AUTOMATIC SENSE PREDICTION OF EXPLICIT DISCOURSE CONNECTIVES IN TURKISH WITH THE HELP OF CENTERING THEORY AND MORPHOSYNTACTIC FEATURES

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

SAVAŞ ÇETİN

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

FEBRUARY 2018

Approval of the thesis:

AUTOMATIC SENSE PREDICTION OF EXPLICIT DISCOURSE CONNECTIVES IN TURKISH WITH THE HELP OF CENTERING THEORY AND MORPHOSYNTACTIC FEATURES

submitted by **SAVAŞ ÇETİN** in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Science, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

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

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

Examining Committee Members:

Prof. Dr. Ümit Deniz Turan
The Department of English Language Teaching, Anadolu University _____

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

Assist. Prof. Dr. Burcu Can Buğlalılar
Department of Computer Engineering, Hacettepe University _____

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

Assist. Prof. Dr. Umut Özge
Cognitive Science Department, 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: SAVAŞ ÇETİN

Signature :

ABSTRACT

AUTOMATIC SENSE PREDICTION OF EXPLICIT DISCOURSE CONNECTIVES IN TURKISH WITH THE HELP OF CENTERING THEORY AND MORPHOSYNTACTIC FEATURES

ÇETİN, SAVAŞ

M.S., Department of Cognitive Science

Supervisor : Prof. Dr. Deniz Zeyrek Bozşahin

February 2018, 53 pages

Discourse connectives (and, but, however) are one of many means of keeping the discourse coherent. Discourse connectives are classified into groups based on their senses (expansion, contingency, etc.). They describe the semantic relationship of two discourse units. This study aims to build a machine learning system to predict the sense of explicit discourse connectives on the Turkish Discourse Bank data, which is manually gold-annotated. To do so, this study examines the effect of several features: i.e. transitions of Centering Theory and morphosyntactic characteristics of main verbs of the arguments in a discourse relation. The results imply that Centering Theory, morphosyntactic features and their combinations affect each class of sense in a different way. When the base score is calculated with only the connective feature, the addition of Centering Theory features seems to have increased the predictions scores for Comparison and Expansion classes. Also, Tense, Aspect and Modality features are observed to slightly affect the Temporal class in a positive way.

Keywords: explicit discourse relations, supervised learning, Turkish Discourse Bank, automatic sense prediction, Centering Theory

ÖZ

TÜRKÇE'DE AÇIK BAĞLAÇLARIN MERKEZLEME TEORİSİ VE MORFO-SENTATİK ÖZELLİKLER YARDIMI İLE OTOMATİK OLARAK BELİRLENMESİ

ÇETİN, SAVAŞ

Yüksek Lisans, Bilişsel Bilimler Programı

Tez Yöneticisi : Prof. Dr. Deniz Zeyrek Bozşahin

Şubat 2018 , 53 sayfa

Söylem bağlaçları (ve, ama, ancak vb.), söylemi tutarlı halde tutmanın birçok yönteminden biridir. Söylem bağlaçları, anlamları bakımından sınıflandırılmışlardır. İki söylem ünitesinin aralarındaki anlamsal ilişkiyi tanımlamaktadırlar. Bu çalışma, bir makine öğrenmesi sistemi geliştirerek, Türkçe Söylem Bankası'ndaki (TDB) açık söylem bağlaçlarının anlamlarını belirlemeyi hedeflemektedir. Bu hedefi gerçekleştirmek için çeşitli özelliklerin etkileri incelenmiştir. Bu özellikler, Merkezleme Teorisi'nin geçişleri ve söylem bağlantılarındaki ünitelerin morfo-sentaktik yapılarıdır. Sonuçlar, Merkezleme Teorisi özelliklerinin, morfo-sentaktik yapıların ve bunların birleşimlerinin her bir anlam sınıfını farklı yönde etkilediğini öne sürmektedir. Sadece söylem bağlaçlarıyla hesaplanan taban skor göz önünde bulundurulunca Merkezleme Teorisinin eklenmesi, Karşılaştırma ve Açımlama sınıflarının tahmin skorlarını arttırmıştır. Ayrıca, Zaman, Görünüş ve Kiplik özelliklerinin Zamansal sınıfını olumlu yönde etkilediği gözlemlenmiştir.

