The different forms of these pronouns when inflected for case are given in Table 5.1.

Table5.1: Forms of demonstrative pronouns in Turkish

Case	**Singular**	**Plural**
Absolute	*bu*	*bunlar*
	şu	*şunlar*
	o	*onlar*
Accusative	*bunu*	*bunları*
	şunu	*şunları*
	onu	*onları*
Genitive	*bunun*	*bunların*
	şunun	*şunların*
	onun	*onların*
Dative	*buna*	*bunlara*
	şuna	*şunlara*
	ona	*onlara*
Locative	*bunda*	*bunlarda*
	şunda	*şunlarda*
	onda	*onlarda*
Ablative	*bundan*	*bunlardan*
	şundan	*şunlardan*
	ondan	*onlardan*

The main difference between these three pronouns is described as a difference in proximity. In the simplest sense, closer objects are referred with *bu*, farther objects are referred with *şu* and objects that are furthest away are referred with *o*. However, *şu* is often conceived to be accompanied by an ostensive gesture of pointing. Göksel and Kerslake (2005) also state that *şu* implies that the referent is newly introduced, whereas *bu* does not, and they cannot be substituted for each other. Also, the referent of *şu* may succeed it after a colon. In cases where a previously mentioned concrete item that is out of sight for both the speaker and the hearer is referred to, *o* is used. If an object in context is to be topicalized, then either *bu* or *o* can be used.

The use of demonstrative pronouns *bu* ('this) and *şu* ('this/that) in Turkish discourse were investigated by Turan (1997) on a total of 56 instances of these pronouns in the Bilkent University Electronic Database. The probable antecedents of these pronouns are identified as the previous sentence (80)[2], the previous clause (81) and the previous NPs (82).

(80) [Bugün dinletilmeye yeltenilen müzik, yüz ağartacak değerde olmaktan uzaktır]. **Bunu** açıkça bilmeliyiz. Ulusa ince duyguları, düşünceleri anlatan yüksek deyişleri, söyleyişleri toplamak, onları bir an önce günün son müzik kurallarına göre işlemek gerekir.

[The music that is being attempted to be make us listen, is far from being worth honorary]. **This** should be clearly known. High praises, utterances which express gracious emotions, thoughts to the nation should be gathered, they should be processed according to latest musical convetions as soon as possible.

(Turan, 1997, p. 202, ex.3)

(81) [Atatürk bu işe, yalandan bir ulusallık görüntüsü vermeyi], **bunun** arkasına saklanmayı, çelişkili ve küçültücü bir davranış sayıyordu.

[Atatürk] considered [giving this job a fake nationalism appearance], hiding behind **this**, as contradictory and demeaning.

(Turan, 1997, p. 202, ex.4)

(82) Arkeologların en yakıcı sorunlarından biri buluntuları tarihlemektir. [Değişik birikintileri] dikkatle inceleyerek, **bunların** hangi sırayla yığıldıklarını hesaplamak olanaklıdır. Belirli bir bölgede kullanılan [el yapımı eşyaların biçimi] zaman içinde evrimleşmiştir, **bunlar** kronolojik sıraya döküldükleri zaman bulundukları ortamın tarihinin belirlenmesine yardımcı olabilirler.

One of the most prominent problems of archeologists is dating findings. By carefully examining [different conglomerations], calculating the order of **their** accumulation is possible. [The shape of hand made goods] that are used in a specific region have evolved over time, **these** when chronologically timed may help in dating the medium/stage that they are found in.

(Turan, 1997, p. 202, ex.5)

Turan (1997) identifies the Turkish demonstrative *bu/bunlar* 'this/these' to be anaphoric, and *şu/şunlar* 'that/those' to be cataphoric. In her study, Webber's (1988c, 1991) right frontier rule denoting the boundary of text *this* and *that* can access is also accepted for Turkish *bu* ('this) pronoun and redefined for *şu* ('this/that) as its right sibling in the discourse tree structure. It is suggested that connectives such as *bu nedenle* 'for this reason', *buna karşın* 'despite this', *bundan dolayı* 'owing to this', etc. may be investigated with the right frontier rule as well.

[2] In the examples (80-82) the demonstrative anaphor is shown in bold face and the antecedent is given inside square brackets.

5.2 An Overview of Corpus Annotation Studies on Reference Relations for Languages other than Turkish

Large-scale studies, such as corpora research, have strived to shed light into the nature of reference relations, including anaphoric relations. Important studies, in which corpora were annotated for anaphora, discourse deixis and also with coreference, include Eckert and Strube (2000), Viera, Salmon-alt, & Gasperin (2005), Byron (2002), Botley (2006), Navaretta and Olsen (2007), Poesio and Artstein (2008), Recasens and Marti (2010), Dipper and Zinsmeister (2009, 2010) and Lee and Song (2010). Each of these studies is reviewed briefly below. This section concludes with a table summarizing these studies, the corpus they created and the type(s) of anaphor annotated.

Eckert and Strube (2000) annotate pronouns and demonstratives in spontaneous spoken English dialogues in the Switchboard corpus (Linguistic Data Consortium 1993). They provide a classification system for types of pronouns and demonstratives found in spoken language, where they differentiate *individual anaphors* (i.e. anaphors with NP antecedents), *discourse deixis* (i.e. reference to abstract objects: sentential and VP-antecedents), *vague anaphors* (i.e. no clearly defined linguistic antecedent, not a referent to a sentence or a VP, but a referent to a general discourse topic.), and *inferrable - evoked pronouns* (i.e. a particular usage of the third person plural pronoun *they*, in which there is no explicit antecedent but there is often an associated singular NP denoting an institution, e.g. a country, school, hospital, etc. Hence the antecedent is inferred from the institution as the authority or population/members of the institution). They test a resolution algorithm based on Strube (1998) to classify pronouns and demonstratives, as well as to co-index anaphors with NP and sentential antecedents. They use predicate information, NP form and dialogue structure for the anaphora resolution process. The algorithm makes use of the fact that demonstratives prefer discourse-deictic antecedents (i.e. discourse deictic anaphora) and pronouns prefer NP-antecedents (i.e. individual anaphora). Individual anaphora are assumed to specify entities already present in the discourse model, whereas discourse deictic anaphora are assumed to be used to create new referents. The possible lists of referents for each anaphor type are kept separate. The algorithm achieves 66.2% precision and 68.2% recall[3] for individual anaphors, and 63.6% precision and 70% recall for discourse-deictic anaphors using hand-simulation (i.e. it is not actually implemented).

In Vieira et al. (2005), demonstrative noun phrases in French and Portuguese written texts are annotated in the multilingual MLCC corpus with the aim of designing a tool for definite and demonstrative noun phrase reference resolution. The relations are classified as direct coreference, indirect coreference and other anaphora. The demonstratives and their antecedents are then classified for some syntactic features (i.e. as NPs, parts of sentences, full sentences and antecedents longer than full sentences). Basic semantic features of head nouns of demonstratives and the antecedents are also classified (i.e. concrete or abstract nouns), as well as hypernymy, synonymy, discourse deixis, etc. They conclude that the resolution of demonstrative NPs is mainly context dependent, where textual chunks are identified as antecedents for more than 80% of the cases. Moreover, among these cases, more than half have NP antecedents. Concrete demonstratives are observed to take concrete NPs as antecedent for more than 90% of the cases, whereas abstract demonstratives are resolved to have NP antecedents

[3] Precision is defined as the number of correct results returned divided by the total number of results returned. Recall is defined as the number of correct results returned divided by the total number of relations in the data.

in only 50-70% of the cases (depending on the annotators and language). A specific distribution of syntactic and semantic features are observed for demonstrative NPs. They suggest that this specific distribution of features require a different treatment for demonstrative NPs than other anaphoric expressions such as pronouns or definite expressions.

Byron (2002) describes a technique to resolve pronominal reference to individual and abstract entities, called PHORA. She identifies the differences between personal pronominal reference resolution and demonstrative reference resolution. According to this demonstrative pronouns have clausal antecedents or non-subject NP antecedents. For personal pronouns a speaker can quickly compute agreement features, however since demonstrative pronouns have semantically complex referents, this is not possible. Demonstrative pronouns tend to refer to an entity not in focus. Finally, when the predication does not constrain the pronoun, personal pronouns prefer individual referents, whereas demonstrative pronouns prefer abstract referents. Semantic filtering is applied to complement salience calculations, enabling the resolution of less salient abstract entities such as actions, propositions and kinds. This technique is evaluated on ten problem-solving dialogs taken from the TRAINS93 corpus, having 557 utterances and 180 pronouns. The full PHORA model using a different search order for demonstrative pronouns, achieves an accuracy of 72%, whereas the baseline model which treats personal and demonstrative pronouns alike displays 37% accuracy.

Botley (2006) investigates indirect anaphora by annotating demonstrative pronouns in three English corpus samples (spoken discourse, news, literature) containing 100,000 words each. Five features of the demonstratives are annotated: recoverability of the antecedent (directly, indirectly, non-recoverable, not-applicable), direction of reference (anaphoric, cataphoric, exophoric/deictic), phoric type (referential, substitutional, non-phoric), syntactic function (noun modifier, noun head, not-applicable), and antecedent type (nominal, propositional/factual, clausal, adjectival, no antecedent). Only indirectly recoverable demonstratives are examined in this study. The indirectly recoverable demonstratives are further examined according to their subtypes (i.e. labelling, as in Francis (1994)), situation reference and textual/discourse deictic).

Navarretta and Olsen (2007) annotate abstract pronomial anaphora (third person singular personal and demonstrative pronouns) in Danish and Italian texts, where each text contains about 60K words. The study decribes an extended annotation scheme for coreference called DAD. The functions differentiated for the pronouns are expletive (pleonastic), cataphoric (i.e. pronoun precedes the antecedent), deictic (i.e. pronoun refers to something in the physical world), individual anaphoric, individual vague anaphoric (i.e. the antecedent is implicit in the discourse), abstract anaphoric, textual deictic (i.e. pronoun refers to textual elements), abstract vague anaphoric (i.e. abstract antecedent is implicit in the discourse), abandoned (i.e. the utterance is unfinished so the context cannot be inferred). For the antecedents the classifications include NP antecedent, non-NP antecedent, eventuality, fact-like, speech-act, question and proposition.

This study provides insights as to what language-particular differences may be present for languages other than English for abstract anaphora. In Danish and Italian all occurrences of singular third personal and demonstrative pronouns can potentially be abstract anaphora. In English only the pronouns *it, this, that* can be abstract anaphors. However, in Danish one of the two pronouns for abstract anaphora *det* (it/this/that) is ambiguous in its pronominal type. The information about stress can help distinguish the personal pronoun *det* 'it' and the demonstrative pronoun *det* 'this/that'. In Italian the personal pronouns *lo*, *ne* and *ci* (can be clitics or independent pronouns), demonstrative pronouns *questo* 'this', *quello* 'that', *ciò* 'this/that' and zero anaphors occur as abstract anaphora.

In English normally only demonstrative pronouns can be used for clausal antecedents, whereas in written Danish the most frequent abstract pronoun *det* is ambiguous for pronominal type, and for Italian both personal and zero anaphora can have clausal antecedents. Another difference for Danish is that the demonstrative *dette* generally marks antecedents in the latter subclause, rather than the whole preceding clause.

Navarretta and Olsen (2007) also stress the importance of the distance between the anaphor and the antecedent (i.e. anaphoric distance) as a factor affecting saliency of entities, and mark this information for abstract anaphora.

A total of 2502 pronouns are annotated, where 569 of them identified to be abstract anaphora and 1393 are identified to be individual (NP) anaphora. Their results indicate that there are differences in abstract anaphora use in Danish and Italian with respect to English. One difference pointed out is that Danish demonstrative pronouns are favored when there are verbal phrase antecedents than when there are clausal antecedents.

Poesio and Artstein (2008) annotate the reference status of NPs and pronouns on the Arrau corpus, which is an aggregated corpus containing task-oriented dialogues from the Trains-91 and Trains-93 corpus, narratives from the English Pear Stories corpus, newspaper articles from the Wall Street Journal portion of the Penn Treebank, and mixed text from the Gnome corpus. The reference statuses are annotated as *anaphoric*, *discourse-new*, and *non-referring*, as well as classify their semantic types as *person, animate, concrete, space, time*, etc.

All noun phrases are treated as markables which can be anaphoric or serve as antecedents (or both), and all clauses are treated as potential antecedents for discourse deixis. Each NP is marked with the attributes gender, grammatical function, number, person, category (marks animacy, abstract/concrete distinction), reference (marks if NP is anaphoric, discourse-new, non-referential). About 9K markables are annotated, where 3.8K are identified as coreferent which participate in an anaphoric chain. Poesio and Artstein (2008) stress that the linguistic aspects of anaphora are not yet completely understood and its annotation is an open problem. However, they suggest that shortcomings in this respect can only be overcome by further annotation efforts.

One of the most recent studies, which is of Recasens and Marti (2010) annotates Spanish and Catalan text in the AnCora corpora composed of newspaper and newswire articles (400K words each), with coreference information for pronouns, full NPs and discourse segments. A classification for the referentiality of the entities is made (i.e. named entities, non-named entities, specific entities, lexicalized entities and non-referential entities). Also homophoricity (i.e. "proper-noun-like and generic definite NPs that refer to something in the cultural context or world view" (Recasens & Martí, 2010, p. 328)) of the entities are identified. Coreference links are distiguished as identity, discourse deixis (further classified as token, type, proposition) and predicative (further classified as definite/indefinite) links.

Dipper and Zinsmeister (2009) annotate the semantic types of anaphora in German for the pronoun *dies* 'this' in the Europarl corpus. Since abstract object anaphora resolution cannot use grammatical restrictions (because the antecedent is not nominal and the anaphor is usually neuter singular), they assume that in addition to saliency, semantic restrictions should be considered. They use the anaphor's semantic type to restrict the semantic type of the antecedent. The antecedent is located using a paraphrase test. They identify the semantic types of event, process, state, circumstance, modal, opinion/claim, generic, fact and proposition, as well as some defining features including world-dependent, time-dependent, dynamic, telic, and modal. This is done by a replacement test, which replaces the pronoun with a suitable NP in order to determine its semantic type. 48 instances of the demonstrative *dies* 'this' were annotated in 32 texts in an initial study. An 85% agreement was observed for the antecedent span selection by two independent annotators, whereas α=0.52 agreement was observed for identifying the semantic type of the antecedent. In a further annotation of 17 texts after a discussion period an agreement of α=0.60 was achieved. For the semantic type of the anaphor α=0.37 agreement was observed initially, and α=0.66 after the discussions.The anaphor's grammatical role was identified as the subject in 79% of the cases, where in the remaining cases it was found as an object.

Another related study was done by Lee and Song (2010) for Korean where 1235 demonstratives were annotated in spoken and written corpora, both of which were about 20K *eojeols* (corresponding to words in English). This study provides guidance in terms of investigating the distribution of lexical and syntactic features of demonstrative anaphora with respect to demonstrative type. They differentiate between exophoric and endophoric reference functions of demonstratives as suggested by Halliday and Hasan (1976). Similar to Botley and McEnery (2001) and Botley (2006), the demonstratives were annotated with the features lexical category, endophoricity (anaphor, cataphor), exophoricity (context-based/situational, deictic), syntactic category of the antecedent (nominal, clausal, sentential), phoric type (reference, substitution, non-phoric) and semantic function of the antecedent (entity, event, proposition). The lexical categories identified included adnoun (i.e. forms corresponding to this+N, that+N), pronoun (i.e. forms corresponding to this, it, that, these, they), locative pronoun (i.e. forms corresponding to here, there, over there), and exclamation (i.e. forms used as intensifiers, hedges, personal habitual noise).

The Korean demonstratives, like Turkish, are of three distinct forms: proximal (*i* – seems to correspond to *bu* in Turkish), speaker-centered distal (*ku* – seems to correspond to *o* in Turkish) and speaker-hearer centered distal (*ce* – seems to correspond to *şu* in Turkish), instead of having two different forms (i.e. proximal, distal) as in English. The study reveals that the proximal demonstrative *i* is more frequent than the speaker-centered distal demonstrative *ku*, which are both much more frequent than the speaker-hearer centered distal *ce*. In fact, *ce* is not observed in the written corpus. It is also observed that adnominal demonstratives are more frequent than pronouns or locative pronouns. The preferred demonstratives to refer to clausal or sentential elements were observed to be *i* 'this' and *ku* 'it'. In fact these demonstratives refer to clausal/sentential elements more frequently than nominal elements.

Dipper and Zinsmeister (2010) survey abstract anaphora annotation efforts and try to combine together common features as a standard for such annotation. They propose that reference corpora should minimally annotate prototypical pronominal realizations of anaphors such as demonstrative pronouns in terms of form. In terms of semantics of the anaphora, minimally the distinction between concrete and abstract should be made. They note that many different types of information have been used for a more fine-grained labeling, such as speech acts, eventualities, factualities, type-token distinction. For the antecedent's form, minimal annotation of clauses or verbal heads is proposed.

Table 5.2 shows a general picture of all these studies according to the corpora and kind of anaphors they annotated.[4] All of these studies are a step towards understanding the cross-linguistic features involved in anaphora and coreference resolution, as well as, language-specific features. In order to enhance this insight, studies need to be conducted for other languages as well. This is the motivation for the present thesis as well. As we will show in Section 5.4, we have annotated a small portion of the TDB for anaphors as a preliminary step in understanding the dynamics of anaphors in Turkish. This small-scale study led to an automatic identification of a specific kind of discourse connecting device (i.e. phrasal expressions such as *bu nedenle* 'for this reason').

Table5.2: Corpora annotated with discourse deixis and coreference.

Study	**Corpus**	**Type(s) of Anaphor Annotated**
Eckert and Strube (2000)	English dialogs	pers. & dem. pr.
Viera et al. (2005)	Portuguese 50 dem. NPs French 50 dem. NPs	dem. full NP dem. full NP
Byron (2002)	English problem-solving dialogs from TRAINS93 corpus	all pronouns
Botley (2006)	English (300K) spoken discourse news literature	*this* *that* *these* *those*
Poesio and Artstein (2008)	Arrau Corpus mixed texts (95K)	NPs and pronouns
Navarretta and Olsen (2008)	Danish texts (60K) Italian texts (55K)	pers. & dem. pr. (zero) pers. & dem. pr.
Recasens and Marti (2010)	Catalan (400K) newspaper/newswire articles Spanish texts (400K) newspaper/newswire articles	(zero) pers. & dem. pr. full NP (zero) pers. & dem. pr. full NP
Dipper and Zinsmeister (2009)	Europarl corpus (32 German texts)	*this* (Ger. dies)
Lee and Song (2010)	Korean spoken and written corpora (20 K)	dem. pr.

[4] The table is a combined version of table 1 from Recasens (2008), p. 75 and table 1 from Dipper and Zinsmeister (2010), in the Appendix.

5.3 Annotation and Resolution Studies on Reference Relations for Turkish

There have been some previous studies on anaphora done for Turkish, most of which involve pronominal anaphora and some involve zero anaphora. More recent studies have concentrated on computational approaches for the resolution of anaphora. These include Tın and Akman (1994), Yüksel and Bozşahin (2002), Yıldırım, Kılıçaslan, and Aykaç (2004), Küçük (2005), Tüfekçi and Kılıçaslan (2005), Tüfekçi, Küçük, Turhan Yöndem, and Kılıçaslan (2007), Küçük and Turhan Yöndem (2007), Yıldırım and Kılıçaslan (2007), Yıldırım, Kılıçaslan, and Yıldız (2009), and Kılıçaslan, Güner, and Yıldırım (2009), as explained below. The formulations and findings in linguistic studies in Section 5.1.2 above have proven useful to further computational approaches to anaphora resolution in Turkish, where the theoretical claims are supported.

In the first part of our overview of these work, we look at studies merely defining a methodology to resolve anaphora. In the second part, we look at the work on using methodologies developed to obtain corpus-based classification results for anaphora resolution, or to develop some NLP applications, such as in Küçük and Yazıcı (2008, 2009) and Can et al. (2008, 2009). Table 11 at the end of this subsection, displays a general overview of computational studies on Turkish anaphora resolution. The findings and methods employed in the studies described in this section will guide our analysis of Turkish demonstrative anaphora in the TDB, which will be described in Section 5.4 and our preliminary effort to resolve demonstrative pronouns, which will be described in Section 5.5.

5.3.1 Studies on Turkish Anaphora Resolution Describing a Methodology

Tın and Akman (1994) is one of the earliest computational approaches on Turkish pronominal anaphora resolution. The approach used in this study is based on situation theory. The study views anaphora resolution as the task of forming a cognitive structure and defining its relationship with previously formed structures. A situation-theoretic computational medium called BABY-SIT allows the use of contextual information, and the computation over the situations is done using constraints. The resolution of zero/pronominal anaphora is explained within this environment on some example sentences.

Rules are defined with respect to anaphora type (i.e. zero anaphora or pronominal anaphora). For zero anaphora if the anaphor is the subject of an embedded sentence, then the antecedent is the subject of the main sentence or the non-subject NP of the main sentence which precedes the anaphora. Otherwise, if the zero anaphora is possessor of a genitive construction, then the antecedent is the subject NP or the non-subject NP preceding the anaphora. For the pronominal anaphora, the anaphoric expression is the non-subject NP and the antecedent is any NP c-commanding it. Otherwise, if the anaphora is the possessor of a genitive construction in an embedded structure, then the antecedent is a non-subject NP preceding the anaphora, in which case it is referred to be a free anaphor.

Tüfekçi and Kılıçaslan (2005) develop a syntax-based pronoun resolution system for 3rd person singular pronominal anaphora (i.e. *o* 'he/she/it', *onu* 'him/her/it', *onun* 'his/her/its') and the reflexive *kendi* ('himself/herself/itself') to NP antecedents in Turkish based on Hobbs' Naïve Algorithm (Hobbs, 1978). This algorithm traverses full parse trees starting from the pronoun up and searches left-to-right breadth-first in the subtrees dominated by S, NP and VP nodes for an appropriate antecedent. The original Hobbs' Naïve algorithm is reformulated in this study for Turkish as some new rules are incorporated to encompass Turkish syntax. The thematic hierarchy proposed in Yıldırım et al. (2004) is used for Turkish. The study shows how the modified Hobb's Naïve algorithm can be applied for Turkish in some example sentences.

Two other studies apply the theoretical framework of Centering Theory to pronominal anaphora resolution in Turkish. Yıldırım et al. (2004) uses the principles of Centering Theory to resolve pronominal anaphora in Turkish, while Yüksel and Bozşahin (2002) employ the findings of both Binding Theory and Centering Theory to generate anaphora. We will first look at Yıldırım et al. (2004) and briefly describe the Centering Theory framework. Yüksel and Bozşahin (2002) will be described in the next section as they evaluate their system through corpus-based experimentation.

Yıldırım et al. (2004) explores pronominal anaphora resolution in Turkish using Centering Theory. Centering Theory models a discourse segment as a sequence of utterances U_i, i=1, 2,..n. A partially ordered list of possible antecedents for a given utterance makes up the list of *forward-looking centers*, $C_f(U_n)$. The highest ranked element in this list is the *preferred center*, C_p, which is to be the primary focus of the following discourse. The entity currently being the focus after U_n is interpreted is known as the *backward looking center* of U_n, $C_b(U_n)$ in this model. Yıldırım et al. (2004) propose the following thematic hierarchy for Turkish to be used for ordering the list in the Centering Model: *agent-time-duration-location-instrument-manner-benefactive-theme (or patient)-source-goal*. Four transition states are defined in their algorithm for pronoun resolution. If the backward looking center is the same as of the previous utterance (i.e. $C_b(U_n) = C_b$ $(U_n\text{-}1)$) and it is in fact the preferred center (i.e. $C_b(U_n) = C_p(U_n)$), then *Continue* transition is employed; otherwise if it is not the preferred center, then *Retaining* transition is employed. If the backward looking center is the not the same as that of the previous utterance (i.e. $C_b(U_n)$!= C_b $(U_n\text{-}1)$) but it is the preferred center (i.e. $C_b(U_n) = C_p(U_n)$), then *Smooth Shifting* transition is done; otherwise if it is not the preferred center, then *Rough Shifting* transition is employed. Their system first tags all the strings in a given sentence using a POS tagger and then retrieves the position information for the strings using a parser. The tagged strings are input into the Anaphora Resolvent, which employs the centering-based model described above.

5.3.2 Studies on Turkish Anaphora Resolution Presenting Corpus-Based Results or NLP Applications

Different from the other studies previously mentioned, Yüksel and Bozşahin (2002) describe a system for generating anaphora and pronouns which are contextually appropriate. However, the rules they employ are similar to the principles used for pronoun resolution. Another difference is that they evaluate their system through corpus-based experimentation. In this section we will describe the work related to anaphora resolution studies in Turkish which present corpus-based evaluations.

Yüksel and Bozşahin (2002) use the findings of Binding Theory and Centering Theory and describe rules to generate reference in Turkish. They plan local reference by binding rules and nonlocal reference by both binding rules and centering rules. Separate rules for reciprocals, reflexives and pronouns have been defined amounting to a total of 15 rules. Reference planning is conducted in four stages, where the first two stages consist of marking anaphors and pronouns to decide if an NP should be a referring expression or a pro-form. In these stages local reference rules are applied. In the third stage surface pro-form generation is done by checking if any of the pronouns should be dropped. The final stage involves exception marking. The reference planner uses the backward looking centers and the preferred centers of Centering Theory. Turkish case frames are the output from the planner and they are passed on to a surface form generator. The study defines 3 kinds of rules as *overt realization rule, drop rule* and *exception rule*. The first one contains the rules for realizing a full NP as an overt anaphor or pronoun. The drop rule, encapsulates the rules for realizing the full NP, a zero pronoun or zero anaphor, and is used in the surface pro-form generation stage. Finally, the exception rule specifies the situations when an NP cannot be realized as a zero-form or pro-form in the last stage. The system described obtains ~70% success rate in pronoun generation. Comparative tests between the implemented system and native speakers show that the system produces appropriate output. However, it is seen that native speakers also use other sources of information in reference planning. Incorporation of other information sources (e.g. lexical information between non-nominal entities, mental model of the deictic reference) to be used by new rules is suggested to increase success.

Küçük (2005) describes a knowledge-poor pronoun resolution system which resolves third-person personal and reflexive pronouns in Turkish. Overt pronouns referring to proper names and zero pronouns at subject positions are manually marked in the input text as a preprocessing step. The resolution system then splits the input text into sentences, extracts the pronouns to be resolved, creates a list of candidate antecedents and finally determines the antecedent of each pronoun by applying certain constraints and preferences. These constraints and preferences are determined by an empirical analysis on a Turkish child narrative, and verified by a questionnaire conducted on native Turkish speakers. Due to the results of the empirical analysis, the study limits the search space for the antecedent to the current sentence and the preceding three sentences. The person names in this search space are extracted to form a candidate list using a Turkish person names dictionary. Candidates for plural pronouns are identified using set-generation. Finally the antecedent of a given pronoun is determined by the application of the constraints and preferences.

The constraints applied to the antecedent candidate list are: number agreement, reflexive pronoun constraint, and the personal pronoun constraint. The number agreement constraint ensures that the pronoun and the antecedent agree in number. The reflexive pronoun constraint is what Küçük calls "an adaptation of c-command constraints used in many anaphora resolution algorithms", and selects the closest candidate to the pronoun as the antecedent. The personal pronoun constraint employed in Küçük (2005) eliminates the sentence containing the personal pronoun as a possible search scope for its antecedent. After the application of the constraints, if there are multiple possible antecedents for a given pronoun then preferences are calculated and a final antecedent is determined by choosing the one with the highest aggregate preference score. The preferences used in this step include quoted/unquoted, recency, nominative case, first NP, nominal predicate, repetition, punctuation, antecedent of zero pronouns preferences. The first preference ensures that the pronoun and its antecedent are either both in quoted or unquoted text. Closer antecedents are given precedence according to the recency preference. Nominative case preference prefers proper nouns in nominative case as antecedents assuming that such nouns usually employ subject positions in the sentence. Another preference used gives precedence to the NPs that are the first phrase in their containing sentence. Also, NPs in the nominal predicates are given precedence. NPs repeated in the text are also preferred as antecedents. Another preference is applied to NPs succeeded by a comma, which is assumed to increase the salience of the NP. Finally, a zero pronoun preference is given to antecedent candidates that are determined as antecedents of zero pronouns in previous sentences.

Two sample texts; one narrative from the METU Turkish Corpus (MTC) (having 20 reflexive, 170 personal pronouns) and a child narrative (having 15 reflexive and 190 pers. pronouns) are used to test the system. The recall and precision values of the two tests were 85.2% recall, 88% precision and 73.6% recall, 90.9 % precision, respectively.

In a follow up study, Küçük and Turhan-Yöndem (2007), third-person personal and reflexive anaphoric pronouns are automatically identified using a decision-tree learning approach (Quinlan's C4.5 implemented using Weka J48 classifier (Witten & Frank, 2005)) using linguistic features determined by a corpus examination. The features used in this study are the surface form of the candidate, being part of an idiom, having a preceding *ki* ('that'), preceding a noun phrase, having the specific patterns identified for pronominal anaphora in the corpus, being the last phrase in the sentence, succeeding *benim* ('mine'), succeeding *ilk* ('first'), or *saat/gece* ('time/night' as in 'at ten/ten at night' when used before *onda*), or preceeding *bir/iki/üç/../on* ('one/two/three/.../ten'), succeeding *saat* ('time/clock' when used before *ona*), conforming to the patterns some given patterns for onu implying homonymy with on ('ten'). These features were extracted in order to classify the anaphora as *idiomatic* (i.e. non-anaphoric uses of the third person pronoun *o* 'he/she/it' as part of an idiom), *cataphoric, lexical noun phrase anaphoric* (i.e. demonstrative + NP uses such as *o adam* 'that man'), *definitely non-anaphoric*, and *pronominal anaphoric*.

The system architecture implemented consists of a *candidate extractor*, which provides pronominal anaphor candidates from each entence in the input text using surface forms. The output of this module is input into a *feature extractor*, which determines the feature values for the candidates. The features are the results of applying certain rules such as determining if the candidate is preceeding an NP. They are devised to distinguish between the different classes of anaphora. The feature extractor module provides input for a DTClassifier (Decision Tree Classifier). Evaluation on two child narratives revealed 97.8% and 98% classification accuracy for the two samples.

Kılıçaslan et al. (2009) explored a learning - based pronoun resolution system for overt and zero Turkish 3rd person personal, locative, reciprocal, reflexive pronouns. Different from the previously described studies, this study employs the feature semantic type of the antecedent and compares the performances of several different classification algorithms. A corpus consisting of a compilation of 20 Turkish child stories is annotated with the features of case, grammatical role, overtness, type (i.e. personal, locative, reciprocal, reflexive), semantic type (animal, human, place, abstract object, physical object), person and number, position, true antecedent position, and referential status. Five different algorithms are used for the classification, namely naïve bayes, k-nearest neighbor (kNN), decision tree, support vector, and voted perceptron. The study compares the performances of these different algorithms, as well as investigating the contributions of the various features defined and their possible value distributions in terms of overtness, and pronoun type. The highest accuracy obtained using the different classifiers is 82%, where an f-measure of 0.74 is found.

