A STUDY ON ALTERNATIVE LEXICALIZATIONS IN TURKISH DISCOURSE BANK

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

BY

FİKRET GÜNAY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN THE DEPARTMENT OF COGNITIVE SCIENCE

SEPTEMBER 2015

A STUDY ON ALTERNATIVE LEXICALIZATIONS IN TURKISH DISCOURSE BANK

Submitted by FİKRET GÜNAY in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Science, Middle East Technical University** by,

Prof. Dr. Nazife Baykal _____

Director, **Informatics Institute, METU**

Prof. Dr. Cem Bozşahin _____

Head of Department, **Cognitive Science, METU**

Prof. Dr. Deniz Zeyrek Bozşahin _____

Supervisor, **Cognitive Science, METU**

Examining Committee Members:

Prof. Dr. Mustafa AKSAN _____

English Language and Literature, Mersin University

Prof. Dr. Deniz Zeyrek Bozşahin _____

Cognitive Science, METU

Assist. Prof. Dr. Cengiz Acartürk _____

Cognitive Science, METU

Prof. Dr. Hüseyin Cem Bozşahin _____

Cognitive Science, METU

Assist. Prof. Dr. Murat Perit Çakır _____

Cognitive Science, METU

Date: _____

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: Fikret GÜNAY

Signature : _____

ABSTRACT

A STUDY ON ALTERNATIVE LEXICALIZATIONS IN TURKISH DISCOURSE BANK

Günay, Fikret

MS, Department of Cognitive Science

Supervisor: Prof. Dr. Deniz ZEYREK BOZŞAHİN

September 2015, 61 pages

Discourse relations connect two pieces of discourse and represent a relationship between these two arguments. Discourse relations can be expressed both explicitly and implicitly. The objective of the present thesis is to identify alternative lexicalizations (ALTLEXs) in Turkish (which is a type of implicit relations) in Turkish Discourse Bank, or TDB by means of a corpus-based approach. The thesis contributes to our understanding of Turkish discourse by revealing a set of ALTLEXs. Three methods are employed: a) An annotation process of ALTLEXs is undertaken in TDB. In this procedure, first, 10% of the entire TDB (20 files, approximately 20000 words) are doubly annotated; then, the discovered ALTLEXs are searched and annotated in the entire TDB. Inter-annotator agreement (IAA) is calculated to check the reliability of annotations. b) A lexico-syntactic classification of Turkish ALTLEXs is done, where the ALTLEXs are classified into three groups; i.e. the closed class, the partially open class, and the open ended category. c) Since the open-ended category had too few instances, an automatic extraction method is developed to extract more possible open-ended ALTLEXs. Using all these methods, the thesis finds a total of 94 types (297 tokens) of ALTLEXs in Turkish. This set of ALTLEXs will contribute to the enrichment of TDB with more annotations and help pave the way to new research.

Key words: discourse connective, alternative lexicalization, ALTLEX implicit discourse relation, explicit discourse relation, Turkish Discourse Bank

ÖZ

TÜRKÇE SÖYLEM BANKASINDAKİ BAĞLAÇSILARIN ÇALIŞMASI

Günay, Fikret

Yüksek Lisans, Bilişsel Bilimler Bölümü

Tez Yöneticisi: Prof. Dr. Deniz ZEYREK BOZŞAHİN

Eylül 2015, 61 sayfa

Söylem bağıntıları iki öğeyi birbirine bağlar ve bu iki öğe arasındaki ilişkiyi gösterir. Söylem bağıntıları gizli ya da açık olarak ifade edillir. Bu tezin amacı, Türkçe'de gizli söylem bağıntı çeşitlerinden biri olan bağlaçsıların Türkçe Söylem Bankası'nda derlem çalışmasıyla tespit edilmesi ve tanımlanmasıdır. Bu çalışma, bağlaçsı çeşitlerini ortaya çıkararak Türkçe'deki söylem kavramına katkıda bulunmaktadır. Bağlaçsıların tanımlanması için üç yöntem kullanılmıştır. a) Türkçe Söylem Bankası'ndaki (TSB) bağlaçsılar işaretlenmiştir. Bu aşamada, TSB'nin %10'unundan oluşan kısımda (20 dosya = yaklaşık 20000 kelime) bağlaçsılar işaretlenmiş ve bulunan bağlaçsılar tüm TSB'de işaretlenmiştir. Ayrıca, işaretlemelerin güvenilirliğini ölçmek için işaretleyiciler arası uyum hesaplanmıştır. b) Bağlaçsılar için sözlüksel ve sözdizimsel sınıflandırma yapılmıştır, Türkçe'deki bağlaçsılar üç gruba ayrılmıştır; kapalı, kısmen açık, açık bağlaçsılar. c) Açık bağlaçsı sayısı çok az olduğundan dolayı, daha çok açık bağlaçsının tanımlanması için bir Java kodu geliştirilmiştir. Bu yöntemlerle, toplam 94 tür/297 türce bağlaçsı tanımlanmıştır. Bu bağlaçsılar, Türkçe Söylem Bankası'ndaki işaretlemelerin atrmasına ve yeni araştırma alanlarına yol açacaktır.

Anahtar Kelimeler: söylem bağıntısı, bağlaçsı, gizli söylem bağıntısı, açık söylem bağıntısı, Türkçe Söylem Bankası

DEDICATION

To my beloved husband…

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Prof. Dr. Deniz Zeyrek Bozşahin for her endless support and valuable commends. This thesis would not be possible without her support. I would like to thank Işın Demirşahin and Ayışığı Başak Sevdik Çallı for their valuable comments and contributions to this study.

I owe special thanks to Deniz Hande Çakmak for her help especially in the annotation process and Murathan Kurfalı for his help especially in calculating inter annotator agreement results. They had always helpful answers to my endless questions.

I wish to express my gratitude to my parents Nurhayat and İsmet Arslan, and my sister

Zehra Arslan, and my brother Servet Arslan for their support, trust and encouragement throughout my life.

I would like to express my deepest and sincere gratitude to my excellent husband Murat Günay for his love, understanding, encouragement and patience. I am grateful to him for his help and guidance in the automatic extraction method used in the thesis. Without his support, this study would never have ended.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

ALTLEX: Alternative Lexicalization

AOs: Abstract Objects

CLExp: Complete List of Explicit Connectives

CDTB: Chinese Discourse Tree Bank

EntReL: Entity Relation

HDRB: Hindi Discourse Relation Bank

IAA: Inter-Annotator Agreement

MTC: METU Turkish Corpus

LT: Language Technology

NoReL: No Relation

PDiT: Prague Discourse Treebank

PDTB: PENN Discourse Tree Bank

TDB: Turkish Discourse Bank

CHAPTER 1

INTRODUCTION

This chapter includes the description of discourse and discourse relations in general. Implicit and explicit discourse relations are analyzed with examples both in English and Turkish. The Penn Discourse Treebank finds three types of implicit discourse connectives, i.e., alternative lexicalization (ALTLEX), entity relation (Entrel), and no relation (NoRel). Each of these is introduced, with particular emphasis on ALTLEXs.

1.1 Discourse, Discourse Relations, Discourse Connectives

Discourse is concerned with the relationship of sentences to each other with respect to the contexts in which they are used. It grew out of work in different disciplines in the 1960s and early 1970s, including linguistics, semiotics, psychology, anthropology and sociology (McCharthy, 1991). The relationship of the sentences can be made explicit with a discourse marker, or a discourse connective. Discourse markers are defined as a pragmatic class, i.e. “lexical expressions drawn from the syntactic classes of conjunctions, adverbials, and prepositional phrases” (Fraser, 1999: p. 931). According to Fraser (1999), with certain exceptions, discourse markers signal a relationship between the segments they introduce, which is the S2, and the prior segment, i.e. the S1. According to a recent work by Pitler et al. (2008), discourse relations between textual units are considered the key for the ability to properly interpret or produce discourse.

In Turkish, discourse connectives can be conjunctions (çünkü (because), ama (but), ve (and)), discourse adverbials (üstelik (additionally)) and connectives with a deictic item (buna rağmen (despite this)). This thesis is about discourse relations and how they are signaled by means of linguistics expressions other than the canonical

connectives (i.e. ALTLEXs) such as “buna rağmen” (despite this) in Turkish. According to Prasad et al. (2007: p.1), “an important aspect of discourse understanding and generation involves the recognition and processing of discourse relations”. For the Penn Discourse Tree Bank (PDTB) research group, discourse connectives are treated as discourse level predicates that take two abstract objects such as events, states, and propositions as their arguments. Zeyrek and Webber (2008) also assert that from a semantic perspective, a discourse connective is a predicate taking as its arguments, abstract objects (propositions, facts, events, descriptions, situations, and eventualities). From our perspective, the idea is that discourse relations connect two segments of discourse no matter how these segments are named, i.e. as Sentence1 (S1) - Sentence2 (S2) (Fraser, 1999), or Argument1 (ARG1)- Argument2 (ARG2).

1.2 Explicit and Implicit Discourse Relations

In PDTB, discourse connectives include:

I) explicit discourse connectives, which are drawn from well-defined syntactic classes: subordinating conjunctions (e.g., because, when, since, although), coordinating conjunctions (e.g., and, or, nor) and adverbials (e.g., however, otherwise, then, as a result, for example) (Prasad et al., 2007).

II) Implicit discourse relations, which signal a relation between adjacent sentences where the relation is not expressed with an explicit discourse connective (Prasad et al., 2007).

Zeyrek and Webber (2008) confirm that an explicit connective is realized in the form of a lexical item (e.g. ama (but), çünkü (because), ve (and), etc.) or a group of lexical items (e.g. hem…hem…(both… and), sonuç olarak (as a result), ne var ki (however), etc.), while an implicit connective can be inferred from adjacent text spans that realizes abstract objects (AOs) and whose AOs are taken to be related.

In Turkish, explicit discourse connectives are identified by analyzing three syntactic categories: (I) Coordinating conjunctions (II) Subordinators (III) Discourse adverbials (or anaphoric connectives). All these discourse connectives have two and only two arguments, which are labeled as ARG1 and ARG2. ARG2 is always the argument which is bound to the connective, while ARG1 is the other argument (Zeyrek and Webber, 2008). Example (1) includes a sample from Turkish Discourse

Bank for an explicit connective (namely, a conjunction), where the connective is in bold.

1) Yapılarını kerpiçten yapıyorlar, ama sonra taşı kullanmayı öğreniyorlar. Mimarlık açısından çok önemli, **çünkü** bu yapı malzemesini başka bir malzemeyle beraber kullanmayı, ilk defa burada görüyoruz. (Zeyrek & Webber: 2008: p.3)

   "They constructed their buildings first from mudbricks but then they learnt to use the stone. Architecturally, this is very important **because** we see the use of this construction material with another one at this site for the first time."

Example (2) includes an explicit connective (namely, a subordinating conjunction), karşın (even though).

2) Mehpare Hanım gibi piyano çalan, kitap okuyan birkaç genç hanımın **olmasına karşın** çoğunluk yerleşik zevklere sahipti.

   "**Even though** there are several young women as Mrs. Mehpare who plays a piano and reads books, the majority has ordinary pleasures.

In this thesis, we will use METU Turkish Discourse Bank (Zeyrek, et al. 2009) as the data, which follows PDTB, an influential corpus for English annotated at the level of discourse. Samples (3), (4), (5) show explicit connectives from PDTB.

3) **Since** McDonald's menu prices rose this year, the actual decline may have been more. (Prasad et. al.,2007, p.8)
4) The House has voted to raise the ceiling to $3.1 trillion, **but** the Senate isn't expected to act until next week at the earliest. (Prasad et. al.,2007, p.8)
5) In the past, the socialist policies of the government strictly limited the size of new steel mills, petrochemical plants, car factories and other industrial concerns to conserve resources and restrict the profits businessmen could make. **As a result**, industry operated out of small, expensive, highly inefficient industrial units. (Prasad et. al.,2007, p.8)

In addition to explicit discourse connectives, there are implicit discourse relations which can be inferred from related text spans that have coherence relations (Zeyrek et al., 2009). PDTB annotates implicit discourse relations, where the goal of annotating them is to capture relations between abstract objects that are not

realized by an explicit discourse connective in the text and are left to be inferred by the reader (Prasad et al., 2007). In PDTB 2.0, implicit discourse connectives are only annotated on the inter-sentential level in discourse (Prasad et al., 2007). Samples (6) and (7) show implicit discourse relations in English. The PDTB research group asks their annotators to insert an explicit connective for each implicit relation (shown in parentheses in bold):

6) Several leveraged funds don't want to cut the amount they borrow because it would slash the income they pay shareholders, fund officials said. But a few funds have taken other defensive steps. Some have raised their cash positions to record levels. **(because)** High cash positions help buffer a fund when the market falls.
7) The project under construction will increase Las Vegas' supply of hotel rooms by 11,795, or nearly 20%, to 75,500. **(so)** By a rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs.

Samples (8), (9) and (10) are examples for implicit discourse relations in Turkish. There is no explicit discourse connective in these examples, but there is an inference of a relation between two sentences in each example. The inferred relation can easily be made explicit by means of lexical expressions, which are shown in bold for each example. The examples are from Annotation Guidelines for Implicit Relations in Turkish (Zeyrek et al., ms).