Anahtar Kelimeler: açık söylem bağıntıları, gözetimli öğrenme, Türkçe Söylem Bankası, otomatik anlam belirlenmesi, Merkezleme Teorisi

ACKNOWLEDGMENTS

First and foremost, I wish to express my deepest gratitude to my advisor Prof. Dr. Deniz Zeyrek Bozşahin. Without her guidance, this idea could never be more than a proposal. And I would not find the will, neither the courage, to continue on this path without her understanding, deep knowledge, suggestions and help.

I should express my thanks to Ali Doğan for his continuous encouragement and support throughout this process. He was always there to consult with on any issues I faced.

I wholeheartedly thank Murathan Kurfalı as he was always a question away from me when I encountered anything unexpected. His support will never be forgotten.

Last but not the least, I should thank Nihal Meriç Atila, Gökçe Gündoğdu, Cengiz Dikme, Gamze Uslu, Pınar Orbay and Milan Soucek who made everything else flawless. Thanks to them, I could find the chance to focus on my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Discourse, which refers to organized bodies of text, tends to be coherent (Hobbs, 1978). This suggests that discourse is not a bunch of *units* brought together randomly. All and every meaningful unit is in a relation with the other units in the discourse. These relations are necessary for discourse to be well-structured and meaningful. The relations among the units can be revealed via several means: discourse connectives, lexical cohesion, ellipsis, anaphora and coreference, centers of attention, and etc. Below, representative examples of each of these devices are provided.

a Discourse connective (Pitler et al., 2008)

(1) He is very tired because he played tennis all morning.

The perceived causality relation between these sentences is provided with the connective 'because'.

b Lexical cohesion (Halliday & Hasan, 2014)

(2) a. There's a boy climbing the old elm.
b. That tree isn't very safe.

Here, the cohesion is supplied lexically as 'that tree' is superordinate for 'the old elm'.

c Ellipsis (Sanders & Maat, 2006)

(3) a. All the children had an ice-cream today.
b. Eva chose strawberry.
c. Arthur had orange and Willem too.

When sentence 3b is seen or heard after 3a, one can easily guess that Eva chose a strawberry flavored ice-cream. No conversant in this conversation would think it is about anything else than ice-cream under the given conditions.

d Anaphora (Sanders & Maat, 2006)

(4) Jan lives near the park. He often goes there.

Anaphora lets the conversants understand what 'he' and 'there' stand for as they are related to their antecedents in what they refer to.

e Centers of attention (M. A. Walker et al., 1998)

Centering Theory (Grosz & Sidner, 1986) is based on the notion of 'centers of attention' in discourse. There are three major centering transitions referring to changes in attentional state, which is a property of the discourse itself. *Continue*, *Retain* and *Shift*. *Continue* is used to express that the center of the discourse unit is continuing in the next unit. *Retain* indicates that the center will be changed, and *Shift* demonstrates that a shift of center has taken place.

(5) a. Jeff helped Dick wash the car.
b. He washed the windows as Dick waxed the car.
c. He soaped a pane.

The center of discourse is the entity realized as 'Jeff' and it doesn't change throughout the whole discourse (*Continue*). It has been suggested that Continue is a strong sign of coherence when compared to the other transitions (Gordon et al., 1993).

Among these coherence devices, this study will focus on discourse connectives. Discourse connectives are "words or phrases that connect or relate two coherent sentences or phrases and indicate the presence of discourse relations" (Ramesh & Yu, 2010). We can reanalyze the example here from Pitler et al. (2008) (p. 87); in example 7, when the order of the text pieces connected by *because* is changed, coherence is lost; readers would not infer the causality relation any more:

(6) He is very tired because he played tennis all morning.

(7) # He is very tired because he will play tennis tomorrow.

Discourse connectives can either be explicit, i.e. overt (as in example 6 above) or concealed as can be seen from the following sentence. These have been known as implicit discourse relations:

(8) He is very tired; he played tennis all morning.

The semantic relation between text pieces is always inferred. That is to say, discourse can be well-formed and coherent without an explicit connective, though certain relations tend to require a connective so that the units will not be incoherent:

(9) He is not very strong, but he can run amazingly fast.

(10) * He is not very strong, he can run amazingly fast.

In summary, discourse relations can be signaled with a connective to make the relation salient. Implicit relations are the ones that can be interpreted by the context of the utterance without the use of a connective (Pitler et al., 2008). Discourse relations are often named with the name of the sense of the relation. Example 6 is a causal relation while 9 is a contrastive one.