The study observes several conclusions regarding the classifiers and the anaphora resolution process. First of all, the performance of a model using a non-linear classifier is always better than with a linear one. Second, it is observed that the expressiveness of the model is directly correlated with the success rate of the resolution. Third, null pronouns are resolved better than overt pronouns and finally, reflexive, reciprocal and locative pronouns are resolved better than personal pronouns. Evaluation of the contribution of each feature used shows that the distance feature is very useful, while the person-number feature is not useful.

Yıldırım and Kılıçaslan (2007) provide a machine learning approach to personal pronoun resolution in Turkish. They present a corpus based learning approach using a decision-tree classifier with ensemble learning (boosting) reinforcement. The study uses the features of grammatical role with the ranking used in Yıldırım et al. (2004), case information such as accusative, nominative, overtness of pronouns, animacy type and the distance between the anaphor and the candidate antecedent. The performance results reveal 0.70 recall, 0.75 precision and 0.72 f-measure scores, indicated as being reasonably acceptable.

Two other studies Yıldırım, Kılıçaslan, & Yıldız (2007) and Yıldırım et al. (2009) present a decision-tree and a rule-based learning model for pronominal anaphora resolution in Turkish. Ten child stories which contain more than 500 pronouns were used in this study. An annotation tool was used to assign semantic information to entities. Similar features to the Yıldırım and Kılıçaslan (2007) study were used. Namely, case, grammatical role, overtness, pronoun type (personal/reflexive/locative), semantic type, case and grammatical role of the referring expression, distance, number/person agreement features were used. The distance and grammatical role features were discovered to be the most important features in the classification process. Their decision tree model resulted in 51.9 % recall, 73.2 % precision with an f-score of 60.8%, whereas their rule-based model showed 55.8% recall, 67.3% precision with an f-score of 61%.

Other studies which may be related to anaphora resolution in Turkish include Küçük and Yazıcı (2008, 2009) and Can et al. (2008, 2009). In Küçük and Yazıcı (2008), a fuzzy conceptual data model is used to extract salient objects from Turkish political news text in order to identify coreference chains using a heuristic rule based on recency. In Küçük and Yazıcı (2009) named entity recognition of person, location, organization names, numeric and temporal expressions is done in the transcriptions of news videos, using a rule-based named entity recognizer (NER) for semantic retrieval. Such studies may aid in discovering antecedent candidates and coreference chains. The Can et al. (2008, 2009) studies, on the other hand, investigate query-based information retrieval (IR), new event detection (NED) and topic tracking (TT) scenarios, which have similar techniques to those used for anaphora resolution. For the IR study, they investigate the effectiveness of different stemmers as well as query lengths in a collection of newspaper texts. NED and TT are performed on a larger news story collection of about 200K stories amounting to 350MB, where the texts are annotated for news category and topic.

Table5.3: An Overview of computational studies on Turkish anaphora.

Study	**Corpus / Medium**	**Type(s) of Anaphor Annotated**	**Theory**	**NLP / ML Methods Used**
Tın and Akman (1994)	BABY-SIT	Zero / overt pronominal	Situation theory	-
Yüksel and Bozşahin (2002)	Turkish novels, technical text	reflexives, reciprocals,	Binding Theory and Centering Theory	Reference Planning with Prolog

Table 5.3 (continued)

Study	**Corpus / Medium**	**Type(s) of Anaphor Annotated**	**Theory**	**NLP / ML Methods Used**
		pronouns (generation)		
Kılıçaslan et al. (2009)	Turkish child stories (20 text)	3rd-pers. overt/zero personal, locative, reciprocal, reflexive pr.	-	Instance Based (kNN), Naïve Bayes, Decision tree (J48), Voted Perceptron, Support Vector (C-SVC)
Yıldırım et al. (2004)	A discourse fragment	Zero/overt pronominal	Centering theory	-
Yıldırım et al. (2007, 2009)	Child stories (10 text, >500 pr.)	pronominal	-	Decision tree (J48) and rule-based alg. (PRISM, PART)
Yıldırım and Kılıçaslan (2007)	Turkish text	Pers. pr.	-	Learning-based approach: decision-tree class. (J48) with ensemble learning (boosting)
Tüfekçi and Kılıçaslan (2005, 2007)	10 toy sentences	3rd-pers. sg. pr. and reflexive kendi (NP antecedents)	-	Hobbs Naïve Alg.
Küçük (2005)	Two sample texts (MTC and a narrative, ~400 pr.)	3rd-pers. personal and reflexive pr.	-	Knowledge-poor sys.
Küçük and Turhan-Yöndem (2007)	Sentences from MTC (7401 o occurences), Two child narratives	3rd-pers. personal and reflexive pr.	-	Decision-tree (Quinlan's C4.5 by J48)
Küçük and Yazıcı (2008)	Three sample political news texts (MTC)	Pronominal and coreference of NPs	Fuzzy conceptual data model	-

Table 5.3 (continued)

Study	Corpus / Medium	Type(s) of Anaphor Annotated	Theory	NLP / ML Methods Used
Küçük and Yazıcı (2009)	Transcriptions of 16 news videos from TRT	Named entities (person, location, organization names, numeric and temporal expr.)	-	Rule-based NER, semantic retrieval

The lessons to take home from the overview presented in this section can be summarized as follows: Most of the studies described initially make a corpus-based analysis in order to understand the phenomena at hand and develop a methodology, or determine a set of features to be used in a classifier, etc. Then a general system architecture of candidate generation, feature selection followed by a rule-based elimination system or a classifier is observed. Mainly syntactic features are employed in Turkish pronominal anaphora resolution. These include grammatical role, part of speech, distance between the antecedent and the anaphor, position of the anaphor and the relative position of the antecedent, person/number agreement and type of anaphora.

5.4 An Analysis of Turkish Demonstrative Anaphora in the TDB

In order to obtain a better view of abstract object anaphora in Turkish, we analyzed (Sevdik Çallı, 2012) demonstrative anaphora including bare demonstrative uses and demonstrative+NP uses on a 20K subpart of the TDB. Antecedents of demonstrative anaphora including abstract object references were identified in 10 texts of approximately 2000 words each consisting of texts from the genre novel. All uses of Turkish demonstrative pronouns (i.e. *bu* 'this' / *şu* 'this/that' / *o* 'that'), including bare demonstrative usages and demonstrative+NP usages, have been identified with their antecedents. The antecedents have also been identified as being abstract objects, concrete objects, or exophoric (where the reference is to text-external material). In (83) below, reference to an abstract object where bunun 'this+Gen' refers to "getting sick" is shown. In the examples that follow (in this and the following section), the anaphor will be rendered in bold, while the antecedent will be presented between square brackets.

(83) Şemsî Ahmed Paşa onu ayakta karşılayarak, başına gelen talihsiz kazaya çok üzüldüğünü, eğer [hasta olursa] **bunun** sorumluluğunun kendisinde olduğunu söyledi.

Şemsî Ahmed Pasha greeting him on his feet, said he was very sorry for the unfortunate accident that happened (to him), if [he were to get sick] the responsibility of **this** would be on himself.

(00001231.txt)

A concrete reference to his father's study by the demonstrative+NP this room is given in (84).

(84) Üst katta [babasının çalışma odasına] girdi. Onun ölümünden beri ilk kez girdiği **bu_odayı** ağır bir koku kaplamıştı.

He entered [his father's study] upstairs. A heavy stench had filled **this_room**, which he entered for the first time after his death.

(00001131.txt)

The referent for those devious pains in (85) is marked as exophoric since it cannot be found in the text.

(85) Her yere kendisiyle birlikte taşımaz mı içindeki **o_sinsi_acıları**?

Does he not carry **those_devious_pains** with himself everywhere?

(00001131.txt)

Apart from the demonstratives, explicit third person pronouns have also been identified and resolved. The reason for identifying the third person pronouns is that in Turkish the demonstrative *o* 'that' is homonymous to the third person pronoun. In order to find distinguishing features for these two uses, third person pronouns have also been resolved as exemplified in (86).

(86) Gözlerine bakamazdım ben [insanların]. Korkaktım ben. Ben **onlardan** korkardım, kızgınken bile.

I couldn't look into [people's] eyes. I was a coward. I was afraid of **them**, even when I was angry.

(00001131.txt)

These included references to proper nouns, as well as reference to NPs. All analysis was done manually by a single annotator (the author herself) for this work. A total of 682 instances of demonstrative anaphora were identified. Table 5.4 shows the distribution of demonstrative anaphora, as well as 3rd person pronoun *o* in this sample.

Table5.4: Distribution of Turkish Demonstrative Anaphora in TDB 1.0

	Abstract	**Concrete**	**Pers. Prn.**	**Exophoric**	**Total**
bu	61	43	0	11	115
bu+NP	45	92	0	67	204
şu	1	0	0	0	1
şu+NP	1	2	0	21	24
o	6	27	126	47	206
o+NP	17	60	0	55	132
Bare Demonstrative Total	68	70	126	58	322
Demonstrative + NP Total	63	154	0	143	360
Total	131	224	126	201	682

Usages of *bu*, *şu*, *o* as abstract anaphora was identified in 131 cases versus 224 concrete anaphora uses, 126 references by personal pronouns, 201 of the cases were identified as exophoric uses, where the referent was not mentioned in the text. The exophoric cases and personal pronoun uses make up about 48% of all anaphoric uses. Of the remaining 355 cases, 138 instances are pure demonstrative anaphora, i.e. referencing by bare demonstrative uses, without NP complements (either abstract or concrete referrents) as in (48); the rest are demonstrative NP uses as in (49) and (50). Out of a total of 138 abstract and concrete bare anaphora 104 (75%) involve the demonstrative *bu*, 33 (24%) involve *o* and only 1 (1%) is using *şu*. The abstract references occur more with the use of the demonstrative bu, i.e. 61 abstract uses versus 43 concrete uses. On the other hand, concrete references occur more with the demonstrative o, i.e. 27 concrete uses versus 6 abstract uses. Out of a total of 217 abstract and concrete demonstrative NP anaphora, 137 (63%) were discovered for *bu*, 3 (1%) observed for *şu*, and 77 (36%) involve *o*. However, it can be said that demonstrative NP anaphora favors concrete referents, as they are more than twice as much as abstract referents (i.e. 154 concrete cases versus 63 abstract cases). Of all the demonstrative anaphora cases observed 201 had exophoric referents. Most of these anaphora were simply just unmentioned in text, for example referring to the particular day of the event described. Some included ostension, whereas some were specialized uses (28 cases) as in the use of *o kadar* 'so much'. Other special uses involved *bu kadar* 'this much' (9 cases) and *o zaman* 'that time' (6 cases). Some others were vague references via a personal pronoun (47 cases) to an unmentioned salient person. There were also a total of 11 cataphoric instances observed, where the referent was found after the anaphor (*bu*:3, *şu*:2, *o*:6).

Some preliminary conclusions can be drawn about Turkish demonstrative anaphora observed in novels. It is observed that about 1/3$^{\text{rd}}$ of all the demonstrative anaphora in Turkish novels consists of exophora, whereas the rest is endophoric (i.e. within text) uses. Within the endophora, the most frequently used demonstrative in Turkish is found to be bu and it is also the most preferred demonstrative for AO reference, where *şu* and *o* are rarely used. On the other hand, demonstrative use of *o*, is preferred for referencing concrete objects. There is also substantial use of *o* as a personal pronoun, dominating all its other uses in terms of frequency.

This analysis was an empirical study about Turkish demonstrative pronouns introduced in Section 5.1 above and our quantitative findings support the theoretical explanations provided. For example, we have observed that *o* is more frequently employed for concrete object reference when used as a demonstrative. This finding supports Göksel and Kerslake's explanation that concrete items that are out of sight are referred to with *o*, as in written text items are mostly expected to be out of sight of the hearer and the speaker. In a similar fashion, we can interpret our finding that şu is most frequently used exophorically to support the claim that *şu* is conceived to be accompanied by an ostensive gesture of pointing. As our resource consists of written data, such ostension cannot be observed and hence may be rendering the uses exophoric.

The findings of this empirical study will be first validated by an annotation agreement study presented below in the next section. After our results are validated as acceptable, we can use these findings, as well as the findings and methods of previous studies discussed in Sections 5.1 and 5.3 in our Turkish demonstrative pronoun resolution experiment which will be presented below in Section 5.5.

5.4.1 The Control of the Annotations by Two Secondary Annotators and an Evaluation of the Disagreements

Above we introduced an empirical study on 10 texts. The annotations in these 10 texts have been checked by two other annotators in order to validate the results obtained. One annotator checked all the annotations, and the other only checked half of them. These secondary annotators were given a complete list of the annotations of the primary annotator, as well as the 10 texts. The list was organized according to the appearance order of the search tokens in the texts. The texts provided were searchable, and the annotators were advised to search and highlight the main search token as either *bu*, *şu*, or *o*. In the annotation list, the text file from which the annotation is from was specified. The annotator was advised to open the related text, search for the token and check if the reference of the demonstrative was identified correctly. If they had any disagreements, they were asked to write the correct referent for the given demonstrative. If they had other notes, they could write them in the note section of the related annotation. The instructions were given to the annotators in a written format and explained.

The agreements with the two annotators were evaluated using the exact and partial match criterion (Miltsakaki et al., 2004a). If the secondary annotator agreed with the resolution, than the agreement is recorded as 1, if there is partial agreement or no agreement then it is recorded as 0. Agreement is calculated for each secondary annotator by the number of exact matches found in the total number of annotations initially annotated and also given as a percentage (See Table 5.5).

Table5.5: Exact Match Agreement

	Non - exophoric	**Exophoric**	**Total Match**	**Total # of Ann.**[5]	**%**
Annotator 1	489	203	692	695	99.56
Annotator 2	274	103	377	380	99.21

The disagreements with Annotator 1 were for 3 non-exophoric cases, where there was partial match; and with Annotator 2 were for 3 exophoric cases, where there was no agreement. The first one of the disagreements with Annotator 1 was for the resolution of "*bundan*" ('of this), where the primary annotator resolved it as "*bedeninde yaralar olduğundan*" ('of the fact that there were scars on his body), i.e. as an abstract object, a fact; Annotator 1 resolved it as "*bedenindeki yaralardan*" ('of the scars on his body), i.e. as a concrete object (87).

[5] The number of annotations is slightly different from the initial study (it was 682 in total) because the primary annotator found some missing demonstratives that had not been annotated and added annotations for these. The secondary annotators checked the resolutions of these modified annotations (which had a total of 695 annotations).

(87) Ne var ki bedenindeki yaraları Hekim Bosnalı Kerim Bey de görmüş ve böyle bir şeyle ilk kez karşılaştığını söylemişti. Ante'nin ricası üzerine **bundan** kimseye söz etmemişti.

Even so Doctor Bosnian Kerim Bey had seen the scars on his body and said he was seeing such a thing for the first time. On Ante's request he had not told anybody **of this**.

Primary Annotator: "*bedeninde yaralar olduğundan*" ('of the fact that there were scars on his body)

Annotator 1: "*bedenindeki yaralardan*" ('of the scars on his body)

(00001231.txt)

The second disagreed case involved a similar resolution difference, where the primary annotator again resolved the anaphor as an abstract object and Annotator 1 resolved it as a concrete object (88).

(88) Gezi boyunca sigara içmeyeceğim[6], ama belki, paltomun büyük cebinde naylon tapası çakıyla kesilmiş bir şişe ucuz şarap; **o** da görüntüyü iyice sindirmek ve yaşanan ânı kalıcı kılmak için.

I won't smoke during the trip, but maybe, (drink[6]) the bottle of cheap wine in my big pocket whose plastic cork has been cut up by a pocket knife; **that**'s just to absorb the scenery completely and make the moment permanent.

Primary Annotator: paltomun büyük cebinde naylon tapası çakıyla kesilmiş bir şişe ucuz şarabı içersem ('if I drink a bottle of cheap wine in my coat's big pocket whose plastic cork has been cut up by a pocket knife)

Annotator 1: paltomun büyük cebinde naylon tapası çakıyla kesilmiş bir şişe ucuz şarap ('a bottle of cheap wine in my big pocket whose plastic cork has been cut up by a pocket knife)

(00003121.txt)

The third disagreement was in fact not a disagreement but, a difference in the selected text span of the referent (89).

(89) .. masamıza gelen garson kıza, iki bourbon viski ısmarladık. "Biliyor musun ben **bu** bourbon viskiyi daha çok seviyorum."

..we ordered two bourbon whiskeys to the waitress who came to our table. "You know I like **this** bourbon whiskey more."

Primary Annotator: iki bourbon (cinsi olan) viski ('two bourbon (kind of) whiskeys)

Annotator 1: burbon viski ('bourbon whiskey)

Annotator 1's Note: There should not be a number here. This means bourbon whiskey. He likes not scotch whiskey, but bourbon.

[6] *İçmek* in Turkish is used for both smoking (as in cigarettes) and drinking (as in wine). Hence, there is ellipsis in this example and the verb drinking is not explicitly mentioned for the wine.

(00006231.txt)

The two annotators meant the same thing, but made slightly different annotations. The disagreements with Annotator 2 were on 3 exophoric cases (90-92). Annotator 2 did not think they were exophoric and found referents.

(90) .. aydınlık bir gün bugün. .. **Bu zamansız güneş** onu bile yaşam sevinciyle doldurmuş.

..its a bright day today. .. **This untimely sun** has filled even him with joy of life.

Primary Annotator: exophoric

Annotator 2: zamansız bir kış güneşi ('an untimely winter sun)

(00003221.txt)

(91) .. Bir dönemdir **bu**; numara ya da büyük adlar verilmesi gerekmeyen bir dönem.

..**It**'s an era; an era which does not need to be given any number or great names.

Primary Annotator: exophoric

Annotator 2: bir dönem ('an era)

Annotator 2's Note: which does not need to be given any number of great names

(00003221.txt)

(92) Onun adını neden versinler ki! Ne kötü bir beklemekti **bu**.

Why should they give his name! What a terrible waiting **this** was.

Primary Annotator: exophoric

Annotator 2: Az sonra otomatın sesi. Sonra merdivenlere sürten kalabalık ve ivecen ayak sesleri. Sonra güm güm kapı. Cass, bir sigara. Hemen tüy, demişti Ekrem ('A little later the sound of the buzzer. Then a crowd scraping against the stairs and impatient foot steps. Them bam bam door. Cass, a cigarette. Get away immediately, had said Ekrem)

Annotator 2's Note: Birilerinin gelip onu almasını ('for someone to come and pick him up)

(00003221.txt)

Overall, the control of the annotations by two other annotators resulted in very high agreement. This gave confidence in the annotations and the disagreement instances were used to finalize the annotations as agreed upon resolutions. The high agreement observed validates our manual annotations and thus the findings based on the results of these annotations. In empirical studies such as the one presented above in Section 5.4, it is important to ensure the validity of the annotations so that the results and the conclusions derived from these results are reliable. Since manual annotations harbor the human error factor, these kind of control studies as the one presented here, ensure that this error is negligible or acceptable. Achieving high agreement in our manual annotations, we can use our findings to develop further experiments such as the one below in the next section with confidence.

5.5 An Experiment Using ML Techniques for Demonstrative Pronoun Resolution in Turkish

In this section, we present an experiment using the findings corroborated in the previous section for Turkish demonstrative pronouns and the findings of prior studies on Turkish pronouns and pronoun resolution. This test case experiment involves applying several machine learning (ML) techniques to resolve demonstrative pronouns in the Metu-Sabancı Turkish Treebank (Atalay et al., 2003; Oflazer et al., 2003), in order to gain a basic understanding of how these techniques are applied and how some standard features aid in anaphora resolution of demonstrative pronouns in Turkish. First, the referents of the Turkish demonstrative pronouns *bu* 'this', *şu* 'this/that', *o* 'that' were annotated in the Treebank. Then, part of the annotations are used along with some features such as person-number, case, grammatical role, part-of-speech (POS), type of demonstrative, etc. were used to train the system. Finally, the performance of the trained system is tested on the rest of the annotations. This training is done for twelve different classification algorithms.

5.5.1 Experimental Setup

Two sets of experiments are conducted on each of the twelve classification algorithms used, including two baseline algorithms. First, a percentage split approach is used to split the data into training and test sets. In the second set of experiments, ten-fold cross validation is used to ensure that the results were representative of what independent test sets would yield.

5.5.2 Data Annotation

The METU-Sabancı Turkish Treebank (MST)[7] was used to annotate the demonstrative pronouns. This is a morphologically and syntactically annotated treebank corpus of 7262 grammatical sentences, whose sentences are taken form METU Turkish Corpus. The structure of MST is based on XML. It has part-of-speech (POS), person-number, case, possessive suffix, tense-aspect-mood (TAM), etc. information for each word.

The treebank contains 676 tokens of *bu* 'this', 36 tokens of *şu* 'this/that' and 518 tokens of *o* 'that'. However, some of these are determiners (such as in *bu numara* 'this number'), non-anaphoric uses such as idiomatic cases (as in "Halamın çocukları da *o kadar* yaramaz ki anlatamam", 'I can't explain *how* naughty my aunt's children are'), and some references may be out of scope, (i.e. those involving anaphors at the very beginning of the documents whose reference is not included in the document). The determiners are excluded by taking only that *bu/şu/o* identified as demonstrative pronouns (with the tag "DemonsP") in the treebank, as well as excluding those tagged as "DemonsP" having the grammatical relation "DETERMINER". The rest are excluded in manual annotation.

For the data preparation step, Matlab 2008b was used for extracting the necessary information from the Treebank xml files, generation of candidates, and preparation of the final data file to be input to the classifier. All the demonstrative pronouns (demP) in the Treebank were identified and their antecedents were annotated manually. The antecedents may be single words, or

[7] Referred to as MST or Treebank, from here on.

identified as spans of words, which may also be whole sentences or groups of sentences. If the demP was non-referrent (which may happen when there is simple ostension, or the antecedent cannot be resolved, or out of scope, as explained above), it was not annotated. A total of 253 demonstratives were annotated (*bu*: 202; *şu*: 15; *o*: 36).

5.5.3 Identifying Features

Using learning-based methods requires the data to be represented as a set of features. In this study, 13 features were utilized for this purpose, as listed in Table 5.6. These are selected as a combination of features used in some previous methods, as well as some additional features such as candidate length (i.e. Cand_length). Four of the features are properties of the demP, seven of them are related to the candidate antecedent and two are properties of the (demP, candidate) pair.

All the previously mentioned computational studies on pronominal resolution in Turkish use some similar versions of these features in their systems. For example, Yıldırım and Kılıçaslan (2007), Yıldırım et al. (2007), and Kılıçaslan et al. (2009) all use case, grammatical role, type, person/number, referential status and position, which is used to calculate distance, information. They also have a semantic type feature, which was not available in our Treebank data, and an overtness feature, which is used to handle zero pronouns. Since, this study focuses on only the overt use of pronouns; this feature was not included in the feature set.

The features of a (demP, candidate) pair were identified from the annotations of the Treebank available or calculated in the candidate list generation procedure. The details of this procedure are explained in the following.

Table5.6: List of Features Used

Feature	**Explanation**
Dem_Number	the person-number information of the demonstrative
Dem_Case	the case information of the demonstrative
Dem_GR	the grammatical role of the demonstrative
Dem_type	the bu/ şu/ o type of the demonstrative
Cand_POS	the POS information of the candidate
Cand_Number	the person-number information of the candidate
Cand_PossSuf	the possessive suffix information of the candidate
Cand_Case	the case information of the candidate
Cand_GR	the grammatical role of the candidate
Cand_length	the length (in word counts) of the candidate span
Cand_distance	the distance (in word counts) btw. the candidate span and the demonstrative
Cand_type	the word/ sequence/ phrase/ sentence type of the candidate
Ref_status	is the candidate the true antecedent of the demonstrative (yes/no)

5.5.4 Candidate Generation and Preparation of the Training Data

In order to prepare the data fed into the learning algorithms in the form of feature vectors, first a candidate generation procedure is performed and a list of (demP, candidate) pairs is generated. Then, the corresponding features of each pair are added aggregated to form the feature vectors. These feature vectors are then fed into the learning algorithms as input data.

For each demonstrative pronoun in the Treebank, a candidate list is generated composed of different combinations of words, sequence of words, or whole sentences. Looking back 10 sentences, sentences are added to the list of candidate antecedents in their entirety. Then for the sentence containing the demP, all the words up to the pronoun are taken as separate candidates. Finally, the part of the sentence up to the demP is taken as a whole sequence and added as a candidate. If it is not already in the candidate list, the true antecedent span of the demP is also added to the candidate list.

After all the candidate spans are generated for all annotated demonstratives, feature vectors are populated for each (demP, candidate) pair. The features of the last word of a candidate are selected for candidates that are sentences or sequences of words, i.e. which are not single words themselves. The last feature, the referential status is identified as "Yes" if the candidate is in fact the true antecedent of the given demP; otherwise, it is set as "No".

5.5.5 Classification Algorithms Used

Two sets of experiments are run on the Weka tool (Version 3.6.4) for a group of twelve learning algorithms. These are selected as similar to Kılıçaslan et al. (2009) and Yıldırım et al. (2007) studies; as Naïve Bayes algorithm, k-Nearest Neighbor classifier, Decision-Tree classifier, Voted Perceptron and finally the PART algorithm. Two baseline algorithms are also run to compare the performance of the classifiers. Baseline algorithms used are ZeroR and OneR, respectively. ZeroR performs a majority class prediction if the attribute value is nominal, or predicts the average value if it is numeric. OneR method chooses a single attribute which classifies the data best.

k-Nearest Neighbor Classification: This algorithm is implemented using the IBk classifier in Weka. It is a lazy basic instance-based learner which finds the training instance closest in Euclidean distance to the given test instance and predicts the same class as this training instance. It looks at k nearest neighbors. In this study k is taken as 1 and 11 to have two separate classifiers.

Decision-Tree Classification: This learning algorithm is a "divide-and-conquer" approach. The decision nodes compare an attribute value with a constant, compare two attributes with each other, or use some function of one or more attributes. Classifying a test instance involves routing down the tree according to attribute values to a leaf node. Hence, the instance is classified as of the class of the leaf.

An implementation of Quinlan's the C4.5 decision tree (Quinlan, 1993) - predefined in Weka as J48 classifier- is used in three versions in the experiments. The two versions use a pruned and unpruned tree with deafult settings, while the third is set as a pruned tree with boosting. Weka's AdaBoostM1 algorithm is used with J48 for this latter version.

Voted Perceptron: This neural network algorithm is a weighted version of the kernel perceptron classifier. It involves the contribution of each weight vector by a certain number of votes. These votes are determined by the measure of the number of successive trials after the inception of a weight vector, in which it is kept unchanged due to the fact that it correctly classified subsequent instances. This algorithm is also implemented using Weka's built-in voted perceptron function. Two versions are run; one with exponent set to 1.0 and the other with exponent set to 2.0.

PART Classification: This is a separate and conquer algorithm which obtains rules from partial C4.5 decision trees. The built-in version in Weka is used as is.

5.5.6 Experimental Results

A first set of experiments conducted using a percentage split and a second set of experiments, using ten-fold cross validation technique was conducted for the group of twelve learning algorithms previously described.

The percentage split experiments split the input data from 66% to determine a training set and a test set. This is the default setting in the Weka environment (i.e. two-thirds training set and one-third test set), where the instances are ordered randomly. The results of these experiments are given in Table 5.7. The v and * symbols next to a number represents significantly better or worse performance, respectively; whereas an absence of a symbol represents similar performance with Baseline-1 (ZeroR).

Table5.7: Percent Split Classification Results of the Experiments
(66% Training, 33% Test)

Exp No	**Classifier**	**Parametric Variations**	**Percentage Split Classification Results (66%)**				
			Acc.	**P**	**R**	**F**	**Kappa**
1	Baseline-1 (ZeroR)	Majority based	94.34	-	-	-	-
2	Baseline-2 (OneR)	Class distributionbased	94.70v	0.77v	0.08v	0.14v	0.14v[8]
3	Naive Bayes	Normal distribution	94.72	0.68v	0.13v	0.21v	0.20v
4	Naive Bayes	Kernel estimator	94.67	0.74v	0.08v	0.15v	0.14v
5	J48	Pruned	94.90	0.63	0.31v	0.35v	0.33v
6	J48	Unpruned	94.69	0.54v	0.49v	0.51v	0.48v
7	J48	Boosted and unpruned	93.69	0.44v	0.42v	0.43v	0.39v
8	IBk	k=1	89.06*	0.28v	0.56v	0.37v	0.32v

[8] The Kappa statistic provided here is the built-in Kappa measure of the Weka environment. They report it as Cohen's Kappa (Cohen, 1960) in Witten and Frank (2005).

Table 5.7 (continued)

Exp No	**Classifier**	**Parametric Variations**	**Percentage Split Classification Results (66%)**				
			Acc.	**P**	**R**	**F**	**Kappa**
9	IBk	k=11	93.97	0.47v	0.46v	0.46v	0.43v
10	Voted Perceptron	Exponent = 1	94.62	0.75v	0.07v	0.12v	0.11v
11	Voted Perceptron	Exponent = 2	87.57	0.48v	0.21v	0.20v	0.17v
12	PART	Default	95.04	0.59v	0.46v	0.51v	0.48v

The results show similar performance for all algorithms except Baseline-2 (OneR) and IB1, in terms of accuracy. IB1 performs worse than ZeroR, whereas OneR performs significantly better than all other algorithms, demonstrating its main development aim that "very simple structures underlie most of the practical datasets being used" (Witten & Frank, 2005, p. 139). However, the F-measures show similar performance for pruned J48 and voted perceptron (e=1) with ZeroR, and better performance for all the rest of the classifiers. The action taken by the ZeroR method is to predict the class value "no", since it is the most frequent attribute in our data set with more negative examples than positive ones. Hence, in reality the baseline sets the performance a very low standard for our purposes. Overall, the F-measure results display this poor performance as can also be observed from the opposing performances in terms of precision and recall, except for the more balanced but still low results of unpruned J48 (both boosted and non-boosted) and PART.

Table5.8: 10-fold Cross Validation Classification Results of the Experiments

Exp No	**Classifier**	**Parametric Variations**	**Percentage Split Classification Results (66%)**				
			Acc.	**P**	**R**	**F**	**Kappa**
1	Baseline-1 (ZeroR)	Majority based	94.34	-	-	-	-
2	Baseline-2 (OneR)	Class distribution based	94.74v	0.87v	0.08v	0.15v	0.14v[9]
3	Naive Bayes	Normal distribution	94.53	0.58v	0.13v	0.21v	0.19v
4	Naive Bayes	Kernel estimator	94.58	0.67v	0.08v	0.15v	0.14v
5	J48	Pruned	95.00	0.43	0.33	0.37	0.35
6	J48	Unpruned	94.38	0.50v	0.47v	0.49v	0.46v
7	J48	Boosted and unpruned	93.76	0.45v	0.41v	0.43v	0.39v

[9] The Kappa statistic provided here is the built-in Kappa measure of the Weka environment. They report it as Cohen's Kappa (Cohen, 1960) in Witten and Frank (2005).