8) Susamış görünüyorsun. **(Öyleyse)** Dolapta bira var.
   "You look as if you are thirsty. **(If so)** There is a beer in the fridge.
9) Geçen hafta her gün okula gittim. **(Sadece)** Hasta olduğum gün gitmedim.
   "Last week, I went to school every day. **(Only)** I did not go the day I was ill.
10) Saat 12.30. Cengiz hoca odasında yok. **(Demek ki)** Yemeğe gitmiş.
    "It is 12:30. Mr. Cengiz is not in his room. **(It means that)** He went out for lunch."

1.3 Alternative lexicalizations

PDTB classifies implicit relations (or connectives) into three; Alternative Lexicalization (ALTLEX), Entity Relation (EntRel) and No Relation (NoRel) (Prasad et al., 2007). This thesis only concerns the ALTLEX group but for the sake of completeness, we will define all three types below.

-According to PDTB research group; “ALTLEX is where a discourse relation is inferred, but insertion of an implicit connective leads to redundancy in its expression due to the relation being alternatively lexicalized by some other expression” (Prasad et al., 2007).

> 11) And she further stunned her listeners by revealing her secret garden design method: Commissioning a friend to spend “five or six thousand dollars . . . on books that I ultimately cut up.” **(ALTLEX) After that**, the layout had been easy. (Prasad et. al., 2007: p.22).

-ENTREL shows an entity relation or an inference of discourse relation of further information about an entity in the previous sentence (Prasad et. al., 2007).

> 12) **Hale Milgrim**, 41 years old, senior vice president, marketing at Elecktra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern. **(EntRel) Mr. Milgrim** succeeds David Berman, who resigned last month. (Prasad et. al., 2007: p.23).

-NOREL includes neither an explicit discourse connective, ALTLEX nor an entity-based relation. If there is no relation between two sentences then it is tagged as Norel.

> 13) Jacobs Engineering Group Inc.’s Jacobs International unit was selected to design and build a microcomputer-systems manufacturing plant in County Kildare, Ireland, for Intel Corp. Jacobs is an international engineering and construction concern. **(NoRel)** Total capital investment at the site could be as much as $400 million, according to Intel. (Prasad et. al., 2007: p.25).

Methodologically, PDTB annotates explicit connectives first, then implicit connectives. After finishing implicit connectives, annotators realize that in many cases, they are not able to supply an implicit connective. Reasons includes “there is a relation between the sentences for which I can think of a connective, but it doesn’t soundgood”. Such cases are annotated as ALTLEX (Prasad et al., 2010). Prasad et al. (2010) explain the annotation procedure of ALTLEXs as follows; “while annotating implicit connectives, annotators were unable to insert a connective despite the inference of a discourse relation because there was a perceived redundancy after insertion of the connective, and thus the connective did not meet the fluency criteria. Although no explicit connective was present to relate the two sentences, some other expression appeared to be doing the job” (Prasad et al., 2010).

Samples (14), (15) and (16) are examples for ALTLEXs from PDTB,

14) Ms. Bartlett's previous work, which earned her an international reputation in the non-horticultural art world, often took gardens as its nominal subject. **Mayhap this metaphorical connection made** the BPC Fine Arts Committee think she had a literal green thumb (Prasad et. al., 2007:p.23).

15) But a strong level of investor withdrawals is much more unlikely this time around, fund managers said. **A major reason** is that investors already have sharply scaled back their purchases of stock funds since Black Monday (Prasad et. al., 2010).

16) Now, GM appears to be stepping up the pace of its factory consolidation to get in shape for the 1990s. **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 (Prasad et. al., 2010).

In these examples (14, 15, 16), the phrases "one reason is, a major reason, Mayhap this metaphorical connection made" are taken to denote the relation and are marked as ALTLEX.

Below, samples (17), (18), and (19) show ALTLEXs in Turkish which we have identified; "bir başka deyişle (in other words), bu benzerliklerin yanında (In addition to these similarities), bu bahaneyle (under this excuse)";

17) "Çocuklar ihtiyaçları göz önüne alındığında çok hararetli tüketicilerdir. **Bir başka deyişle**, çocuklara birşey satmak isteniyorsa onların görebileceği ve alabileceği yerlere konulmalı ve canı sıkılmış bir çocuğun eğlenmesini sağlamak gerekli."

"Taking into account their needs, children are extreme consumers. **In other words**, if one wants to buy something to children, it should be put in the place where they can see and reach, and it is required to entertain a child who is bored."

18) Çatalhöyük ile Aşıklı arasında kültür olarak da birtakım benzerlikler var. **Bu benzerliklerin yanında** farklılıklar da var.

“There are some similarities between Çatalhöyük and Aşıklı culturally. **In addition to these similarities**, there are also differences.

19) Halil her gün şarap satın almaya gidiyor. **Bu bahaneyle** de de tek dostu olan Ante'yle konuşuyordu.

“Halil went to buy beer every day. **With this excuse**, he talked to Ante, his only friend.

Turkish Discourse Bank or TDB is an annotated corpus where discourse relations are annotated; so far, only those relations that are expressed with explicit discourse connectives have been annotated. After annotating explicit discourse connectives, the research group aims to annotate ALTLEXs. The current thesis will therefore serve as a starting point for enriching TDB with the ALTLEX category.

1.4 Significance of the study

The study of discourse has become relevant when it was noticed that language studies should not be restricted to the grammatical analysis of the language systems, rather actual language use in the social context (Dijk, 1983). Within the perspective of early linguistic studies, which focused mainly on phonology, morphology and syntax, little attention was paid to discourse particles (Yılmaz, 2004). Both in Turkish and in English, the study of discourse and specifically discourse connectives have largely followed structural studies on the major components of language (phonology, morphology, syntax). In Turkish, except a few studies such as Uzun (1995), which is a study on Orhon inscriptions, and studies on modern Turkish such as Ruhi (1994), Özbek (1995), Yılmaz (2004), Zeyrek & Webber (2008) among others, studies on discourse markers/connectives are still rare. Moreover, to the best of our knowledge, implicit relations have not been studied separately from explicit connectives in sufficient detail. Therefore, this thesis will fill an important gap by investigating alternative lexicalizations in Turkish theoretically and taxonomically.

Recently, discourse studies have been gaining ground in language technology. “Given that language carries information in its structures – morphological, phonological, syntactic, etc., it is fitting that Language Technology (LT) can exploit these structures in two ways. On the one hand, LT can operate on the units provided by structure” (Webber et al., 2012: p.437)). As Webber et al. (2012) noted, early discourse studies lack the huge amounts of text which are used for empirical

language studies (e.g., annotated corpus). With the help of growing amounts of data, the contribution of discourse to language technology will rise. Our study may be the first to raise awareness and pave the way to LT studies concerning ALTLEXs in Turkish.

In this thesis, we primarily aimed to identify possible ALTLEXs in TDB and annotate them. While annotating these devices, PTDB principles regarding ALTLEXs are taken to be the guide, that is, if there is a semantic relation between two adjacent sentences, and insertion of a connective makes the expression redundant, then we annotate it as an ALTLEX. We do not claim that ALTLEXs can only be found only between adjacent sentences. This is only a limitation we have used in our study.

1.5 Aims of the study

This study is a corpus-based study and has three major aims:

1) To identify the typology of Turkish ALTLEXs from TDB by means of manual corpus annotation,
2) To find out how to classify ALTLEXs in an appropriate way,
3) To extract possible open ended ALTLEXs automatically.

Regarding the aims of this thesis, first, the differences between ALTLEXs and explicit discourse connectives are analyzed. For this purpose, in TDB, 20 randomly chosen files (approximately 20000 words) which cover all types of genres is selected and manually annotated for ALTLEXs and their arguments.

Before the annotation procedure, a full list of explicit discourse connectives is formed by the author, based on Göksel and Kerslake (2004), Lewis (1985) and TDB. This list guides our ALTLEX detection and annotation process; while looking for ALTLEXs it is usually hard to determine if a lexical expression is an ALTLEX or an explicit discourse connective. Our guiding rule is that, if the expression is not in the explicit connectives list then it may be taken as an ALTLEX.

Regarding the typology of ALTLEXs, Prasad et al. (2010) provide the following classification for English ALTLEXs;

- Syntactically admitted, lexically frozen;

E.g. quite the contrary, for one thing, as well, too, soon, eventually, thereafter, even, especially, actually, still, only, in response

- Syntactically free, lexically frozen;

E.g. What's more, never mind that, to begin with, so, another, further, as in, so what if, best of all

- Syntactically and lexically free.

E.g. That compares with, after these payments, that would follow, the plunge followed, until then, the increase was due mainly to, that is why, once triggered

Aravind Joshi (2010) also suggests a typology:

- closed class ALTLEXs,
- partially open class ALTLEXs,
- open-ended ALTLEXs.

Although both classifications are based on lexico-semantic categories, we find Joshi's categorization (2010) simpler. Hence, it will be preferred in this thesis.

The outline of the thesis is as follows: In Chapter 2, a literature review of coherence, discourse and ALTLEXs are introduced. Chapter 3 presents the methods of the study both for identifying ALTLEXs and extracting possible open ended ALTLEXs. In Chapter 4, the results of the study are provided, and the discussions of the results are provided in Chapter 4. Chapter 5 presents the conclusions of the thesis.

CHAPTER 2

LITERATURE REVIEW

This chapter includes the theoretical framework for the current study. First, the concept of discourse and discourse connectives are defined with Turkish and English examples. Next, corpus efforts concerned with discourse are reviewed, primarily focusing on ALTLEXs and Turkish discourse. Finally, the possible contribution of to language technology in corpus studies is mentioned.

2.1 How Do We Infer Coherence

Discourse is not a list of random utterances, but it shows connectedness of utterances, and the aim of discourse studies is to find out how this connectedness is formed in discourse (Sanders and Maat, 2006). For Halliday and Hasan (1976), reference, substitution, ellipsis, conjunction and lexical cohesion are the types of cohesion which describe text connectedness. However, Yavuz (2011) reveals that in some of the Turkish discourse studies, the existence of conjunctions is controversial for linguists. Uzun's (1995) study on the Orhon Inscriptions, where she analyzed old Turkish language, finds that there exist very few conjunctions in old Turkish. On the other hand, Korkmaz (2005) analyzes connective devices in Turkish, and argues that Turkish has connectives. She classifies connectives into three, as the borrowed connectives (e.g. bilakis (to the contrary), adeta (as if)), blended forms (e.g. Turkish and Farsi blends, e.g. demek ki (as a result)), and conjunctions of Turkic origin (e.g. ayrıca (in addition), ancak (however))). The sentences below represent an example of a connect,ve of Turkish origin;

Tahsilimi yarıda bırakırsam beni evlatlıktan reddedecekmiş. **Üstelik** de maldan mülkten mahrumiyet (Kocagöz, 1964).

"If I leave my education, he will disinherit me. **Furthermore**, he will deprive me of propery."

A notable aspect of Kokmaz (2005) is that, what we call ALTLEXs are referred as connectives, e.g., bununla birlikte (in addition to this), öncelikle (first of all), özellikle (specifically), bundan dolayı (because of this), başka bir deyişle (in other words).

In traditional accounts, conjunction relates the arguments in discourse with explicit cohesive conjunctions (Halliday and Matthiessen, 2013). Example (20) includes a conjunction which relates two arguments;

> 20) Ewa walked into town, **because** she wanted an ice cream (Sanders and Maat, 2006).

Coherence can also be achieved by various other means. For example reference creates referential chains to create links between elements (Halliday and Matthiessen, 2013). Example (21) shows a referential link between "the park" and "there";

> 21) Jan lives near **the park**. He often goes **there** (Sanders and Maat, 2006).

A lexical item in text is replaced by substituting an item. Example (22) shows the substitution of "ice-cream" with "one".

> 22) Daan loves strawberry **ice-creams**. He has one every day (Sanders and Maat, 2006).

Ellipsis includes the omission of a lexical item if it is predictable in the prior utterances, as in example (23):

> 23) All the children had an **ice-cream** today. Eva chose **strawberry**. Arthur had **orange** and William too (Sanders and Maat, 2006).

Lexical cohesion is yet another aspect of coherence; it reveals the lexical aspect of coherence while others reveal the grammatical aspect of coherence (Sanders and Maat, 2006). Lexical cohesion comprises two elements which share a lexical field, and this is achieved by repetition, synonymy, hyponymy and collocation (Halliday, 1985). Example (24) presents an instance of lexical cohesion. This example shows both the repetition of "wriggle" and the lexical relation between "boys" and "girls".

24) Why does this little **boy** wriggle all the time? **Girls** don't wriggle (Halliday and Hasan, 1976: p.285).

Anaphoric reference reveals a relationship between lexical items in discourse, and there are direct and indirect anaphora types. Indirect anaphora constructs bridging inferences in discourse (Irmer, 2009), which make the text coherent, and the sequence of lexical items in a text makes the message predictable (Singer et. al., 1992). Example (25) shows a bridging inference from Turkish, where "çatı (the roof)" is understood ad the "roof of the house".