1.1 Aim and Scope

In this work, the ultimate aim is to automatically predict the sense of explicit discourse connectives in Turkish. To do this, we will rely on (a) the semantic information conveyed through morphosyntactic features, such as polarity and tense and (b) the transitions of Centering Theory (Brennan et al., 1987). These features are given to a Maximum Entropy model, which is a supervised machine learning system, to predict the sense labels of explicit discourse connectives. We use the Stanford Classifier (Manning & Klein, 2003), which is a feature-based modeling where each feature is assigned a weight depending on the training data. Then, these weights are used to classify the test data. Whether these features increase the accuracy of the prediction score of the classifier is determined by comparing the results with the prediction score calculated depending only on the connective. (Zeyrek & Kurfalı, 2017). We hypothesize that, to the extent we can increase the performance of the system, the morphosyntactic features and the Centering Theory transitions are meaningful linguistic factors in discourse.

While the morphosyntactic features we use have been used in previous work on discourse parsing, the incorporation of transitions from Centering Theory is, to the best of our knowledge, novel.

This work is conducted solely on explicit discourse relations, leaving implicit relations out of scope. The data is Turkish Discourse Bank version 1.1 (Zeyrek & Kurfalı, 2017), a corpus annotated at the discourse level following the principles of Penn Discourse Treebank (Prasad et al., 2007). The data is annotated in terms of the explicit connective and the phrases or sentences that are related with the connective. These text parts are referred to as Arg1 and Arg2. An example is provided below, where the connective is underlined, Arg1 is shown in italics, Arg2 in bold fonts.

(11) *Ben İngilizce bilmiyordum, İngilizce ismini şu anda hatırlamıyorum*, ama **Nurhan tayf analizlerinden söz eden İngilizce bir fizik kitabı edinmişti**.

I didn't know English, I can't remember its English name now but **Nurhan acquired an English physics book mentioning about the spectrum analysis**.

("COMPARISON: Pragmatic Contrast", fileNO: 00050220 in TDB)

The arguments of a discourse connective always have an abstract object interpretation, which are propositions, properties, states of affairs and facts without any spatio-temporal location (Asher, 2012). They can be a tensed or non-tensed clause or a group of clauses (Prasad et al., 2007).

As Graesser et al. (2011) states "discourse comprehension is a very rich, multilevel cognitive activity." Resolving the sense of an explicit discourse connective is one of the major steps of discourse parsing, which first became prominent by Marcu (1997) as an automatic approach to discourse analysis. To the best of our knowledge, the step that this work will attempt to handle; i.e. automatic sense prediction of explicit connectives, has not been performed on a Turkish corpus. Thus, the main contribution of this work will be to fill a gap in Turkish discourse parsing.

1.2 Outline

This thesis is composed of six chapters: Introduction, Literature Review, Data and Feature Set, Methodology, Results, Discussion and Conclusion.

Chapter 2, Literature Review, elaborates on theories and previous work related to this study. It gives detailed information on discourse banks, discourse parsing, previous studies on discourse parsing and Centering Theory.

Chapter 3, Data and Feature Set, gives details on Turkish Discourse Bank 1.1, the features to be used for this study and limitations of this work as well as the modifications made on the data.

Chapter 4, Methodology, gives explanations on the Maximum Entropy classifier and how its results are calculated. Then, this chapter mentions the process of data preparation and how it is annotated for the features explained in Chapter 3.

Chapter 5, Results, is divided into nine sections, each of which stands for a feature set tested on the classifier. Then, Chapter 6, Discussion, summarizes the results for each class level sense; Finally, Chapter 7, Conclusion, summarizes and concludes the thesis.

CHAPTER 2

LITERATURE REVIEW

In this chapter, two discourse banks, namely Penn Discourse Treebank and Turkish Discourse Bank, their approaches to annotation and general principles are summarized. A general explanation on discourse parsing is provided and previous studies on explicit and implicit relations are overviewed. Finally, Centering Theory and its major notions are explained.

2.1 Discourse Banks

This section introduces the PDTB framework as well as the Turkish Discourse Bank in detail.

2.1.1 Penn Discourse Treebank(PDTB)

Penn Discourse Treebank (PDTB), which is composed of 1 million word Wall Street Journal (WSJ) Corpus, is a discourse-level annotated corpus. Discourse annotations are added to the sentence-level syntactic annotation of Penn Treebank or PTB (Miltsakaki et al., 2004).