Table 5.8 (continued)

Exp No	Classifier	Parametric Variations	Percentage Split Classification Results (66%)				
			Acc.	**P**	**R**	**F**	**Kappa**
8	IBk	k=1	89.12*	0.28v	0.56v	0.37v	0.32v
9	IBk	k=11	93.88	0.46v	0.38v	0.41v	0.38v
10	Voted Perceptron	Exponent = 1	94.51	0.74v	0.03	0.06	0.06
11	Voted Perceptron	Exponent = 2	88.07	0.39	0.18	0.17v	0.13
12	PART	Default	94.96	0.57v	0.46v	0.51v	0.48v

The same twelve classification algorithms in the first experiment set were also run with 10-fold cross validation as a second set of experiments. The results are presented in Table 5.8. This second set of experiments were conducted to ensure that the results were representative of what independent test sets would yield. These results display a different related performance between the learning algorithms. It is seen that both Naive Bayes classifiers, voted perceptron with exponent = 1 and the PART algorithm perform similarly with OneR results. Furthermore, these significantly outperform Baseline-1 (ZeroR), J48 classifications, IB11 and voted perceptron with exponent = 2. The IB1 classifier is displayed as the most poor classifier of the given data. In terms of precision and recall all algorithms perform better than Baseline-1, but present a similar imbalance as in the first experiment set. However, J48 and PART algorithms display the best F-measure and Kappa results.

5.5.7 Evaluation and Discussion

The results reveal good anaphora resolution accuracy for the given input data with high classification accuracy. However, there are many issues present here, that require a more thorough investigation.

First of all, lower Kappa statistics in both sets of experiments indicate performance slightly better than chance. However, results of other studies such as Kılıçaslan et al. (2009) also have low Kappa statistics. It should also be noted that since these are learning algorithms, Kappa results as high as those seen in human judgments is merely the ultimate goal, whereas higher than zero results indicate that the algorithms perform better than chance and expectations should be set accordingly.

Other findings can be extracted by looking at the performances and actions of the separate algorithms in classifying true antecedents. In the first set of experiments the OneR algorithm outperforms all others by selecting Cand_length feature as its decision node and setting a length criteria of 38.5 words in order to distinguish between true antecedent candidates and others. This may be interpreted to show that the candidate length feature carries some significance in resolving demonstrative anaphora in Turkish. 94% of the results are correctly classified using this feature in both percentage split and cross validation experiments. Hence, it may be inferred that longer candidate lengths are preferred as true antecedents. However, this is probably due to the fact that our candidate list contains too few variations of length. It may be the case that the longer length examples are in the list solely because they were added as true antecedent instances at a later step. Another and maybe more prominent reason may be the negative/positive instance distribution of the data. The positive examples only form 6% of the data and the correct classification results are probably due to the negative instances.

Very low recall values for the positive classification by individual classifiers is a strong indicator of this. It means that the correctly classified positive instances are very small compared to all the positive instances in the data. The higher precision indicates that if an instance is classified as positive, then it is classified correctly and is truely positive. However, with very low frequency of positive instances, this is not a real indicator of successful classification.

The IB1 results show that, although about 11% of the instances are incorrectly classified, more positive instances are correctly classified than IB11 case, where only about 6% of all instances are incorrectly classified. Hence, the nearest neighbor algorithm which looks a single step, seems to make better sense for demonstrative anaphora resolution than one which looks a distance of k=11.

Comparing the parameter variations of the J48 algorithm shows that the unpruned version, correctly classifies more positives and incorrect classification of positives is also not as much. Boosting version seems to increase the performance of a pruned algorithm but still is outperformed by the unpruned version in terms of correctly classifying positive instances. The PART algorithm, which also uses J48, performs similar to the unpruned version.

Pure data analysis of the annotations for this experiment shows that the true antecedents tend to have the grammatical roles of sentence> object> subject> modifier in terms of frequency (i.e. antecedents are more frequently entire sentences than objects of sentences, etc.). The frequent reference to entire sentences as antecedents suggest that the antecedents are mostly abstract objects. It shows that have been successful in eliminating most of the concrete uses of demonstrative pronouns in this annotation by eliminating determiner uses. Hence, the findings provided here can be beneficial for our phrasal expression identification study presented in Chapter 6, as phrasal expressions also refer to abstract objects.

On the other hand, demP has the most frequent grammatical ordering of object> subject> possessor> sentence (i.e. demPs are more frequently in object position in a sentence), although no correspondence can be observed between the grammatical roles of the antecedent and the grammatical roles of the demP. Similar examination of candidate case shows the most frequent true antecedent case to be equative, followed by nominative; whereas the demonstratives seem to favor the nominative case, followed by accusative case. Candidate POS tags reveal that antecedents are mostly verbs or nouns, as expected. True candidates also seem to prefer 3sg person-number the most. Average true antecedent span is found to be 15.44, where a maximum span of 426 words is observed for one instance. Finally antecedent distance has an average value of 10.88, where the maximum distance is 137.

This was a preliminary attempt on Turkish demonstrative pronoun resolution. Besides the investigations of general demonstrative pronoun reference behavior, the performance of different learning algorithms were compared. Although high accuracy results were observed, concerns which may be influencing this effect have been identified. As a future work direction, it seems two of the most important steps are to provide a better candidate list and to balance the positive to negative instance ratio in the data set provided for the classifiers. This goal may be obtained by increasing the variations in candidate type but finding systematic ways to filter out candidates to have a more balanced distribution of postive/negative examples in the data set. The pure data analysis may prove useful in this respect, where filters are to be applied to the initial candidate list to select the best candidates. Hence a grammatical role filter may be utilized for the candidate, as well as POS tag filters.

Although in this attempt spans of text are identified resolving demonstrative pronouns, there are implications for the identification of phrasal expressions, where expressions are identified to be discursive or not: 1.) the positive/negative instance ratio should be balanced as much as possible 2.) Syntactic information such as POS or grammatical role filters can be used to filter candidates. 3.) A balanced precision and recall outcome (hence higher F-measure) is desired rather than very high precision and low recall, or vice versa. 4.) It is important that the Kappa values are above the zero limit to ensure better than chance results and of course, the higher the Kappa, the better. 5.) Accuracy alone is not an indication of good results, as an imbalanced distribution of data can display high accuracy even for simple majority class prediction.

One remark to note here is that the Treebank provides gold standard data for information such as POS tags, whereas the TDB is not annotated with such information. In an effort to identify phrasal expressions, in order to produce such syntactic information, the TDB has to be parsed with an automatic syntactic parser. Although an effort can be made to align the Treebank information with the TDB since they are both built on the MTC, this will not provide the resource needed to study discourse. The Treebank is merely a collection of separate sentences and even though its gold standard parses proved it to be a valuable resource for identifying demonstrative pronouns, it is not a suitable resource to investigate discourse relations as these by definition relate more than one sentence.

5.6 Conclusions

In this Chapter, we have examined demonstrative pronouns in Turkish as preliminary work to identify Turkish phrasal expressions in the next chapter. Phrasal expressions harbour demonstrative pronouns as deictic elements and involve abstract object anaphora. Hence, in this chapter we have first reviewed the concepts of anaphora and deixis focusing on abstract object anaphora, and introduced the basics of Turkish pronouns focusing on demonstrative pronouns in Section 5.1. Then, in Section 5.2 we overviewed corpus-based annotation studies on reference relations for languages other than Turkish. Reference relation studies for Turkish were reviewed in Section 5.3. It was observed that mainly syntactic features are employed in Turkish pronominal anaphora resolution. However, most of these studies performed sentence-level resolution and it remains to be seen if syntax is adequate to resolve/identify phrasal expressions which require resolution beyond the sentence-level.

In Section 5.4, we presented an analysis of Turkish Demonstrative anaphora in the TDB along with a control study to ensure the correctness of our generalizations. Finally in Section 5.5, we presented an experiment to resolve demonstrative pronouns in Turkish. This enabled us to test the general architecture and observed and the features utilized in prior resolution studies. Some of the main points learned include the importance of positive-negative instance balance, possibility of using filters to narrow down the search space, as well as aid in the balance of the instances, and key ideas in interpreting the results of the system, such as aiming for a high f-measure arising from a balanced precision-recall result. The know-how obtained from the findings in this chapter can be applied to the next chapter where we identify discursive uses of phrasal expressions.

CHAPTER 6

PHRASAL EXPRESSIONS IN TURKISH

This chapter is about phrasal expressions in Turkish, which were briefly introduced in Chapter 2 (Section 5). Phrasal expressions are discourse relational devices which have a deictic demonstrative counterpart combined with a subordinating conjunction, e.g. *buna rağmen* 'despite this'. In this chapter we first provide some examples of what phrasal expressions are, along with a discussion of the decision to annotate them together with connectives, in relation to the approaches taken by other studies (e.g. PDTB, PCC, PDiT). Then, in the rest of the chapter we focus on phrasal expressions in Turkish providing examples from the TDB. Finally, we provide an application utilizing the phrasal expression annotations in the TDB: a model for automatically extracting discourse uses of phrasal expressions from any written text in Turkish, which will also provide a means towards full coverage of the TDB, as it would enable us to discover those phrasal expressions that have not been annotated in the first version of the TDB. Before developing our automatic identification model, however, we first provide an overview of other studies that have tackled the task of identifying discourse relations in general.

6.1 Phrasal Expressions as a Kind of Discourse Relational Devices

In TDB 1.0 one of the annotation decisions taken along the way was to annotate phrasal expression variations as a type of discourse relational devices. Initially the annotations had started with a predetermined set of discourse connectives. Throughout the preliminary annotations, it was seen that there were many phrasal variations derived from connectives. In order to capture the productivity these phrasal variations bring forward in achieving discourse coherence, all these variations were annotated while annotating explicit discourse connectives. These phrasal expressions included a deictic counterpart, which involved demonstrative reference. The deictic item in such phrasal expressions can access the inference in the prior discourse (Webber et al., 2003). They refer to the first argument of the relation which the phrasal expression connects to the second argument. Since the arguments in the TDB are taken as abstract objects, the deictic pronouns also refer to abstract objects. This first argument may be a nonadjacent text span to the sentence containing the connective (Zeyrek & Webber, 2008). Halliday and Hasan (1976) report such prepositional expressions with a reference item to be a subtype of conjunctive expressions denoting a conjunctive relation. Similarly, Göksel and Kerslake (2005, p. 512) classify "postpositional phrases with a demonstrative pronoun as its complement" such as *bununla birlikte* (in spite of this), *onun için* (for that reason) as discourse connectives having adverbial form. Our phrasal expressions contain a subordinating connective and a deictic element. Examples (93) and (96) show the use of a subordinator connective, examples (94) and (95) show the use of phrasal expressions with the same subordinator and a deictic element (e.g. a demonstrative pronoun).

(93) [Köpekler], *bahçenin öbür tarafinda olmalarına* rağmen **havlamaya başlamışlardı**.

[The dogs], despite *being on the other side of the yard*, **had started barking**.

(00001131.txt)

(94) Tabii, *eroinin alındığı günlerin sayısı da artmaya başlar ve eroin kullanımı günlük hale gelir*. Buna rağmen **bağımlı hala eroine bağımlı olmadığını, istediği an bırakabileceğini sanır**.

Of course, *the number of days heroin is injected increases and heroin use becomes daily*. Despite this **the addict still believes he is not addicted to heroin, he can quit whenever he wants**.

(00095133.txt)

(95) *Bizi var eden şeylerden biri sevgi ve aşk diye düşündüm*. Bunun için, **aşk ortak kavram olabilir dedim**.

I thought that one of the things that make us be is love and passion. Because of this, **I said passion may be a common concept**.

(10650000.txt)

(96) *Birkaç yıl sonra artık develerle gitmeyeceklerini düşündüğüm* için **onları fotoğraflamak ve bu yolculuğu yapmak için yola çıktım**.

Because *I thought they wouldn't go with camels in a couple of years*, **I set out to photograph them and to make this journey**.

(10650000.txt)

In the PDTB (Prasad et al., 2010) and also in the Hindi Discourse Relation Bank (HDRB) (Kolachina et al., 2012), expressions such as *after that* are annotated as *Alternative Lexicalizations* (AltLex), as part of the annotation of implicit connectives. Hence, they were not systematically annotated (Webber, et al., 2011). In the Postdam Commentary Corpus, phrasal connectives were annotated as part of complex connectives (Stede & Heintze, 2004). In the TDB, since they were conveniently retrieved while searching for the related subordinator connective in the annotation tool (Aktaş, et al. 2010) and due to their highly productive nature in the language, phrasal expressions were annotated together with explicit connectives so as not to miss an important property of Turkish (Zeyrek et al., 2013). This was akin to Halliday and Hasan (1976, p.75) stating their reasoning in including phrasal expressions into their conjunctives classification as "This also avoids making an awkward and artificial distinction between pairs of items such *as as a result* and *as a result of this*; both of these are interpreted in the same way, as conjunctives". In the future, when other AltLex's are identified, phrasal expressions will be separated from explicit connectives through post processing.

(97) They fought a battle. *After that*, it snowed.

(Halliday and Hasan, 1976, p. 229, ex. [5:I] g.)

(98) He was very uncomfortable. *Despite this*, he fell asleep.

(Halliday and Hasan, 1976, p. 230, ex. [5:2] g.)

In fact, Halliday and Hasan (1976) say that it is the reference item in instances like *after that* (97) or *despite this* (98) that relates the two sentences together. However, three reasons are provided for their inclusion in the classification of conjunctions: 1.) These adjuncts also can present cohesive relations on their own, without the reference item; 2.) Some current conjunctive adverbs have reference items in their origins, although they have been unified into their current forms (e.g. *therefore, thereby*); 3.) Many conjunctive expressions occurring as adjuncts (which may or may not be followed by a preposition) have several almost synonymous forms with/without a demonstrative (e.g. *as a result (of that), instead (of that))*. Thus, Halliday and Hasan's (1976) criterion for being a conjunctive adjunct requires that if there is a conjunctive semantic relation then any expression signaling this relation (with or without a reference item) is considered to be conjunctive.

A general assumption in most studies is that discourse connectives are from a closed-set of expressions, thus annotations are usually based on a list of explicit connectives (Al-Saif & Markert, 2010; Prasad et al., 2008; Stede & Neumann, 2014). Even in studies annotating alternative lexicalizations, these expressions are considered *non-connectives* (Kolachina et al., 2012). Studies such as Prasad et al. (2010) have attested this view of putting restrictions on discourse connectives, saying discourse connectives (hence discourse relations) can be realized in other ways as well.

As we have already mentioned, special constructions in the forms *This/that is why/ when/ how/ before/ after/ while/ because/ if/ etc.*, *The reason/result is* and *What's more* have been called alternative lexicalizations in Prasad et al. (2010). In fact, as explained in Chapter 3 (Section 1), in the PDTB, during the annotation of explicit connectives, in cases where the forms such as *this* and *that* were used to refer to clausal textual spans from the preceding discourse (i.e. they were discourse deictic expressions), annotators were guided to treat it as an exception to the requirement that 'each argument of a discourse relation must contain a verb' and select the deictic item as the argument (see ex. (98)) (Miltsakaki et al., 2004a). The assumption made was that a comprehensive event reference annotation of *this* and *that* would cover these types of adverbials (Webber et al., 2011). Furthermore, sentence-initial prepositional phrases with a deictic argument (*such as for that reason*, or *by then*) were not annotated as explicit discourse connectives, as they were not listed in the predetermined explicit connective set. In the process of implicit connective annotations, annotators were advised to annotate them as an alternative lexicalization if they believed that such an expression caused redundancy when an implicit connective was inserted. As explained in Prasad et al. (2014), due to the limitation of resources and the belief that possible future annotation of deictic coreference would resolve these items, these were not systematically annotated. In the same re-evaluation article for the PDTB, an argument against restricting the set of expressions signaling discourse relations is presented and it was pointed out that many such tokens were under-annotated as AltLex expressions having consequences for the application of machine-learning techniques (c.f. Prasad et al., 2014, pp. 925-927, for a more detailed re-evaluation).

The AltLex expressions annotated in the PDTB were broken down into three main classes by syntactic and lexical flexibility: 1.) Syntactically admitted, lexically frozen 2.) Syntactically free, lexically frozen 3.) Syntactically and lexically free (Prasad et al., 2010). The first class consisted of expressions fitting the connective definition of the PDTB but were not included in the list of connectives, hence were additional connectives identified through the AltLex annotations. The second class consisted of expressions such as *what's more*, which were syntactically parsed differently than the connectives, but are lexically frozen. The final class (forming the majority of the AltLex expressions) freely modifiable but contained a fixed core phrase such as *consequence of, attributed to* and obligatory and optional elements (e.g. *a major reason is, the reason is, a possible reason for the increase is*).

Annotation of alternative lexicalizations have not been done for Turkish, thus, a similar categorization is not possible currently, to see where the phrasal expressions relate to the alternative lexicalizations in Turkish. This is planned for future work. The phrasal expressions annotated in the TDB can have the forms *bu nedenle/sebeple/amaçla* 'for this reason/purpose' from a set of core expressions (Prasad et al., 2010), they can be inflected (e.g. *bu nedenle* 'for this reason', *bu nedenlerle* 'for these reasons') and morphologically they resemble the structure of subordinators (e.g. the phrasal expression *bu-na rağmen* 'despite this' vs. the subordinator -nA *rağmen* 'despite').

Traugott (1997) describes a four stage (Stage 0 – Stage 3) grammaticalization process for discourse markers, where a change can be thought of as grammaticalization if there is an original construction with a lexical item as constituent, and at a later stage fixed forms serve grammatical functions of sentence adverbial or discourse marker. Although various scholars characterize phrasal expressions as explicit connective devices (e.g. Göksel & Kerslake, 2005) it is not clear whether phrasal expressions are grammaticalizing into discourse connectives. Whether the Turkish subordinator constructions of a morphological suffix + subordinator form (e.g. –nA rağmen 'despite') are an intermediary stage in this grammaticalization or not requires detailed linguistic analysis. Traugott's derivations by grammaticalization start at the full lexical noun level (called Stage 0) and continue to discourse marker (in Stage 3), where the expression is used sentence-initially and 'serve pragmatically to evaluate the relation of the up-coming text to that which precedes, and does not evaluate the proposition itself" (*ibid.*: p. 13). The phrasal expressions in Turkish are used sentence-initially and sentence-medially. They are also fixed in their form, although they can be inflected for number. But, there is no bonding of the phrase as in cases like *indeed* in English (from 'in deed'). There is no phonological reduction as in *have to* reducing to *hafta* in English. These and other structural and pragmatical unidirectional shifts (e.g. decategorialization, generalization of meaning, increase in pragmatic function, subjectification) correlated with grammaticalization needs investigation. Nevertheless, their function as a discourse relational device is agreed upon.

Stede (2012) consideres connectives to belong to a closed-set, stating that they are neither easily invented nor are they ever inflected. However, his definition of *cue phrases* (or *discourse markers*) correspond to what we call *phrasal expressions* in the TDB, where they are defined as other lexical items signalling coherence relations. They are stated to be "more open to lexical modification or extension", e.g. *for this reason, for this important reason, for all these reasons* (Stede, 2012, p. 98). Due to their productive nature, cue phrases are excluded from the definition of connectives in his categorization, whereas connectives having multiple tokens are included as *complex connectives* (e.g. *even though, on the other hand*).

In a recent study by Rysová and Rysová (2014), forms equivalent to our phrasal expressions are annotated in the PDT as secondary connectives. The (universal) secondary connectives are defined as multiword expressions which clearly signal discourse relaitons but do not belong to the generally accepted syntactic categories for discourse connectives (i.e. conjuctions and particles) (Rysová & Rysová, 2015). The distinction Rysová and Rysová (2014) make to include such expressions into the connective category is their context dependence, where only context-independent forms (e.g. *for these reasons, because of this*) are considered to be universal secondary connectives. The context-dependent forms, referred to as non-universal connecting phrases, are expressions such as *because of this increase/situation/etc., the reason of his late arrival, etc*. Furthermore, they argue that some previously universal secondary connectives have grammaticalized and are now used as primary connectives (e.g. there + the preposition *fore* in Old English and Middle English, becoming the connective *therefore*). They report similar historical processes observed for Czech *proto* (where the preposition pro 'for' and the pronoun to 'this' is combined), Dutch *daarfoor*, German *dafür* and Danish *derfor*. They suggest that current universal secondary connectives may become grammaticalized to form primary connectives in the future, and that the borderline between primary connectives and secondary connectives is permeable. We agree with this suggestion as it is inline with our intuitions of phrases grammaticalizing or lexicalizing over time, in both English and Turkish. For example, we know that in English the phrase "in its stead" meaning "in its place", where a now obsolete use of the word *stead* means "locality, place"[1], became the connective we now know today as "instead". Another example is the lexicalization of *ne için* 'for what' as a single word and becoming *niçin* 'for what/why', which is a frequently used question word now in Turkish. Furthermore, a lexicalized use of *onun için* 'for that/that's why' in a Balkanian Turkish song as *onçin* makes us wonder if this phrasal expression can be lexicalized into a single word form in the future daily Turkish (see an excerpt of the song in (99) below[2]). Of course, such examples for Turkish phrasal expressions require detailed linguistic analysis.

(99) gidın sorun yengesine
gidin sorun yengesine
go ask her brother's wife
hati kadın niçin aglay
hati kadın niçin ağlar
hati woman why does (she) cry
...
hati kadın onçin aglay
hati kadın onun için ağlar
hati woman cries for that

To sum up, it is seen that in general, other studies have either not annotated phrasal expressions at all, or annotated them under a non-connective classification (such as the AltLex category of PDTB), except for one recent study by Rysová and Rysová (2014), which annotates them as secondary connectives. Some studies excluded them due to their productive nature (as in

[1] Definition of stead taken from Merriam Webster Dictionary.

[2] The first line presents lyrics in the original Balkanian Turkish, the second line presents them in modern day Turkish and the third line provides the English translation.

Stede, 2012), while others failed to annotate them systematically due to limited resources or a priori restriction of a connective list (as in the PDTB, and HDRB). Although categorized under implicit connectives, as alternative lexicalizations, it seems some other studies (such as CDTB: Y. Zhou & Xue, 2014) did systematically annotate them. Irrespective of their classification as explicit discourse connectives, there is a unanymous understanding that these expressions do signal coherence/cohesive relations.

In this thesis we are interested in discursive uses of phrasal expressions, where the deictic item refers to a previously mentioned abstract object. However, some phrasal expressions can be ambiguous in that they can also refer to concrete objects behaving non-discursively. In the previous chapter we analyzed the distribution of demonstrative anaphora in the TDB and found that demonstrative + NP uses (which included phrasal expressions) favor reference to concrete objects. However, our annotations of phrasal expressions in the TDB show that these expression forms more frequently make abstract references. The challenge we face is extracting only the abstract object references made by the phrasal expression forms.

In the previous chapter we have also conducted a machine learning experiment to resolve demonstrative anaphora in Turkish. This experiment enabled us to determine some guiding pointers to use in an automatic identification system for phrasal expressions, such as the importance of balancing the positive/negative instance distribution, applying a filter for this purpose, or the approaches to take in interpreting the results. In this chapter we will utilize the findings in the previous chapters to develop an automatic identification for discursive uses of phrasal expressions in Turkish.

As a first step, detailed examples of the ways phrasal expressions are used in the TDB are provided in the next section, where we try to determine differences in discursive vs. non-discursive uses in order to guide the way to developing a model to automatically extract discursive ones.

6.2 Phrasal Expressions in the TDB and Exploratory Statistics

In this thesis, we focus on these phrasal expressions and try to tease apart discursive uses from non-discursive occurences. In this way, we hope to develop a model to extract such phrasal expressions from any written text in Turkish. This would also provide a means towards fuller coverage of the TDB, as it would enable us to discover those phrasal expressions that have not been annotated in the first version of the TDB. Hence, one of the main aims of this thesis is to automatically detect and retrieve phrasal expressions serving a discursive function.

The challenge in this quest is that, just doing a simple token search to match connective forms is not enough. Some phrasal expressions are not always used as discourse connectives, sometimes they are used non-discursively. Although most phrasal expressions in Turkish are found to be discursive, a few such as *o zaman*, *onun için* and *bunun için* are at times used non-discursively in text. These unannotated occurrences are valuable in terms of finding ways to distinguish discourse uses. We have identified all the occurrences of the phrasal expressions annotated (i.e. discursive) in the TDB and observed the unannotated (non-discursive) uses (see Table 6.1 below). For example, the search token *o zaman* is present in the TDB in 136 occurrences, however, only 84 of these have been determined to be discursive and annotated in the TDB.

Table6.1: Discursive and Non-discursive uses of phrasal expressions in the TDB

Phrasal Expression	English Translation	Discursive Use	Non-Discursive Use
bu amaçla	for this purpose	11 (3)[3]	0
bu nedenden ötürü	because of this reason	1 (0)	0
bu nedenle	for this reason	108 (67)	2 (0)
bu nedenlerle	for these reasons	3 (2)	0
bu sayede	by this means	5 (1)	0
bu sebeple	for this reason	1 (1)	0
bu sebepten dolayı	due to this reason	1 (1)	0
bu yüzden	this is why	58 (29)	2 (1)
buna ek olarak	in addition to this	1 (1)	0
buna karşılık	despite this	23 (19)	0
buna karşın	despite this	19 (12)	0
buna örnek olarak	as an example to this	1 (0)	0
buna rağmen	despite this	12 (8)	0
bundan dolayı	due to this	3 (3)	0
bundan önce	before this	1 (0)	0
bundan ötürü	because of this	8 (6)	0
bundan sonra	after this	13 (6)	17 (5)
bunlara rağmen	despite these	1 (0)	0
bunların sonucunda	as a consequence of these	1 (0)	0
bunun aksine	contrary to this	1 (0)	0
bunun/bunların ardından	subsequent to this	1 (1)	1 (0)
bunun/bunlar için	for this	30 (14)	6 (3)
bunun içindir ki	it is for this	1 (1)	0
bunun neticesinde	as a result of this	1 (1)	0
bunun sonucunda	as a consequence of this	3 (3)	0
bununla birlikte	along with this	2 (2)	0
ne zaman.. *o* zaman	when .. that is when	4 (3)	0
o halde	in that case	4 (3)	0
o nedenle	for that reason	9 (6)	0
o yüzden	that is why	8 (4)	0
o zaman	that is when	84 (39)	48 (8)
ondan sonra	after that	8 (4)	0
onun için	for that	15 (12)	15 (12)
önce .. ondan sonra	before.. after that	1 (0)	0
şu halde	in this/that case	1 (1)	0
TOTAL		444 (253)	91 (29)

[3] Numbers inside the parantheses represent capitalized uses within the total use, indicating sentence-initial use.

(100) *Sırıtınca beni küçümsediğini anlardım tabii*, o zaman **sinirlenirdim**.

When you grinned I knew you were looking down on me of course, then **I would get mad**.

(00010111.txt)

(101) Melendiz'e bakan tarafta duvar şimdi yok; ama o zaman var mıydı, bilmiyoruz.

There is no wall on the side looking at Melendiz now; but we do not know if there was one at that time.

(00013112.txt)

Examples of discursive and non-discursive uses for *o zaman* are given in (100) and (101). In (100), the phrasal expression *o zaman* refers to "*the time when you smirked and I knew you were looking down on me*", which is "when I would get mad". However, in (101) it refers to an unknown date in the past, not an event or an eventuality.

Similarly, *onun için* is also frequently seen in non-discourse uses, where onun ('his/her) is used as a personal pronoun taking a genitive case marking and complementing the postposition *için* ('for) as in (102). In fact, this is as frequent as its discursive use (103) in the TDB.

(102) Ölmek doğmak kadar doğaldı onun için.

Dying was as natural as being born for him/her.

(00058211.txt)

(103) ... *çocuklar ise hiçbir düşüncelerini kendilerine saklamazlar hemen söylerler*, onun için **onlarla konuşurken daima ilginç bir şeyler duyabilirsiniz**.

...children however, do not keep any of their thoughts to themselves they say them right away; that is why **when you talk to them you can always hear interesting things**.

(00047124.txt)

There are also cases where the expression connects a noun phrase (NP) to a verb phrase (VP). These were also considered as non-discursive uses, as in (104). In this example, *Bunun için* refers to the ANAP-DYP government coalition that AKP will support from the outside, which is an NP.

(104) Şener, [AKP'nin dışarıdan destekleyeceği ANAP-DYP hükümeti] ile ilgili senaryonun anımsatılması üzerine "Bunun için Başbakan'ın görevini bırakması gerekir...." dedi.

Şener, when reminded of the scenario about AKP supporting an ANAP-DYP government coalition from outside, said "For this, the Prime Minister has to leave his post..."

(10030000.txt)

6.2.1 Exploratory Statistics for Phrasal Expressions in the TDB

In this section we provide some exploratory statistics for the phrasal expressions annotated in the TDB. We explore features such as differences in sentence-initial and sentence-medial uses, or the argument configurations of discursive uses, that could help in distinguishing discursive vs. non-discursive uses of phrasal expressions.

An analysis of the capitalized first letter of phrasal expressions (sentence-initial uses) show that sentence-initial uses of these expressions form more than half (~56%) of the discursive uses (i.e. 253 capitalized forms out of 444 as seen in Table 6.1). On the other hand, only ~30% of the non-discursive uses appear sentence-initially (i.e. 29 out of 91). Most of the other uses have been identified to be sentence-medial. However, *o zaman* (6), *bu yüzden* (2), *bu nedenle* (1) have sentence-final discursive uses; *o zaman* (6) and *onun için* (2) have sentence-final non-discursive uses, as well, where counts are indicated inside the parenthesis. As sentence-initial adverbials are able to link entire groups of sentences to each other (Stede, 2012), in an automatic argument span selection task (or segmentation task), it would be easier to identify sentence-medial uses as syntactic features could be of more help. Both arguments of a sentence-medial connective will likely be in the same sentence, whereas in sentence-initial uses the Arg1 would be in another sentence (Prasad et al., 2010). Although this does not aid in identifying discursive connectives, such information will be helpful in future argument span identification studies.

In order to get a better understanding of how phrasal expressions are used discursively so that discursive uses can be teased apart from non-dicursive uses, the argument configurations of the phrasal expressions annotated in the TDB and the use of modifiers with these expressions were also analyzed. For example, the phrasal expression *o zaman* is observed in four different configurations. It is mostly observed in a [arg1][conn][arg2] configuration (56 discursive uses), followed by 20 occurences in which it is embedded inside the second argument as [arg1][arg2[conn][arg2]. The parallel uses with another connective or uses with a modifier may also aid in the extraction of discursive uses. In 7 of the instances, there is a connective-final use as [arg1][arg2][conn] and in 4 instances it is used in parallel with *ne zaman* with the configuration [conn1][arg1][conn2][arg2]. In 7 occurrences it is seen together with another intervening discourse connective. These connectives are *önce, ve, aksine, yani, çünkü* and *ama*. The modifiers used with this phrasal expression are *işte, ancak, da* and *belki*, where all except *da* appear before the connective.