25) Evin duvarları düzdü. Çatı eğikti. (Zeyrek et. al, ms)

"The walls of the house were straight. The roof was slanted."

2.2 Discourse Relations

The studies about coherence relations and discourse markers have gained an increasing interest in the current linguistic studies with the rise of corpus studies (Prasad et. al., 2007: Zeyrek et. al., 2009: Das and Taboada, 2013: Taboada, 2009). As has already been indicated, there are different labels for discourse connectives; discourse markers (Fraser, 1999), discourse particles (Siegel, 2002), discourse connectives (Prasad et. el., 2007: Zeyrek et. al., 2009), coherence relations (Halliday, 1985).

Regarding Turkish, Kerslake (1996) is one of the first studies which emphasize the functional classification of Turkish discourse connectives. This functional classification is based on Halliday and Hasan (1976), where discourse relations are classified into two groups as internal and external conjunctive relations. Halliday and Hasan (19876) define external conjunctive relations as "being inherent in the phenomena that language is used to talk about", and internal conjunctive relations as "being inherent in the communication process, in the forms of interaction between speaker and hearer". The examples below show the difference between internal and external conjunctive relations.

a) She was never really happy here. <u>So</u> she's leaving.
b) She'll be better off in a new place. <u>So</u> she's leaving (Kerslake, 1996).

The sentence in (a) explains a causal relation between two sentences which is external; however, in (b), there is a there is an inference of a relation which is internal.

Examples (26) and (27) show what Kerslake (1996) have called internal and external conjunctive relations. In this thesis, we do not distinguish between internal and external conjunctions in Turkish.

26) Her sabah evi topluyor. **Sonra** akşama yemek yapıyor.

"Every morning she tidies the house. **Then** she cooks for the evening."

27) Evini hep toplu tutuyor. **Sonra** güzel yemekler yapıyor.

"She always keeps her house tidy. **And again**, she cooks well."

Example 26 shows an external relations indicating temporal relation between two habitual events, and Example 27 shows an internal relation indicating two statements which are presented by the speaker as two points made in support of a single argument (Kerslake, 1996).

2.3 The Distributions of Explicit and Implicit Relations across Languages

As we have already emphasized, many researchers have realized that coherence does not necessitate explicit markers. For example, Knott and Sanders (1998) provides the following examples:

28) Tim must love that Belgian beer. The crate in the hall is already half empty (Knott and Sanders, 1998).

29) Tim must love that Belgian beer. He's six foot tall (Knott and Sanders, 1998).

Example (28) is coherent and it is easy to infer a relation between two sentences, but Example (29) creates a problem while trying to interpret the relation between two sentences. There are certain signals for a text to be coherent; for example, the second argument is an expansion, justification, or conclusion of the previous argument (Knott and Sanders, 1998). There is an absence of these signals in (29), and it is not possible to infer a relation between two sentences. However, (28) includes an evidence relation even without an explicit connective.

Das and Taboada (2013) assert that early work about discourse markers (Taboada and Mann, 2006) take discourse markers as the only signals for a text to be coherent. Similar to Knott & Sanders (1998) and many others, the current work by Das and Taboada (2013) claims that the text can be coherent in the absence of discourse markers and the absence of a discourse marker in a coherence text is called implicit relation. The idea is that not only explicit discourse relations but also implicit ones make the text coherent. Examples (30), (31) and (32) show instances of implicit discourse relation in English and Turkish:

30) John is tall. **(Implicit: BUT)** Mary is short (Das and Taboada, 2013)

31) Sesi soğuk ve uzaktı. **(Implicit: BU YÜZDEN)** Uygunsuz bir zamanda aramış olduğumu düşündüm. (TDB)

"His voice was cold and distant. **(Implicit: THEREFORE)** I thought I had called at an inappropriate time." (TDB)

32) Yaşamınızın elinizden alındığını düşünüyorsunuz; **(Implicit: AMA)** size yeni bir yaşam sunulduğunu değil. (TDB)

"You think your life has been taken away from you; **(Implicit: BUT)** it is not that you think a new life is being offered to you." (Zeyrek et. al., 2015)

In examples (30, 31, 32), two clauses or two sentences are related without an explicit discourse connective. In example (30), there is an inference of contrast because "being tall" and "short" sufficiently establish a connection between these arguments. In example (31) and (32), there is also an inference of contrast.

2.3.1 Penn Discourse Tree Bank (PDTB)

PDTB is a large corpus which covers manually annotated discourse relations, which comprises files from Wall Street Journal (Prasad et al. 2008). There are two underlying principles used in PDTB: "First, it makes no commitment to any kind of higher-level discourse structure over the discourse relations annotated between individual text spans. Thus, while theory-neutral itself with respect to higher-level discourse structure, the PDTB invites experimentation with approaches to high-level topic and functional structuring" (Prasad, Webber and Joshi, 2014). The second principle is that "the annotation of discourse relations is lexically grounded. Rather than asking annotators to directly classify the sense of relations, which is a difficult task (Stede 2008), annotators were asked to look at lexical items that can signal

discourse relations" (Prasad et. al., 2014). If annotators find a discourse relation between adjacent sentences, then they annotate the arguments of the relation, and senses of the relation (Prasad et. al., 2014).

Methodologically, in PDTB first, explicit connectives are annotated, where the categories of explicit connectives are taken from the previous researchers (Halliday and Hasan 1976; Martin 1992; Knott 1996; Forbes-Riley, Webber, and Joshi 2006). After these explicit connectives, implicit connectives are added to the annotation if annotators find them in the corpus (Prasad et. al., 2014). If there is a relation between two arguments, but there is an absence of explicit discourse connective, then annotators annotate it as an implicit discourse relation. Explicit discourse connectives are annotated one connective at a time throughout the corpus. Implicit discourse relations are annotated document by document (Prasad et. al., 2014).

In PTDB, annotators are told to annotate only the relations between adjacent sentences, excluding those that hold across sentences (Prasad et. al., 2014). Therefore,

ALTLEXs are not fully annotated in PDTB because they are annotated only when the insertion of an explicit connective is redundant (Prasad et. al., 2014). They annotate these implicit discourse relations first by inserting an appropriate explicit discourse connective between the arguments, and then they classify implicit relations as Alternative Lexicalization (ALTLEX), Entity Relation (EntRel) and No Relation (NoRel). Example (33) shows an instance of implicit discourse relation in PDTB:

> 33) Meanwhile, the average yield on taxable funds dropped nearly a tenth of a percentage point, the largest drop since midsummer. **(implicit = in particular)** The average seven-day compound yield, which assumes that dividends are reinvested and that current rates continue for a year, fell to 8.47%, its lowest since late last year, from 8.55% the week before, according to Donoghue's (Prasad et. al., 2007: p.20).

As already mentioned, PDTB annotators notice that sometimes it is impossible to insert an explicit connective between the arguments even though there is a relation between sentences because insertion of an explicit discourse connective makes the meaning redundant, so they annotate it as ALTLEX (Prasad et. al., 2007). Example (34) includes an alternative lexicalization with its senses also annotated. The first sense category (CONTINGENCY) shows the top-level sense of the

relation, the second tag (Cause) indicates the sense class, the third sense (reason) indicates the subtype.

> 34) 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 (Prasad et. al., 2014).

According to Prasade et al. (2014), Example (34) has the sense of the relation "reason", and it is absurd to insert a connective which has a sense of reason such as "because."

The distributions of implicit and explicit relations have been examined in a number of languages. Table 1 below shows this distribution in PDTB (See Section 1.3 for a definition of implicit relation categories):

Table 1: Total Number of Relations Annotated in PDTB (Prasad et. al., 2014).

PDTB Relations	Number of tokens
Explicit	18.459
Implicit	16.224
ALTLEX	624
EntRel	5.210
NoRel	254
Total	40.600

Table 1 reveals that the ALTLEX, Entrel and NoRels outnumber the explicit discourse relations, and this result proves that implicit discourse relations should not be ignored in annotation efforts targeting discourse relations.

It also reveals that not only explicit discourse relations but also implicit discourse relations contribute to the coherence of discourse in a text. Finally, implicit discourse relations and their subtypes (e.g. ALTLEX) should be analyzed to understand how discourse coherence is constructed in a text. These findings from PDTB are the basic motivation for this thesis.

PDTB creates a starting point for other corpus studies. There are corpus studies for other languages based on PDTB such as: Chinese Discourse TreeBank (Xue

2005; Zhou and Xue 2012;), Turkish Discourse Bank (Zeyrek et al. 2009; Zeyrek et al. 2010; Demirşahin et al. 2013; Zeyrek et al. 2013), the Hindi Discourse Relation Bank (Oza et al. 2009; Kolachina et al. 2012; Sharma et al. 2013), the Prague Discourse TreeBank (Poláková et. al., 2013: Rysova, 2012, in Prasad et. al., 2014). These corpora will be analyzed in the next sections separately, focusing on how they deal with ALTLEXs in the respective language.

2.3.2 Chinese Discourse TreeBank

The aim of the Chinese Discourse Treebank (CDTB) is to enrich Penn Chinese Treebank, which includes the annotation of explicit and implicit discourse connectives (Xue, 2005). According to CDTB, as in the PDTB, "all discourse relations are lexically grounded and anchored by a discourse connective", and a discourse connective is defined as a predicate relating two arguments (Xue, 2005). Examples (35, 36) show how an explicit discourse relation occurs in CDTB;

(35) 现代/modern 父母/parent 难/difficult 为/to 的/DE 地方/area 是/be 既/not only 无法/no way 排除/eliminate 血液/blood 中/in 传统/traditional 的/DE 观念/values 又/but Also 要/need 面对/face 新/new 的/DE 价值/values.

"The difficulty of being modern parents not only lies in the fact they cannot get rid of the traditional values flowing in their blood, but they also need to face new values." (Xue, 2005)

(36) 如果/if 改革/reform 措施/measure 不/not 得力/effective(那么/then) 投资者/investor 就/then 有/have 可能/possibility 把/BA 注意力/attention 转向/turn to 新

兴/emerging 市场/market.

"If the reform measures are not effective, confidence crisis still exists, then investors is likely to turn their attention to other emerging market." (Xue, 2005)

Implicit discourse connectives are also annotated in CDTB, where the annotators are asked to insert an explicit discourse connective to find the type of implicit discourse relation. Example (37) shows an implicit discourse relation in CDTB;

(37)

[*arg1* 其中 出口 为 一百七十八点三亿美元
among them export be 17.83 billion dollar
，比 去年 同 期 下降
, compared with last year same period decrease
百分之一点三] [**conn**=而 ；] [*arg2* 进口
1.3 percent ; import
一百八十二点七亿美元 ，增长
18.27 billion dollar , increase
百分之三十四点一] 。
34.1 percent .

"Among them, export is 17.83 billion, and 1.3 percent increase over the same period last year. (Meanwhile) Import is 18.27 billion, which is a 34.1 percent increase." (Xue, 2005)

Because of language specific properties, CDTB follows some different guidelines in the annotation process; in PDTB, intra sentential discourse relations are annotated and to do this, punctuation marks are used. CDTB cannot utilize punctuation marks, as Chinese sentences do not end only with full stops. CDTB annotated both inter-sentential and intra-sentential implicit discourse connectives because of annotating the sentences both with a full stop and comma (Prasad et. el., 2014). This type of annotation affects the percentage distribution of explicit and implicit discourse connectives. In PDTB, explicit discourse connectives occur at a frequency of 46% and implicit discourse connectives occur at a frequency of 54%; however, in CDTB, there are 3.951 relations, and explicit discourse connectives make up 18% of the data; implicit discourse connectives constitute 82% of the data.

In CDTB, there is another difference in the annotation process for implicit discourse relations; in CDTB, insertion of an explicit discourse connective is not possible because of wording of Chinese. Instead of inserting an explicit discourse

connective, annotators are asked to paraphrase the relation, and paraphrasing shows the lexically grounded property of implicit discourse relations (Zhou and Xue, 2012).

To conclude, CDTB is a rich corpus in terms of implicit discourse relations, and it includes the annotation of both inter sentential and intra sentential implicit discourse relations. The results show that implicit discourse relations outnumber explicit discourse relations. These results reveal the role and percentage of implicit discourse relations in discourse once more, so further annotation of implicit relations is essential to have a full coverage of a corpus.

2.3.3 Hindi Discourse Relation Bank (HDRB)

HDRB is a corpus including the annotation of both implicit and explicit discourse relations in Hindi where the annotation process is affected by the linguistic features of Hindi (Prasad, Husain, Sharma and Joshi, 2008). Hindi has a rich morphology and free word order, and HDRB follows the guidelines of PDTB. The annotation process of HDRB is slightly different from PDTB because of some language specific reasons. The annotation process of HDRB includes the annotation of arguments and their spans, and senses as in PDTB, but a difference occurs while annotating explicit discourse connectives. Hindi does not have a comprehensive list of explicit connectives as English, so the first work for annotating explicit connectives is to discover explicit connectives in Hindi. An initial list for explicit connectives is given to annotators, and they improve this list while annotating (Kolachina et. al., 2012). Another difference of HDRB from PDTB is the annotation work-flow. In HDRB, annotators annotate all types of connectives (implicit, explicit, ALTLEX) simultaneously, and this type of work flow has some advantages and disadvantages; it is time saving as all discourse relations in one file are annotated for, but the inter annotator agreement is low (Kolachina et. al., 2012).