The aim is to annotate the argument structure of discourse relations. An argument structure of a discourse relations is minimally made up of one discourse connective (explicit or implicit) and its two arguments that have abstract object interpretation (Asher, 2012). PDTB takes at least one-predicate verb phrases, a single clause, a single sentence, a sequence of clauses and/or sentences, or combinations of both, as arguments of a relation.

PDTB provides annotations on Explicit, Implicit, AltLex, EntRel, and NoRel relations. The examples below for each type of relation are from Prasad et al. (2007):

- Explicit Discourse Relations: These relations are the ones where there is a connective between the discourse units.

  The PDTB corpus has identified three kinds of explicit discourse connectives: (Forbes-Riley et al., 2005):

  - subordinating conjunctions (e.g. ‘because’, ‘although’),
  - coordinating conjunctions (e.g. ‘and’, ‘or’),
  - adverbial connectives (e.g. ‘therefore’, ‘instead’).

(12) Since **McDonald's menu prices rose this year**, *the actual decline may have been more*.

- Implicit Discourse Relations: These relations are the ones that do not have a realized, explicit connective between the arguments. The common annotation practice is to find and insert a connective expression that is the most appropriate representing the inferred relation. Mostly, in the annotation manual, annotators are provided a default explicit connective for each sense in the hierarchy.

  (13) *The projects already under construction will increase Las Vegas's supply of hotel rooms by 11,795, or nearly 20%, to 75,500*. Implicit = 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**.

- Alternative Lexicalization (AltLex): In AltLex case, insertion of an implicit connective leads to a redundancy in the expression as there is already an alternatively lexicalized non-connective expression instantiating the relation.

  (14) 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**.

- Entity Relation (EntRel): This is the case when there is an entity-based relation between the arguments.

  (15) *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**.

- No Relation (NoRel): This case occurs when basically there are no discourse relations between adjacent sentences.

  (16) The products already available are cross-connect systems, used instead of mazes of wiring to interconnect other telecommunications equipment. *This cuts down greatly on labor, Mr. Buchner said*. NoRel **To be introduced later are a multiplexer, which will allow several signals to travel along a single optical line; a light-wave system, which carries voice channels; and a network controller, which directs data flow through cross-connect systems**.

PDTB contains sense labels for discourse relations in a hierarchical classification schema. There are three hierarchical levels of senses in PTDB, which are called class, type and subtype, respectively. All the levels and sense relations are given in Figure 2.1 (Prasad et al., 2007).

2.1.2 Turkish Discourse Bank (TDB)

This study uses the manually annotated data from Turkish Discourse Bank 1.1 (TDB 1.1). TDB is the product of an effort of extending METU Turkish Corpus (MTC) (Say et al., 2002),