Investigating the different co-occurences of configurations of phrasal expressions we are trying to distinguish discursive uses from non-discursive uses and see if there are any features that aid in this distinction. Looking at the configurations of phrasal expressions in this regard, we see that all the 15 annotated examples for *onun için*, display the same configuration. It is always used as [arg1] [conn] [arg2], with a single occurrence of *de* as a modifier for arg1 before the connective. The overall 11 discourse uses of *bu amaçla* are divided into two different configurations: 7 cases show a regular [arg1][conn][arg2], whereas 4 cases display the connective final [arg1][arg2][conn] sequence. The only occurrence of *bunun ardından* displays what we can call the regular connective configuration (i.e. [arg1] [conn] [arg2]). The only instance of *bu nedenden ötürü* also displays the regular configuration. *Bunun aksine*, however, presents a [arg1][arg2 [conn]arg2] configuration where it is placed in the middle of arg2 for its only occurrence in the TDB. *Bu nedenlerle* shows the same sequencing in one occurrence, but presents the regular configuration in the two other uses, with a single use along with the modifier *bütün*. Again, most of the uses (100 instances) of *bu nedenle* are in the regular configuration, but in 13 cases, it is embedded inside arg2, in one other case it is used connective-finally, and in another it is embedded inside arg2 along with arg1 as [arg2 [arg1][conn]arg2]. All the discourse uses of *o yüzden* and most of the *bu yüzden* instances (56 of them) appear in the regular configuration. *Belki de, sadece* (prepositionally), *de* and *mi* (postpositionally) can modify these discursive uses.

However, since in an initial identification of an existence of a discourse relation for phrasal expressions we do not have the arguments identified, we cannot benefit from argument configurations. This information can be utilized in automatic argument assignment, which is among the intended future work. Nevertheless, the use of modifiers with the phrasal expression may be explored in terms of suggesting a discursive use.

In what follows, first an overview of related studies involving automatic identification of discourse structures will be presented. Then, our data and methodology in identifying phrasal expressions will be explained in detail.

6.3 Studies on Automatic Identification of Discourse Relations

In recent years, there has been a growing interest in studies involving automatic identification for both the existence of discourse relations and labeling these discourse relations. Among some of these are Marcu and Echihabi (2002), Burstein et al. (2003), Hutchinson (2003), Wellner et al. (2006), Pitler and Nenkova (2009), Louis, Joshi, and Nenkova (2010) which investigate labeling discourse relations/segments. Studies which determine discursive uses of discourse connectives include Marcu (2000), Hutchinson (2004), Wellner et al. (2006), Pitler and Nenkova (2009), Louis and Nenkova (2010) and Polepalli Ramesh et al. (2012). Other studies such as Soricut and Marcu (2003) identify elementary discourse units, whereas Wellner and Pustejovsky (2007) and Elwell and Baldridge (2008) identify discourse relation arguments by detecting their heads. Torabi Asr and Demberg (2013), on the other hand, provide a probabilistic measure to identify the strength of linguistic cues for discourse relations, which they suggest can be used in an automatic identification task. An overview of these studies is presented in Table 6.2.

Marcu (2000) uses surface-based algorithms to identify discursive uses of cue phrases, segment sentences into clauses, label rhetorical relations and produce rhetorical structure trees. The study uses punctuation to segment sentences into EDUs which are related to each other by the connectives. For example, an occurence of *Although* in sentence-initial position and a comma in the same sentence is used to segment the sentence into two parts (EDUs) which are related by the connective. A total of 2100 instances (1197 discursive) were annotated with discourse-related features (occurrences of punctuation, functional role; position of the marker, right boundary, link direction, rhetorical relation, types of textual units, nucleus/satellite, clause distance, sentence distance, distance to salient unit). This information was used to determine regular expressions for each cue phrase to retrieve discursive uses and determine clause-like unit boundaries from unannotated text. Marcu (2000), states that knowledge lean, shallow parsing is not sufficient to disambiguate ambiguous markers. According to Marcu disambiguating the different relations conveyed by a given connective requires a complete semantic analysis where the intentions of the writer must be understood. This is deemed impossible using shallow parsing. For this reason, Marcu (2000) used mostly unambiguous cue phrases. Almost 81% accuracy and about 90% precision was achieved for both discourse marker identification and unit boundary segmentation.

Table6.2: Overview of Studies for Automatic Identification of Discourse Relations

Study	**Corpus**	**Discourse Structure Identified**	**Theory**	**NLP/ML Methods Used**
Marcu (2000)	Brown Corpus	Disc. use of cue phrases, rhetorical relation labeling	RST	Surface feats (punc., func.role; pos., right bound, rhetorical rel., types of text. units, nucleus/satellite, dist.)
Marcu and Echihabi (2002)	LDC corpora + BLIPP	Disc. Rel. labeling	-	Naive Bayes classifiers Lex. feats (cue phr. and patterns), most repr. word pairs
Soricut and Marcu (2003)	RST Disc. Treebank, Penn Treebank	Identify EDUs	RST	Lexical and syntactic features
Burstein et al. (2003)	Student Essays	Label disc. segments	RST	Decision-tree with boosting; lexical, syn., gram. feat. (e.g. punc)
Hutchinson (2003)	BNC	Disc. Rel. Labeling		Textual co-occurence, syntax (parse trees)
Hutchinson (2004)	WWW	Disc. vs. Non-disc. relations	-	Surface forms, Syn. parses
Wellner et al. (2006)	GraphBank	Disc. vs. Non-disc. relations, Disc. Rel. labeling	-	Max. Ent. classifier; Syn. parses, events (modal + temporal parses), gram.feats, word-pair sim., dist.
Wellner and Pustejovsky (2007)	PDTB	Identify argument heads of disc. conns.	-	Log-likelihood ranking; dep. parses, dist. and POS limit
Elwell and Baldridge (2008)	PDTB	Identify argument heads of disc. rels.	-	Max. Entropy ranker; morph. and syn. feats.
Pitler and Nenkova (2009)	PDTB	Disc. vs. Non-disc. relations, Disc. Rel. labeling	-	Lexical form + Syn. parses
Louis, Joshi and Nenkova (2010)	OntoNotes, PDTB	Implicit Disc. rel. Labeling	-	Linear SVM classifier; Syntax + Entity-related feats.

Table 6.2 (continued)

Study	Corpus	Discourse Structure Identified	Theory	NLP/ML Methods Used
Louis and Nenkova (2012)	PDTB	Disc. vs. Non-disc. relations	-	HMM; co-occurence of syntactic patterns (parse tree prod., phr. node seq.)
Polepalli Ramesh et al. (2012)	BioDRB, PDTB	Disc. vs. Non-disc. disc. conns.	-	Supervised ML (support vector machines, conditional random fields)
Torabi Asr and Demberg (2013)	PDTB	Strength of ling. cues for disc. rel. (prob. of a rel./cue)	Bayes' thm.	-

Marcu and Echihabi (2002) train a family of naive bayes classifiers on a large set of examples generated automatically from an unannotated corpus (which they call the Raw corpus) of ~40M English sentences formed by aggregating several corpora obtained from the Linguistic Data Consortium and the BLIPP corpus, which is a corpus of ~1.8M automatically parsed English sentences (Charniak, 2000). Focusing on four types of relations (CONTRAST, CAUSE-EXPLANATIONEVIDENCE (CEV), CONDITION, and ELABORATION), the study develops techniques explained below to label the relations that hold between input sentence pairs even if the discourse relations are not explicit and achieves up to 93% accuracy.

In Marcu and Echihabi (2002), the hypothesis is that lexical item pairs can provide clues about the discourse relations that hold between the text spans in which the lexical items occur. In order to train the classifiers, they automatically construct datasets using simple rules such as extracting sentence pairs that have the keyword "But" at the beginning of the second sentence to collect CONTRAST relation examples and extracting sentences that contain the keyword "because" to collect examples of CAUSE-EXPLANATION-EVIDENCE relations. In this way, cue phrases and patterns were used to extract samples of the four selected relations from the raw corpus. About 3.9M examples for the CONTRAST relation, ~900K examples for the CEV relation, ~1.2M examples for the CONDITION relation and ~1.8M examples for ELABORATION relation were extracted. Also, non-relation samples (1 million examples) were extracted by randomly selecting non-adjacent sentence pairs that are at least 3 sentences apart in a given text, and cross-document nonrelations (1 million examples) were extracted by randomly selecting two sentences from distinct documents. Removing the cue phrases used to extract the sample discourse relations, a word-pair based classifier is trained for each discourse relation pair in the samples. They report that each classifier outperforms the 50% baseline in distinguishing a specific one of the discourse relations. A six-way classifier trained similarly displays a 49.7% performance, where the baseline is 16. 67% for labeling all relations as CONTRAST.

In a second set of experiments, the discourse relation samples were extracted from the BLIPP corpus. This sample set had fewer examples (~186K CONTRAST, ~45K CEV, ~56K CONDITION and ~33K ELABORATION relations examples). Again, nonrelation samples (58K) and cross-document nonrelation samples (58K) were added. Using the parse trees of the

BLIPP corpus, what they called *most representative words* of each sentence were selected by extracting nouns, verbs and cue phrases. Training classifiers for each discourse relation with this new dataset dataset achieves up to 82% accuracy. However, it is shown that using only the most representative word pairs as features, only 100K examples are needed to achieve the same performance of training over 1M examples using all word pairs as features.

A final experiment is done on a manually annotated corpus of RST discourse trees (Carlson et al., 2001). The four discourse relations in question were automatically extracted from this corpus yielding 238 test cases for CONTRAST, 307 test cases for CEV, 125 test cases for CONDITION and 1761 test cases for ELABORATION relations. Binary classifiers were retrained on the Raw corpus without removing the cue phrases. The results show that CONTRAST and CEV, CONTRAST and CONDITION, CEV and CONDITION relations can be distinguished from each other in this way better than the baseline, ELABORATION relation cannot be distinguished so well from other relations. One important achievement of these classifiers is that they can label originally unmarked (implicit) relations correctly.

Soricut and Marcu (2003) identify elementary discourse units (edus) and build sentence-level parse trees using syntactic and lexical information on the RST Discourse Treebank (RST-DT), where the syntactic information is gathered from the associated syntactic trees of the Penn Treebank. Lexical and syntactic features were used to determine the probability of inserting discourse boundaries to identify edus and build sentence-level RST-style discourse trees. A training set of 5809 instances and a test set of 946 instances were used. The 110 different rhetorical relations in the RST-DT were compacted to 18 labels, where levels of granularity enabled to pinpoint the levels where a difficulty was observed in assigning a label. A two phase approach consisting of discourse segmentation and discourse parsing was applied.Discourse segmentation was further divided as sentence segmentation and sentence-level discourse segmentation. The probabilistic method determines the probability of inserting a discourse boundary using both lexical and syntactic features. The corpus is used to estimate if a boundary should be inserted between two words. It is said that without lexicalization syntactic context is too general and fails to determine discourse boundaries. For the discourse parsing the parsing model assigns probability to every potential candidate pair. They achieve up to 85% precision and recall for their models. Their experiments show that there is a correlation between syntax and discourse at the sentence-level.

Burstein et al. (2003), trained a machine-learning system to identify discourse elements in student essays. The system segments student essays into discourse spans in linear order, assuming that each span has an overall essay-specific communicative goal. A decision-tree machine learning algorithm with boosting was used for this purpose. The feature set included rhetorical relations and their status (as either nucleus or satellite), discourse marker words, terms and syntactic structures (e.g. infinitive clauses) functioning as discourse markers, lexical items for general essay and category-specific language, syntactic structure and grammatical features including sentence mechanics (e.g. sentence number, sentence position, sentence-final punctuation). Additionally, a probabilistic-based discourse analyzer weighing label sequence probabilities employing a simple noisy-channel model and two language models; one using local dependencies among labels and aanother a finite-state network of grammatical label sequences was developed. The two models were trained on a set of essays with human-annotated labels using maximum likelihood techniques, where the decision-based model outperformed the probabilistic model, while both systems performed better than baseline in labelling essays.

Hutchinson (2003) automatically classifies the sense of discourse markers using their textual configurations of co-occurences with each other achieving about 75% accuracy. 61 markers were automatically assigned to their respective categories of negative polarity, temporal, additive, causal and hypothetical (as in Knott, 1996). In the automatically parsed British National Corpus (BNC), markers attached to S or VP nodes and markers co-occurring in the same host clause without an intervening marker were considered, where considering k=8 nearest neighbours provided the highest accuracy.

Hutchinson (2004) tries to automatically identify dicourse connectives in the World Wide Web (WWW). As a first step discourse markers were gathered from the WWW by searching for their surface forms in a search engine to collect documents. These documents are parsed to retrieve sentences containing the surface forms of the connectives, repetitions were removed. Parse trees of candidate sentences are used to extract the cases where the candidate immediately precedes the S node in the tree. These are identified as structural connectives with accuracy of 84-91%, where *and, after, as long as, assuming that* and *every time* are tested.

Wellner et al. (2006) automatically classify the type of discourse coherence relation and identify whether any discourse relation exists on two text segments using a set of linguistic features as input to a Maximum Entropy classifier. They achieve 81% accuracy on relation type identification and 70% accuracy on identifying the existence of a discourse relation. Tokenization, sentence tagging, POS tagging, and shallow syntactic parsing was done automatically on 135 documents of the GraphBank. The grammatical relations used were head word, modifier and relation type. Modal parsing and temporal parsing was performed over automatically identified events. They also use word-pair similarities. It is shown that cue word and proximity features, as well as syntactic features and event attributes have the most impact on classifying discourse relations.

Wellner and Pustejovsky (2007) using features derived from parse representation to automatically identify the argument heads of discourse connectives with a log-linear ranking model on the PDTB. In order to constrain the candidate space, they employ a proximity limit and limit the types of POS to verbs, nouns and adjectives. It is observed that dependency parse features improve the results. Upto 76% accuracy for Arg1 and 95% accuracy for Arg2 is presented.

Elwell and Baldridge (2008) develop models for specific connectives and types of connectives, as well as make use of additional features that provide greater sensitivity to morphological, syntactic, and discourse patterns, and less sensitivity to parse quality to automatically identify the arguments of discourse connectives. The argument identification is modeled as identifying the heads of candidate arguments and choosing the best one. They use a Maximum Entropy ranker on the PDTB. For each connective or connective type (i.e. coordinating connectives, subordinating connectives and adverbial connectives) separate models, as well as interpolated models are trained using all instances. Morphological stemming and syntactic features are used, where about 78%, 82% and 94% accuracy is achieved for connective, Arg1 and Arg2 identification, respectively using gold standard parses. With automated parses they achieve about 74%, 80% and 90% accuracy for the three spans.

One other related work is by Pitler and Nenkova (2009), which uses syntactic features to distinguish discourse vs. non-discourse uses of discourse connectives, as well as disambiguate between different senses of discourse connectives. The PDTB corpus, which contains annotations of ~18K instances of 100 explicit discourse connectives, is used to train and test a maximum entropy classifier using ten-fold cross-validation in order to distinguish between discourse and non-discourse uses of the connectives. The annotated instances of explicit connectives were taken as positive examples and unannotated PDTB texts with the same strings were taken as negative examples. The syntactic features extracted from the PDTB gold standard parses were self category, parent category, left sibling category, right sibling category and the information Right Sibling Contains a VP and Right Sibling Contains a Trace. The baseline was determined using only the connective string as a feature, providing an f-score of 75.33% and accuracy of 85.86%. Using the previously determined syntactic features f-score was 88.19% and accuracy was 92.25%. Using both the connective string and syntactic features increased the f-score to 92.28% and accuracy to 95.04%. The same features were also used to disambiguate four senses of explicit connectives using ten-fold cross-validation with a Naïve Bayes classifier. The baseline of just the connective string feature turned up 93.67% accuracy, whereas using the syntactic features increases the accuracy to 94.15%, which is reported to be the performance ceiling as the human annotators had 94% inter-annotator agreement on sense annotations.

Louis et al. (2010) use entity-related features from OntoNotes to predict implicit discourse relations in the PDTB. The grammatical role, given/new attribute, POS, modifiers, topicalization and number features of referring expressions in a sentence are used, where instances of the training data are sentence pairs. The negative/positive distribution of examples were equalized using random down-sampling of negative instances. A linear SVM classifier was trained for each relation type. Better than baseline performance is observed using only entity-related features, however performance is lower than only using basic word pairs as features.

Louis and Nenkova (2012) identify the existence of a coherence relation between sentences using syntactic patterns. A local model uses co-occurence of structural feautres in adjacent sentences, whereas a global model uses the same features for clusters of sentences. Syntactic patterns are determined using parse tree productions and sequence of phrasal nodes with POS tags. They show upto 90% accuracy for their global Hidden Markov Model using sequences of POS-tagged phrasal nodes. The models are also tested aside content and entity-grid models, where the syntax is shown to complement content and entity methods.

Polepalli Ramesh et al. (2012) automatically detect discourse connectives in biomedical text using the BioDRB corpus and the PDTB by utilizing domain adaptation techniques. Supervised ML approaches (probabilistic modeling framework of CRF and SVMs) are used with lexical and punctuation patterns, where 0.76 F-score is reported with a hybrid domain adaptation using syntactic features. However, in-domain CRF-based classifier using syntactic features also performs similarly (0.75 F-score). The study further reports that 76% of connectives are ambiguous.

Torabi Asr and Demberg (2012) propose a measure to define the strength of linguistic cues for discourse relations. The measure tries to reflect how well a discourse marker makes the discourse relation explicit in text, where the probability of a relation given a cue is estimated using the Bayes' theorem. On 30 relation senses in the PDTB, cue strengths for 95 connectives were identified by defining the most reliable cue for a given relation and the strongest marked relation. They suggest the use of this strength measure to automatically identify discourse relations, where every phrase or word in a discourse relation are counted as cues.

To summarize, an overall evaluation of previous studies show that the most effective features observed for the automatic detection of discursive relations are lexical form and syntactic properties of discourse connectives (or cue phrases). Similar tendency is observed for sense labeling. Using syntactic parse trees and dependency information seems to provide benefits. We can also see that connective specific or sense specific approaches have displayed better results, as expected. Another feature exploited was co-occurences of word pairs, which proved effective as seen in Hutchinson (2003), Wellner et al. (2006) and Louis and Nenkova (2012). In light of these findings and our theoretical motivations as stated in Chapter 1 of this thesis, we take a similar route and exploit the use of morphological, syntactic and dependency parses, alongside lexical form. In what follows, we develop a classification model on a separate training set and try to identify if our test set has any deictic phrasal expressions serving as discourse connectives.

6.4 A Method to Automatically Extract Discursive Phrasal Expressions in Turkish

In this section we present a first effort to automatically extract discursive uses of phrasal expression in Turkish. We develop a model which utilizes lexical form, and features derived from morphological, syntactic and dependency parses trained on a pre-separated training set from the TDB with a Maximum Entropy classifier employing a Decision Tree algorithm with boosting. Then, we test our model against a distinct test set and observe better than baseline performance. However, we argue that due to a positive/negative instance imbalance, the results are inflated. We propose a cascaded approach to resolve this issue, where a lexical form filter is applied as a first step and the classifier is applied in a second step. In the cascaded method, a perfect recall and a decrease in the total number of false predictions is reported. In the following we present a detailed account of our data preparation procedures and methodology.

6.4.1 Data Preparation

Since the TDB is not annotated with syntactic information, in order to obtain this information we utilized the freely available ITU Turkish Natural Language Processing Pipeline (ITU NLP Pipeline) (Eryiğit, Nivre, & Oflazer, 2008; Eryiğit, 2014). The ITU NLP Pipeline processes a given text through the following NLP components in sequential order: a tokenizer, a normalizer, a morphological tagger, a named entity recognizer and finally a dependency parser. The final output is in Conll format (Bucholz & Marsi, 2006), producing word form, lemma, coarse grained part of speech (*POS*) tag (*CPOS*), fine grained POS tag (*POS*), a set of syntactic and morphological features (*FEATS*), head of the token (*HEAD*), and dependency relation to the head (*DEPREL*). The pipeline was applied to the TDB text split into sentences.[4]

[4] The sentence-divided form of the TDB was obtained from Ahmet Faruk Acar, which was built as part of his thesis data (Acar, 2014).

(105)

(a)

ID	*Word form*	*Lemma*	*CPOS*	*POS*	*FEATS*	*HEAD*	*DEPREL*
3	_	kaç	Verb	Verb	Pos	4	DERIV
4	kaçmak	_	Noun	Inf1	A3sg|Pnon|Nom	5	OBJECT
5	istiyordum	iste	Verb	Verb	Pos|Prog1|Past|A1sg	0	PREDICATE

(b)

Word form	*Lemma*	*CPOS*	*POS*	*FEATS*	*DEPREL*
kaçmak	kaç	Verb_Verb	Noun_Inf1	A3sg| Pnon| Nom	OBJECT
Head Word form	Head Lemma	Head CPOS	Head POS	Head Feats	
istiyordum	iste	Verb	Verb	Pos| Prog1| Past| A1sg	

The separate tokens for derived words as in (105a) were combined into a single token (105b) by merging their CPOS, POS and FEATS features so as not to loose the information of derivation from a root. The *HEAD* links were resolved and added directly as the *Head Word form* and its related features except the dependency relation of the head as in (105b). Then, the data was aligned with the annotated TDB data by offset identification for each token. However, the alignment procedure proved difficult as the normalizer and named entity recognizer in the ITU NLP pipeline corrected spelling or combined what it recognized as named entities into single tokens. Such corrected words could not be aligned, but they did not pose a threat to our study as the phrasal expressions that we are mainly interested in were left unchanged. Two more features were added to the aligned data, namely *isPhrasalForm* and *isDConnForm* by comparing the tokens to lists of phrasal expression forms and discourse connective forms. Hence, if the forms were a match, the corresponding feature was set to 1, otherwise it was defaulted to 0. The alignment of the data was used to add the feature *isDconn*, denoting if the current token has been identified as a discourse connective in the TDB, by comparing the offset indices. This was added to be used as a class identifying feature for the classification. One final feature added was *isPhrasal*, indicating if the corresponding token was annotated as a phrasal expression in the TDB. Thus, a total of 15 features were defined for each token and these tokens with added feature sets formed the instances of our dataset.

For the purposes of training and testing a classifier, this dataset was split into two sets: a training set and a test set. The split was performed by maintaining the genre distribution of the TDB in both sets, where the training set was formed of 2/3rds of the data and the test set was the remaining 1/3rd portion. The number of files taken for each set and their percentages according to the genre distribution, as well as the genre distribution and file counts of the TDB is given in Table 6.3 below. The test set consisted of 66 files amounting to 157,531 instances of tokens with added feature sets, and the training set consisted of 131 files, amounting to 311,611 instances.

Table6.3: Genre Distribution and File Counts of the Training and Test Sets with respect to TDB

Genre	**TDB**	**%**	**Training**	**%**	**Test**	**%**
Novel	29	14.72%	19	14.50%	10	15.15%
Story	28	14.21%	19	14.50%	9	13.64%
Research/Survey	13	6.60%	9	6.87%	4	6.06%
Article	9	4.57%	6	4.58%	3	4.55%
Travel	5	2.54%	3	2.29%	2	3.03%
Interview	2	1.02%	1	0.76%	1	1.52%
Memoir	6	3.05%	4	3.05%	2	3.03%
News	105	53.30%	70	53.44%	35	53.03%
TOTAL	**197**	100.00%	**131**	100.00%	**66**	100.00%

6.4.2 Methodology

In the first round of experiments, a Maximum Entropy classifier equivalent[5] (LogitBoost with DecisionStump) in the Weka environment[6] (Hall et al., 2009) was trained using the training set described in Section 6.4.1 and was evaluated on the test set. In order to provide a comparison, a Naive Bayes classifier was also trained and evaluated on this data set. Two baseline classifiers were defined, where *Baseline1* was set to evaluate all instances to be non-phrasal, and *Baseline2* was set to evaluate all instances to be phrasal expressions. As a first experiment, all the previously described features, except the *isDconn* feature, were used to classify the instances in the test set as either a phrasal expression (*isPhrasal* = 1) or a non-phrasal token (*isPhrasal* = 0), where the isPhrasal feature is chosen as the class feature. Table 6.4 presents the results of the experiment.

[5] The Logistic Regression and Maximum entropy have been shown to be equivalent models for classification (Mount, 2011; Qian, 2013). LogitBoost (Friedman, Hastie, & Tibshirani, 2000) is an additive logistic regression algorithm readily available in Weka.

[6] Weka Version 3.7.11 was used.

Table6.4: Results of Different Classifiers for the Classification of Discursive Phrasal Expressions

Classifier	**P**	**R**	**F**	**Acc. (%)**	**Total True**	**Total False**	**TP**	**FN**	**FP**
Baseline1 (all non-disc.)	0	0	0	99.83	157260	271	0	271	0
Baseline2 (all discursive)	0.005	1.00	0.01	0.47	271	157260	271	0	157260
Naive Bayes	0.72	0.79	0.75	99.91	157388	143	213	58	85
LogitBoost (i=10)	0.75	0.99	0.85	99.94	157439	92	269	2	90
LogitBoost (i=20)	0.77	0.98	0.86	99.95	157447	84	265	6	78

It can be seen that, even *Baseline1* has an accuracy of 99.83%. This is due to the fact that the total number of positive instances in our data set is quite low with respect to the negative instances. Hence the very low accuracy of *Baseline2*. The Naive Bayes algorithm produces a relatively higher accuracy than *Baseline1* and classifies 213 of the phrasal expressions correctly as discursive, missing 58 of them and producing 85 false positives (i.e. instances that are actually non-discursive were labeled as discursive: FP). The LogitBoost algorithm does an even better job and correctly identifies 269 of the phrasal expressions, only missing 2 but producing 90 FP results. Increasing the iteration count of this algorithm only slightly improves the accuracy by increasing the number of correctly identified non-discursive uses, but decreases the true positive (i.e. instances that are discursive phrasal expressions and are marked as such: TP) count.

Table6.5: Feature-based Performance

Feature	**P**	**R**	**F**	**Acc. (%)**	**Total True**	**Total False**	**TP**	**FN**	**FP**
Word Form	0.96	0.19	0.32	99.86	157309	222	51	220	2
Head Word Form	0.70	0.19	0.29	99.85	157289	242	50	221	21
isPhrasal Form	0.74	1.00	0.85	99.94	157433	98	270	1	97
Other [7]	0	0	0	99.83	157260	271	0	271	0

[7] The other features (Lemma, CPOS, POS, Feats, DepRel, Head Lemma, Head CPOS, Head POS, Head Feats, isDconnForm) applied separately were unable to classify the phrasal expressions. All instances were predicted to be non-phrasal.

In order to specify which features are more valuable in distinguishing the phrasal expressions, in a second set of experiments we utilized the features in isolation to identify phrasal expressions. The results given in Table 6.5 show that only 3 features were successful in this identification: Word form, Head Word form and isPhrasalForm. The other features were unable to classify any instances as phrasal expressions on their own.

Table6.6: Models Based on Best-Performing Features

Feature[8]	**P**	**R**	**F**	**Acc. (%)**	**Total True**	**Total False**	**TP**	**FN**	**FP**
Best 3	0.75	0.99	0.85	99.94	157438	93	269	2	91
Best 7	0.75	0.99	0.85	99.94	157439	92	269	2	90

Table6.7: Performance of n-gram Models

Classifier	**n-gram**	**P**	**R**	**F**	**Acc. (%)**	**Total True**	**Total False**	**TP**	**FN**	**FP**
Logit Boost i=10	2-gram	0.75	0.99	0.86	99.94	157440	91	269	2	89
Logit Boost i=10	3-gram	0.75	0.99	0.86	99.94	157440	91	269	2	89
Logit Boost i=10	4-gram	0.75	0.99	0.86	99.94	157440	91	269	2	89
Logit Boost i=10	5-gram	0.75	0.99	0.86	99.94	157440	91	269	2	89

We also investigated the features used in the best resulting classification so far, which is the LogitBoost with i=20. Only half of the original features were used in this classification, namely Word form, Lemma, Head lemma, Head Feats, Head CPOS and isPhrasalForm. These 7 features were combined (Best7) to see if they were enough to classify phrasal expressions in a third set of experiments. Furthermore, the three best performing features in isolation were combined as Best3. The results show that the Best7 is as good as LogitBoost (i=10) and even the Best3 is almost as good (See Table 6.6). This suggests that in the aforementioned setups half of the features are enough to get the best classification, but the question of if the other features could aid in better classification still stands.

In the interest of identifying more phrasal expressions, we try an n-gram approach, where

[8] Best 3: Only the features Word form, Head Word Form and isPhrasalForm are used along with the class feature isPhrasal. Best 7: Only the features Word form, Lemma, Feats, Head Lemma, Head Feats, Head CPOS, isPhrasalForm are used along with the class feature is Phrasal.

n-1 instances following a particular instance are added to the current instance along with their features. In this way, 2-gram, 3-gram, 4-gram and 5-gram datasets were obtained. The classification with 2-grams was slightly better than the 1-gram case, however increasing the n-gram size did not produce any improvements (as seen in Table 6.7). What's more, the 2-gram classifier did not utilize any of the additional features of the 2-grams in its decisions.

In light of the experiments conducted so far, we propose a cascaded approach, which first eliminates instances from the dataset in an effort to balance the positive and negative examples, and then does a classification after the elimination. An initial elimination can be done by using the lexical form of the phrasal expressions and eliminating the obvious non-phrasal expressions. Since we already have an *isPhrasalForm* feature, which was also highly used by the previous classification methods, a filter based on this feature can be applied to identify instances that are certainly not phrasal expressions. Then the classifier can be used to distinguish discursive uses from non-discursive ones.

Table6.8: Results of the Cascaded Model

Classifier	**P**	**R**	**F**	**Acc. (%)**	**Total True**	**Total False**	**TP**	**FN**	**FP**
LogitBoost (i=10)	0.75	0.99	0.85	99.94	157439	92	269	2	90
Phrasal Form Filter + LogitBoost (i=10)	0.77	1.00	0.87	99.95	157447	84	270	1	83

The proposed cascaded system is applied to the training set as follows: 1.) The filter to extract all instances of tokens matching the list of phrasal forms is applied to the training set. 2.) The classifier is trained on the training set. 3.) The same filter is applied to the test set. 4.) The trained classifier is tested on the filtered test data. The result displays an improvement over the unfiltered method as seen in Table 6.8. It enables full recall, at the same time improving precision. Hence, the FN and FP counts are decreased, enabling the total number of correctly classified instances to increase.

6.4.3 Evaluation and Discussion

We have presented a novel effort to identify discursive uses of phrasal expressions in Turkish text. In our models, we exploited the use of lexical forms, as well as a combination of morphological, syntactic and dependency features obtained not from gold standard data, but from automatically parsed data. *Our model* using the whole feature set displayed better performance than both baseline models and the Naïve Bayes model.

We determined the best features for the task and our *Best3* and *Best7* models with only 3 and only 7 features, respectively, performed almost as good as our original model. The phrasal form was found to be the best identifier among the other information. Although phrasal form achieves almost perfect recall on its own, addition of the other features contribute to the precision. Just using lexical form does not eliminate the non-discursive uses, which is what we desire in the first place. Although it seems to provide good results, it retrieves the discursive and non-discursive uses together.

Inspired by the effect of the phrasal form and in an effort to balance the distribution (positive vs. negative) of the instances (i.e. eliminate non-discursive uses at the same time), we developed our *cascaded model*, which outperformed all previous models achieving full recall, better precision and f-score providing better accuracy.

Using the *cascaded model* we observe only one false negative (i.e. instance that is actually a phrasal expression but is identified as non-discursive: FN). This FN instance has the word form *ne*. In order to understand why our model missed this instance, let us examine its occurences. The features[9] of the missed occurence and other occurences of this word form both in the test set and the training set are provided in Table 6.9.