As for the types of explicit connectives, HDRB has divergence from PDTB as well; in addition to subordinating conjunctions, coordinating conjunctions and adverbials, HDRB has three other classes of explicit connectives: sentential relatives, subordinators and particles (Oza et. al., 2009). Sentential relatives are relative pronouns which relate the main clause and the relative clause; subordinators includes postpositions, verbal particles and suffixes which relates two arguments with non-finite clause; particles are used for showing the discourse relation between

two sentences which is marked by a particle (Oza et. al., 2009). Example (38) and (39) show explicit discourse connectives in HDRB:

(38) [दानवी लहरों के कारण अंडमान के पश्चिमी तट पर तटीय वनस्पती पूरी तरह बर्बाद हो गयी।] इसके अलावा {मूंगे की चट्टानों को भी नुकसान हुआ है।}

"Dropping all his work, he picked up the bird and ran towards the dispensary **so that** it could be given proper treatment."

(39) [सारा काम छोड़कर वह उस चिड़िया को उठाकर दवा घर की ओर भागा] जिससे {उसका सही इलाज किया जा सके।}

"The coastal vegetation on the west coast of the Andaman has been completely destroyed due to wild waves]. **In addition**, {the coral reefs have also been damaged." (Kolachina et. al., 2012)

Implicit connectives are annotated in HDRB. As in PDTB, annotators insert an appropriate explicit connective when they find out a relation between two arguments. They annotate implicit connectives only between adjacent sentences (Kolachina et. al., 2012). Example (40) shows an implicit connective in HDRB;

(40) {इस गेम के सारे खिलाड़ी सचिन तेन्दुल्कर से भी महान हैं।} Implicit=इसलिये [इनको क्लीन बोल्ड करना किसी के बस की बात नहीं।]

"All players in this game are greater than even Sachin Tendulkar. **Implicit=therefore** It is not possible for anyone to get them clean bowled." (Kolachina et. al., 2012).

ALTLEXs are annotated when there is no explicit connective. The insertion of an explicit connective makes the relation redundant. Example (41) shows an ALTLEX in HDRB;

(41) {बांग्लादेश में कानून-व्यवस्था की हालत में सुधार हुआ है।} AltLex [इसी वजह से भारत ने सम्मेलन में शामिल होने का फ़ैसला किया है।]

"Bangladesh's judiciary has seen an improvement. That is why India has decided to participate in the conference."

Table 2 shows the distribution of implicit and explicit relations in HDRB.

Table 2: Distribution of Discourse Relations in HDRB (Oza et. al., 2009)

Relations	HDRB Types	HDRB tokens
Explicit	49	189(31%)
Implicit	35	185(31%)
ALTLEX	25	37 (6%)
Entrel	NA	140(23%)
NoRel	NA	51 (9%)
Total	109	602(100%)

Table 2 shows the distribution of discourse connective types and tokens in HDRB, where the percentages of explicit and implicit connectives is the same as in PDTB. The other fact about HDRB is that Hindi is a morphologically rich language, even though they do not have a comprehensive explicit connective list to be annotated. The Hindi research group finds 49 explicit connective types. PDTB includes 100 different types of explicit connectives even though it has a comprehensive explicit connective list (Oza et. al., 2009). Another remarkable point especially for the current thesis is the proportion of ALTLEXs in HDRB; The ALTLEX percentage is 6% while the ALTLEX proportion in PDTB is 1%. These percentages show that ALTLEXs need further investigation in Turkish, too. This will provide us with a cross-linguistic comparison.

2.3.4 Prague Discourse TreeBank (PDiT)

Prague Discourse TreeBank (PDiT) is a corpus project which aims to annotate discourse relations in Czech language. This project provides a new layer for the

existing corpus of Czech language: Prague Dependency TreeBank, which includes the annotation of morphology, surface syntax and underlying syntax (Mladová et. al., 2009). PDiT aims to annotate discourse connectives not only because connectives are important for discourse coherence, but also because connectives are the most significant markers on the surface both for humans and machines. The treebank also includes the annotation of argument spans for connectives and senses, textual coreference, and bridging anaphora. It follows the basic guidelines of PDTB (Poláková et. al., 2013). For annotating discourse relations, textual coreference and bridging anaphora, the research group uses a highly customizable tree editor TrED, and annotations are done manually. At first, PDiT annotators annotate explicit discourse connectives using syntactic trees annotated earlier in Prague Dependency TreeBank. Inter-sentential relations are annotated, but intra-sentential relations are annotated only if their discourse semantics are different from the grammatical interpretation. PDiT includes an automatic procedure for extracting discourse structures with the help of syntactic trees (Poláková et. al., 2013).

PDiT also includes the annotation of ALTLEX and EntRel, but the research group argues that implicit connectives are problematic. They conduct an experimental annotation of 100 sentences for implicit connectives. They find that inter-annotator agreement is low; i.e., the annotators agree only in 49% on the type of implicit connectives (Poláková et. al., 2013).

Example (42) and (43) show explicit connectives in Czech language;

(42)Nevysílají české Události právě pro ty banality. **Protože** právě jejich znalost by na Slovensku mohla dělat neplechu.

“They do not broadcast Czech Události (Events) just for those banalities. **Because** it is precisely their knowledge that could bring about mischief in Slovakia.” (Poláková et. al., 2012: p.12).

(43) Naší oporou by mělo být i fantastické domácí publikum. **Vždyť** ATT máme kapacitu stadionu 5000 míst a dva týdny před ligou už jsme prodali 3000 permanentek.

“**Our support** should be also a fantastic home audience. Indeed ATT, we have the capacity of 5000 seats and we have already sold 3000 season tickets two weeks before the league.” (Poláková et. al., 2012: p.16).

Example (44) and (45) show two ALTLEXs in PDiT,

(44) Hráč brazilského týmu napadl v dnešním utkání svého protihráče. **To je důvod, proč** nebude hrát příští tři zápasy.

"The Brazilian football player attacked his opponent in today's match. **This is the reason why** he will not play in the next three matches."(Rysova, 2012).

(45) Gyula Horn se vyslovil pro možné zavedení majetkové daně. **Zdůvodnil** to tím, že utahování opasků se nemůže vztahovat pouze na lidi žijící ze mzdy.

"Gyla Horn agrees with the possible establishing of the property tax. He **gave the reason that** tightening of belts cannot be applied only to people living on wages." (Rysova, 2012)

In PDiT, there are 306 ALTLEX tokens in total. The research group also analyzes the ALTLEXs in a detailed way and come up with a typology on the basis of Czech (Rysova, 2012).

2.3.5 Turkish Discourse Bank (TDB)

Turkish discourse Bank is a corpus project which comprises Turkish texts from different genres. It is a subcorpus of METU Turkish Corpus (MTC), including ~ 400,000-words.

To reiterate, TDB takes into account discourse connectives as being related with two abstract objects from a semantic perspective (Zeyrek et. al., 2009). TDB 1.0 annotates only explicit discourse connectives. Explicit discourse connectives have three different syntactic categories; coordinating conjunctions, subordinators and discourse adverbials or anaphoric connectives. These connectives take two arguments as ARG1 and ARG2 (Zeyrek et. al., 2009). As in PDTB, TDB the uses minimality principle for annotations, which means annotating the spans of connectives and arguments of the connectives as minimally as it is sufficient for describing the discourse relation (Zeyrek et. al., 2013).

The annotation process consists of three processes; firstly, three annotators annotate the connectives and the arguments. The second step is to measure the agreement among the annotators by using an inter-annotator agreement tool. The third step is to resolve the disagreements. Connectives (CONN), modifiers (MOD), first and second argument of the connectives (ARG1, ARG2), shared material of the

connectives (SHARED), supplementary material (SUPP) are annotated (Demirşahin et. al., 2012). In TDB 1.0, senses are not annotated but this work is underway (Zeyrek et. al., 2015). Example (46) and (47) show explicit discourse relations in Turkish:,

(46) Kemal, bir yandan askeri bir savaş verirken **öte yandan** yerli işbirlikçilerle –ki bunların başında da basın- savaşmak zorunda kalmıştır.

"Kemal, while on the one hand fighting a military war, **on the other hand** (he) had to fight with local accomplices –which mainly included the media (Zeyrek et. al., 2010)."

(47) Akıntıya kapılıp umulmadık bir geceyi bölüştü benimle ve bu kadarla kalsın istedi belki. Eda açısından olayın yorumu bu kadar yalın olmalı. **Ama** eger böyleyse benim için yorumlanması olanaksız bir düşten başka kalan yok geriye şimdi.

"She was drifted with a current and shared an unexpected night with me and perhaps she wanted to keep it this much only. From the perspective of Eda, the interpretation of the incident should be that simple. **But**, if this is the case, now there is nothing left behind for me but a dream impossible to interpret (Aktaş et. al., 2010)."

As we have already indicated, Alternative lexicalizations have not yet been annotated in TDB.

2.4 Language Technology Applications

Language carries information with the help of morphology, syntax, phonology, semantics, and language technology (LT) uses these structures to extract information about language. For example, LT uses syllable structure, word structure, or syntactic trees to extract information about the language (Webber et. al., 2012). However, discourse structures and dialogue structures are less used structures in LT because there is no efficient study about discourse structures when it is compared to morphological and syntactic structure studies. Discourse studies lack enough amounts of electronic data for LT, and there are not enough annotated corpora for applying LT (Webber et. al., 2012).

Discourse studies also show that discourse has a structure, and this structure facilitates using Language technology applications (Webber et. al., 2012). Discourse

structures have different features such as topics, functions, eventualities and discourse relations, which are used to extract information from discourse. There are algorithms for discourse structure, discourse segmentation, discourse chunking and discourse parsing. Discourse parsing includes summarization, information extraction, essay analysis and scoring, sentiment analysis, and assessing the naturalness and coherence of automatically generated text (Webber et. al., 2012).

2.5 Previous Work in Turkish LT Applications

For Turkish, language technology studies mostly comprise syntactic, morphological, semantic and phonological studies, but discourse studies are rare when compared to these fields. However there are morphological analyses: For example, Oflazer and Kuruöz (1994) devise a method of automatic text tagging for Turkish,. This study includes a POS tagger for Turkish, which comprises a two level morphological specification of Turkish: an automatic text tagger and a morphological disambiguator. Automatic text tagging is done by annotating the lexical and part of speech features of words in the corpus; this tagging method eases the parsing procedure, which is essential for ambiguity resolution (Oflazer and Kuruöz, 1994). Another study for Turkish includes syntactic parsing and lexicalization in Turkish (Eryiğit, Nivre and Oflazer, 2008). The authors conclude that morphological information of a language increases parsing accuracy, that is, the more morphology is studied in a language the more the morphological parser is accurate. Aksan and Mersinli (2011) develop a morphological tagging module (Nooj) for Turkish. The Nooj module includes modeling and tagging processes of derivational and inflectional affixes of Turkish. These studies (among others) show that Turkish has syntactic parsers, and POS taggers, but the field lacks an analysis of discourse structures to be used in language technology and discourse parsers.

One important study for English (which is also relevant to the current thesis) includes automatic discourse connective detection in biomedical text (Ramesh et. al., 2012). For this study, the first step is to annotate discourse connectives; the second step is to develop supervised machine learning approaches for automatically identification of discourse connectives. This study reveals that discourse connectives can be extracted automatically by using supervised machine learning approaches (Ramesh et. al., 2012). This is a challenging work, because using simple lexical features based on a connective matching system does not give accurate results (Ramesh et. al., 2012). This study shows that annotation is essential for discourse

parsing, but it is not the only requirement for automatic extraction of discourse connectives.

By annotating ALTLEXs, the current thesis will help TDB to reach a higher level of discourse annotation coverage, which will ultimately yield an automatic detection of discourse connectives (which is a first step in discourse parsing) in Turkish.

So far, we have given a snapshot of what discourse relations involve and described the corpora annotated at the discourse level following the PDTB principles and briefly mentioned the necessity of annotated discourse corpora for LT applications. In the next chapter, we describe the methodology of the current thesis.

CHAPTER 3

METHODOLOGY

This chapter describes the methods used to reach the aims of the thesis. First, the annotation procedure for ALTLEXs in TDB is described. Then, how a typology of ALTLEXs in Turkish has emerged is explained. Finally, the method used for automatic extraction of possible ALTLEXs is presented.

3.1 Annotation Workflow for ALTLEXs

Our annotation procedure for ALTLEXs has three steps. The first step is to form a list of ALTLEXs in Turkish because there is not an available list providing instances of ALTLEXs in Turkish. The second step includes preparing an annotation guideline for ALTLEX annotation. The third step includes the inter-annotator agreement process. Each of these steps is explained below.