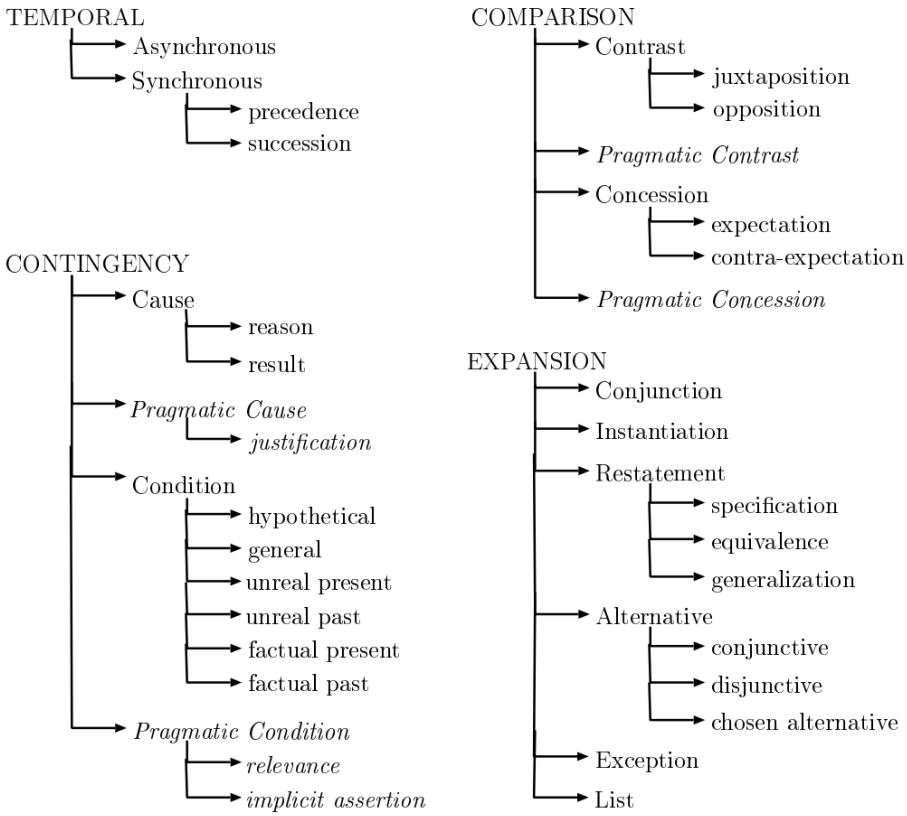

Figure 2.1: Hierarchy of Sense tags in PDTB 2.0

a 2-million word electronic resource, "from a sentence-level language resource to a discourse-level recourse by annotating its discourse connectives, and their arguments" (Zeyrek & Webber, 2008). TDB is a 400.000-word subcorpus of MTC, which followed the annotation style of Penn Discourse TreeBank (PDTB) (Robaldo et al., 2008). There are 197 text from various genres written between the years 1990-2000.

TDB mainly follows PDTB style of annotation in what to annotate. There are five types of discourse relations in TDB though three of them can be seen as subcategory of the other two. The main two categories are Explicit and Implicit relations. The other three are Alternative Lexicalization, Entity Relation and No Relation.

- Explicit Relations: These relations are the ones where there is a connective between the abstract objects.

  (17) *Henüz çok iyi öğrenememiştim New York metrosunu* ama **gene de her gece gideceğim yere varabiliyordum**.
  I haven't learnt about New York subway much but **I was able to arrive at where I wanted to every night**.
  ("COMPARISON: Concession", fileNO: 00002113 in TDB)

The connectives used to signify explicit relation are categorized under five (Zeyrek & Webber, 2008):

I. coordinating conjunctions such as *çünkü* (because), *ama* (but);

II. paired coordinating conjunction such as *hem* ... *hem* (both ... and), *ne* ... *ne* (neither ... nor);

III. simplex subordinators (converbs), which are suffixes forming adverbial phrases such as *-DHK* in geldiğinden (because/since he has come);

IV. complex subordinators that are made up of more than one components: usually a postposition and a suffix on the verb of the subordinate clause such as *-mA rağmen* in *gelmesine rağmen* (even though he came);

V. anaphoric connectives such as *ne var ki* 'however'.

- Implicit Relations: These relations are the ones that do not have a realized, explicit connective between the arguments. The common practice is to find and insert a connective expression that is the most appropriate to the inferred relation.

  (18) *Sesi soğuk ve uzaktı*. IMPLICIT: BU NEDENLE **Uygunsuz bir zamanda aramış olduğumu düşündüm**.
  His voice was cold and afar. IMPLICIT: THUS **I came to a thought that I called him in an inconvenient time**.
  ("CONTINGENCY: Cause: result", fileNO: 00005121 in TDB)

- Alternative Lexicalization, Entity Relation and No Relation: These are the relations where one cannot simply insert an implicit connective between the arguments when an explicit connective is not present.

  – Alternatve Lexicalization (AltLex): In AltLex case, insertion of an implicit connective leads to a redundancy in the expression as there is already an alternatively lexicalized non-connective expression to employ the conduct the relation between the arguments.

    (19) *1998-2002 arasında oynanan müşterek bahis oyunlarında yapılan yasal kesintilerin tutarı 1.5 milyar doları geçti*. Buna karşılık, **at yarışı oynayanlara, bahis için verdikleri paranın yalnızca yarısı ikramiye olarak döndü**.
    The amount of official reductions in mutual gambling games played between 1998-2002 exceeded 1.5 billion dollar. As opposed to this, **only half of the money paid for betting was returned as prize to the people who gambled on horse racing**.
    ("COMPARISON: Contrast: juxtaposition", fileNo: 10330000 in TDB)

  – Entity Relation (EntRel): This is the case when there is an entity-based relation between the arguments.

    (20) *Aşıklı Höyük bu yerleşimlerden biri*. **Aksaray ilinin Kızılkaya Köyü'nün yakınında, Melendiz Nehri'nin kıyısında yer alıyor**.
    Aşıklı Höyük is one of these settlements. **It is situated near Kızılkaya Village of Aksaray city, on the bank of Melendiz River**.
    (fileNo: 00013112 in TDB)

– No Relation (NoRel): This case occurs when basically there are no discourse relations between adjacent sentences.