Table6.9: Occurences of *ne* in the Dataset

Word Form	**CPOS**	**POS**	**Feats**	**DepRel**	**Head Word form**	**isPhrasal**	**Data set**	**# of Occ.**
ne[10]	Adj	Adj	_	MWE	zaman	1	Test	1
ne	Pron	Ques	A3sg|Pnon|Nom	MWE	zaman	0	Test	6
Ne	Pron	Ques	A3sg|Pnon|Nom	MWE	zaman	0	Test	1
ne	Adj	Adj	_	MODIFIER	zaman	0	Train	6
ne	Pron	Ques	A3sg|Pnon|Nom	MWE	zaman	0	Train	12
ne	Adj	Adj	_	MWE	zaman	0	Train	1
Ne	Adj	Adj	_	MWE	zaman	0	Train	4
Ne	Adj	Adj	_	MODIFIER	zaman	1	Train	1

From this table, it can be understood that the reason this occurence of *ne* 'what', which is a part of the phrasal expression *ne zaman.. o zaman* (when.. when), is missed by the classifier is that the exact same instance with the same values for its features appears in the training set as non-discursive. Moreover, there is no other exact occurrence with the same feature values. In fact, the only other discursive occurence of this word form is the capitalized form *Ne* (shown as the last instance in Table 6.9). What's more, the DepRel of this other discursive instance has been defined as MODIFIER, whereas our missed instance is identified as MWE by the parser.

How could this instance be correctly predicted in such a circumstance? This question can be tackled by first looking at the actual example that this instance corresponds to (106). Then comparing it with the non-discursive occurence example with the same feature values in the training set (107).

[9] Only the non-matching features are presented in the table. The feature not shown in the table were the same for all occurences of the Word form.

[10] Missed instance, predicted to be non-discursive.

(106) "Ama en fenası," demişti Hikmet Bey, "ne zaman *bir muzırlık yapsa* o zaman **böyle gülüyordu**".

"But the worst is," said Mr. Hilkmet, "whenever *he acts mischievously* then **he laughed like this**."

(00062211.txt)

(107) .. Türk uçakları ne zaman Kıbrıs semalarında uçacak?

When will Turkish planes fly in the Cyprus sky?

(10670000.txt)

The problem arises from the parallel construction of the phrasal expression *ne zaman... o zaman* and the fact that the phrasal form matching does not consider this parallelism. If we had eliminated the non-paralel uses of *ne zaman* in the filtering step, than there would not be any confusion in the classifier. One solution might be to use n-grams with values of *n* that can grasp the second part of the expression, or search for the parallel phrasal forms in a way that the second part is matched with the first part. Such a search, of course, would require a limit to the intervening token count, which would also be used as the value of *n* for the n-gram approach. Since the frequency of such parallel uses is low (i.e. only 5 parallel phrasal expressions found to be discursive in the TDB), this is left for future work.

Although not specifically designed for parallel constructions, overall our cascaded approach achieves good results without gold standard data parses to build the feature set. This enables our model to be applied to any free Turkish text with ease.

It should be noted that, another contribution of this work is the parser information obtained in the dataset preparation step for the TDB. In this step, we identified morphological, syntactic and dependency features for the corpus, which can also be used in other applications. Since the data is aligned according to character offset, this information enhances the TDB and provides additional value.

One possible limitation of our models is that we may miss possible variations that are not in our list of word forms for phrasal expressions. Hence, this method is not expected to identify novel phrasal expression forms. However, it readily extracts the syntactic features mainly used in identifying them. Furthermore, it provides a means to effortlessly enhance the coverage of the TDB since it can easily be applied to the other subcorpora of the METU Turkish Corpus readily available, without the time and effort overhead of the completely manual annotation. A method to use for such a new annotation study can be to first train *our cascaded system* on TDB 1.0 as a whole, prepare and pass the corpus or subcorpus to be annotated through the system and get predictions for the phrasal expressions. Then two human judges/annotators can go over the predictions to decide if they are indeed discursive uses as a group. Since the system achieves near perfect recall, it might only miss a few phrasal expressions and the overhead would only be due to false positives, which need to be eliminated by the annotators. Overall, such a system would provide benefits in terms of time and effort to annotate a corpus completely manually from scratch. An even higher benefit can be observed for cases where non-discursive uses are relatively high with respect to discourse uses as in the cases of the phrasal expressions *bundan sonra, bunun için, onun için* and *o zaman*. A similar situation is observed for explicit discourse connectives, where only about 40% of all the search token occurences were observed to be discursive (see Appendix C, Table C.1). Hence, our cascaded approach could serve as an initial attempt to identify discourse connectives automatically. This idea will be investigated in the next section.

6.5 An Initial Attempt for the Automatic Identification of Discourse Connectives

In this part of the study, we try to apply the cascaded approach used for automatic identification of phrasal expressions to the automatic identification of explicit discourse connectives. The first step for this procedure is to apply a form filter to the dataset, to eliminate word forms that do not coincide with discourse connective forms. The discourse connective forms annotated in the TDB (provided in Appendix B, Table B.1) are used for this filtering. Then the system needs to be trained on the training data, which has the feature class isDconn this time, indicating if the instance is a discourse connective or not. The trained system is finally used to make predictions about which instances in the test data represent discourse connectives. However, before applying our cascaded system, first let us see how well a stand alone classifier identifies the connectives. The results for a Naïve Bayes classifier and LogitBoost are given in Table 6.10, where the LogitBoost classifier provides better accuracy but near zero recall. The 2-gram approach provides a slightly better result, however no improvements is observed for 3-gram, 4-gram or 5-gram approaches. Again, increasing the iteration count of the classifier also provides only slight improvement.

Table6.10: Results of Different Classifiers for the Classification of Discursive Connectives

Classifier	**n-gram**	**P**	**R**	**F**	**Acc. (%)**	**Total True**	**Total False**	**TP**	**FN**	**FP**
Naïve Bayes	1-gram	0.210	0.700	0.330	97.42	153472	4059	980	407	3652
Logit Boost	1-gram	0.067	0.001	0.001	99.11	156131	1400	1	1386	14
Logit Boost i=10	2-gram[11]	0.560	0.006	0.013	99.12	156137	1385	9	1378	7
Logit Boost i=20	2-gram	0.590	0.014	0.028	99.12	156150	1381	20	1367	14

[11] The same results were observed with 3-gram, 4-gram and 5-gram approaches.

Table6.11: Results of the Cascaded Model Applied to Discourse Connective Identification

Classifier	P	R	F	Acc. (%)	Total True	Total False	TP	FN	FP
Disc. Conn. Form Filter + LogitBoost (i=10)	1.00	0.07	0.12	99.18	156235	1296	91	1296	0
Disc. Conn. Form Filter + LogitBoost (i=30)	0.97	0.08	0.15	99.19	156252	1279	111	1276	3
Disc. Conn. Form Filter + LogitBoost (i=50)	0.91	0.09	0.16	99.19	156250	1281	118	1269	12
Disc. Conn. Form Filter + LogitBoost (i=100)	0.84	0.09	0.17	99.19	156248	1283	128	1259	24

The *cascaded approach* provides better accuracy and higher recall then the standalone methods (see Table 6.11), and slight improvements in the TP counts are achieved by increasing the iteration count to 30, after which the accuracy starts to decrease in terms of correctly classified instances in total. Nevertheless, most of the connectives are predicted to be non-discursive.

Just as Stede (2012, p.93) states when comparing two studies of discourse segmentation, lower results can be explained by one study operating on hand-crafted syntax trees, whereas the other used a dependency parser operating under what he calls 'real-life'conditions; our results are affected by not having gold standard syntactic labels. A quick examination reveals that there was a problem with the automatic addition of the *isDconn* class feature due to misalignments in the character offsets of the tokens retrieved from the NLP pipeline. As mentioned before (in Section 6.4.1), the difficulties in offset alignment are caused by the spell checker and normalizer in the pipeline, which correct the word forms. Since the pipeline only returns corrected forms without the original format, the offset alignment can be problematic.

In the phrasal expression identification, due to the small number of expressions at hand, the procedure of adding the *isPhrasal* feature was performed manually. However, for the discourse connectives, it was preferred that the procedure be done automatically, since the total number was not as manageable. Hence, the incorrectly aligned word forms caused discourse connectives to be missed when adding the *isDconn* feature. A correction of this problem in future work will reveal the true affect of *our cascaded system* in the automatic identification of discourse connectives. The application of the NLP pipeline without spell-correction and normalization, or with the original word forms kept in the results can be investigated. This is left as future work as it requires that the text of the TDB be reanalyzed by the NLP pipeline.

CHAPTER 7

CONCLUSION

Here are three sentences presented in a sequence.

(108) Today I went to school. My clothes were dirty. I put on my coat.

At first read, it seems they are unrelated, even incoherent, just mere expressions of three facts. However, our minds are able to provide background information for the situations expressed in these sentences and relate them to each other. More easily so, when we add the following sentence.

(109) Today I went to school. My clothes were dirty. I put on my coat. I did not take my coat off in front of Mr. A. It was hot inside.

Now, some relations between these sentences can be inferred. Although still not explicit, we can infer that there is a relation between the fact that my clothes were dirty and the event of me putting my coat on. Moreover, these two situations may be linked with why I did not take my coat off in front of Mr. A. Here, the repetition of the phrase my coat is of course an overt signal, as well as the contrasting verbs *put on* and *take off*. Now consider the following form:

(110) Today I went to school. My clothes were dirty, so I put on my coat. For this reason I did not take my coat off in front of the *professor* A, although it was hot inside.

In this case, we can clearly observe *lexical cohesion* between the words *school* and *professor*, infer that the whole sequence of events took place at school, where I went and the professor is also likely to be. Furthermore, why I put on my coat is explicitly linked to the fact that my clothes were dirty by the discourse connective *so*. What's more, the same reason also prevents me from taking off my coat in front of the professor, contrary to the fact that general knowledge suggests that I take the coat off since it was hot inside. Moreover, this time the reasoning is overt with the aid of the phrasal expression *for this reason*, and the contrast is explicitly stated through the use of the discourse connective *although*. Clearly, our minds are able to interpret the sentences in (108) and (109), providing the necessary information to link them together in a coherent manner, using past experience, knowledge of the world, etc. However, the process in this case is totally implicit to us and we do not have much idea as to how our minds go about such a task. Furthermore, two minds can structure two completely different backgrounds or contexts for the use of these sentences together. This would bring forth two different interpretations and defeat a main purpose of language, i.e. communication. One actor tells something encoded in a given language and the other actor (receiver) decodes this encoded message to understand what that something is. If the receiving end does not decode the message in the intended way (i.e. infers a different interpretation), then the message is not successfully delivered. Thus, it is easier to identify the workings of this coherence construction when there are explicit signals in the text. In the case of (110), the relations between the sentences are explicitly stated, hence the receiver can correctly decode the message accordingly.

In order to reflect or share what is in our cognition, we use language. To convey our thoughts in language in a comprehensible manner we can use sentences, or even sequences of sentences, or any other longer stretches means of discourse. In the same way that we can tie concepts, ideas, and facts together in our minds, the reflection of our thoughts and ideas in language can be tied to each other through discourse relations (or coherence relations). Use of explicit discourse connectives makes the relations within and between sentences explicit. Through these explicit signals, we can start to uncover the workings of discourse competence. One frequently utilized method to understand discourse is corpus based analysis.

7.1 Summary and Conclusions

In this thesis we presented a quantitative and qualitative assessment of the Turkish Discourse Bank and showed that the TDB can be used as a reliable discourse resource for studying explicitly signaled Turkish discourse relations. Hence the TDB can be utilized to investigate Turkish discourse phenomena for linguistic studies, cognitive science studies, as well as natural language engineering studies and more. In what follows, we will summarize the evaluations and contributions of this thesis.

7.1.1 Reliability Study

In this thesis we provided a quantitative and qualitative assessment of the TDB in order to establish the reliability of this discourse resource for Turkish. Our quantitative evaluation consisted of basic descriptive statistics and in depth statistical measures to ensure reliability of the annotations in this annotated corpus. We utilized nearly all the available reliability measures starting with the most frequently used measure in other such annotation studies, i.e. the inter-annotator agreement. The inter-annotator agreements showed that 20 search tokens displayed perfect Kappa agreement for both arguments, whereas 15 others displayed perfect agreement for one of their arguments, and substantial agreement (i.e. $0.60< K\text{-value}<0.80$) for the other argument. Only 5 connectives showed less than substantial agreement, where the reasons for disagreements were identified as a) missing annotations (i.e. differences in discourse-non-discourse distinction), (b) partial overlap (for either Arg1, Arg2, or both), (c) no overlap, (d) lack of adequate annotation guidelines (i.e. underdetermined cases in the guidelines), (e) failure to comply with the guidelines (annotator error), (f) other (technical errors such as span selection, etc.).

We also used the intra-annotator agreement measure to establish the integrity of the annotations over time, which as far as we know has only been used in one prior study (Sporleder and Lascarides, 2008). 606 relations were re-annotated by the author and their agreement with prior annotations were compared. About half of the re-annotated connectives displayed perfect Kappa agreement for both argument spans and 13 other connectives displayed perfect Kappa agreement for one of their spans. Only three connectives displayed lower than 0.60 Kappa values for one of their argument spans, which were either due to the re-annotation being more in line with the gold standards or the small number of annotations compared for that particular connective. Overall, the intra-annotator agreement was high, displaying that the annotations maintained their integrity over time.

Another measure utilized was the agreements with the gold standard data (i.e. gold standard agreements), which displayed content validity. 11 connectives displayed perfect agreement with the gold data for all their annotators and 26 connectives display either perfect or substantial agreement. Five connectives, displayed substantial agreement for all annotators for one of their arguments, while the other had perfect agreement with the gold standards. Three connectives had substantial agreement for both arguments for all annotators. Another seven connectives displayed gold standard agreement ranging from poor to substantial.The main reasons for these were final updates to the guidelines (i.e. decisions to include/exclude certain usages), as well as lower agreement values due to partial overlaps resulting from the difficulties in selecting the first arguments of discourse adverbials such as *ayrıca*, *dahası* and *sonuçta*.

We used two approaches in the calculation of the agreements, where the first one involved investigating the span agreement differences between relations commonly identified as discursive (called the *Common* method). The other method (called the *Overall* method) also considered the discursive vs. non-discursive relations identification differences by including all annotations even if a given relation was identified as discursive by only one of the annotators. The Overall method also included annotations with discontinuous spans. Comparisons of these two approaches for the reliability measures calculated, enabled us to pin point sources of disagreements more accurately.

Finally, we proposed calculating some extra evaluators (i.e. precision, recall and f-measure) that are originally used in evaluating information retrieval systems, in order to assess an annotator's reliability when comparing independent annotations with the gold standards. We used these evaluators intending to capture if the annotator annotated the same connective instances to be discourse connectives as the gold standard. The results coincides with some of the results obtained from the gold standard agreements of the Overall approach, suggesting that the P, R and F-measure can be used as complementary measures for the Common method, in order to obtain an approximation to the Overall method.

We tried to be as explicit and transparent in our findings as possible and account the reasons for the disagreements observed. We benefited from the two-way methodology used in reliability calculations and suggest it as a method to use in other assessment studies for annotated corpora. Our conclusive suggestions for reliability evaluations of annotated corpora is as follows: First of all, we cannot stress the importance of any type of reliability calculation enough. It allows to understand the annotated data, where high reliability assures the integrity of the resource, but on the other hand identifying the sources of disagreements or patterns in the disagreements allows to identify certain linguistic characteristics about the data as well. Second, we suggest to calculate all three types of reliability measures: inter-annotator agreements, intra-annotator agreements and gold-standard agreements. All of these provide the opportunity to look at different aspects of the disagreements, aiding in identifying the sources. Moreover, we suggest the calculation of extra evaluators of precision, recall and f-measure, borrowed from information retrieval in order to assess an annotator's agreement with the gold data in terms of identifying discursive uses. However, these evaluators can be used as complementary measures and not on their own. Furthermore, we suggest that researchers try to understand the data and learn from the disagreements and not just evaluate their results as above a certain threshold or not. Finally, as in Spooren and Degand (2010), we advise explicitness and transparency, so that the resource can be utilized in other studies and applications accordingly.

7.1.2 Demonstrative Pronouns in Turkish and Resolution of Demonstrative Anaphora

Another contribution of this thesis is the analysis of demonstrative pronouns in Turkish as preliminary work to identify Turkish phrasal expressions. Analyzing demonstrative anaphora we made some preliminary conclusions. It was observed that about 1/3rd of all the demonstrative anaphora in Turkish novels consists of exophora, whereas the rest is endophoric (i.e. within text) uses. Within the endophora, the most frequently used demonstrative in Turkish is found to be *bu* and it is also the most preferred demonstrative for AO reference, where *şu* and *o* are rarely used. On the other hand, demonstrative use of *o*, is preferred for referencing concrete objects. There is also substantial use of *o* as a personal pronoun, dominating all its other uses in terms of frequency.

Phrasal expressions harbour demonstrative pronouns as deictic elements and involve abstract object anaphora. Some phrasal expressions can be ambiguous in that they can also refer to concrete objects behaving non-discursively. Our analysis on the distribution of demonstrative anaphora in the TDB showed that demonstrative + NP uses (which included phrasal expressions) favor reference to concrete objects. However, our annotations of phrasal expressions in the TDB show that these expression forms more frequently make abstract references (91 concrete object references and 444 abstract object reference, See Table 17). One reason for these findings could be that in our analysis the abstract object references observed for demonstrative + NP uses were due to the phrasal expressions. However, a follow-up analysis shows that only 11 out of the 63 abstract uses of demonstrative + NP are phrasal expressions. There are also 5 phrasal expression forms which are non-discursive among the concrete anaphora. Hence, a further analysis is needed to determine which other demonstrative + NP expressions refer to abstract objects. A conclusion that can be drawn is that phrasal expressions do not act like the general category of demonstrative + NP uses. The challenge faced with phrasal expressions is extracting only the abstract object references made by the phrasal expression forms. Further study on other abstract demonstrative + NP anaphora is required to draw any other relation to the phrasal expressions.

Review of studies on Turkish pronominal anaphora showed that mainly syntactic features are employed in Turkish pronominal anaphora resolution. However, most of these studies performed sentence-level resolution and the adequacy of syntax to resolve/identify phrasal expressions which require resolution beyond the sentence-level is a question to answer. We cannot answer this question for resolution of phrasal expression with the findings in this thesis. However, further discussion is provided regarding the use of syntac for the identification of phrasal expressions in the next section.

We also presented an experiment to resolve demonstrative pronouns in Turkish. This enabled us to test the general architecture and observed and the features utilized in prior resolution studies. Some of the main points learned included the importance of positive-negative instance balance, possibility of using filters to narrow down the search space, as well as aid in the balance of the instances, and key ideas in interpreting the results of the system, such as aiming for a high f-measure arising from a balanced precision-recall result. The know-how obtained from theses findings were applied to the identification of discursive uses of phrasal expressions.

7.1.3 Phrasal Expressions

In this thesis we also focused on identifying the discursive uses of phrasal expressions annotated systematically in the TDB. We built models to automatically extract discursive uses of phrasal expressions given any Turkish text and we tested and evaluated this model using the TDB as our gold standard data. The challenge involved effectively disambiguating discursive uses of these phrasal expressions. We presented a novel effort to identify discursive uses of phrasal expressions in Turkish text. We exploited mainly the use of lexical forms. We also used a combination of morphological, syntactic and dependency features as previous studies in this field have done for languages other than Turkish. In our study, the additional morphological, syntactic and dependency features were obtained not from gold standard data (because TDB is not annotated with these features), but from automatically parsed data. Our model using the whole feature set displayed better performance than both baseline models and the Naïve Bayes model. We further developed a *cascaded model*, which achieves full recall, high precision and f-score providing high accuracy. This cascaded model can be utilized to effortlessly enhance the coverage of the TDB by applying it to the other subcorpora of the MTC and manually adjudicating the predictions of the system. This automated application of our cascaded model would be beneficial in terms of both time and effort with respect to manual annotation, as we have demonstrated on our test data. One possible limitation of our models is that we may miss variations that are not in our list of word forms for phrasal expressions. Hence, this method is not expected to identify novel phrasal expression forms. However, our cascaded model successfully distinguishes discursive uses of the phrasal expressions in a given set.

We have also shown that lexical word form is the main requirement for identifying the discourse uses of phrasal expressions in Turkish, as the best features in our models are found to be the word form and head word form. Hence, our observations of words enable us to reach discourse-level structures. Regarding the question of the adequacy of syntax to identify phrasal expressions, we can say that lexical form, syntax (e.g. part-of-speech) and morphology (e.g. case, tense, person agreement) is adequate to provide acceptable, but not perfect results. Further study is needed to see the affect of using semantics for the task of phrasal expression identification.

7.1.4 Towards the Identification of Discourse Connectives

We have also developed a preliminary model to extract discursive uses of explicit Turkish discourse connectives, which is a first for Turkish. Our cascaded approach provides better accuracy and higher recall then the standalone methods. However, due to the misalignments in the output of the automatic morphological and syntactic parser pipeline and the annotated data, most discourse connectives were erroneously predicted to be non-discursive. This can be attributed to the fact that the TDB does not have gold standard morphological and syntactic parses. A correction of this problem in future work will reveal the true effect of our cascaded system in the automatic identification of discourse connectives.

7.2 Limitations and Implications for Future Work

Our assessment of the TDB 1.0 establishes it as a reliable discourse resource for Turkish to the extent explicit discourse connectives are concerned. We have already demonstrated that this resource can be used in natural language applications. However, we should also acknowledge some of the limitations of this resource. First of all, it is not as large as some comparable corpora for other languages such as English (i.e. 8483 relations vs. 40,600 relations). Currently the TDB is only annotated with explicit discourse connectives but implicit relation annotation is underway. Annotations for the sense of the relations have begun and at the moment is completed for three connectives (*ama*, *fakat* and *yoksa*), which will be made available in future releases. Furthermore, a limitation for NLP applications is the lack of syntactic gold parses of the TDB. Automatic parsing may be utilized as we demonstrated in our model to automatically identify discursive phrasal expressions. However, some difficulties arise as seen in our initial attempt to automatically identify discourse connectives.

Of course all the mentioned limitations make room for future enhancements of the TDB. Hence, future work may try to enrich the TDB data, or enhance the coverage of the TDB. These may be done either manually (e.g. by adding annotations for implicit relations, by adding sense labels to the identified discourse relations, by manually verifying the automatic morphological and syntactic parses obtained in this study to create gold parses, etc.) or automatically (e.g. automatically disambiguating discursive uses of phrasal expressions or discourse connectives in additional subcorpora, automatically identifying novel forms of discourse connectives, etc.). Further enhancements to benefit discourse studies may include coreference annotation on the corpus.

The TDB can also be used as a resource for psycholinguistic or cognitive science studies. It can be used to obtain data for experimental paradigms for example in identifying the implicitness of discourse connectives as in a study conducted by Torabi Asr and Demberg (2012), or in self-paced reading tasks, or in eye-tracking studies as in Prévot, Pénault, Montcheuil, Rauzy, and Blache (2015), etc. In conclusion, we hope the work presented in this thesis has established the TDB to be a basis for many future studies perhaps some currently inconceivable, from many diverse areas of research.

REFERENCES

Acar, A. F. (2014). *Discovering the Discourse Role of Converbs in Turkish Discourse*. MS Thesis. METU, Ankara.

Retrieved from http://etd.lib.metu.edu.tr/upload/12616941/ index.pdf

Aktaş, B., Bozşahin, C., & Zeyrek, D. (2010). Discourse Relation Configurations in Turkish and an Annotation Environment. *In Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 202–206). Stroudsburg, PA, USA: Association for Computational Linguistics.

Al-Saif, A., & Markert, K. (2010). The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC'10). European Languages Resources Association (ELRA).

Al-Saif, A., & Markert, K. (2011). Modelling Discourse Relations for Arabic. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 736–747). Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4), 555–596.

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer Academic Publishers.

Atalay, N., Oflazer, K., & Say, B. (2003). The annotation process in the Turkish treebank. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora - LINC*. Budapest, Hungary.

Banguoğlu, T. (1990). *Türkçenin Grameri [Grammar of Turkish]* (3rd ed.). Türk Tarih Kurumu Basım Evi: Atatürk Kültür, Dil ve Tarih Yüksek Kurumu, Türk Dil Kurumu Yayınları.

Bayerl, P. S., Gut, U., Lüngen, H., & Karsten I., P. (2003). Methodology for Reliable Schema Development and Evaluation of Manual Annotations. In *the Workshop Notes for the Workshop on Knowledge Markup and Semantic Annotation, Second International Conference on Knowledge Capture* (K-CAP 2003).

Bejček, E., Hajičová, E., Hajič, J., Jinová, P., Kettnerová, V., Kolávrová, V., … Zikánová, Š. (2013). Prague Dependency Treebank 3.0. Charles University in Prague, MFF, ÚFAL. Retrieved from http://ufal.mff.cuni.cz/pdt3.0/

Bird, C., Nagappan, N., Gall, H., Murphy, B., & Devanbu, P. (2009). Putting it all together: Using socio-technical networks to predict failures. In *Software Reliability Engineering, 2009. ISSRE'09. 20th International Symposium on* (pp. 109–119).

Botley, S., & McEnery, A. M. (2001). Demonstratives in English: a corpus based study. *Journal of English Linguistics*, 29(1), 7–33.

Botley, S. P. (2006). Indirect anaphora: Testing the limits of corpus-based linguistics. *International Journal of Corpus Linguistics*, 11(1), 73–112. doi:10.1075/ijcl.11.1.04bot

Buchholz, S., & Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)* (pp. 149–164). New York City: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W06-2920

Buch-Kromann, M. (2006). *Discontinuous Grammar: A Dependency-based Model of Human Parsing and Language Learning*. Copenhagen Business School.

Buch-Kromann, M., & Korzen, I. (2010). The Unified Annotation of Syntax and Discourse in the Copenhagen Dependency Treebanks. In *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 127–131). Stroudsburg, PA, USA: Association for Computational Linguistics.

Buch-Kromann, M., Korzen, I., & Müller, H. H. (2009). Uncovering the "lost" structure of translations with parallel treebanks. In F. Alves, S. Göpferich, & I. Mees (Eds.), *Methodology, Technology and Innovation in Translation Process Research* (Vol. 38, pp. 199–224).

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, 18(1), 32–39. doi:10.1109/MIS.2003.1179191

Byron, D. K. (2002). Resolving Pronominal Reference to Abstract Entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 80–87). Philadelphia.

Can, F., Koçberber, S., Bağlıoğlu, Ö., Kardaş, S., Öcalan, H. Ç., & Uyar, E. (2009). New Event Detection and Topic Tracking in Turkish. *Journal of the American Society for Information Science and Technology*, 61(4), 802–819.

Can, F., Koçberber, S., Balçık, E., Kaynak, C., Öcalan, H. Ç., & Vursavaş, O. M. (2008). Information retrieval on Turkish texts. *Journal of the American Society for Information Science and Technology*, 59(3), 407–421.

Can, F., Nuray, R., & Sevdik, A. B. (2004). Automatic performance evaluation of web search engines. *Information Processing and Management*, 40(3), 495–514.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2), 249–254.

Retrieved from http://dl.acm.org/ citation.cfm?id=230386.230390

Carlson, L., Marcu, D., & Okurowski, M. E. (2001). Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16* (pp. 1–10). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115 /1118078 .1118083

Chamberlain, J., Kruschwitz, U., & Poesio, M. (2009). Constructing an Anaphorically Annotated Corpus with Non-experts: Assessing the Quality of Collaborative Annotations. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources* (pp. 57–62). Stroudsburg, PA, USA: Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. doi:doi:10.1177/001316446002000104

Demirşahin, I., Sevdik Çallı, A. B., Ögel Balaban, H., & Zeyrek, D. (2012). Turkish Discourse Bank: Ongoing Developments. In S. Demir, I. D. El-Kahlout, & M. U. Doğan (Eds.), *Proceedings of the First Workshop On Language Resources and Technologies for Turkic Languages* (pp. 15–19). Istanbul, Turkey: European Language Resources Association (ELRA).

Retrieved from http://www.lrec-conf.org/ proceedings/lrec2012/ workshops/02. Turkic Languages Proceedings.pdf

Demirşahin, I., Yalçınkaya, İ., & Zeyrek, D. (2012). Pair Annotation: Adaption of Pair Programming to Corpus Annotation. In *Proceedings of the Sixth Linguistic Annotation Workshop* (pp. 31–39). Jeju, Republic of Korea: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W12-3605

Dipper, S., & Zinsmeister, H. (2009). Annotating Discourse Anaphora. In *Proceedings of the Third Linguistic Annotation Workshop, Association of Computational Linguistics* (pp. 166–169). Suntec, Singapore.

Dipper, S., & Zinsmeister, H. (2010). Towards a standard for annotating abstract anaphora. In *Proceedings of the LREC 2010 workshop on Language Resources and Language Technology Standards* (pp. 54–59). Valletta, Malta.

Eckert, M., & Strube, M. (2000). Dialogue Acts, Synchronizing Units, and Anaphora Resolution. *Journal of Semantics*, 17(1), 51–89. doi:10.1093/jos/17.1.51

Elwell, R., & Baldridge, J. (2008). Discourse Connective Argument Identification with Connective Specific Rankers. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing* (pp. 198–205). Washington, DC, USA: IEEE Computer Society. doi:10.1109/ICSC.2008.50

Enç, M. (1986). Topic Switching and Pronominal Subjects in Turkish. In D. Slobin & K. Zimmer (Eds.), *Studies in Turkish Linguistics*. Amsterdam: John Benjamins.

Erguvanlı Taylan, E. (1986). Pronominal versus Zero Representation of Anaphora in Turkish. In D. Slobin & K. Zimmer (Eds.), *Studies in Turkish Linguistics*. Amsterdam: John Benjamins.

Eryiğit, G. (2014). ITU Turkish NLP Web Service. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Gothenburg, Sweden: Association for Computational Linguistics.

Eryiğit, G., Nivre, J., & Oflazer, K. (2008). Dependency parsing of turkish. *Computational Linguistics*, 34(3), 357–389. doi:http://dx.doi.org/10.1162/coli.2008.07-017-R1-06-83

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London / California / New Delhi / Singapore: Sage Publications.

Fleiss, J. L. (1971). Measuring nominal scale agreement among raters. *Psychological Bulletin*, 76(5), 378–382.

Fodor, J. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Francis, G. (1994). Labelling discourse: an aspect of nominal group lexical cohesion. In M. Coulthard (Ed.), *Advances in Written Text Analysis* (pp. 83–101). London/New York.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive Logistic Regression: a Statistical View of Boosting. *The Annals of Statistics*, 28(2), 337–407. Retrieved from http://projecteuclid.org/download/pdf_1/euclid.aos/1016218223

Ghosh, S., Johansson, R., Riccardi, G., & Tonelli, S. (2011). Shallow Discourse Parsing with Conditional Random Fields. In *Proceedings of 5th International Joint Conference on Natural Language Processing* (pp. 1071–1079). Chiang Mai, Thailand: Asian Federation of Natural Language Processing.