3.1.1 Identifying ALTLEXs in Turkish

Before starting the annotation of ALTLEXs, it is essential to prepare a well-defined list of explicit connectives in Turkish. This list is prepared by the author by using Göksel and Kerslake (2004), Lewis (1985), and the existing explicit connective list of TDB annotations. This list will be referred to as the Complete List of Explicit connectives (CLExp). We realize that this list may not be fully complete and may include certain items that others could call a connective; but since it covers more than one source it is called complete for them purposes of this thesis. Further research may reveal more explicit connectives and more ALTLEXs and the division line between them. The CLExp contains 118 types of explicit connectives in Turkish. Our guiding rule is that if the annotator finds an expression relating two adjacent sentences, and if this expression is not in the CLExp, then it may be annotated as an ALTLEX.

3.1.2 Annotation Steps for ALTLEXs

TDB includes different genres of texts, e.g. fiction, interviews, memoirs, news articles, etc. (Zeyrek et al, 2009). It consists of 197 files in these different genres. Annotation of ALTLEXs starts with the selection of files to be annotated. For the purposes of the current thesis the files are randomly selected making sure they reflected the distribution of all types of genres in TDB. This subcorpus of ALTLEXs constitutes 10% of the entire corpus (20 files, 10x2000 words= approximately 20000 words). The rest of the thesis is built on this subcorpus, which we call the ALTLEX subcorpus.

TDB implicit annotation guidelines are used during the annotation process of ALTLEXs in Turkish (Zeyrek et al. ms). TDB guidelines follow PDTB and have some restrictions on annotation, which we also followed:

- The relations which do not occur in adjacent sentences are not annotated.
- The relations which occur between sentences which are in different paragraphs are not annotated.
- Interjections are not annotated.

Therefore, we only searched ALTLEXs between adjacent sentences with separated with a comma, a colon and a semi colon.

Our annotation process of ALTLEXs is;

- The first step is the annotation of ALTLEX tokens in the selected files. The argument spans are not annotated.
- The second step is the annotation of 25% of the ALTLEX corpus by a secondary annotator.
- The third step is to calculate inter-annotator agreement (IAA) for reliability.

Annotation is done file by file and sentence by sentence by the primary annotator (the author of the current thesis). During the annotation process, an expert (Deniz Zeyrek Bozşahin) is consulted when the primary annotator is in doubt whether something is an ALTLEX or not. In this way, we formed a list of ALTLEXs in Turkish to the extent TDB ALTLEX subcorpus allows. This first list extracted from the ALTLEX subcorpus and comprises 180 tokens/ 34 types.

Then, at a second round, the ALTLEXs identified in the first round are searched and annotated in all the files in the corpus by the primary annotator. In this case, the primary annotator went through the whole corpus file by file annotating all the tokens of the 34 types in the first list. While annotating this list, the annotator found new ALTLEXs in the corpus which are additional to the first 34 types. As a result, these new ALTLEXs are added to the ALTLEX list and the type of ALTLEXs increased to 52 types. However, in the scope of this thesis, only 180 tokens/34 types are annotated in TDB, the remaining 18 types are merely listed and used in ALTLEX analysis in an attempt to form a typology.

In addition, sense annotation is done for ALTLEXs using the PDTB sense hierarchy (Prasad et. al., 2007). The PDTB sense hierarchy is provided in Figure 1 (Prasad et. al., 2007)

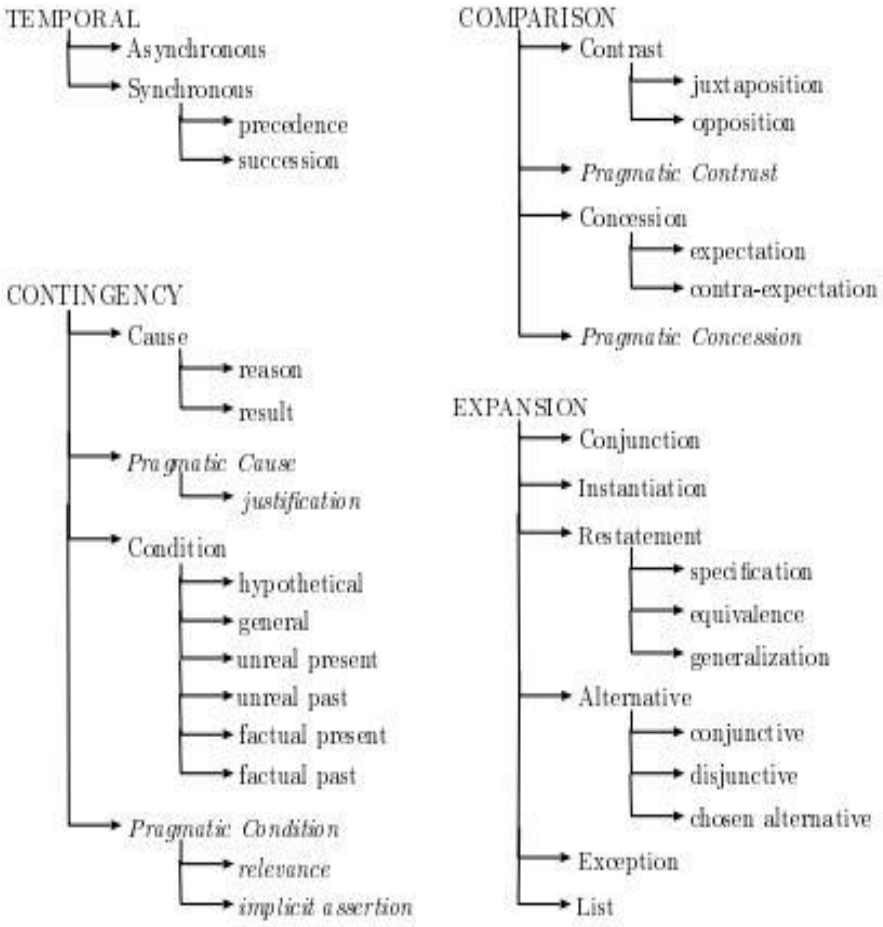

Figure 1: Hierarchy of sense tags

In the PDTB sense hierarchy, there are 4 top-level classes. Types and subtypes are also used, where the subtype is the deepest sense in the hierarchy. In our case, the deepest sense is annotated as well; to the extent it is possible. For example;

> (48) "Çağdaş danstaysa bu alanda biraz daha ilerdeyiz. **Öncelikle (ALTLEX= Expansion: Restatement: specification)** bunun bir sistem olarak kabul edilmesi, bu üç özelliğin uyum içinde çalıştırılması gerekiyor."
>
> "In this field, we are ahead in modern dance. First of all (ALTLEX= Expansion: Restatement: specification), it is necessary to accept it as a system, and to activate these three different features in harmony."

Class = Expansion
Type = Restatement
Subtype = Specification

The connective device öncelikle (first of all) is morphologically complex, derived from the root "önce": önce-lik-le 'first-derivative suffix-derivative suffix'. In other words, there is a degree of grammaticalization in its use as a connective device. Since it does not exist in the CLExp, we annotated is as an ALTLEX. It could have been categorized simply as a connective that has grammaticalized. We need further research to separate ALTLEXs from grammaticalized connectives in Turkish.

3.1.3. The Process of Inter-Annotator Agreement (IAA)

To ensure annotation reliability, a secondary annotator annotated 5 files (25% of the ALTLEX subcorpus) after a training procedure on one file. The secondary annotator is also a graduate of METU Cognitive Science Department working on corpus annotation.

IAA is measured between the primary annotator and the secondary annotator using the EXACT function in Excel (Yalta, 2008). We calculated IAA,

a) for ALTLEX (there is/there is not ALTLEX)
b) for senses at three levels following the PDTB sense hierarchy.

The EXACT function compares two annotations. If both annotators entered the same value, then the result is encoded as TRUE, and if the values are different, then the result is encoded as FALSE. After labeling the results as TRUE and FALSE, these Boolean values are converted to 0 (zero) and 1 (one). If the result is TRUE then it is converted to 1, and if the result is FALSE then it is converted to 0. Finally, the percentage is calculated as the IAA result. Higher agreement results will show the accuracy of ALTLEX annotation. In cases of disagreements, the primary annotator and the secondary annotator discussed their annotations and resolved the disagreements.

3.2 Typology of ALTLEXs

According to Prasad, Joshi and Webber (2010), ALTLEXs have three different categorizations: syntactically admitted, lexically frozen; syntactically free, lexically frozen; syntactically and lexically free. This categorization includes both lexical and syntactic analysis of ALTLEXs. However in a set of slides by Aravind Joshi (2010), there is a suggestion about ALTLEX categorization as being of three types: closed class ALTLEXs, partially open class ALTLEXs and open ended ALTLEXs. This categorization is also based on lexico-syntactic analysis but it is simpler; for these reasons it is used in this thesis. The one by Prasad et al (2007) requires a deeper syntactic analysis, and TDB does not contain syntactic annotations. Table 3 describes the classification of ALTLEXs in PDTB according to the criteria suggested by Joshi (2010).

Table 3: ALTLEX Classification in PDTB (Joshi, 2010)

Closed Class ALTLEXs	Partially open ALTLEXs	Open ended ALTLEXs
After that, after this, That's why, That is why; This is why, This means, That means, Beyond that, That was followed by, etc.	Probably the most egregious example is, Trouble (with that) is, The idea (behind that) is, The problem (regarding that) is, The reason (for that) is, The result (of that) is, etc.	That compares with, After these payments, That would follow, The increase was mainly due to, Once triggered, etc.

Table 3 is used as a reference for the classification of ALTLEXs in Turkish, and a similar table is formed in Table 9 (see Section 4.4). Our findings indicate that the number of closed class ALTLEXs and partially open ALTLEXs outnumber open ended ALTLEXs. This analysis suggests that open ended ALTLEXs require further analysis, and we did this by an automatic extraction method for spotting more possible ALTLEXs in Turkish. This is explained in the next section.

3.3 Automatic Extraction of Possible ALTLEXs from TDB

As it is stated in the previous section, the list of ALTLEXs in Turkish includes few open ended ALTLEXs. To have a fuller list of possible open ended ALTLEXs in Turkish, an automatic extraction of ALTLEXs is done.

This automatic extraction is done by using punctuation as a restriction. Firstly, in the entire TDB, the first three words after a comma, a colon and a semi colon are extracted. Then, a list of key words is formed by using the list of ALTLEXs which we obtained after our corpus annotation effort (see Appendix B for the list of key words). These key words are searched in the first three words of all TDB. Only three words are searched because the open ended list of the first round of annotations includes ALTLEXs with maximum three words.

In the automatic extraction part, the first step is to decide which tool is to be used, and then how to extract automatically possible open ended ALTLEXs. The Java programming language is used in the Eclipse environment. Java includes classes, and this code includes one class including a main part and one function (Arnold et. al., 1992). In the main part, file chooser is used to choose files from the corpus. Then input box is used for word search in the corpus files. Each sentence in the text files is separated by the split function. By using another loop, the sentences are divided into words. After that, for loop is used to obtain the first three words of the sentences. Finally a word search function is developed, and the key words are searched in this word search function. As a result of this search, we extracted a list of possible open ended ALTLEXs.

To summarize, in this chapter, we have presented the methodology we used in our corpus-based approach to ALTLEXs. In the next chapter, the results of IAA are presented. We also provide the analysis of ALTEX types into Joshi's (2010) categories. The results are also discussed in detail.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter introduces the results of the IAA for ALTLEXs in Turkish, and the IAA for three classes of senses. It also includes a list of ALTLEXs in Turkish, with different types of ALTLEXs.

4.1 IAA Results

Table 4 presents the IAA results of ALTLEX annotation, as well as sense annotation in terms of class, type and subtype as discussed in Chapter 3.2.2.

Table 4: IAA Results for ALTLEX and Sense Annotation

ALTLEX/not ALTLEX	0.73
CLASS of Sense	0.65
TYPE of Sense	0.56
SUBTYPE of Sense	0.26

The IAA statistics is done by means of the Exact measure as already described in Section 3.1.3. For measuring reliability of the ALTLEX/not ALTLEX difference, our annotation constitutes a selection of textual spans for ALTLEXs, so we calculated agreement for each ALTLEX as a percentage of the spans that exactly or partially match. For any ALTLEX token, agreement is labeled as 1 when two annotators annotate the same span as ALTLEX, and agreement is labeled as 0 when two annotators annotate non-identical selection for the ALTLEX token (Miltsakaki et. al., 2004). According to Table 4, two annotators agree on whether something is an AltLex or not in 73% of the time. The agreement is not as high as .80 as it is usually expected. But Spooren & Degand (2010: p.256) argue that in discourse annotation,

0.70 is an acceptable measure as follows: “The reasons for nevertheless choosing a standard of .70 are twofold: On the one hand the the task of coding coherence relations is fundamentally determined by its reliance on contextual interpretation. Therefore, we believe a standard of .80 is unrealistic because it is too high.” (Spooren and Degand, 2010: p.256)

By their measure, our results can be accepted well. However, In terms of class of sense, we are slightly lower than the accepted .70. The reason can be that the under-determinacy of language due to different mental interpretations of ALTLEXs gives rise to disagreements in IAA (Spooren and Degant, 2010). In terms of the class of sense and subtype of sense, our results are even lower because to decide the deeper sense tags are difficult to interpret, and it appears that our annotators need a more transparent hierarchy for sense tags. In PDTB, disagreements are resolved in a way that if deeper senses are disagreed, then the higher agreed sense is annotated in the resolution of disagreements (Prasad et. al., 2008). This way is efficient to solve disagreements, but it also shows that deeper senses in the hierarchy may be difficult to annotate even for PDTB annotators. In future studies, whether the PDTB hierarchy involving top level, class and subtype levels fully appeal to the annotators’ intuitions should be examined cross-linguistically.

4.2 The Discussion of IAA Results

Table 5 presents the IAA results of ALTLEX annotation, as well as sense annotation in terms of class, type and subtype as discussed in Chapter 3.2.2.

Table 5: Exact Match Results for ALTLEX Tokens

# of Disagreed tokens	12
# Agreed tokens	34
TOTAL # of ALTLEX Tokens	46

Table 5 shows the exact match results for the ALTLEX tokens, where we achieved 73.9% (46/34*100) agreement between the primary and secondary annotator on 46 ALTLEX tokens. The result shows that there is a good match between two annotators according to Sporeen & Degan (2010).

Disagreements are resolved after IAA results are obtained. There are 12 tokens of ALTLEXs which are disagreed between two annotators. The breakdown of these disagreements is shown in Table 6.

Table 6: The Disagreement Types for ALTLEX Tokens

Missing Annotations	5
Partial Overlap	7
Total	12

The disagreements shown in the Table 6 are resolved after a discussion session between two annotators. All of the annotations in this list are eventually decided to be ALTLEXs except for the token “ne de olsa” (after all), as it does not relate two different abstract objects in the example where we identified it, i.e., “Ne de olsa adressiz bir meslek (“after all an unaddressed job”)”. In Table 6, missing annotations stands for the disagreements where one ALTLEX is annotated, but the second occurrence of the same ALTLEX cannot be identified. For example, “bir diğer deyişle (in other words)” is annotated in a file as ALTLEX, but the second occurrence of “bir diğer deyişle (in other words)” in another file is not annotated because one of the annotators does not notice it. These cases are labeled as missing annotations. Partial overlap stands for annotations in which two annotators annotated some parts of one ALTLEX. For example, one annotator annotates “Ante’nin ricası üzerine” as ALTLEX, and the other annotates only a part of this expression, e.g. “üzerine” as ALTLEX.

Table 4 above provides the IAA results for the senses of ALTLEXs in three levels as class, type and subtype (Prasad et. al., 2007). As already described, the primary annotator annotated 20 random TDB files (which we have been calling the TDB ALTLEX Subcorpus) for senses at three levels but only 25% of this subcorpus (approximately 50000 words) was annotated by the secondary annotator to calculate the IAA. We calculated agreement for the senses at each level using the exact match criterion. For any ALTLEX sense, agreement is labeled as 1 when two annotators annotate the sense as exactly the same; agreement is labeled as 0 when two annotators annotate non-identical senses for an ALTLEX token. According to Table 4 above, two annotators achieve 65% agreement in the top class level of senses, and the agreement decreases as one goes deeper in the PDTB sense hierarchy. As we

have already mentioned, a decrease in IAA as one goes into deeper senses occur in PDTB sense annotation, too (Prasad et. al., 2008: 5). For Turkish, the annotation of top-level senses would need more work to increase the IAA to at least .70. We think that this increase will increase as more annotations are performed.

After two annotators annotate the senses, all disagreements are resolved by a discussion among the annotators. The distribution of the class of the senses as well as the associated IAA results is given in Table 7.

Table 7: The Distribution of ALTLEX senses and IAA Results (in parenthesis)

	# of tokens (IAA Results)
Contingency	9 (31%)
Expansion	5 (17%)
Temporal	15 (52%)
Total	29 tokens

Due to the small size of the double annotations regarding sense, in Table 7, we do not have the fourth top-level sense, i.e. COMPARISON. However, the primary annotator has identified this top-level sense in the ALTLEX subcorpus. Examples 48 and 49 show ALTLEXs with the COMPARISON sense:

(48) Bu gençliğe güvensizlik değil. **Sadece (Comparison: Contrast: Correction)** ikinci meclisin olmamasından kaynaklanıyor.

"This is not lack of confidence in youth. **Only**, it is because of the lack of the second assembly."

(49) Ancak yangın söndükten bir süre sonra tüm çözüm önerileri unutuldu **Sadece (Comparison: Pragmatic Contrast)** bölge halkı her zaman tedirgin yaşadı.

"All solution suggestions are merely forgotten after the fire went out**. Only** the community always lives doubtfully."

The connective device "sadece" (only) does not exist in the CLExp which we have used as a guide. For this reason, it has been annotated as an ALTLEX. This word is essentially an adverb but in discourse, it has come to be used as a connective

device; i.e. it has grammaticalized. Further research will establish firmly whether “sadece” is an ALTLEX or not.

4.3 List of ALTLEXs in Turkish

Table 8 below shows the list of identified ALTLEXs for Turkish. To reiterate, the ALTLEXs were annotated at two rounds. In the first round, ALTLEXs were annotated in randomly selected 20 files from TDB. The primary annotator went through all the files sentence by sentence and identified the ALTLEXs. In the second round, the entire TDB was annotated by using the list from the first round as search items. In this round, more ALTLEXs were identified. All the identified ALTLEX types and tokens are provided in Table 8. We emphasize that some of these connecting devices may be half way between an ALTLEX and a discourse connective; i.e. expressions that are in the process of grammaticalization. Those connecting devices are marked with a star, showing that their status should be further examined to understand whether they are ALTLEXs, explicit connectives, or any other kind of discourse connecting device.

Table 8: The List of Turkish ALTLEXs in TDB

First Round	Types	Tokens
	Başka bir açıdan	Başka bir açıdan
	Adeta*	Adeta
	Aynı zamanda	Aynı zamanda
	Bir de	Bir de
	Bir kere*	Bir kere
	Böyle bir gerekçeyle	Böyle bir gerekçeyle
	Böylelikle *	Böylelikle
	Bu açıdan	Bu açıdan
	Bu bahaneyle	Bu bahaneyle
	Bu bakımdan	Bu bakımdan
	Bu itibarla	Bu itibarla
	Büyük ihtimalle	Büyük ihtimalle
	Derken*	Derken
	Diyelim*	Diyelim
	Düşünün *	Düşünün
	En azından	En azından
	Esasen *	Esasen
	Hiç olmazsa	Hiç olmazsa
	Kaldı ki *	Kaldı ki
	Ne de olsa *	Ne de olsa
	Olsa olsa *	Olsa olsa
	Böyle olunca *	Böyle olunca

	Öncelikle * Öte yandan * Öyle ki * Öylece * Özellikle * Bundan böyle Bundan ötürü Bunun üzerine Hani*	Öncelikle Öte yandan Öyle ki Öylece Özellikle Bundan böyle Bundan ötürü Bunun üzerine Hani
	… yanında	Bu benzerliklerin yanında, Bu büyük fotoğrafların yanında, bunun yanında
	Bir diğer deyişle	Bir diğer deyişle, Bir başka ifadeyle
Total	33	38

4.4 The Classification of ALTLEXs in Turkish

Table 9 below classifies Turkish ALTLEXs into three groups as Joshi (2010) suggests in a set of slides (for English). Again, the connecting devices with a star indicate that their status as an ALTLEX is not certain and needs further research.

- The closed classes ALTLEXs are frozen utterances and they do not undergo any changes in the corpus such as “ne de olsa (after all) öncelikle (first of all), özellikle (especially)”.
- The partially open ALTLEXs are not frozen but they are partially open to lexical changes such as bu bahaneyle (with this excuse), bunun yanında (besides this) ”

  These lexical changes occur by taking different lexical items in the phrase that we have called an ALTLEX. Partially open ALTLEXs mostly consist of deictic items.

- The open ended ALTLEXs occur freely but they also have a fixed part; the other parts freely modify this fixed part. For example “bu benzerliklerin yanında (in addition to these similarities)” is an open ended ALTLEX which can be freely modified as “bu amaçların yanında (in addition to these aims) bu farklılıkların yanında (in addition to these differences), where the word “yanında” remains fixed.

Table 9: The Classification of Turkish ALTLEXs

	Closed Class ALTLEX Tokens	Partially Open Class ALTLEX Tokens	Open-ended ALTLEX Tokens
	hiç olmazsa (1) Öylece* (1) bir kere* (2) Diyelim* (2) Sonunda (2) olsa olsa* (2) Hani* (3) Esasen*(3) en azından(3) Adeta* (4) kaldı ki *(4) Derken *(5) bu bakımdan (6) ne de olsa* (8) Böylelikle (9) aynı zamanda (10) Öncelikle* (10) bir de (11) Özellikle* (11) Sadece *(11) Bu açıdan (17)	(büyük) ihtimalle (1) bu bahaneyle(1) bununla birlikte (2) bunun yanında (2) düşünün*(2) bu itibarla(2) bundan böyle bundan ötürü(2) Başka bir açıdan (2) öyle ki* (5) öte yandan (10) bunun üzerine (16)	Bu büyük fotoğrafların yanında (1) bu benzerliklerin yanında (1) bir başka ifadeyle(1) böyle bir gerekçeyle (1) bir diğer deyişle (6)
Total	21 Types/ 125 Tokens	12 Types/45 Tokens	5 Types/10 Tokens

The classification in Table 9 reveals that the open ended class needs further analysis because there are too few of them; i.e., only 13% of the list contains open-ended ALTLEXS. This result suggests that the open-ended ALTLEXs need to be identified by a different means. As already explained in Section 3.3, we attempted to automatically extract more examples of possible open-ended ALTLEXs from TDB using the ALTLEXs in the 13% of the list given in Table 11.

The above classification of ALTLEXs shows that the ALTLEX subcorpus version 1 does not have many types that fit under the open ended class. We identified 21 in the closed class, 12 tokens in the partially open class and only 5 tokens in the open ended class (Table 9). The raw numbers are presented together with their frequencies in Table 10;

Table 10: The Frequencies of the ALTLEX Tokens and types in ALTLEX subcorpus (version 1)

Closed Class	6.22 % (125 Tokens /21 Types)
Partially Open	4.69% (45 Tokens/ 12 Types)
Open-ended	2 % (10 Tokens/5 types)
Total	12.91% (180 Tokens/38 Types)

Table 10 reveals that there are 38 types/ 180 token ALTLEXs annotated in TDB in the annotation process. This table also reveals that the more an ALTLEX is closed, the more frequent it is. The frequency of the open ended ALTLEXs in our corpus is very low. To obtain more possible tokens of the open ended ALTLEXs an automatic method is used. As a result of this method 89 new open-class ALTLEX tokens were obtained. In this way, we formed a second version of the ALTLEX subcorpus.

4.5 Open- Ended ALTLEXs in Turkish

Table 11 includes samples of possible open ended ALTLEXs in TDB; this list is formed by using an automatic extraction method in Java programming by using the comma, semi-colon and colon punctuation marks between two sentences (see Section 3.3).

Table 11: Token samples from the Open-Ended ALTLEXs in Turkish (obtained after eliminating certain expressions manually)

Ana amacı	Bu gelişmenin ardından
Balkan savaşlarından sonra	Bu inanışa göre
Binlerce yıl öncesinde	Suikasttan hemen sonra
Bir an önce	Şişmanlığın birçok nedeni
Bir süre sonra	Şişmanlığın tek nedeni
Bir zaman önce	Uzun tartışmalar sonucu
Birinci neden	Üzüntüsünün başlıca nedeni
Birkaç zamandan beri	Verilen izinler nedeniyle
Böyle bir durumda	Yakın zamana kadar
Böyle hareketler karşısında	Yapılan incelemeler sonucunda
Bu arayışlar sırasında	Yapılan oylama sonucu
Bu cinayet nedeniyle	Yaşanan koşuşturmanın ardından
Bu çağda	Yazarın amacı
Bu doğrultuda	Yenilginin birkaç nedeni
Bu dönemde	
Bu durumda	

This list is not a full list of possible open-ended ALTLEXs we have extracted; the initial and longer list currently consists of 89 tokens/ 42 types (Appendix A). This list is obtained after eliminating certain expressions. This is because the Java program we used also extracts expressions which are not ALTLEXs because the code lacks sound linguistic parameters. This is why we did a manual elimination of the initial list with the help of an experienced annotator and an expert. For example we eliminated clauses such as “beni tanıdıktan sonra” (after she met me) because such clauses are already annotated in the TDB as discourse relations anchored by the subordinator “sonra” (after). We also eliminated expressions with connective modifiers e.g. “üç gün sonra” (three days later) where “üç gün” (three days) is annotated as a modifier of the connective “sonra”. Note that we have kept expressions such as “yıllar sonra” in the ALTLEX list due to the partially open nature of this expression, e.g. one may find “günler/aylar/haftalar sonra” (after many days/months/weeks) in Turkish. However, this is a possible point of discussion; this expression may be a modified connective, or an alternative lexicalization.