Göksel, A., & Kerslake, C. (2005). *Turkish: a comprehensive grammar*. London, New York: Routledge.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, 48(1), 163–89. doi:10.1146/annurev.psych.48.1.163

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions and the structure of discourse. *Computational Linguistics*, 12(3), 175–204.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1), 10–18. doi:10.1145/1656274.1656278

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1), 77–89. doi:10.1080/19312450709336664

Hedberg, N., Gundel, J. K., & Zacharski, R. (2007). Directly and Indirectly Anaphoric Demonstrative and Personal Pronouns in Newspaper Articles. In *Proceedings of the the Sixth Discourse Anaphora and Anaphora Resolution Colloquium (DAARC 2007)*

Hobbs, J. R. (1978). Resolving Pronoun References. *Lingua. (44)*, 311-338

Hobbs, J. R. (1985). *On the Coherence and Structure of Discourse*. Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University.

Retrieved from http://www.isi.edu/ ~hobbs/ocsd.pdf

Hobbs, J. R. (1993). An approach to the structure of discourse. In D. Everett & S. G. Thomason (Eds.), *Discourse: Linguistic, Computational, and Philosophical Perspectives*.

Hovy, E., & Lavid, J. M. (2010). Towards a "Science" of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation*, 22(1).

Hutchinson, B. (2003). Mining the Web for Discourse Markers. In *the Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA). (pp.407-410). Lisbon, Portugal.

Hutchinson, B. (2004). Acquiring the Meaning of Discourse Markers. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Article No 684. Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.3115/1218955. 1219042

Kamp, H., & Reyle, U. (1993). *From discourse to logic: introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory. Part 1*. Kluwer Academic.

Kılıçaslan, Y., Güner, E. S., & Yıldırım, S. (2009). Learning-based pronoun resolution for Turkish with a comparative evaluation. *Computer Speech and Language*, 23, 311–331. doi:10.1016/j.csl.2008.09.001

Knott, A. (1996). *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Unpublished Ph.D. dissertation, University of Edinburgh, UK. Retrieved from https://www.era.lib.ed.ac.uk/bitstream/handle/1842/583/1996-alik.pdf

Kolachina, S., Prasad, R., Sharma, D. M., & Joshi, A. (2012). Evaluation of Discourse Relation Annotation in the Hindi Discourse Relation Bank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)* (pp. 823–828).

Korzen, I., & Buch-Kromann, M. (2011). Anaphoric relations in the Copenhagen dependency treebanks. In S. Dipper & H. Zinsmeister (Eds.), *Beyond Semantics Corpus-based Investigations of Pragmatic and Discourse Phenomena* (Vol. 3, pp. 83–98). Bochum: Ruhr-Universität Bochum, Sprachwissenschaftliches Institut.

Krippendorff, K. (1995). On the reliability of unitiizing continuous data. *Sociological Methodology*, 25, 47–76.

Krippendorff, K. (2004). *Content analysis: an introduction to its methodology* (2nd ed.). Thousand Oaks, California: Sage.

Krippendorff, K. (2011). *Computing Krippendorff ' s Alpha Reliability. Annenberg School for Communication Departmental Papers (ASC)* (pp. 1–9). Pennsylvania. Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers

Küçük, D. (2005). *A Knowledge-Poor Pronoun Resolution System for Turkish.* MS Thesis. METU, Ankara.

Küçük, D., & Turhan Yöndem, M. (2007). Automatic identification of pronominal Anaphora in Turkish texts. In *22nd International Symposium on Computer and Information Sciences, 2007 (ISCIS 2007)* (pp. 1–6). doi:10.1109/ISCIS.2007.4456858

Küçük, D., & Yazıcı, A. (2008). Identification of Coreferential Chains in Video Texts for Semantic Annotation of New Videos. In *Proceedings of the International Symposium on Computer and Information Sciences (ISCIS)* (pp. 1–6). Istanbul, Turkey.

Küçük, D., & Yazıcı, A. (2009). Employing Named Entities for Semantic Retrieval of News Videos in Turkish. In *Proceedings of the International Symposium on Computer and Information Sciences (ISCIS)* (pp. 1–6). Güzelyurt, Northern Cyprus.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *International Biometric Society*, 33(1), 159–174.

Lascarides, A., & Asher, N. (2007). Segmented Discourse Representation Theory : Dynamic Semantics with Discourse Structure. In H. Bunt & R. Muskens (Eds.), *Computing Meaning: Volume 3* (pp. 87–121). Springer.

Lee, S.-H., & Song, J. (2010). Annotating Korean Demonstratives. In *Proceedings of the Fourth Linguistics Annotation Workshop, ACL 2010* (pp. 162–165). Uppsala, Sweden.

Leech, G. (2005). Adding linguistic annotation. In M. Wynne (Ed.), *Developing linguistic corpora: a guide to good practice* (pp. 17–29). Oxford: Oxbow Books. Retrieved from http://ahds.ac.uk/linguistic-corpora/

Leech, G., Garside, R., & Bryant, M. (1994). CLAWS4 : The tagging of the British National Corpus the design of the grammatical tagger (CLAWS4). In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics* (pp. 622–628).

Lewis, G. (1967). *Turkish Grammar*. London: Oxford University Press.

Lin, Z., Ng, H. T., & Kan, M.-Y. (2011). Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 997–1006).

Louis, A., Joshi, A., & Nenkova, A. (2010). Discourse indicators for content selection in summarization. In *Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 147–156). Tokyo: Association for Computational Linguistics. Retrieved from http://portal.acm.org/citation.cfm?id=1944533

Louis, A., Joshi, A., Prasad, R., & Nenkova, A. (2010). Using Entity Features to Classify Implicit Discourse Relations. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 59–62). Stroudsburg, PA, USA: Association for Computational Linguistics.

Retrieved from http://dl.acm.org/ citation.cfm?id=1944506.1944516

Louis, A., & Nenkova, A. (2010). Creating Local Coherence : An Empirical Assessment. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL* (pp. 313–316). Los Angeles, California: Association for Computational Linguistics.

Louis, A., & Nenkova, A. (2012). A coherence model based on syntactic patterns. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1157–1168). Jeju Island, Korea: Ass.

Lyons, J. (1977). *Semantics*: London / New York. Cambridge University Press.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8(3), 243–281.

Marcu, D. (2000). The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Computational Linguistics*, 26(3), 395–448. doi:10.1162/089120100561755

Marcu, D., Amorrortu, E., & Romera, M. (1999). Experiments in Constructing a Corpus of Discourse Trees. In *Towards Standards and Tools for Discourse Tagging* (pp. 48–57).

Marcu, D., & Echihabi, A. (2002). An Unsupervised Approach to Recognizing Discourse Relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 368–375). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1073083.1073145

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 313–330.

Meyer, T., & Webber, B. (2013). Implicitation of Discourse Connectives in (Machine) Translation. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)* (pp. 1–8). Sofia, Bulgaria.

Miltsakaki, E., Dinesh, N., Prasad, R., Joshi, A., & Webber, B. (2005). Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT2005)* (pp. 1–12).

Miltsakaki, E., Prasad, R., Joshi, A., & Webber, B. (2004a). Annotating Discourse Connectives And Their Arguments. In *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation.* (pp.9-16).

Miltsakaki, E., Prasad, R., Joshi, A., & Webber, B. (2004b). The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)* (pp. 2237–2240). Lisbon, Portugal: European Language Resources Association.

Mírovský, J., Mladová, L., & Zikánová, Š. (2010). Connective-based Measuring of the Inter-annotator Agreement in the Annotation of Discourse in PDT. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 775–781). Stroudsburg, PA, USA: Association for Computational Linguistics.

Mount, J. (2011). *The equivalence of logistic regression and maximum entropy models* (pp. 1–8). Retrieved from http://www.win-vector.com/dfiles/LogisticRegressionMaxEnt.pdf

Navarretta, C., & Olsen, S. (2007). Annotating abstract pronominal anaphora in the DAD project, In *the Proceedings of LREC-08.* (pp. 2046-2052).

Oflazer, K., Say, B., Hakkani Tür, D. Z., & Tür, G. (2003). Building a Turkish Treebank. In A. Abeille (Ed.), *Building and Exploiting Syntactically-annotated Corpora*. Kluwer Academic Publishers.

Oza, U., Prasad, R., Kolachina, S., Sharma, D. M., & Joshi, A. (2009). The Hindi Discourse Relation Bank. In *Proceedings of the Third Linguistic Annotation Workshop* (pp. 158–161). Stroudsburg, PA, USA: Association for Computational Linguistics.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1), 71–106.

Partee, B. H., Wall, R. E., & Meulen, A. Ter. (1990). *Mathematical Methods in Linguistics* (Vol. 30, pp. 14–16). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Passonneau, R. J., & Litman, D. J. (1993). Intention-based Segmentation: Human Reliability and Correlation with Linguistic Cues. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics* (pp. 148–155). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/981574.981594

Pitler, E., & Nenkova, A. (2009). Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 13–16). Suntec, Singapore: ACL and AFNLP.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., & Joshi, A. (2008). Easily Identifiable Discourse Relations. In *Proceedings of COLING 2008 Posters* (pp. 87–90).

Poesio, M., & Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias (Eds.), *Proceedings of the Sixth Language Resources and Evaluation Conference (LREC 2008).*

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., & Ducceschi, L. (2013). Phrase Detectives: Utilizing Collective Intelligence for Internet-scale Language Resource Creation. *ACM Trans. Interact. Intell. Syst.*, 3(1), 3:1–3:44. doi:10.1145/2448116.2448119

Poláková, L., Mírovský, J., Nedoluzhko, A., Jínová, P., Zikánová, Š., & Hajičová, E. (2013). Introducing the Prague Discourse Treebank 1 . 0. In *International Joint Conference on Natural Language Processi* (pp. 91–99). Nagoya, Japan.

Polepalli Ramesh, B., Prasad, R., Miller, T., Harrington, B., & Yu, H. (2012). Automatic discourse connective detection in biomedical text. *Journal of the American Medical Informatics Association,* 1–9. doi:10.1136/amiajnl-2011-000775

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (pp. 2961–2969). Marrakech, Morocco: European Language Resources Association (ELRA).

Prasad, R., Joshi, A., & Webber, B. (2010). Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics*: Posters (pp. 1023–1031). Stroudsburg, PA, USA: Association for Computational Linguistics.

Prasad, R., Miltsakaki, E., Joshi, A., & Webber, B. (2004). Annotation and Data Mining of the Penn Discourse TreeBank. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation* (pp. 88–97). Stroudsburg, PA, USA: Association for Computational Linguistics.

Prasad, R., Webber, B., & Joshi, A. (2014). Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation. *Computational Linguistics*, 40(4), 921–950. doi:10.1162/COLI_a_00204

Prévot, L., Pénault, A., Montcheuil, G., Rauzy, S., & Blache, P. (2015). Discourse Structure of Back Covers: A pilot Study. In L. Degand (Ed.), In *Proceedings of TextLink First Action Conference* (p. 54). Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., … others. (2003). The timebank corpus. *Corpus Linguistics*, 2003, 40.

Qian, M. (2013). *The equivalence of Logistic Regression and Maximum Entropy Modeling*. Retrieved from http://web.engr.illinois.edu/~mqian2/upload/research/notes/The Equivalence of Logistic Regression and Maximum Entropy Modeling.pdf

Quinlan, J. R. (1993). C4.5: *Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Recasens, M., & Martí, M. A. (2010). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation, 44*(4), 315–345.

Redeker, G., Berzlánovich, I., & Vliet, N. Van Der. (2012). Multi-Layer Discourse Annotation of a Dutch Text Corpus. In N. C. (Conference C. and K. C. and T. D. and M. U. D. and B. M. and J. M. and A. M. and J. O. and S. Piperidis (Ed.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2820–2825). İstanbul: European Language Resources Association (ELRA).

Reidsma, D., & Carletta, J. (2008). Reliability Measurement without Limits. *Computational Linguistics*, 34(3), 319–326. doi:10.1162/coli.2008.34.3.319

Rysová, M., & Rysová, K. (2014). The Centre and Periphery of Discourse Connectives. In W. Aroonmanakun, P. Boonkwan, & T. Supnithi (Eds.), In *the Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (PACLIC28)* (pp. pp. 452–459). Phuket, Thailand: Department of Linguistics, Faculty of Arts, Chulalongkorn University.

Rysová, M., & Rysová, K. (2015). On Definition of Discourse Connectives - Primary vs. SecSecond Connectives (Based on a Corpus Probe). In L. Degand (Ed.), In *the Proceedings of TextLink First Action Conference: Posters* (pp. 41–43). Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Say, B., Zeyrek, D., Oflazer, K., & Özge, U. (2004). Development of a Corpus and a Treebank for Present-day Written Turkish. In K. İmer & G. Doğan (Eds.), *Current Research in Turkish Linguistics, (Proceedings of the 11th International Conference on Turkish Linguistics, ICTL 2002)* (pp. 183–192). Northern Cyprus: Eastern Mediterranean University Press.

Sazedj, P., & Pinto, H. S. (2005). Time to evaluate: Targeting annotation tools. In *Proceedings of Knowledge Markup and Semantic Annotation at ISWC* (pp. 37–48).

Sevdik Çallı, A. B. (2012). Demonstrative Anaphora in Turkish: A Corpus Based Analysis. In S. Demir, I. Durgar El-Kahlout, & M. U. Doğan (Eds.), *Proceedings of the First Workshop On Language Resources and Technologies for Turkic Languages* (pp. 33–38). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/workshops/02.Turkic Languages Proceedings.pdf

Sharma, H., Dakwale, P., Sharma, D. M., Prasad, R., & Joshi, A. K. (2013). Assessment of Different Workflow Strategies for Annotating Discourse Relations: A Case Study with HDRB. In *CICLing* (1)'13 (pp. 523–532).

Siegel, S., & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences* (2nd ed.). New York: McGraw-Hill.

Soricut, R., & Marcu, D. (2003). Sentence Level Discourse Parsing using Syntactic and Lexical Information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003), Volume 1*.

Spooren, W. (2004). On the use of discourse data in language use research. In H. Aertsen, M. Hannay, & R. Lya (Eds.), *Words in their places: A festschrift for J. Lachlan Mackenzie* (pp. 381–393). Amsterdam: Faculty of Arts, VU.

Spooren, W., & Degand, L. (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2), 241–266. doi:DOI: 10.1515/cllt.2010.009

Sporleder, C., & Lascarides, A. (2008). Using Automatically Labelled Examples to Classify Rhetorical Relations: An Assessment. *Natural Language Engineering*, 14(3), 369–416. doi:10.1017/S1351324906004451

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation* (pp. 96–102). Stroudsburg, PA, USA: Association for Computational Linguistics.

Stede, M. (2012). *Discourse Processing*. (G. Hirst, Ed.). Morgan & Claypool Publishers. doi:10.2200/S00354ED1V01Y201111HLT015

Stede, M., & Heintze, S. (2004). Machine-Assisted Rhetorical Structure Annotation. In *Proceedings of the 20th International Conference on Computational Linguistics (Coling 2004)* (pp. 425–431). Geneva, Switzerland.

Stede, M., & Neumann, A. (2014). Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of the Language Resources and Evaluation Conference (LREC)* (pp. 925–929). Reykjavik.

Strube, M. (1998). Never Look Back: An Alternative to Centering. In C. Boitet & P. Whitelock (Eds.), *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, (COLING-ACL '98)* (pp. 1251–1257). Université de Montréal, Quebec, Canada.: Morgan Kaufmann Publishers / ACL.

Şirin, U., Çakıcı, R., & Zeyrek, D. (2012). METU Turkish Discourse Bank Browser. In S. Demir, I. Durgar El-Kahlout, & M. U. Doğan (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)* (pp. 2808–2812). Istanbul, Turkey: European Language Resources Association (ELRA).

Tın, E., & Akman, V. (1994). Situated Processing of Pronominal Anaphora. In *Proceedings of Second Conference for Natural Language Processing (KONVENS'94)* (pp. 369–378). Vienna: University of Austria.

Torabi Asr, F., & Demberg, V. (2012). Implicitness of Discourse Relations. In *Proceedings of COLING 2012: Technical Papers* (pp. 2669–2684). Mumbai.

Torabi Asr, F., & Demberg, V. (2013). On the Information Conveyed by Discourse Markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)* (pp. 84–93). Sofia, Bulgaria: Association for Computational Linguistics.

Traugott, E. C. (1997). The role of the development of discourse markers in a theory of grammaticalisation. In *Proceedings of the International Conference on Historical Linguistics XII* (pp. 1–23).

Turan, Ü. D. (1995). *Null Vs. Overt Subjects in Turkish Discourse: A Centering Analysis*. Phd Thesis. University of Pennsylvania.

Turan, Ü. D. (1997). Metin İşaret Adılları: Bu, Şu ve Metin Yapısı [Demonstrative Pronouns: This, That and Discourse Structure]. In *XI. Dilbilim Kurultayı: Bildiriler* (pp. 201–212). ODTÜ.

Tüfekçi, P., & Kılıçaslan, Y. (2005). A Computational Model for Resolving Pronominal Anaphora in Turkish Using Hobbs' Naïve Algorithm. In *Proceedings of World Academy of Science, Engineering and Technology (PWASET)* (pp. 13–17).

Tüfekçi, P., Küçük, D., Turhan Yöndem, M., & Kılıçaslan, Y. (2007). Comparison of a Syntax-Based abd a Knowledge-Poor Pronoun Resolution Systems for Turkish. In *International Symposium on Computer and Information Sciences (ISCIS 2007).*

Viera, R., Salmon-alt, S., & Gasperin, C. (2005). Coreference and Anaphoric Relations of Demonstrative Noun Phrases in Multilingual Corpus. In A. Branco, T. McEnery, & R. Mitkov (Eds.), *Anaphora Processing: Linguistic, Cognitive and Computational Modeling* (pp. 385–401). Amsterdam: John-Benjamins Pub. Co.

Vliet, N. Van Der, Berzlánovich, I., Bouma, G., Egg, M., & Redeker, G. (2011). Building a discourse-annotated Dutch text corpus. In *S. Dipper & H. Zinsmeister (Eds.), Beyond Semantics, Bochumer Linguistische Arbeitsberichte* 3 (pp. 157–171).

Webber, B. L. (1988). *Discourse Deixis and Discourse Processing*. Technical Report. Department of Computer & Information Science, University of Pennsylvania. Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1642&context=cis_reports

Webber, B. L. (1988a). Discourse Deixis: Reference to Discourse Segments. In J. R. Hobbs (Ed.), *Proceedings of the 26th annual meeting on Association for Computational Linguistics* (pp. 113–122). ACL.

Webber, B. L. (1988b). Tense As Discourse Anaphor. *Computational Linguistics*, 14(2), 61–73. Retrieved from http://dl.acm.org/citation.cfm?id=55056.55061

Webber, B. L. (1991). Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2), 107–135.

Webber, B. (2004). D-LTAG : extending lexicalized TAG to discourse. *Cognitive Science*, 28, 751–779. doi:10.1016/j.cogsci.2004.04.002

Webber, B., Egg, M., & Kordoni, V. (2011). Discourse structure and language technology. *Natural Language Engineering*, 18(04), 437–490. doi:10.1017/S1351324911000337

Webber, B., Stone, M., Joshi, A., & Knott, A. (2003). Anaphora and Discourse Structure. *Computational Linguistics,* 29 (4), 545-587. doi:10.1162/089120103322753347

Wellner B., & Pustejovsky. (2007). Automatically identifying the arguments of discourse connectives. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP-07)* (pp. 92–101). Prague, Czech Republic.: Association for Computational Linguistics.

Wellner, B., Pustejovsky, J., Havasi, C., Rumshisky, A., & Sauri, R. (2006). Classification of Discourse Coherence Relations : An Exploratory Study using Multiple Knowledge Sources. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue (SIGDIAL)* (pp. 117–125). Sydney, Australia: Association for Computational Linguistics.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3), 165–210.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)* (2nd ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Yalçınkaya, Ş. İ. (2010). *An inter-annotator agreement measurement methodology for the Turkish Discourse Bank (TDB).* MS Thesis. METU, Ankara. Retrieved from http://etd.lib.metu.edu.tr/upload/12612643/index.pdf

Yıldırım, S., & Kılıçaslan, Y. (2007). A Machine Learning Approach to Personal Pronoun Resolution in Turkish. In *Proceedings of 20th International FLAIRS Conference, FLAIRS-20* (pp. 269–270). Key West, Florida.

Yıldırım, S., Kılıçaslan, Y., & Aykaç, R. E. (2004). A computational model for anaphora resolution in Turkish via centering theory: an initial approach. In *Proceedings of the International Conference on Computational Intelligence* (pp. 124–128).

Yıldırım, S., Kılıçaslan, Y., & Yıldız, T. (2007). A Decision Tree and Rule - based Learning Model for Anaphora Resolution in Turkish, In *the Proceedings of the 3rd Language and Technology Conference (L&TC'07)*: Human Language Technologies as a Challenge for Computer Science and Linguistics. October 5-7, 2007, Poznañ, Poland. (pp. 89-92).

Yıldırım, S., Kılıçaslan, Y., & Yıldız, T. (2009). Human Language Technology. Challenges of the Information Society. In Z. Vetulani & H. Uszkoreit (Eds.), *Human Language Technology: Challenges of the Information Society* (pp. 270–278). Berlin, Heidelberg: Springer-Verlag. doi:10.1007/978-3-642-04235-5_23

Yüksel, Ö., & Bozşahin, C. (2002). Contextually Appropriate Reference Generation. *Natural Language Engineering*, 8(1), 69–89.

Zeyrek, D., Çakıcı, R., Sevdik Çallı, A. B., & Demirşahin, I. (2015). Turkish Discourse Bank (TDB). (L. Degand, Ed.)*Poster Presentation at the First Textlink Action Conference*. Louvain-la-Neuve, Belgium: Université Catholique de Louvain.

Zeyrek, D., Demirşahin, I., Sevdik Çallı, A. B., & Çakıcı, R. (2013). Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue and Discourse*, 4(2), 174–184. doi:10.5087/dad.2013.208

Zeyrek, D., Demirşahin, I., Sevdik Çallı, A. B., Ögel Balaban, H., Yalçınkaya, İ., & Turan, Ü. D. (2010). The Annotation Scheme of the Turkish Discourse Bank and an Evaluation of Inconsistent Annotations. In *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 282–289). Stroudsburg, PA, USA: Association for Computational Linguistics.

Zeyrek, D., Turan, Ü. D., Bozşahin, C., Çakıcı, R., Sevdik Çallı, A. B., Demirşahin, I., … Ögel, H. (2009). Annotating Subordinators in the Turkish Discourse Bank. In *Proceedings of the Third Linguistic Annotation Workshop* (pp. 44–47). Stroudsburg, PA, USA: Association for Computational Linguistics.

Zeyrek, D., & Webber, B. (2008). A discourse resource for Turkish: annotating discourse connectives in the METU corpus. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR6)*. Hyderabad, India.

Retrieved from http://aclweb.org/ anthology//I/I08/I08-7009.pdf

Zhou, L., Li, B., Gao, W., Wei, Z., & Wong, K. (2011). Unsupervised Discovery of Discourse Relations for Eliminating Intra-sentence Polarity Ambiguities. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 162–171). Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Zhou, Y., & Xue, N. (2012). PDTB-style Discourse Annotation of Chinese Text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1* (pp. 69–77). Stroudsburg, PA, USA: Association for Computational Linguistics.

Zhou, Y., & Xue, N. (2014). The Chinese Discourse TreeBank: a Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 48(4), 1-35. doi:10.1007/s10579-014-9290-3

APPENDIX A

TWO SAMPLE XML RELATIONS CREATED BY DATT

(From the file 00001131_agreed_ama.xml)

```
<Relation  note="" type="EXPLICIT">
    <Conn>
        <Span>
            <Text>ama</Text>
            <BeginOffset>281</BeginOffset>
            <EndOffset>284</EndOffset>
        </Span>
    </Conn>
    <Mod/>
    <Arg1>
        <Span>
            <Text>Dışa  karşı güçlüydü</Text>
            <BeginOffset>260</BeginOffset>
            <EndOffset>279</EndOffset>
        </Span>
    </Arg1>
    <Arg2>
        <Span>
        <Text>içe,  kendi yüreğine yıkılmak üzereydi</Text>
        <BeginOffset>285</BeginOffset>
        <EndOffset>322</EndOffset>
        </Span>
    </Arg2>
    <Supp1/>
    <Supp2/>
</Relation>
```

(From the file 00001231_agreed_ve.xml)

```
<Relation  note="" type="EXPLICIT">
    <Conn>
        <Span>
            <Text>ve</Text>
            <BeginOffset>12329</BeginOffset>
            <EndOffset>12331</EndOffset>
        </Span>
    </Conn>
    <Mod/>
    <Arg1>
        <Span>
            <Text>beklemenin</Text>
            <BeginOffset>12318</BeginOffset>
            <EndOffset>12328</EndOffset>
        </Span>
    </Arg1>
    <Arg2>
        <Span>
            <Text>aramanın</Text>
            <BeginOffset>12332</BeginOffset>
            <EndOffset>12340</EndOffset>
        </Span>
    </Arg2>
    <Supp1/>
    <Supp2/>
    <Shared>
        <Span>
            <Text>onu</Text>
            <BeginOffset>12314</BeginOffset>
            <EndOffset>12317</EndOffset>
        </Span>
    </Shared>
    <Supp_Shared>
        <Span>
            <Text>Sabah  uyanır uyanmaz Beril&apos;i bulması gerektiğini düşündü</Text>
            <BeginOffset>11787</BeginOffset>
            <EndOffset>11843</EndOffset>
        </Span>
    </Supp_Shared>
</Relation>
```

APPENDIX B

DISCOURSE CONNECTIVE FORMS IN THE TDB

Table B.1: Search tokens and discourse connective forms annotated in TDB 1.0

Types	Forms	Modified Forms
aksine	aksine, bunun aksine	tam aksine
ama	ama	-
amaçla	bu amaçla	belki de bu amaçla, bu amaçla da
amacı ile	amacı ile	-
amacıyla	amacıyla	amacıyla da
ancak	ancak	-
ardından	ardından, bunun ardından, ilk olarak...ardından	ardından da, hemen ardından
aslında	aslında	-
ayrıca	ayrıca	ayrıca da
beraber	beraber	-
beri	beri	-
birlikte	birlikte, bununla birlikte	-
böylece	böylece	-
bu yana	bu yana	-
çünkü	çünkü	-
dahası	dahası, dahası var	-
dolayı	dolayı, bundan dolayı, bu sebepten dolayı	bu sebepten dolayı da, bundan dolayı da
dolayısı ile	dolayısı ile	-
dolayısıyla	dolayısıyla	dolayısıyla da
ek olarak	buna ek olarak	-
fakat	fakat	-
fekat	fekat	-
gene de	gene de	-
gerek	gerek...gerekse	gerek...gerekse de
gibi	gibi	sanki...gibi, tıpkı...gibi, aynı...gibi, gibi de
ha	ha...ha	-
halbuki	halbuki	-
halde	halde, o halde, şu halde, aksi halde	aksi halde de

Table B.1 (continued)

Types	**Forms**	**Modified Forms**
hem	hem...hem, hem...hem...hem	hem de, hem...hem de, hem...hem de sanki
için	için, bunun için, onun için, için...için	belki de...için, için mi, için de, sırf bunun için, için değil, için değil...için, için olmalı, için olacak, belki... İçin, için olsa gerek, biraz da...için, yalnızca...için, çoğu kez...için, bunun için de
içindir	içindir, içindir ki, bunu içindir ki	-
iken	iken	-
ister	ister...ister, ister...isterse	-
kadar	kadar	taa ki...kadar, daha...kadar, kadar da
karşılık	karşılık, buna karşılık	-
karşın	karşın, buna karşın, herşeye karşın	-
mesela	mesela	-
ne	ne...ne, ne...ne...ne	ne...ne de, ne...ne...ne de
ne ki	ne ki	-
ne var ki	ne var ki	-
nedeni ile	nedeni ile	-
nedeniyle	nedeniyle	-
nedenle	bu nedenle, o nedenle	bu nedenle de
nedenlerle	bu nedenlerle, yukarıdaki nedenlerle	bu nedenlerle olsa gerek, bütün bu nedenlerle
neticede	neticede	-
neticesinde	neticesinde, bunun neticesinde	-
önce	önce, bundan önce, önce...ardından, önce...arkasından, önce...sonradan, önce...şimdi, önce...artık	hemen önce, daha önce, ilk önce...ardından, önce...şimdi ise, uzun yıllar önce, önce...şimdi de
örneğin	örneğin	-
örnek olarak	örnek olarak, buna örnek olarak	çarpıcı örnek olarak
ötürü	ötürü, bundan ötürü, bu nedenden ötürü	bundan ötürü de
oysa	oysa, oysa ki	-
rağmen	rağmen, buna rağmen, bunlara rağmen, herşeye rağmen	bütün bunlara rağmen
sayede	bu sayede	-
sayesinde	sayesinde	-
sebeple	bu sebeple	-
söz gelimi	söz gelimi	-
sözgelimi	sözgelimi	-

Table B.1 (continued)

Types	**Forms**	**Modified Forms**
sonra	sonra, bundan sonra, ondan sonra, önce...sonra, ilk...sonra, ilkinde...sonra, ilkin...sonra, ilk önce...sonra, şimdi...sonra, başta...sonra, önce...ondan sonra	sonra da, daha sonra, daha sonra da, önce...sonra da, önce...daha sonra, önce...daha sonra da, iki gün sonra, üç gün sonra, ancak bundan sonra, biraz sonra, yirmi dakika sonra, bir hafta sonra, kısa bir süre sonra, yarım saat sonra, 16 yıl sonra, kaç yıl sonra, neden sonra, az sonra, biraz sonra, bir süre sonra, özellikle...sonra, aylar sonra, yıllar sonra, ancak...sonra, üç ay sonra, yedi sene sonra, 2 yıl sonra, birkaç gün sonra, bir yıl sonra, 8 saat sonra, 24 saat sonra, 19 yıl sonra, iki hafta sonra, bir süre sonra, 15 yıl sonra, daha sonra ise, kısa bir süre sonra, beş yıl sonra, 20 gün sonra, sadece üç ay sonra, iki gece sonra, iki hafta ya da bir ay sonra, 15 - 20 gün sonra, 3 saat sonra, kısa süre sonra, 25 yıl sonra, 25 sene sonra
sonuç olarak	sonuç olarak	sonuç olarak da
sonuçta	sonuçta	-
sonucunda	sonucunda, bunun sonucunda, bunların sonucunda	şüphesiz...sonucunda, çoğu kez...sonucunda
taraftan	diğer taraftan, öte taraftan	-
tersine	tersine	tam tersine
ve	ve	eğer..ve, ve de, ve hatta
veya	veya, ya...veya	-
veyahut	veyahut	yahut da
ya	ya, ya...ya	-
ya da	ya da, ya...ya da, ya...ya...ya da	-
yahut	yahut	-
yalnız	yalnız	-
yandan	bir yandan, bir diğer yandan, beri yandan, diğer yandan, öte yandan, bir yandan...bir yandan, bir yandan...diğer yandan, bir yandan...öbür yandan, bir yandan...öte yandan	bir yandan da, öte yandan da, bir yandan...diğer yandan da
yine de	yine de	-
yoksa	yoksa	-

Table B.1 (continued)

Types	**Forms**	**Modified Forms**
yüzden	bu yüzden, o yüzden	bu yüzden de, o yüzden de, belki de bu yüzden, biraz da bu yüzden, sadece bu yüzden, acaba bu yüzden, bu yüzden mi
yüzünden	yüzünden	
zaman	zaman, o zaman, ne zaman...o zaman	zaman da, işte o zaman, işte...zaman, ancak...o zaman, belki o zaman, o zaman da
zamanda	bir zamanda, aynı zamanda	-

APPENDIX C

STATISTICS FOR EACH SEARCH ITEM IN THE TDB

Table C.1: Annotation Counts, Usage Percentages and Inter-annotator Kappa Agreements for Independent Annotators in TDB 1.0

Search Item	Discourse Connective		Other uses		Gloss	# of ant.	# of rels.	Arg1	Arg2
	# of ann.	%	# of tokens	%					
aksine[1,2,3]	13[4]	61.9	8	38.1	contrary to	1	-	GA	GA
ama	1024	90.94	102	9.06	but, yet	3	161	0.75	**0.81**
						3	77	0.71	0.85
						2	214	**0.93**	**0.94**
						3	133	0.01	-
amaçla	11	68.75	5	31.25	with this aim of	3	11	0.75	**0.89**
amacıyla	64	83.12	13	16.88	with the aim of	3	69	0.71	**0.85**
amacı ile	1	50.00	1	50.00	with the aim of	3	-	NA	NA
ancak	419	79.81	106	20.19	however	1	-	GA	GA
ardından	71	34.30	136	65.70	after	2	83	**0.87**	**0.88**
aslında	81	63.78	46	36.22	in fact	2	80	0.65	0.68
ayrıca	108	86.40	17	13.60	in addition	3	118	0.6	0.76
beraber	6	15.38	33	84.62	along with	1	-	GA	GA
beri	4	4.94	77	95.06	since (temporal)	2	-	NA	NA
birlikte	33	9.09	330	90.91	together / though, nevertheless	1	-	GA	GA
böylece	85	87.63	12	12.37	thus	2	95	**0.83**	**0.98**
bu yana	10	13.70	63	86.30	since this time (temporal)	2	11	**1.00**	**1.00**
çünkü	300	98.36	5	1.64	because	3	307	**0.86**	**0.91**
dahası	10	76.92	3	23.08	furthermore	3	12	0.74	**0.89**
dolayı	21	36.21	37	63.79	owing to	3	30	0.69	0.76

[1] GA indicates that the annotations were created by the group annotation procedure for which agreement cannot be calculated; NA shows that inter-coder reliability was not calculated because there were too few annotations.