With the addition of this new set of open-class ALTLEXs, the ALTLEX subcorpus version2 emerged, with 94 types and 297 tokens.

Our method yielded many multi-word expressions which do not have any discourse connecting function, which we eliminated manually. Note that, such expressions are also named as n-grams, or lexical bundles. However, in this thesis, our aim is not to reveal any multi-word expression/lexical bundles/n-gram in Turkish, but rather reveal those that function as ALTLEXs in discourse. That is why we called them open ended ALTLEXS. In addition, it is an empirical question whether ALTLEXs are always multiword expressions/lexical bundles or n-grams; this is a further reason why we have opted to use a more neutral term, i.e. open ended expressions.

4.6 Towards a Projection of ALTLEX Types and Tokens in TDB

TDB 1.0 annotates only explicit connectives and some alternative expressions, namely those derived from a subordinator connective involving a deictic element referred to as "phrasal expressions" (e.g. buna rağmen (despite this), bunun için (for this) etc.). In a subsequent annotation effort, implicit relations and entity relations have been annotated in 19 files (10 %) of TDB (Zeyrek et al. 2015). The distribution of these relations (from Zeyrek et al. 2015 study) is shown in Table 12. Table 12 also shows the frequency of ALTLEXs, which we have annotated in this thesis.

Table 12: The Distribution of Explicit and Implicit Relations in 10% of TDB

Zeyrek et al. (2015)			The current study	TOTAL
Explicit	Entrel	Implicit	ALTLEX	
739 (42.4%)	566 (32.5%)	372 (21.5%)	72 (4%)	1749 (100%)
Total Explicit: 42.4%	Total Implicit with ALTLEXs: 57.6%			

The distribution of explicit and implicit relations with the ALTLEX analysis reveals that implicit relations outnumber the explicit relations so implicit relations need further analysis. This work helps us to come up with a general distribution of Turkish discourse relations. By the help of this analysis, a cross linguistic analysis of discourse relations for different languages is done, and Table 13 reveals this cross linguistic analysis. Regarding the explicit-implicit distribution, Turkish presents a different picture than the other languages with more implicits (excluding the ALTLEXs) than explicits. This may be due to the fact that TDB annotates intra-sentential explicit connectives but not intra-sentential implicit relations. Regarding ALTLEXs, Turkish is quite close to the other languages analyzed, with only 4% ALTLEXs.

Table 13: A Comparison of the Distribution of Discourse Relations across Languages

Relations	PDTB Tokens	HDRB Tokens	TDB Tokens
Explicit	18,459 (45%)	189(31%)	739 (42%)
Implicit	16,224(40%)	185(31%)	372 (21.5%)
ALTLEX	624(1.4%)	37(6%)	72(4%)
EntRel	5,210(13%)	140(23%)	566(32.5%)
NoRel	254(0.6%)	51(9%)	NA
Total	40,600(100%)	602(100%)	1749(100%)

In the course of our annotation process we have come to realize that ALTLEXs may also occur in different forms such as “bir başka ifadeyle; başka bir deyişle” (in another expression; in other words). We suggest that they are all derivates of “bir diğer deyişle” (in other words). In this thesis “bir diğer deyişle” (in another

expression) is called a type; the derivatives are referred to as “token”. Example 50 is an instance of “bir diğer deyişle” (in another expression/in other words).

> (50) Katılımın gerçek ve sahte biçimleri bulunmaktadır. Bir diğer deyişle ciddi etkili ve bağımsız" halk katılımı ile "yönlendirilmiş tümüyle hayali veya sembolik" sahte katılım birbirinden ayrilmaktadır
>
> “There are real and fake participations. In other words; “serious affective and independent” public participation diverges from “imaginary or symbolic” fake participation.

Table 13 shows us that the distribution of discourse relations for each language is different, but the aim of this cross-linguistic analysis is to reach universal results about discourse relations. In the future, this study should be done with larger corpora because the size of corpus in this analysis is not large enough.

To summarize, in this chapter, we have given the results of our reliability analysis, and presented an ALTLEX list for Turkish and the classification of ALTLEXs into three. We have also presented the list of open ended ALTLEXs and described how we created this list. In the next chapter, we summarize our work and draw some conclusions.

CHAPTER 5

CONCLUSION

In this thesis the first aim was to identify the typology of ALTLEXs. This study reveals it in three ways:

a) By annotating ALTLEXs
b) by classifying ALTLEXs into three classes (closed, partially open, open-ended)
c) by automatically extracting a possible list of open-ended ALTLEXs.

This chapter provides a summary of these steps and concludes the thesis.

5.1 The Annotation Procedure of ALTLEXs in Turkish

For the purposes of this thesis an ALTLEX subcorpus is formed by selecting 20 files from TDB. ALTLEXs are annotated by the primary annotator in these 20 files (10% of TDB). A list of ALTLEXs (34 types/108 tokens) is formed as a result of this annotation procedure. The types of this list is searched in all TDB and annotated in all TDB files. While annotating these search items, a new set of new ALTLEXs (18 types/ 28 tokens) are found in TDB and they are added to the list of ALTLEXs. As a result, ALTLEX list of Version 1 contains 52 types (208 tokens) of ALTLEXs identified in TDB. Table 14 provides the overall picture resulting from our annotation effort in the creation of the ALTLEX subcorpus (version 2):

Table 14: The Overall Numbers of ALTLEXs in TDB

First Round of Annotations	34 Types/180 Tokens
Second Round of Annotations	18 Types/ 28 Tokens
Automatically Extracted Open-ended ALTLEXs	42 Types/ 89 Tokens
Total	94 Types/ 297 Tokens

5.2 A typology of ALTLEXs

Regarding the linguistic classification of ALTLEXs, Joshi's (2010) three-way categorization was used due to its simplicity. His classification includes a closed class, a partially open class, and an open ended class. We classified the ALTLEXs we have identified during the annotation procedure with respect to this three-way categorization and revealed the distribution of the three classes (See Table 9 in Chapter 4.4).

We saw that the frequencies of the subclass of ALTLEXs differ. Most importantly the frequency of the open ended ALTLEXs are higher than the closed class of ALTLEXs.

To conclude the results of this study are a first step to contribute to the enrichment of TDB with implicit discourse relations. The ALTLEX tokens identified in this current study can be used in future linguistic investigations as well as in future automatic systems.

5.3 The Automatic Extraction of Possible Open Ended ALTLEXs

An automatic extraction method is used to identify possible open ended ALTLEXs. The Java Program is used to extract for this purpose. The Comma, colon and semi colon are used as descriptors. Three words following these punctuation marks are extracted using key words, which are formed according to the ALTLEX list in Table 8. As a result, 130 tokens of possible open ended ALTLEXs were obtained. However, not all these expressions were ALTLEXs. Non-ALTLEXs, which were multi-word expressions with no discourse connecting function, were manually eliminated by the primary annotator, and then checked by an experienced annotator and an expert. After discussion sessions, 89 tokens/42 types were identified as open ended ALTLEXs.

We suggest that although this technique is not perfect and not linguistically informed, it is useful as a first step. It is not a fully sufficient because it only uses punctuation marks between two sentences as in the manual annotation of the corpus. The results show that an efficient ALTLEX extraction needs a detailed lexico-syntactic analysis and possibly a MWE extraction method.

5.4 Lessons Learnt from ALTLEXs in Turkish and Questions that Arise

There are 94 types/297 tokens of ALTLEXs identified as a result of the current thesis. These overall ALTLEXs help us to deduce some facts about Turkish ALTLEXs. First of all, unlike PDTB, while annotating ALTLEXs, we did not include clausal tpes of expressions such as “beni tanıdıktan sonra” (after she met me). This is because TDB systematically annotates such expressions as the ARG2 of the subordinator connective, “sonra” (after). It would be redundant and counter-intuitive to annotate such clauses as ALTLEXs.

Secondly, we find that a substantial portion of the ALTLEXs are those that contain deictic expressions, such as bu (this), şu (that), o (that). A total of 37 tokens of “phrasal expressions” (as the TDB calls them) are annotated in the ALTLEX subcorpus. We think that this number may increase with more annotations.

Thirdly, our investigations have pointed out that ALTLEXs differ in their lexical/syntactic forms, and we chose to classify Turkish ALTLEXs following Joshi (2010). Our investigations have pointed out that the open ended ALTLEXs, i.e. those that are more flexible in lexical/syntactic changes, are quite frequent: we detected 42 types/89 tokens only by searching them between adjacent sentences delimited by a comma, semicolon and colon. We limited our search to expressions of three words. But this raises the question of when do open-ended ALTLEXs occur? This question is also asked by Joshi (2010). From the Turkish side, we can give the example of “çarpıcı örnek olarak” (as a striking example), which is attested in TDB. Here, the ALTLEX (örnek olarak) is modified by the adverb “çarpıcı” (striking). That is, a possible closed class ALTLEX (örnek olarak) becomes modified and can very well be int[1]erpreted as a “modified ALTLEX”.[1]

As for their senses, we have found that the ALTLEXs may carry all the top-level 4 sense classes of the PDTB hierarchy, just like explicit connectives do. This is interesting because then the question arises, what guides language users in their choice of an explicit connective or an ALTLEX. Further research is needed to find possible reasons. Joshi (2010) reports that the subsense cause-reason is not attested as an adverbial ALTLEX English. He gives the following example from PDTB where cause-reason is only expressed by “because” (in its discourse adverb function):

“Why was containment so successful? Because it has bipartisan support.”

[1]I owe these ideas to my advisor, Deniz ZEYREK.

He then asks, “Is this specific to English or a linguistic universal? Hindi, Czech, Turkish, Italian, Arabic?”. In the current thesis, only a small portion of the TDB, i.e. ALTLEX subcorpus was annotated for senses. In that portion, we did not find any adverb ALTLEX with the cause-reason sense. Further annotation will answer Joshi’s question (2010) regarding whether cause-reason is universally impossible or not.

An important point that we came across during the course of this thesis is that it is difficult to determine the division line between an explicit connective and an ALTLEX, particularly when the connecting device has just one word or two words. In some cases, the connecting device appeared somewhere between an ALTLEX and an explicit connective, possibly in its way to grammaticalization as an explicit connective. On the other hand, we were able to determine the open ended ALTLEXs (i.e. those with three words) with more certainty. For future studies, it may be safer to limit ALTLEXs to multi-word expressions that have a discourse connecting function.

5.5 Limitations of the Study and Further Research

Our study is limited with the number of files we have annotated for the purposes of this thesis (i.e. 20 files, ~20000 words). To obtain an expanded list of ALTLEXs, the whole of TDB or other corpora need to be annotated.

Our study has not accomplished a complete sense annotation of the ALTLEXs. Only the primary annotator annotated the ALTLEXs (in 20 files) for their sense in three levels according to the PDTB sense hierarchy. In the future, a secondary annotator should annotate the whole ALTLEX subcorpus for reliability and validity.

Another limitation has to do with the automatic extraction method used to determine open-ended ALTLEXs. The words in the expressions that we extracted are limited to three, that is, possible ALTLEXs up to three words are extracted in this study. There may be other open- ended ALTLEXs which comprise four or more words and this requires further work.

In this thesis, we only looked for ALTLEXs in written language and between adjacent sentences. Future studies should also look for ALTLEXs across sentences in both written and spoken language.

Lastly, as we have pointed out in various places in the thesis, our starting point in determining ALTLEXs is based on previous work on Turkish (Göksel &

Kerslake, 2004 and Lewis, 1985). Expressions not listed as an explicit connective in these works were taken to be possible ALTLEXs. This is a limitation on its own. We have not carried out a detailed analysis on whether the spotted ALTLEXs are really ALTLEXs or whether they are grammaticalized connectives but such words/expressions were shown where relevant. For future studies, it may be safer to limit ALTLEXs to multi-word expressions that have a discourse connecting function.

Despite these limitations, this thesis finds a list of open ended expressions which appear to be ALTLEXs. (Appendix A).

What we have not done in this thesis is that we have not discussed why a language user prefers an ALTLEX instead of an explicit connective that carries the same sense. This question is also asked by Prasad et al. (2010) for English. Therefore, future studies may concentrate on this fact and investigate the answers in psycholinguistic studies.

REFERENCES

Aktaş, B., Bozşahin, C., and Zeyrek, D. (2010). Discourse Relation Configurations in Turkish and an Annotation Environment. In Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV).

Aksan, M., & Mersinli, Ü. (2011). A corpus based NooJ module for Turkish. InProceedings of the Nooj 2010 International Conference and Workshop. Komotini (pp. 29-39).

Arnold, K., Gosling, J., Holmes, D., & Holmes, D. (1996). The Java programming language (Vol. 2). Reading: Addison-wesley.

Das, D., & Taboada, M. (2013). Explicit and implicit coherence relations: A corpus study. In Proceedings of the 2013 annual conference of the Canadian Linguistic Association.

Demirsahin, I., A. Ozturel, C. Bozsahin, and D. Zeyrek. (2013). Applicative structures and immediate discourse in the Turkish Discourse Bank. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pages 122–130, Sofia.