[2] The number of annotators specifies if the annotations were done by 3 independent annotators (represented with 3), or by pair annotation procedure (represented with 2). The agreement for the pair annotation is calculated by taking the pair of annotators as a single annotator and comparing their agreement with the independent annotator.

[3] The Kappa values given are for the second approach.

[4] ann.: annotations, annot.: annotators, # of rel. comp.: number of relations compared.

Table C.1 (continued).

Search Item	**Discourse Connective**		**Other uses**		**Gloss**	**# of ant.**	**# of rels.**	**Arg1**	**Arg2**
	# of ann.	%	# of tokens	%					
dolayısı ile	1	50.00	1	50.00	in consequence of, consequently	3	-	NA	NA
dolayısıyla	66	79.52	17	20.48	in consequence of, consequently	3	67	0.73	**0.90**
ek olarak	1	33.33	2	66.67	in addition to (this)	3	-	NA	NA
fakat	80	89.89	9	10.11	but/yet	3	88	0.69	**0.81**
fekat	3	100.0	0	0.00	but/yet	2	-	NA	NA
gene de	26	96.30	1	3.70	still	1	-	GA	GA
gerek	2	1.64	120	98.36	both..and	3	-	NA	NA
gibi	228	15.17	1275	84.83	as	2	418	0.57	0.43
ha.. ha	2	50.00	2	50.00	either..or	3	-	NA	NA
halbuki	17	94.44	1	5.56	however	1	-	GA	GA
halde	61	87.14	9	12.86	inspite of, inspite of (this/that)	2	61	0.77	**0.83**
hem	41	20.81	156	79.19	at the same time	3	108	0.74	**0.82**
hem.. hem	41	32.54	85	67.46	both..and		-	Sng	Sng
için	1102	51.40	1042	48.60	for, so as to, for (this/that), for..for	3	377	0.77	**0.85**
içindir	4	66.67	2	33.33	because of (this/that)	2	4	0.27	0.56
iken	22	100.0	0	0.00	when	1	-	GA	GA
ister	6	12.50	42	87.50	either..or	3	6	**1.00**	**1.00**
kadar	159	15.39	874	84.61	as well as, until	2	170	0.73	**0.82**
karşılık	28	40.58	41	59.42	despite	3	-	GA	GA
karşın	71	62.83	42	37.17	despite	3	53	0.71	0.75
mesela	13	65.00	7	35.00	to exemplify	3	15	0.65	0.77
ne..ne	44	26.99	119	73.01	neither..nor	3	56	**0.88**	**0.88**
ne ki	14	87.50	2	12.50	howbeit	1	-	GA	GA
ne var ki	32	94.12	2	5.88	even so	1	-	GA	GA
nedeni ile	3	37.50	5	62.50	due to the reason	2	-	NA	NA
nedeniyle	42	19.09	178	80.91	due to the reason	2	67	**0.97**	**0.99**
nedenle	117	97.50	3	2.50	for (this/that) reason	2	117	**0.95**	**0.98**
nedenlerle	4	30.77	9	69.23	for (these) reasons, for the reasons above	2	**4**	**0.93**	**1.00**
neticede	1	100.0	0	0.00	eventually	2	-	NA	NA
neticesinde	1	50.00	1	50.00	as a result of (this)	2	-	NA	NA
önce	134	25.19	398	74.81	prior to	2 2	40 49	**0.84** 0.58	**0.81** 0.58

Table C.1 (continued).

Search Item	Discourse Connective		Other uses		Gloss	# of ant.	# of rels.	Arg1	Arg2
	# of ann.	%	# of tokens	%					
örneğin	64	77.11	19	22.89	for example	3	68	**0.84**	**0.84**
örnek olarak	2	50.00	2	50.00	to illustrate	3	-	NA	NA
ötürü	11	55.00	9	45.00	due to	2	11	**1.00**	**1.00**
oysa	136	99.27	1	0.73	however	3	139	0.71	**0.87**
rağmen	77	56.62	59	43.38	despite, although	3	86	0.63	0.64
sayede	5	100.0	0	0	thanks to (this/that)	2	5	**1.00**	**1.00**
sayesinde	3	11.54	23	88.46	since (causal)	2	-	NA	NA
sebeple	1	50.00	1	50.00	for (this/that) reason	2	-	NA	NA
sonra	713	56.81	542	43.19	after	3	332	0.67	0.74
						2	237	**0.80**	**0.81**
						2	159	0.74	0.79
						2	61	0.65	0.69
sonucunda	12	25.00	36	75.00	result of	3	21	0.56	0.48
sonuç olarak	5	100.0	0	0.00	as a result	3	6	0.69	**0.90**
sonuçta	10	55.56	8	44.44	finally	3	13	0.51	0.62
söz gelimi	6	75.00	2	25.00	for instance	1	-	GA	GA
sözgelimi	1	50.00	1	50.00	for instance	1	-	GA	GA
taraftan	3	20.00	12	80.00	on the other hand	3	4	**0.84**	0.64
tersine	11	40.74	16	59.26	in contrast	3	16	0.49	0.68
ve	2111	28.20	5375	71.80	and	3	293	0.74	**0.86**
						2	1394	**0.85**	**0.87**
						2	132	0.73	**0.90**
						2	337	0.76	**0.83**
veya	40	21.28	148	78.72	or	3	46	**0.82**	**0.85**
veyahut	4	66.67	2	33.33	or	3	-	NA	NA
ya	2	0.36	550	99.64	or	3	9	0.55	**1.00**
ya..ya	6	9.09	60	90.91	either..or		-	Sng	Sng
ya da	139	33.74	273	66.26	or	3	70	0.69	0.77
						2	87	**0.93**	**0.95**
yahut	3	50.00	3	50.00	or	1	-	GA	GA
yalnız	12	9.76	111	90.24	it is just that	1	-	GA	GA
yandan	70	68.63	32	31.37	on the one hand	3	104	0.46	0.56
yine de	65	97.01	2	2.99	still	1	-	GA	GA
yoksa	75	72.82	28	27.18	otherwise	3	82	0.78	0.78
yüzden	66	97.06	2	2.94	due to	2	67	**0.87**	**0.94**
yüzünden	5	7.25	64	92.75	since	2	11	**0.82**	**0.88**
zaman	159	30.52	362	69.48	when	2	170	**0.84**	**0.88**
zamanda	39	46.43	45	53.57	at the same time, at a time when	2	45	**0.85**	0.77
TOTAL	**8483**	**39.07**	**13227**	**60.93**					

Table C.2: Kappa Agreements of Annotators with the Gold Standards in TDB 1.0

Search Item	**Annotator Code**	**Arg1**	**Arg2**
aksine[5]		GA[6]	GA
ama	Ann1[7]	0.78	**0.83**
	Ann2	0.78	**0.87**
	Ann3	0.45	0.48
	Ann4	0.78	**0.90**
	Ann5	**0.89**	**0.92**
amaçla	Ann1	**0.95**	**0.95**
	Ann5	0.75	**0.91**
	Ann2	0.71	**0.91**
amacıyla	Ann1	**0.87**	**0.90**
	Ann5	**0.84**	**0.93**
	Ann2	0.70	**0.86**
amacı ile		NA	NA
ancak*[8]	Ann6	**0.99**	**0.99**
ardından	Ann5	**0.83**	**0.82**
	PA1	**0.94**	**0.94**
aslında	Ann5	0.51	0.60
	PA1	**0.83**	**0.89**
ayrıca	Ann6	0.57	**0.82**
	Ann2	0.69	**0.84**
	Ann4	0.58	0.77
beraber		GA	GA
beri		NA	NA
birlikte		GA	GA
böylece	Ann5	0.75	**0.91**
	PA1	**0.89**	**0.93**
bu yana	Ann5	**0.95**	**0.92**
	PA1	**0.95**	**0.92**
çünkü	Ann1	**0.87**	**0.95**
	Ann5	**0.89**	**0.91**
	Ann2	**0.91**	**0.93**
dahası	Ann1	0.75	0.79
	Ann5	0.64	**0.86**
	Ann2	0.64	0.75
dolayı	Ann1	**0.89**	**0.86**
	Ann5	0.68	0.74
	Ann2	**0.88**	**0.96**

[5] Kappa values given are for the Overall approach; Bold values represent above 0.80 (perfect) agreement.

[6] GA indicates that the annotations were created by the group annotation procedure for which agreement cannot be calculated; NA shows that inter-coder reliability was not calculated because there were too few annotations; Sng: Single agreement calculated for repeated uses.

[7] Ann<X> represents independent annotators, whereas PA<X> respresents pair annotators (e.g. Ann1, PA1, respectively).

[8] Connectives marked with * denote that there is only one independent or pair annotation available, where the gold standards were finalized by group decision on these annotations.

Table C.2 (continued)

Search Item	**Annotator Code**	**Arg1**	**Arg2**
dolayısı ile		NA	NA
dolayısıyla	Ann1	**0.85**	**0.95**
	Ann5	0.79	**0.95**
	Ann2	**0.82**	**0.94**
ek olarak		NA	NA
fakat	Ann6	0.72	**0.84**
	Ann2	0.71	**0.87**
	Ann4	0.66	0.76
fekat		NA	NA
gene de*	PA1	**0.93**	**0.98**
gerek		NA	NA
gibi	PA1	0.62	0.56
	Ann5	0.31	0.20
ha.. ha		NA	NA
halbuki		GA	GA
halde	Ann5	0.73	**0.85**
	PA2	**0.89**	**0.92**
hem	Ann1	0.74	0.79
	Ann5	0.72	0.76
	Ann2	0.76	0.79
hem.. hem		Sng	Sng
için	Ann1	0.78	**0.90**
	Ann5	**0.82**	**0.86**
	Ann2	**0.85**	**0.91**
içindir		NA	NA
iken		GA	GA
ister	Ann1	**1.00**	**1.00**
	Ann5	**1.00**	**1.00**
	Ann2	**1.00**	**1.00**
kadar	PA1	**0.85**	**0.93**
	Ann5	0.68	**0.82**
karşılık*	PA1	**0.90**	**0.98**
karşın	Ann1	0.69	**0.81**
	Ann5	**0.80**	0.78
	Ann2	0.77	**0.84**
mesela	Ann1	0.68	0.74
	Ann5	**0.82**	**0.88**
	Ann2	0.75	**0.92**
ne..ne	Ann1	**0.86**	**0.82**
	Ann5	0.77	0.77
	Ann2	**0.82**	0.76
ne ki*	PA1	**0.96**	**1.00**
ne var ki		GA	GA
nedeni ile		NA	NA
nedeniyle	Ann5	0.64	0.66
	PA1	0.67	0.68

Table C.2 (continued)

Search Item	**Annotator Code**	**Arg1**	**Arg2**
nedenle	Ann5	**0.94**	**0.97**
	PA1	**0.99**	**0.99**
nedenlerle		NA	NA
neticede		NA	NA
neticesinde		NA	NA
önce	PA1	0.65	0.66
	Ann5	0.51	0.47
	PA3	**0.80**	0.75
örneğin	Ann1	0.79	**0.83**
	Ann5	**0.87**	**0.91**
	Ann2	**0.89**	**0.91**
örnek olarak		NA	NA
ötürü	Ann5	**1.00**	**0.94**
	PA1	**1.00**	**0.94**
oysa	Ann1	0.77	**0.90**
	Ann6	0.72	**0.88**
	Ann2	**0.81**	**0.92**
rağmen	Ann6	0.63	0.64
	Ann2	0.79	**0.88**
	Ann3	0.63	0.63
sayede	Ann5	0.68	0.66
	PA1	0.68	0.66
sayesinde		NA	NA
sebeple		NA	NA
sonra	Ann1	0.77	**0.83**
	Ann5	0.76	**0.81**
	Ann2	0.78	**0.83**
	PA1	**0.91**	**0.93**
	Ann6	0.67	0.74
	Ann3	0.64	0.69
sonucunda	Ann1	0.58	0.46
	Ann5	0.73	0.59
	Ann2	**0.80**	0.76
sonuç olarak	Ann1	0.65	**0.92**
	Ann5	**0.89**	**0.92**
	Ann2	0.79	**1.00**
sonuçta	Ann1	0.40	0.46
	Ann5	0.38	0.58
	Ann2	0.47	0.47
söz gelimi		GA	GA
sözgelimi		GA	GA
taraftan		NA	NA
tersine	Ann1	0.64	**0.93**
	Ann5	0.53	0.58
	Ann2	0.75	**0.94**
	Ann1	**0.83**	**0.92**

ve

Table C.2 (continued)

Search Item	**Annotator Code**	**Arg1**	**Arg2**
	Ann2	**0.80**	**0.90**
	Ann3	0.76	**0.86**
	Ann5	**0.80**	**0.84**
	PA1	**0.91**	**0.95**
	Ann6	0.57	0.62
veya	Ann5	**0.84**	**0.88**
	Ann2	**0.87**	**0.89**
	Ann3	0.78	0.77
veyahut		NA	NA
ya	Ann1	0.32	0.72
	Ann5	0.60	0.44
	Ann2	0.21	0.72
ya..ya		Sng	Sng
ya da	Ann1	0.65	0.77
	Ann5	**0.84**	**0.90**
	Ann2	0.78	**0.85**
	PA1	**0.96**	**0.98**
yahut		GA	GA
yalnız		GA	GA
yandan	Ann1	**0.81**	**0.92**
	Ann5	0.76	**0.88**
	Ann2	0.35	0.41
yine de		GA	GA
yoksa	Ann3	0.67	0.76
	Ann1	0.76	**0.84**
	Ann4	0.63	**0.83**
yüzden	Ann5	**0.83**	**0.93**
	PA1	**0.94**	**0.99**
yüzünden	Ann5	0.52	0.55
	PA1	0.67	0.64
zaman	Ann5	**0.81**	**0.85**
	PA1	**0.93**	**0.93**
zamanda	PA1	**0.91**	**0.93**
	Ann5	0.79	0.71

APPENDIX D

KAPPA AGREEMENTS FOR COMMON VS. OVERALL APPROACH

Table D.1: Common vs. Overall Inter-annotator Kappa Agreements for Independent Annotators in TDB 1.0

Search Item	**# of Annotators**	**Common**			**Overall**			**Difference of Rels. Compared**
		# of rel.	***Arg1***	***Arg2***	***# of rel.***	***Arg1***	***Arg2***	
aksine	3	-	GA	GA	-	GA	GA	GA
ama	3	61	**0.85**	**0.92**	161	0.75	**0.81**	100
	3	19	**0.94**	**0.91**	77	0.71	**0.85**	58
	2	201	**0.97**	**0.97**	214	**0.93**	**0.94**	13
	3	4	0.69	0.78	133	0.01	-0.03	129
amaçla	3	11	0.69	**0.94**	11	0.75	**0.89**	0
amacıyla	3	64	0.69	**0.93**	69	0.71	**0.85**	5
amacı ile	3	-	NA	NA	-	NA	NA	NA
ancak		-	GA	GA	-	GA	GA	GA
ardından	2	69	**1.00**	**0.99**	83	**0.87**	**0.88**	14
aslında	2	44	**0.81**	**0.85**	80	0.65	0.68	36
ayrıca	3	84	0.66	**0.84**	118	0.60	0.76	34
beraber		-	GA	GA	-	GA	GA	GA
beri	2	-	NA	NA	-	NA	NA	NA
birlikte	3	-	GA	GA	-	GA	GA	GA
böylece	2	93	**0.90**	**0.99**	95	**0.83**	**0.98**	2
bu yana	2	11	**1.00**	**1.00**	11	**1.00**	**1.00**	0
çünkü	3	292	**0.92**	**0.95**	307	**0.86**	**0.91**	15
dahası	3	11	0.71	**0.90**	12	0.74	**0.89**	1
dolayı	3	16	**0.98**	**1.00**	30	0.69	0.76	14
dolayısı ile	3	-	NA	NA	-	NA	NA	NA
dolayısıyla	3	63	0.78	**0.97**	67	0.73	**0.90**	4
ek olarak	3	-	NA	NA	-	NA	NA	NA
fakat	3	55	**0.83**	**0.90**	88	0.69	**0.81**	33
fekat	2	-	NA	NA	-	NA	NA	NA
gene de		-	GA	GA	-	GA	GA	GA
gerek	3	-	NA	NA	-	NA	NA	NA
gibi	2	96	**0.94**	**0.95**	418	0.57	0.43	322

Table D.1 (continued).

Search Item	**# of Annotators**	**Common**			**Overall**			**Difference of Rels. Compared**
		# of rel.	***Arg1***	***Arg2***	***# of rel.***	***Arg1***	***Arg2***	
ha.. ha	3	-	NA	NA	-	NA	NA	NA
halbuki		-	GA	GA	-	GA	GA	GA
halde	2	56	**0.87**	**0.93**	61	0.77	**0.83**	5
hem	3	62	**0.83**	**0.94**	108	0.74	**0.82**	46
hem.. hem		-	Sng	Sng	-	Sng	Sng	Sng
için	3	263	**0.81**	**0.92**	377	0.77	**0.85**	114
içindir	2	2	0.50	**1.00**	4	0.27	0.56	2
iken		-	GA	GA	-	GA	GA	GA
ister	3	6	**1.00**	**1.00**	6	**1.00**	**1.00**	0
kadar	2	114	**0.84**	**0.99**	170	0.73	**0.82**	56
karşılık	3	-	GA	GA	-	GA	GA	GA
karşın	3	32	**0.86**	**0.84**	53	0.71	0.75	21
mesela	3	8	**0.92**	**1.00**	15	0.65	0.77	7
ne..ne	3	40	**1.00**	**0.97**	56	**0.88**	**0.88**	16
ne ki		-	GA	GA	-	GA	GA	GA
ne var ki		-	GA	GA	-	GA	GA	GA
nedeni ile	2	-	NA	NA	-	NA	NA	NA
nedeniyle	2	67	**0.96**	**0.99**	67	**0.97**	**0.99**	0
nedenle	2	117	**0.94**	**0.99**	117	**0.95**	**0.98**	0
nedenlerle	2	4	0.75	**1.00**	4	**0.93**	**1.00**	0
neticede	2	-	NA	NA	-	NA	NA	NA
neticesinde	2	-	NA	NA	-	NA	NA	NA
önce	2	32	**1.00**	**1.00**	40	**0.84**	**0.81**	8
	2	21	**0.84**	**0.88**	49	0.58	0.58	28
örneğin	3	56	**0.87**	**0.92**	68	**0.84**	**0.84**	12
örnek olarak	3	-	NA	NA	-	NA	NA	NA
ötürü	2	11	**1.00**	**0.94**	11	**1.00**	**1.00**	0
oysa	3	100	0.78	**0.91**	139	0.71	**0.87**	39
rağmen	3	71	0.73	0.78	86	0.63	0.64	15
sayede	2	5	**1.00**	**1.00**	5	**1.00**	**1.00**	0
sayesinde	2	-	NA	NA	-	NA	NA	NA
sebeple	2	-	NA	NA	-	NA	NA	NA
sonra	3	72	**0.85**	**0.91**	332	0.67	0.74	260
	2	145	**0.91**	**0.96**	237	**0.80**	**0.81**	92
	2	101	**0.89**	**0.94**	159	0.74	0.79	58
	2	32	**0.89**	**0.98**	61	0.65	0.69	29
sonucunda	3	10	0.78	0.78	21	0.56	0.48	11
sonuç olarak	3	5	0.67	**1.00**	6	0.69	**0.90**	1
sonuçta	3	8	0.70	**0.87**	13	0.51	0.62	5
söz gelimi		-	GA	GA	-	GA	GA	GA
sözgelimi		-	GA	GA	-	GA	GA	GA

Table D.1 (continued).

Search Item	**# of Annotators**	**Common**			**Overall**			**Difference of Rels. Compared**
		# of rel.	***Arg1***	***Arg2***	***# of rel.***	***Arg1***	***Arg2***	
taraftan	3	3	0.55	0.50	4	**0.84**	0.64	1
tersine	3	10	0.77	**1.00**	16	0.49	0.68	6
ve	3	71	0.70	**0.85**	293	0.74	**0.86**	222
	2	322	**0.96**	**0.97**	1394	**0.85**	**0.87**	1072
	2	46	0.78	**0.95**	132	0.73	**0.90**	86
	2	159	0.74	0.79	337	0.76	**0.83**	178
veya	3	36	**0.96**	**0.97**	46	**0.82**	**0.85**	10
veyahut	3	-	NA	NA	-	NA	NA	NA
ya	3	9	**0.80**	**0.87**	9	0.55	**1.00**	0
ya..ya		-	Sng	Sng	-	Sng	Sng	Sng
ya da	3	27	**0.84**	**1.00**	70	0.69	0.77	43
	2	76	**0.97**	**1.00**	87	**0.93**	**0.95**	11
yahut		-	GA	GA	-	GA	GA	GA
yalnız		-	GA	GA	-	GA	GA	GA
yandan	3	60	0.55	0.66	104	0.46	0.56	44
yine de		-	GA	GA	-	GA	GA	GA
yoksa	3	39	**0.88**	**0.98**	82	0.78	0.78	43
yüzden	2	62	**0.91**	**0.99**	67	**0.87**	**0.94**	5
yüzünden	2	9	**1.00**	**1.00**	11	**0.82**	**0.88**	2
zaman	2	134	**0.97**	**0.98**	170	**0.84**	**0.88**	36
								0
zamanda	2	32	**1.00**	**1.00**	45	**0.85**	0.77	13

Table D.2: Common vs. Overall Kappa Agreements of Annotators with the Gold Standards in TDB 1.0

Search Item	Annotator Code	Common		Overall	
		Arg1	Arg2	Arg1	Arg2
aksine[123]		GA	GA	GA	GA
ama	Ann1	**0.84**	**0.92**	0.78	**0.83**
	Ann2	**0.84**	**0.91**	0.78	**0.87**
	Ann3	**0.83**	**0.90**	0.45	0.48
	Ann4	**0.86**	**0.91**	0.78	**0.90**
	Ann5	**0.95**	**0.94**	**0.89**	**0.92**
amaçla	Ann1	**0.92**	**1.00**	**0.95**	**0.95**
	Ann5	0.67	**0.92**	0.75	**0.91**
	Ann2	0.67	**0.92**	0.71	**0.91**
amacıyla	Ann1	**0.85**	**0.95**	**0.87**	**0.90**
	Ann5	**0.81**	**0.95**	**0.84**	**0.93**
	Ann2	0.78	**0.96**	0.70	**0.86**
amacı ile		NA	NA	NA	NA
ancak		GA	GA	GA	GA
	Ann6	**0.99**	**0.99**	**0.99**	**0.99**
ardından	Ann5	**0.99**	**0.97**	**0.83**	**0.82**
	PA1	**0.99**	**0.99**	**0.94**	**0.94**
aslında	Ann5	0.63	0.78	0.51	0.60
	PA1	**0.89**	**0.97**	**0.83**	**0.89**
ayrıca	Ann6	0.66	**0.90**	0.57	**0.82**
	Ann2	0.74	**0.89**	0.69	**0.84**
	Ann4	0.71	**0.81**	0.58	0.77
beraber		GA	GA	GA	GA
beri		NA	NA	NA	NA
birlikte		GA	GA	GA	GA
böylece	Ann5	**0.89**	**1.00**	0.75	**0.91**
	PA1	**0.98**	**1.00**	**0.89**	**0.93**
bu yana	Ann5	**1.00**	**1.00**	**0.95**	**0.92**
	PA1	**1.00**	**1.00**	**0.95**	**0.92**
çünkü	Ann1	**0.91**	**0.98**	**0.87**	**0.95**
	Ann5	**0.94**	**0.95**	**0.89**	**0.91**
	Ann2	**0.96**	**0.97**	**0.91**	**0.93**
dahası	Ann1	**0.81**	**0.94**	0.75	0.79
	Ann5	0.79	**1.00**	0.64	**0.86**
	Ann2	0.79	**0.93**	0.64	0.75
dolayı	Ann1	**1.00**	**1.00**	**0.89**	**0.86**
	Ann5	**0.91**	**0.95**	0.68	0.74
	Ann2	**0.88**	**1.00**	**0.88**	**0.96**

[1] GA indicates that the annotations were created by the group annotation procedure for which agreement cannot be calculated; NA shows that inter-coder reliability was not calculated because there were too few annotations.

[2] Ann<X> represents independent annotators, whereas PA<X> respresents pair annotators (e.g. Ann1, PA1, respectively).

[3] Bold face values indicate >0.80 agreement.

Table D.2 (continued).

Search Item	**Annotator Code**	**Common**		**Overall**	
		Arg1	**Arg2**	**Arg1**	**Arg2**
dolayısı ile		NA	NA	NA	NA
dolayısıyla	Ann1	**0.91**	**0.97**	**0.85**	**0.95**
	Ann5	**0.84**	**0.99**	0.79	**0.95**
	Ann2	**0.85**	**0.99**	**0.82**	**0.94**
ek olarak		NA	NA	NA	NA
fakat	Ann6	**0.83**	**0.89**	0.72	**0.84**
	Ann2	**0.81**	**0.95**	0.71	**0.87**
	Ann4	0.76	**0.90**	0.66	0.76
fekat		NA	NA	NA	NA
gene de	PA1	**0.97**	**1.00**	**0.93**	**0.98**
gerek		NA	NA	NA	NA
gibi	PA1	**0.93**	**0.91**	0.62	0.56
	Ann5	0.55	0.53	0.31	0.20
ha.. ha		NA	NA	NA	NA
halbuki		GA	GA	GA	GA
halde	Ann5	**0.82**	**0.94**	0.73	**0.85**
	PA2	**0.93**	**1.00**	**0.89**	**0.92**
hem	Ann1	**0.89**	**0.98**	0.74	0.79
	Ann5	**0.85**	**0.97**	0.72	0.76
	Ann2	**0.95**	**0.95**	0.76	0.79
hem.. hem		Sng	Sng	Sng	Sng
için	Ann1	**0.82**	**0.95**	0.78	**0.90**
	Ann5	**0.90**	**0.95**	**0.82**	**0.86**
	Ann2	**0.88**	**0.97**	**0.85**	**0.91**
içindir		NA	NA	NA	NA
iken		GA	GA	GA	GA
ister	Ann1	**1.00**	**1.00**	**1.00**	**1.00**
	Ann5	**1.00**	**1.00**	**1.00**	**1.00**
	Ann2	**1.00**	**1.00**	**1.00**	**1.00**
kadar	PA1	**0.92**	**1.00**	**0.85**	**0.93**
	Ann5	0.75	**0.99**	0.68	**0.82**
karşılık		GA	GA	GA	GA
	PA1	**0.89**	**1.00**	**0.90**	**0.98**
karşın	Ann1	0.77	**0.95**	0.69	0.81
	Ann5	**0.83**	**0.83**	**0.80**	0.78
	Ann2	0.74	**0.90**	0.77	**0.84**
mesela	Ann1	**1.00**	**1.00**	0.68	0.74
	Ann5	0.74	**1.00**	**0.82**	**0.88**
	Ann2	**0.93**	**1.00**	**0.75**	**0.92**
ne..ne	Ann1	**1.00**	**1.00**	**0.86**	**0.82**
	Ann5	**1.00**	**0.95**	0.77	0.77
	Ann2	**1.00**	**1.00**	**0.82**	0.76
ne ki	PA1	**1.00**	**1.00**	**0.96**	**1.00**
ne var ki		GA	GA	GA	GA
nedeni ile		NA	NA	NA	NA

Table D.2 (continued).

Search Item	**Annotator Code**	**Common**		**Overall**	
		Arg1	**Arg2**	**Arg1**	**Arg2**
nedeniyle	Ann5	**0.88**	**1.00**	0.64	0.66
	PA1	**0.94**	**1.00**	0.67	0.68
nedenle	Ann5	**0.94**	**0.99**	**0.94**	**0.97**
	PA1	**1.00**	**1.00**	**0.99**	**0.99**
nedenlerle		NA	NA	NA	NA
neticede		NA	NA	NA	NA
neticesinde		NA	NA	NA	NA
önce	PA1	**0.92**	**0.96**	0.65	0.66
	Ann5	**0.87**	**0.93**	0.51	0.47
	PA3	**0.96**	**0.96**	**0.80**	0.75
örneğin	Ann1	**0.84**	**0.94**	0.79	**0.83**
	Ann5	**0.94**	**0.95**	**0.87**	**0.91**
	Ann2	**0.95**	**0.95**	**0.89**	**0.91**
örnek olarak		NA	NA	NA	NA
ötürü	Ann5	-	-	**1.00**	**0.94**
	PA1	-	-	**1.00**	**0.94**
oysa	Ann1	**0.81**	**0.93**	0.77	**0.90**
	Ann6	0.75	**0.94**	0.72	**0.88**
	Ann2	**0.85**	**0.95**	**0.81**	**0.92**
rağmen	Ann6	0.64	0.76	0.63	0.64
	Ann2	**0.87**	**0.93**	0.79	**0.88**
	Ann3	0.72	**0.81**	0.63	0.63
sayede	Ann5	**1.00**	**1.00**	0.68	0.66
	PA1	**1.00**	**1.00**	0.68	0.66
sayesinde		NA	NA	NA	NA
sebeple		NA	NA	NA	NA
sonra	Ann1	**0.83**	**0.95**	0.77	**0.83**
	Ann5	**0.84**	**0.91**	0.76	**0.81**
	Ann2	**0.88**	**0.96**	0.78	**0.83**
	PA1	**0.98**	**0.99**	**0.91**	**0.93**
	Ann6	**0.92**	**0.96**	0.67	0.74
	Ann3	**0.93**	**0.93**	0.64	0.69
sonucunda	Ann1	0.75	0.75	0.58	0.46
	Ann5	**0.92**	**0.83**	0.73	0.59
	Ann2	**1.00**	**0.86**	**0.80**	0.76
sonuç olarak	Ann1	0.67	**1.00**	0.65	**0.92**
	Ann5	**1.00**	**1.00**	**0.89**	**0.92**
	Ann2	**0.83**	**1.00**	0.79	**1.00**
sonuçta	Ann1	0.62	**1.00**	0.40	0.46
	Ann5	**0.87**	**1.00**	0.38	0.58
	Ann2	**0.83**	**0.83**	0.47	0.47
söz gelimi		GA	GA	GA	GA
sözgelimi		GA	GA	GA	GA
taraftan		NA	NA	NA	NA
tersine	Ann1	**0.80**	**1.00**	0.64	**0.93**

Table D.2 (continued).