Demirşahin, I., Sevdik-Çallı, A., Balaban, H. Ö., Çakıcı, R., & Zeyrek, D. (2012). Turkish Discourse Bank: Ongoing Developments. In First Workshop on Language Resources and Technologies for Turkic Languages (p. 15).

Eryiğit, G., Nivre, J., & Oflazer, K. (2008). Dependency parsing of Turkish.Computational Linguistics, 34(3), 357-389.

Fraser, B. (1999). What are discourse markers?. Journal of pragmatics, 31(7), 931-952.

Forbes-Riley, K., B. Webber, and A. Joshi. 2006. Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG. Journal of Semantics, 23:55–106.

Göksel, A., & Kerslake, C. (2004). Turkish: A comprehensive grammar. Routledge.

Gönül, G. (2013). From lexical and conjunctive cohesion to coherence: reading, recalling and comprehending high cohesive and low cohesive clauses.

Halliday, M. A. K. (1985). An Introduction to Functional Grammar. London: Edward Arnold.

Halliday, M. A., & Hasan, R. (1976). Cohesion in English. London: Longman.

Halliday, M. A., & Matthiessen, C. M. (2013). Halliday's introduction to functional grammar. Routledge.

Irmer, M. (2009). Bridging Inferences in Discourse Interpretation (Doctoral dissertation, Dissertation: Universität Leipzig.(Moodle)).

Joshi, A. K. (2010). Discourse Relations: Perhaps, a New Kind of MWE's?. MWE Workshop on Multiword Expressions. Beijing, China.

Kerslake, C. (1996). The role of connectives in discourse construction in Turkish. Modern Studies in Turkish Linguistics: Proceedings of the 6th international onference on Turkish Linguistcs, 77-104.

Knott, A. 1996. A Data-Driven Methodology for Motivating a Set of Coherence Relations. Ph.D. thesis, University of Edinburgh.

Knott, A., & Sanders, T. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. Journal of pragmatics,30(2), 135-175.

Kolachina, S., R. Prasad, D. M. Sharma, and A. Joshi. (2012). Evaluation of discourse relation annotation in the Hindi Discourse Relation Bank. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, pages 823–828, Istanbul.

Korkmaz, Z., & Sözlüğü, G. T. (2005). Bağlaçlar ve Türkiye Türkçesindeki Oluşumları. Türk Dili, S, 638.

Lewis, G. L. (1985). Turkish grammar. Oxford University Press, USA.

McCarthy, Michael. Discourse analysis for language teachers. Cambridge University Press, 1991.

Miltsakaki, Eleni. (2004) "Annotating discourse connectives and their arguments." Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation.

Miltsakaki E, Prasad R, Joshi A, Webber B (2004) Annotating discourse connectives and their arguments. Paper presented at Proceedings of the NAACL/HLT Workshop: Frontiers in Corpus Annotation.

Mladová, L., Zikánová, Š., Bedrichová, Z., & Hajicová, E. (2009). Towards a discourse corpus of Czech. In Proceedings of the Corpus Linguistics Conference, Liverpool, UK

Oflazer, K., & Kuruöz, Ì. (1994). Tagging and morphological disambiguation of Turkish text. In Proceedings of the fourth conference on Applied natural language processing (pp. 144-149). Association for Computational Linguistics.

Oza, U., R. Prasad, S. Kolachina, S. Meena, D. M. Sharma, and A. Joshi. (2009). Experiments with annotating discourse relations in the Hindi Discourse Relation Bank. In Proceedings of the 7th International Conference on Natural Language Processing (ICON), pages 1–10, Hyderabad.

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). Association for Computational Linguistics.

Prasad, R., Husain, S., Sharma, D. M., & Joshi, A. K. (2008). Towards an Annotated Corpus of Discourse Relations in Hindi. In IJCNLP (pp. 73-80).

Özbek, Nurdan. (1995). Discourse Markers in Turkish and English: a comparative study.Unpublished Ph.D. Thesis. University of Nottingham, UK.

PDTB-Group. 2008. The Penn Discourse TreeBank 2.0 Annotation Manual. Technical report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., & Joshi, A. K. (2008). Easily identifiable discourse relations.

Poláková, Lucie, Jínová, Pavlína, Zikánová, Šárka, Bedřichová, Zuzana, Mírovský, Jiří, Rysová, Magdaléna, Zdeňková, Jana, Pavlíková, Veronika, Hajičová, Eva. 2012. Manual for Annotation of Discourse Relations in the Prague Dependency Treebank. Technical Report No. 47, ÚFAL, Charles University in Prague.

Poláková, L., Mírovský, J., Nedoluzhko, A., Jínová, P., Zikánová, Š., & Hajičová, E. (2013).Introducing the Prague Discourse Treebank 1.0. In Proceedings of the 6th international joint conference on natural language processing (pp. 91-99).

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., & Webber, B. L. (2008). The Penn Discourse TreeBank 2.0. In LREC.

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). Association for Computational Linguistics.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., & Webber, B. L. (2007). The penn discourse treebank 2.0 annotation manual.

Prasad, R., Webber, B., & Joshi, A. (2014). Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. Computational Linguistics, 40(4), 921-950.

Ramesh, B. P., Prasad, R., Miller, T., Harrington, B., & Yu, H. (2012). Automatic discourse connective detection in biomedical text. Journal of the American Medical Informatics Association, 19(5), 800-808.

Ruhi, Ş. (1994, August). Restrictions on the interchangeability of discourse connectives: a study on ama and fakat. In *Seventh International Conference on Turkish Linguistics*. 3-6 August 1994, Institute of Oriental Studies, Turcology, Johannes Gutenberg University, Mainz.

Rysova, M. (2012). Alternative Lexicalizations of Discourse Connectives in Czech. In LREC (pp. 2800-2807).

Sanders, T., & Maat, H. P. (2006). Cohesion and coherence: Linguistic approaches. reading, 99, 440- 466.

Say, B., Zeyrek, D., Oflazer, K., and Özge, U. (2002). Development of a Corpus and a Treebank for Present-day Written Turkish. In Proceedings of the Eleventh International Conference on Turkish Linguistics (ICTL 2002).

Sharma, H., P. Dakwale, D. Sharma, R. Prasad, and A. Joshi. (2013). Assessment of different workflow strategies for annotating discourse relations: A case study with

HDRB. In A. Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, LNCS 7816, pages 523–532. Springer.

Siegel, M. E. (2002). Like: The discourse particle and semantics. Journal of Semantics, 19(1), 35-71.

Singer, M., Andruslak, P., Reisdorf, P., & Black, N. L. (1992). Individual differences in bridging inference processes. Memory & Cognition, 20(5), 539-548.

Stede, M. 2008. RST revisited: Disentangling nuclearity. In C. Fabricius-Hansen and W. Ramm, editors, 'Subordination' versus 'coordination' in sentence and text–from a cross-linguistic perspective, pages 33–58. John Benjamins, Amsterdam.

Taboada, M. (2009). Implicit and explicit coherence relations. Discourse, of course. Amsterdam: John Benjamins, 127-140.

Uzun, L. S. (1995). Orhon yazıtlarının metinbilimsel yapısı (Vol. 7)., Simurg.

Van Dijk, T. A., Kintsch, W., & Van Dijk, T. A. (1983). Strategies of discourse comprehension. New York: Academic Press.

Webber, B., Egg, M., & Kordoni, V. (2012). Discourse structure and language technology. Natural Language Engineering, 18(04), 437-490.

Xue, N. (2005). Annotating discourse connectives in the Chinese Treebank. In Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, pages 84–91, Ann Arbor, MI.

Yalta, A. T. (2008). The accuracy of statistical distributions in Microsoft® Excel 2007. Computational Statistics & Data Analysis, 52(10), 4579-4586.

Yavuz, S. (2011). TÜRKİYE TÜRKÇESİ AĞIZLARINDA BAĞLAÇLAR.Diyalektolog, 2(2).

Yilmaz, E. (2004). A pragmatic analysis of Turkish discourse particles: Yani, işte and şey (Doctoral dissertation, MIDDLE EAST TECHNICAL UNIVERSITY).

Yuhas, B. J. (2013). An analysis of discourse markers and discourse labels as cohesive devices in ESL student writing (Doctoral dissertation, Colorado State University).

Zeyrek, D. (2015). Annotating Implicit Discourse Relations in Turkish & The Challenge of Annotating Corrective Discourse Relations. Slides from the Paper presented at IPrA Conference 2015, Antwerp.

Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A., Balaban, H. Ö., Yalçınkaya, İ., and 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 (LAW IV).

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.

Zeyrek, Deniz, and Bonnie L. Webber. "A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus." IJCNLP. 2008.

Zeyrek, D., Turan, Ü., Bozsahin, C., Çakici, R., Sevdik-Çalli, A., Demirşahin, İ., ... & Ögel, H. (2009, August). Annotating subordinators in the Turkish discourse bank. In Proceedings of the third linguistic annotation workshop (pp. 44-47). Association for Computational Linguistics.

Zhou, Y. and N. Xue. (2012). PDTB-style discourse annotation of Chinese text. In Proceedings of the 50th Annual Meeting of the ACL, pages 69–77, Jeju Island.

Zhou, Z. M., Xu, Y., Niu, Z. Y., Lan, M., Su, J., & Tan, C. L. (2010). Predicting discourse connectives for implicit discourse relation recognition. InProceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 1507-1514). Association for Computational Linguistics.

APPENDICES

APPENDIX A: THE LIST OF OPEN ENDED ALTLEX TYPES and TOKENS FROM AUTOMATIC EXTRACTION

	ALTLEX Types	ALTLEX Tokens
	… neden/nedeniyle	birinci neden, bu cinayet nedeniyle, bunun başlıca nedeni, elektrik kesintisi nedeniyle, ikinci neden, karnelerin basılması nedeniyle, iklim değişikliklerinin nedeni, kriz nedeniyle, yukarıdaki nedenlerle, rahatsızlığı nedeniyle, şişmanlığın birçok nedeni, şişmanlığın tek nedeni, üzüntüsünün başlıca nedeni, verilen izinler nedeniyle, yenilginin birkaç nedeni, yüksek kira nedeniyle
	… (-a/e) gore	bu inanışa gore
	…(-den/dan) beri	birkaç zamandan beri
	… (-den/dan) dolayı	bu sebepten dolayı
	…(-nın/nin) ardından	Gül'ün ziyaretinin ardından, haberin çıkması ardından, açılış töreninin ardından, bu gelişmenin ardından, bunun ardından, kavganın ardından, mahkeme kararının ardından, olayın ardından, olayların ardından, öğle namazının ardından, yaşanan koşuşturmanın ardından, ziyaretin ardından
	…(- nın/nin)sonunda	bu konuşmanın sonunda
	… önce	bir zaman önce, bir an önce, her şeyden önce, Seçimlerden önce

	… öncesinde	binlerce yıl öncesinde
	… sonra/ sonrası	akşam yemeği sonrası, balkan savaşlarından sonra, bir süre sonra, bu konuşmadan sonra, dördüncü karardan sonra, erteledikten sonra, holdingten ayrıldıktan sonra, italya' ya döndükten sonra, nice zamanlardan sonra, onların ölümünden sonra, öğle yemeğinden sonra, seçim sonrası, Sezer'in vetosundan sonra, bundan sonra, zaman sonra
	… sonucu	bu tartışma sonucu, uzun tartışmalar sonucu, yapılan oylama sonucu
	…(-den/dan sonra)	kaza sonrası, kitap alışverişlerinden sonra, yıllar sonra
	akşam yemeği sonrası	akşam yemeği sonrası
	ana amacı	ana amacı
	ardından	ardından
	böyle bir durumda	böyle bir durumda
	böyle hareketler karşısında	Böyle hareketler karşısında
	bu arayışlar sırasında	bu arayışlar sırasında
	bu çağda	bu çağda
	bu doğrultuda	bu doğrultuda
	bu dönemde	bu dönemde
	bu durumda	bu durumda
	bu noktada	bu noktada
	bu örneklerin hepsi	bu örneklerin hepsi
	bu sürede	bu sürede
	bu tamir esnasında	bu tamir esnasında
	bu yazıdaki amaç	bu yazıdaki amaç
	bu yolla	bu yolla

	bunlardan birisi	bunlardan birisi
	bunlardan ilki	bunlardan ilki
	bunların başında	bunların başında
	bunun tersi	bunun tersi
	örnek olarak	örnek olarak
	her zamanki gibi	her zamanki gibi
	işte o zaman	işte o zaman
	mizahın amacı	mizahın amacı
	o zamanlar	o zamanlar
	onların amacı	onların amacı
	son zamanlarda	son zamanlarda
	uygunsuz bir zamanda	uygunsuz bir zamanda
	yakın zamana kadar	yakın zamana kadar
	yazarın amacı	yazarın amacı
Total	**42 Types**	**89 tokens**

APPENDIX B: KEY WORDS USED FOR AUTOMATIC EXTRACTION OF OPEN ENDED ALTLEXs

Amaç
Arada
Ardından
Bahane
Beraber
Böyle
Bu
Gerekçe
Neden
Önce
Örnek
Ötürü
Özet
Sebep
Sonra
Sonuç
Şu
Yanında
Zaman