Search Item	Annotator Code	Common		Overall	
		Arg1	Arg2	Arg1	Arg2
	Ann5	**0.90**	**1.00**	0.53	0.58
	Ann2	**0.80**	**1.00**	0.75	**0.94**
ve	Ann1	**0.91**	**0.96**	**0.83**	**0.92**
	Ann2	**0.94**	**0.99**	**0.80**	**0.90**
	Ann3	**0.90**	**0.93**	0.76	**0.86**
	Ann5	**0.94**	**0.96**	**0.80**	**0.84**
	PA1	**0.99**	**1.00**	**0.91**	**0.95**
	Ann6	**0.86**	**0.90**	0.57	0.62
veya	Ann5	**0.96**	**1.00**	**0.84**	**0.88**
	Ann2	**0.98**	**1.00**	**0.87**	**0.89**
	Ann3	**0.96**	**0.96**	0.78	0.77
veyahut		NA	NA	NA	NA
ya	Ann1	0.75	**1.00**	0.32	0.72
	Ann5	**1.00**	**1.00**	0.60	0.44
	Ann2	0.62	**1.00**	0.21	0.72
ya..ya		Sng	Sng	Sng	Sng
ya da	Ann1	**0.83**	**1.00**	0.65	0.77
	Ann5	**0.95**	**0.99**	**0.84**	**0.90**
	Ann2	**0.89**	**0.94**	0.78	**0.85**
	PA1	**0.99**	**1.00**	**0.96**	**0.98**
yahut		GA	GA	GA	GA
yalnız		GA	GA	GA	GA
yandan	Ann1	**0.90**	**0.99**	**0.81**	**0.92**
	Ann5	**0.86**	**0.94**	0.76	**0.88**
	Ann2	0.43	0.56	0.35	0.41
yine de		GA	GA	GA	GA
yoksa	Ann3	**0.80**	**0.93**	0.67	0.76
	Ann1	**0.92**	**0.93**	0.76	**0.84**
	Ann4	0.75	**0.92**	0.63	**0.83**
yüzden	Ann5	**0.89**	**0.99**	**0.83**	**0.93**
	PA1	**0.99**	**1.00**	**0.94**	**0.99**
yüzünden	Ann5	**1.00**	**1.00**	0.52	0.55
	PA1	**1.00**	**1.00**	0.67	0.64
zaman	Ann5	**0.96**	**0.96**	**0.81**	**0.85**
	PA1	**0.99**	**0.98**	**0.93**	**0.93**
zamanda	PA1	**0.97**	**1.00**	**0.91**	**0.93**
	Ann5	**0.97**	**1.00**	0.79	0.71

APPENDIX E

PRECISION AND RECALL VALUES

Table E.1: Precision and Recall Values for Separate Annotators in TDB 1.0

Search Item	Ann. Code	P	R	RinSpan	F-Measure
aksine[1]		GA	GA	GA	GA
ama	Ann1	0.95	0.29	0.99	0.44
	Ann2	0.95	0.14	0.96	0.25
	Ann3	0.97	0.17	0.48	0.28
	Ann4	0.91	0.06	0.92	0.11
	Ann8	1.00	0.10	0.87	0.19
	Ann5	0.98	0.32	0.98	0.48
	Ann7	0.99	0.33	1.00	0.50
amaçla	Ann1	1.00	1.00	1.00	1.00
	Ann5	1.00	1.00	1.00	1.00
	Ann2	1.00	1.00	1.00	1.00
amacıyla	Ann1	0.96	1.00	1.00	0.98
	Ann5	0.98	0.98	1.00	0.98
	Ann2	0.98	1.00	1.00	0.99
amacı ile	Ann1	1.00	1.00	1.00	1.00
	Ann5	1.00	1.00	1.00	1.00
	Ann2	1.00	1.00	1.00	1.00
ancak		GA	GA	GA	GA
ardından	Ann5	0.85	0.96	0.99	0.90
	PA1	0.93	1.00	1.00	0.97
aslında	Ann5	0.94	0.60	0.80	0.74
	PA1	0.95	0.94	0.96	0.94
ayrıca	Ann6	0.99	0.98	0.98	0.99
	Ann2	0.97	0.94	0.98	0.96
	Ann4	0.77	0.79	0.79	0.78
beraber	Ann5	1.00	1.00	1.00	1.00
beri	Ann5	1.00	1.00	1.00	1.00
	PA1	1.00	1.00	1.00	1.00
birlikte	Ann5	1.00	0.52	0.94	0.68
böylece	Ann5	0.89	0.99	0.99	0.94
	PA1	0.89	1.00	1.00	0.94

[1] Ann.: Annotator; P.: Precision; R.: Recall.

Table E.1 (continued)

Search Item	Ann. Code	P	R	RinSpan	F-Measure
bu yana	Ann5	0.91	1.00	1.00	0.95
	PA1	0.91	1.00	1.00	0.95
çünkü	Ann1	0.98	1.00	1.00	0.99
	Ann5	0.99	1.00	1.00	1.00
	Ann2	0.98	0.99	0.99	0.99
dahası	Ann1	0.83	1.00	1.00	0.91
	Ann5	0.82	0.90	0.90	0.86
	Ann2	0.82	0.90	0.90	0.86
dolayı	Ann1	0.88	1.00	1.00	0.93
	Ann5	0.71	0.95	0.95	0.82
	Ann2	0.95	1.00	1.00	0.98
dolayısı ile	Ann1	1.00	1.00	1.00	1.00
	Ann5	1.00	1.00	1.00	1.00
	Ann2	1.00	1.00	1.00	1.00
dolayısıyla	Ann1	0.99	1.00	1.00	0.99
	Ann5	1.00	1.00	1.00	1.00
	Ann2	1.00	1.00	1.00	1.00
ek olarak	Ann1	1.00	1.00	1.00	1.00
	Ann5	1.00	1.00	1.00	1.00
	Ann2	1.00	1.00	1.00	1.00
fakat	Ann6	0.97	0.89	0.89	0.93
	Ann2	0.93	0.95	0.95	0.94
	Ann4	0.90	0.96	0.96	0.93
fekat	Ann5	1.00	1.00	1.00	1.00
	PA1	1.00	1.00	1.00	1.00
gene de		GA	GA	GA	GA
gerek	Ann1	1.00	1.00	1.00	1.00
	Ann5	0.67	1.00	1.00	0.80
	Ann2	1.00	1.00	1.00	1.00
gibi	PA1	0.58	0.98	0.98	0.73
	Ann5	0.45	0.46	0.46	0.45
ha.. ha	Ann1	1.00	1.00	1.00	1.00
	Ann5	1.00	1.00	1.00	1.00
	Ann2	1.00	1.00	1.00	1.00
halbuki		GA	GA	GA	GA
halde	Ann5	0.98	0.92	0.92	0.95
	PA2	0.98	0.93	0.93	0.96
hem	Ann1	0.80	0.99	0.99	0.89
	Ann5	0.81	0.96	0.96	0.88
	Ann2	0.83	0.98	0.98	0.90
hem.. hem		Sng	Sng	Sng	Sng
için	Ann1	0.98	0.32	0.98	0.49
	Ann5	0.92	0.98	0.98	0.95
	Ann2	0.98	0.33	0.99	0.49
içindir	Ann5	1.00	0.50	0.50	0.67
	PA1	1.00	1.00	1.00	1.00

Table E.1 (continued)

Search Item	**Ann. Code**	**P**	**R**	**RinSpan**	**F-Measure**
iken		GA	GA	GA	GA
ister	Ann1	1.00	1.00	1.00	1.00
	Ann5	1.00	1.00	1.00	1.00
	Ann2	1.00	1.00	1.00	1.00
kadar	PA1	0.93	0.99	0.99	0.96
	Ann5	0.96	0.84	0.84	0.90
karşılık		GA	GA	GA	GA
karşın	Ann1	0.89	0.58	0.89	0.70
	Ann5	0.92	0.65	1.00	0.76
	Ann2	0.96	0.62	0.96	0.75
mesela	Ann1	0.92	0.85	0.85	0.88
	Ann5	1.00	1.00	1.00	1.00
	Ann2	0.87	1.00	1.00	0.93
ne..ne	Ann1	0.84	0.98	0.98	0.91
	Ann5	0.80	1.00	1.00	0.89
	Ann2	0.81	0.98	0.98	0.89
ne ki		GA	GA	GA	GA
ne var ki		GA	GA	GA	GA
nedeni ile	PA1	1.00	1.00	1.00	1.00
	Ann5	1.00	1.00	1.00	1.00
nedeniyle	Ann5	0.63	1.00	1.00	0.77
	PA1	0.63	1.00	1.00	0.77
nedenle	Ann5	1.00	1.00	1.00	1.00
	PA1	1.00	1.00	1.00	1.00
nedenlerle	Ann5	0.50	0.50	0.50	0.50
	PA1	0.50	0.50	0.50	0.50
neticede	Ann1	1.00	1.00	1.00	1.00
	Ann5	1.00	1.00	1.00	1.00
neticesinde	Ann5	1.00	1.00	1.00	1.00
	PA1	1.00	1.00	1.00	1.00
önce	PA1	0.71	0.35	0.96	0.47
	Ann5	0.83	0.57	0.65	0.67
	PA3	0.84	0.28	0.95	0.42
	Ann1	0.81	0.59	0.95	0.68
örneğin	Ann1	0.94	0.94	0.94	0.94
	Ann5	0.97	0.98	0.98	0.98
	Ann2	0.95	0.98	0.98	0.97
örnek olarak	Ann1	1.00	1.00	1.00	1.00
	Ann5	1.00	0.50	0.50	0.67
	Ann2	1.00	1.00	1.00	1.00
ötürü	Ann1	1.00	1.00	1.00	1.00
	Ann5	1.00	1.00	1.00	1.00
oysa	Ann1	0.99	0.99	0.99	0.99
	Ann6	0.98	0.93	0.93	0.96
	Ann2	1.00	0.99	0.99	0.99
rağmen	Ann6	0.91	0.95	0.95	0.93

Table E.1 (continued)

Search Item	Ann. Code	P	R	RinSpan	F-Measure
	Ann2	0.99	0.97	0.97	0.98
	Ann3	0.92	0.94	0.94	0.93
sayede	Ann5	1.00	1.00	1.00	1.00
	PA1	1.00	1.00	1.00	1.00
sayesinde	Ann5	1.00	1.00	1.00	1.00
	PA1	1.00	1.00	1.00	1.00
sebeple	Ann5	1.00	1.00	1.00	1.00
	PA1	1.00	1.00	1.00	1.00
sonra	Ann1	0.97	0.35	0.85	0.51
	Ann5	0.96	0.64	0.81	0.77
	Ann2	0.91	0.40	0.97	0.55
	PA1	0.93	0.59	1.00	0.72
	Ann6	0.91	0.17	0.82	0.28
	Ann3	0.87	0.05	0.74	0.10
sonucunda	Ann1	0.73	0.92	0.92	0.81
	Ann5	0.60	1.00	1.00	0.75
	Ann2	0.85	0.92	0.92	0.88
sonuç olarak	Ann1	0.83	1.00	1.00	0.91
	Ann5	1.00	1.00	1.00	1.00
	Ann2	1.00	1.00	1.00	1.00
sonuçta	Ann1	0.55	0.60	0.60	0.57
	Ann5	0.67	0.60	0.60	0.63
	Ann2	0.60	0.60	0.60	0.60
söz gelimi	Ann1	1.00	1.00	1.00	1.00
sözgelimi	Ann1	1.00	0.83	0.83	0.91
taraftan	Ann1	0.75	1.00	1.00	0.86
	Ann5	1.00	1.00	1.00	1.00
	Ann2	0.25	0.33	1.00	0.29
tersine	Ann1	1.00	0.91	0.91	0.95
	Ann5	0.67	0.73	0.73	0.70
	Ann2	0.92	1.00	1.00	0.96
ve	Ann1	0.98	0.31	0.97	0.47
	Ann2	0.97	0.12	0.96	0.21
	Ann3	0.94	0.15	0.92	0.26
	Ann5	0.95	0.54	0.85	0.69
	PA1	0.95	0.68	1.00	0.79
	Ann6	0.96	0.20	0.96	0.33
veya	Ann5	0.87	0.98	0.98	0.92
	Ann2	0.90	0.95	0.95	0.93
	Ann3	0.92	0.88	0.88	0.90
veyahut	Ann1	1.00	1.00	1.00	1.00
	Ann5	1.00	1.00	1.00	1.00
	Ann2	1.00	1.00	1.00	1.00
ya	Ann1	0.89	1.00	1.00	0.94
	Ann5	0.67	1.00	1.00	0.80
	Ann2	0.89	1.00	1.00	0.94

Table E.1 (continued)

Search Item	**Ann. Code**	**P**	**R**	**RinSpan**	**F-Measure**
ya..ya		Sng	Sng	Sng	Sng
ya da	Ann1	0.85	0.36	0.91	0.51
	Ann5	0.92	0.96	0.96	0.94
	Ann2	0.90	0.39	0.98	0.54
	PA1	0.97	0.60	1.00	0.74
yahut	Ann1	1.00	1.00	1.00	1.00
yalnız		GA	GA	GA	GA
yandan	Ann1	1.00	0.97	0.97	0.99
	Ann5	0.97	0.93	0.93	0.95
	Ann2	0.51	0.51	0.51	0.51
yine de	Ann5	0.98	0.97	0.97	0.98
yoksa	Ann3	1.00	0.75	0.75	0.85
	Ann1	0.97	0.87	0.87	0.92
	Ann4	0.91	0.97	0.97	0.94
yüzden	Ann5	0.98	0.97	0.97	0.98
	PA1	0.99	1.00	1.00	0.99
yüzünden	Ann5	0.45	1.00	1.00	0.63
	PA1	0.56	1.00	1.00	0.71
zaman	Ann5	0.95	0.90	0.90	0.92
	PA1	0.94	0.99	0.99	0.97
zamanda	PA1	0.87	1.00	1.00	0.93
	Ann5	0.89	0.79	0.79	0.84

APPENDIX F

RELATION COUNTS WITH RESPECT TO ARGUMENT DISCONTINUITY

Relation Counts with respect to Argument Discontinuity

Table F.1: Relation Counts with Discontinuous Spans

Search Item	Arg1	Arg2	Arg1 & Arg2	Total Discont.	Not Discont.	TOTAL
aksine[1]	-	4	-	**4**	9	**13**
ama	15	12	-	**27**	997	**1024**
amaçla	-	4	-	**4**	7	**11**
amacıyla	22	1	-	**23**	41	**64**
amacı ile	-	-	-	**0**	1	**1**
ancak	1	-	-	**1**	418	**419**
ardından	10	3	-	**13**	58	**71**
aslında	-	13	-	**13**	68	**81**
ayrıca	1	13	-	**14**	94	**108**
beraber	-	-	-	**0**	6	**6**
beri	1	-	-	**1**	3	**4**
birlikte	5	-	-	**5**	28	**33**
böylece	2	10	-	**12**	73	**85**
bu yana	1	-	-	**1**	9	**10**
çünkü	3	6	1	**10**	290	**300**
dahası	-	1	-	**1**	9	**10**
dolayı	3	-	-	**3**	18	**21**
dolayısı ile	-	-	-	**0**	1	**1**
dolayısıyla	2	-	-	**2**	64	**66**
ek olarak	-	-	-	**0**	1	**1**
fakat	1	2	-	**3**	77	**80**
fekat	-	-	-	**0**	3	**3**
gene de	-	4	-	**4**	22	**26**
gerek	-	-	-	**0**	2	**2**
gibi	68	-	1	**69**	159	**228**

[1] Only *Arg1* denotes the relations that only have discontinuous Arg1 span, *Only Arg2* denotes those with discontinuous Arg2 spans, whereas *Arg1 & Arg2* correspond to relations having both of their arguments selected discontinuously. *Total Discont.* refers to total number of relations with any type of discontinuous arguments, and *Not Discont.* refers to number of relations which have only continuous arguments.

Table F.1 (continued)

Search Item	**Arg1**	**Arg2**	**Arg1 & Arg2**	**Total Discont.**	**Not Discont.**	**TOTAL**
ha.. ha	-	-	-	**0**	2	**2**
halbuki	1	-	-	**1**	16	**17**
halde	3	-	-	**3**	58	**61**
hem	3	2	-	**5**	77	**82**
hem.. hem	Sng	Sng	Sng	Sng	Sng	Sng
için	226	4	1	**231**	871	**1102**
içindir	-	-	-	**0**	4	**4**
iken	-	-	-	**0**	22	**22**
ister	-	1	-	**1**	5	**6**
kadar	41	-	-	**41**	118	**159**
karşılık	1	2	-	**3**	25	**28**
karşın	7	3	-	**10**	61	**71**
mesela	-	1	-	**1**	12	**13**
ne..ne	2	-	-	**2**	42	**44**
ne ki	1	-	-	**1**	13	**14**
ne var ki	-	1	-	**1**	31	**32**
nedeni ile	1	-	-	**1**	2	**3**
nedeniyle	14	-	-	**14**	28	**42**
nedenle	-	13	-	**13**	104	**117**
nedenlerle	1	1	-	**2**	2	**4**
neticede	-	-	-	**0**	1	**1**
neticesinde	-	-	-	**0**	1	**1**
önce	17	10	2	**29**	105	**134**
örneğin	-	7	-	**7**	57	**64**
örnek olarak	-	1	-	**1**	1	**2**
ötürü	-	-	-	**0**	11	**11**
oysa	1	4	1	**6**	130	**136**
rağmen	7	2	-	**9**	68	**77**
sayede	-	1	-	**1**	4	**5**
sayesinde	-	-	-	**0**	3	**3**
sebeple	-	-	-	**0**	1	**1**
sonra	49	53	-	**102**	611	**713**
sonucunda	5	-	-	**5**	7	**12**
sonuç olarak	-	2	-	**2**	3	**5**
sonuçta	-	-	-	**0**	10	**10**
söz gelimi	-	-	-	**0**	1	**1**
sözgelimi	-	1	-	**1**	5	**6**
taraftan	-	-	-	**0**	3	**3**
tersine	-	-	-	**0**	11	**11**
ve	31	7	-	**38**	2073	**2111**
veya	1	-	-	**1**	39	**40**
veyahut	-	-	-	**0**	4	**4**
ya	1	-	-	**1**	7	**8**
ya..ya	Sng	Sng	Sng	Sng	Sng	Sng
ya da	5	-	-	**5**	134	**139**

Table F.1 (continued)

Search Item	**Arg1**	**Arg2**	**Arg1 & Arg2**	**Total Discont.**	**Not Discont.**	**TOTAL**
yahut	-	-	-	**0**	3	**3**
yalnız	-	-	-	**0**	12	**12**
yandan	1	5	-	**6**	64	**70**
yine de	2	13	-	**15**	50	**65**
yoksa	-	2	-	**2**	73	**75**
yüzden	2	8	-	**10**	56	**66**
yüzünden	3	-	-	**3**	2	**5**
zaman	11	21	1	**33**	126	**159**
zamanda	2	14	-	**16**	23	**39**
TOTAL	*574*	*252*	*7*	*833*	*7650*	*8483*

APPENDIX G

KRIPPENDORFF'S ALPHA AGREEMENT RESULTS FOR ANNOTATIONS IN TDB 1.0

Krippendorff's Alpha Agreement Results for Annotations in TDB 1.0[1][2][3]

Table G.1: Inter-annotator Krippendorff's Alpha Agreements for Independent Annotators in TDB 1.0

Search Item	**# of Annotators**	**Krippendorff's Alpha**			
		Arg1	**Units**	**Arg2**	**Units**
aksine	3	GA	-	GA	-
ama	3	0.73	2682	**0.80**	1186
	3	0.66	732	0.73	730
	2	**0.93**	1841	**0.94**	1670
	3	0.44	1177	0.42	961
amaçla	3	0.75	204	**0.89**	147
amacıyla	3	0.71	689	**0.85**	558
amacı ile	3	NA	-	NA	-
ancak		GA	-	GA	-
ardından	2	**0.87**	987	**0.88**	679
aslında	2	0.65	1074	0.68	861
ayrıca	3	0.53	3590	0.74	1833
beraber		GA	-	GA	-
beri	2	NA	-	NA	-
birlikte	3	GA	-	GA	-
böylece	2	**0.83**	1878	**0.98**	916
bu yana	2	**1.00**	110	**1.00**	54
çünkü	3	**0.86**	3697	**0.91**	3738
dahası	3	0.74	229	**0.89**	240
dolayı	3	0.69	216	0.76	136

[1] GA indicates that the annotations were created by the group annotation procedure for which agreement cannot be calculated; NA shows that inter-coder reliability was not calculated because there were too few annotations; Sng: Single agreement calculated for repeated uses.

[2] The number of annotators specifies if the annotations were done by 3 independent annotators (represented with 3), or by pair annotation procedure (represented with 2). The agreement for the pair annotation is calculated by taking the pair of annotators as a single annotator and comparing their agreement with the independent annotator.

[3] The Krippendorff's alpha values given are for the Overall approach; Bold values represent above 0.80 agreement.

Table G.1 (continued)

Search Item	# of Annotators	Krippendorff’s Alpha			
		Arg1	Units	Arg2	Units
dolayısı ile	3	NA	-	NA	-
dolayısıyla	3	0.73	1204	**0.90**	764
ek olarak	3	NA	-	NA	-
fakat	3	0.65	1634	0.77	1729
fekat	2	NA	-	NA	-
gene de		GA	-	GA	-
gerek	3	NA	-	NA	-
gibi	2	0.57	3992	0.43	1580
ha.. ha	3	NA	-	NA	-
halbuki		GA	-	GA	-
halde	2	0.77	464	**0.83**	406
hem	3	0.74	885	**0.82**	659
hem.. hem		Sng	-	Sng	-
için	3	0.77	3384	**0.85**	2136
içindir	2	NA	-	NA	-
iken		GA	-	GA	-
ister	3	**1.00**	19	**1.00**	29
kadar	2	0.73	1092	**0.82**	700
karşılık	3	GA	-	GA	-
karşın	3	0.71	570	0.75	378
mesela	3	0.65	243	0.77	242
ne..ne	3	**0.88**	170	**0.88**	137
ne ki		GA	-	GA	-
ne var ki		GA	-	GA	-
nedeni ile	2	NA	-	NA	-
nedeniyle	2	**0.97**	658	**0.99**	376
nedenle	2	**0.95**	1728	**0.98**	1331
nedenlerle	2	NA	-	NA	-
neticede	2	NA	-	NA	-
neticesinde	2	NA	-	NA	-
önce	2	**0.84**	438	**0.81**	165
	2	0.58	449	0.58	282
örneğin	3	**0.84**	1255	**0.84**	1830
örnek olarak	3	NA	-	NA	-
ötürü	2	**1.00**	188	**1.00**	99
oysa	3	0.71	2364	**0.87**	1686
rağmen	3	0.63	1195	0.64	890
sayede	2	**1.00**	66	**1.00**	49
sayesinde	2	NA	-	NA	-
sebeple	2	NA	-	NA	-
sonra	3	0.67	4118	0.74	2811
	2	**0.80**	2456	**0.81**	1531
	2	0.74	1746	0.79	1221
	2	0.65	637	0.69	488

Table G.1 (continued)

Search Item	**# of Annotators**	**Krippendorff's Alpha**			
		Arg1	**Units**	**Arg2**	**Units**
sonucunda	3	0.56	390	0.49	189
sonuç olarak	3	0.69	275	**0.90**	109
sonuçta	3	0.51	452	0.62	110
söz gelimi		GA	-	GA	-
sözgelimi		GA	-	GA	-
taraftan	3	NA	-	NA	-
tersine	3	0.49	118	0.68	107
ve	3	0.74	2448	**0.86**	1884
	2	**0.85**	10030	**0.87**	10149
	2	0.73	1005	**0.90**	900
	2	0.76	2179	**0.83**	1759
veya	3	**0.82**	390	**0.85**	176
veyahut	3	NA	-	NA	-
ya	3	0.55	27	**1.00**	18
ya..ya		Sng	-	Sng	-
ya da	3	0.69	334	0.77	278
	2	**0.93**	521	**0.95**	479
yahut		GA	-	GA	-
yalnız		GA	-	GA	-
yandan	3	0.46	1908	0.56	1279
yine de		GA	-	GA	-
yoksa	3	0.69	757	0.75	585
yüzden	2	**0.87**	792	**0.94**	532
yüzünden	2	**0.82**	151	**0.88**	56
zaman	2	**0.84**	1534	**0.87**	1060
zamanda	2	**0.84**	616	0.77	493

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name	**:**	Sevdik Çallı, Ayışığı Başak
Nationality	**:**	Turkish (TC)
Date and Place of Birth	**:**	01.01.1981
Gender	**:**	Female
Marital Status	**:**	Married

EDUCATION

Degree	**Institution**	**Year of Graduation**
M.S.	Bilkent University Computer Engineering Ankara, Turkey	January 2004
B.S.	Başkent University Computer Engineering Ankara, Turkey	June 2001

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
November 2007 – February 2012	Middle East Technical Unv., ANKARA	Research and Teaching Assistantship.
September 2007 – March 2011	Middle East Technical Unv., ANKARA	Group Member for Turkish Discourse Bank Project (medid.ii.metu.edu.tr) Project supported by TÜBİTAK (The Scientific and Technological Research Council of Turkey - Project No: 107E156).
February 2006 – February 2007	TOBB, ANKARA	Global Standards Department, Assistant Expert.
January 2005 – June 2005	Bilkent University, ANKARA	Research and Teaching Assistantship.
October 2004 – January 2005	Bilkent University, ANKARA	Teaching Assistantship.
October 2004 - January 2005	Middle East Technical Unv., ANKARA	YUUP (DPT Sponsored) - "Visual Archive Systems for E-government": Education work package.
January 2004 – October 2004	Bilkent University, ANKARA	Research and Teaching Assistantship.
June 2000 – July 2000	Marconi Komunikasyon A.S., ANKARA	Summer Practice
July 1999 – August 1999	HAVELSAN, ANKARA	Summer Practice

PUBLICATIONS

Zeyrek, D.; Demirşahin, I.; Sevdik Çallı Ayışığı B. & Çakıcı, R. (2013). "Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language", *Dialogue and Discourse*, 2013, 4, pp.174-184.

Sevdik Çallı, A.B. (2012). "Demonstrative Anaphora in Turkish: A Corpus Based Analysis" (poster). In *the Workshop Abstracts of the 8th Language Resources and Evaluation Conference*, 1st Workshop on Language Resources and Technologies for Turkic Languages. p.21.

Demirşahin, I., Sevdik Çallı, A.B., Ögel Balaban, H., Çakıcı, R., and Zeyrek, D. (2012). "Turkish Discourse Bank: Ongoing Developments" (poster). In *the Workshop Abstracts of the 8th Language Resources and Evaluation Conference*, 1st Workshop on Language Resources and Technologies for Turkic Languages. p.19.

Zeyrek, D., Demirşahin, I., Sevdik Çallı, A.B., Ögel Balaban, H., Yalçınkaya, İ., and Turan, Ü. D. (2010). "The annotation scheme of the Turkish Discourse Bank and an evaluation of inconsistent annotations". In *the Proceedings of the Fourth Linguistic Annotation Workshop*. pp. 282–289.

Zeyrek, D., Turan Ü. D., Bozşahin C., Çakıcı R., Sevdik Çallı A. B., Demirşahin I., et al. (2009). "Annotating Subordinators in the Turkish Discourse Bank". In *the Proceedings of ACL-IJCNLP, Linguistic Annotation Workshop III*. pp. 44–48.

Zeyrek, D., Demirşahin, I., and Sevdik Çallı, A. B. (2008). "ODTÜ Metin Düzeyinde İşaretlenmiş Derlem Projesi Tanıtımı" [METU Turkish Discourse Bank Presentation]. Mersin Üniversitesi Sempozyumu Bildirileri. pp. 544-553. [in Turkish].

Can, F., Nuray, R. and Sevdik, A. B. (2004), "Automatic Performance Evaluation of Web Search Engines", *Information Processing and Management*, Volume 40, Issue 3, May 2004, pp. 495-514.

Sevdik, A.B. (2004). "Extensible Markup Language (XML) in Electronic Government: Some Exemplary Scenarios", MS Thesis, Bilkent University, January 2004.

Kaynar, İ., Satıroğlu, Y., Sevdik, A.B. (2003). "Applying AOSD to Legacy Code." In *the Proceedings of TAOSD2003*, pp.3-7, Ankara, Turkey, May 2003.

Akman, V. and Sevdik, A. B. (2003). "Türk Kadınının Hayatında İnternetin Yeri." [The Place of the Internet in the Lives of Turkish Women]. *Karizma* (13), pp. 92-107 (Ocak/Şubat/Mart 2003) [in Turkish]..

Sevdik, A. B. and Akman, V. (2002). "Internet in the Lives of Turkish Women", *First Monday*, Vol.7, Number 3-4, March 2002.

COMPUTER SKILLS

Programming Languages:

C/C++, Java, Delphi, Pascal, PERL, Python, Prolog, Matlab, SQL, HTML, XML, SOAP.

Tools:

IntelliJ IDEA, Borland JBuilder, JCreator, EditiX XML Editor, XML Spy, Borland C++ Builder, Rational Rose, Borland Delphi, Microsoft Office, SPSS.

Operating Systems:

Windows, UNIX, DOS.

LANGUAGES

- Turkish (Native)
- English (Advanced) (TOEFL IBT:115)
- French (Elementary)

INTERESTS

- Computational Linguistics: Anaphora Resolution, Discourse Connectives and Relations, Anaphora Annotation, Discourse Annotation, Sense Annotation.
- Social Aspects of the Internet, XML and related standards, Electronic Government, Women in Computing.
- Digital photography, music (flute), tennis.

MEMBERSHIPS

- IEEE Member (since January 2000)
- Cognitive Science Society Member (since January, 2011)
- ACL Member (2010)

HONORS AND AWARDS

- Awarded tuition scholarship for PhD study from Middle East Technical University, 2007-2011.
- Awarded tuition scholarship for MS study from Bilkent University, 2001-2004.