(21) *Başka kimse olmadığından iki kadının da yüzü açıktı*. **Halil onları korkutacağı yere geldiğinde donakaldı**.
Faces of both women were open as there were noone else. **Halil, when he arrived at the spot where he would scare them, froze**.
(fileNo: 00001131 in TDB)

After annotating explicit discourse relations in TDB, senses are also annotated on a subcorpus of TDB created from 10% of the TDB corpus in a balanced way regarding the genres of the texts.

As the result of recent annotation efforts, a subcorpus of Turkish Discourse Bank, called TDB 1.1 is created. TDB 1.1 is subcorpus version of the original TDB enriched with sense annotations. TDB 1.1 constitutes the training/test data of the current study therefore is explained in great detail in Chapter 3.

2.2 Discourse Parsing

Given that this work aims to predict the sense of explicit discourse relations through a computational model, this section introduces a feature set, partly based on those which are offered and used in similar works for other languages. Then, a supervised classifier is trained on those features.

Discourse parsing is composed of five components: *connective classifier*, *argument labeler*, *explicit classifier*, *non-explicit classifier* and *attribution span labeler* (Lin et al., 2014):

- *The connective classifier* classifies discourse connectives by distinguishing them from non-discourse connectives. In the PDTB framework, non-discourse connectives are those that do not link abstract objects:

  (22) *Financial planners often urge investors to diversify and to hold a smattering of international securities*. And **many emerging markets have outpaced more mature markets, such as the US and Japan**.

  (23) Political and currency gyrations can whipsaw the funds.

  For example, the discourse parser should classify the ‘and’ in example 22 as a discourse connective because it combines two abstract objects whereas ’and’ is not a discourse connective as it connects two adjectives (*political* and *currency*) in example 23.

- *The argument labeler* is supposed to identify the span of the arguments of a given discourse connective.

  The argument labeler’s task is to find the portion of the sentence which is directly associated with the given relation.

- *The explicit classifier* identifies the sense label of the explicit connectives. It has high importance as discourse connectives can be ambiguous as in the case of this example:

(24) *Microsoft added 2 1/8 to 81 3/4* and **Oracle Systems rose 1 1/2 to 23 1/4**.

Here 'and' is ambiguous between Expansion:Conjunction and Expansion:List senses.

- *The non-explicit classifier* works in the same way as the explicit classifier but on non-explicit relations, which are realized by implicit connectives, AltLex (alternative lexicalization), EntRel (entity relations) and NoRel (no relation). When all the explicit discourse relations are found and annotated, the remaining sentences are extracted from the given text and the parser is fed with the all sentence pairs. Therefore, the non-explicit classifier needs to analyze all these pairs and decide which of those pairs convey a meaning and label them with the correct sense tag. Finally, those which do not convey a sense are annotated as EntRel or NoRel.
- Finally, the *attribution span labeler* decides on the attribution spans. Attribution spans show how a discourse relation and its arguments are attributed (TDB has so far left attribution annotation out of its scope).

Lin et al. (2014)'s parsing algorithm is designed in the same sequential way as PDTB annotators perform annotations. The pipeline from is shown in 2.2.

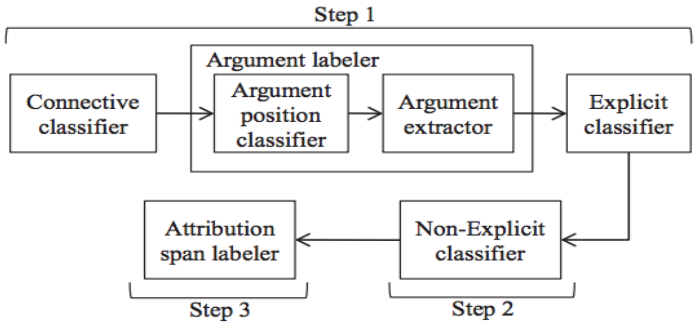

Figure 2.2: Discourse parsing algorithm pipeline from (Lin et al., 2014)

Of these five components, the current study aims to implement *the Explicit classifier*.

Below, previous work on explicit classifiers are provided. As implicit and explicit classifiers are intertwined and they share a great deal of features, implicit classifiers are also mentioned.

2.3 Previous Studies on Discourse Parsing

In this section, previous studies conducted on automatic sense prediction for explicit connectives and implicit relations are provided.

2.3.1 Explicit connectives

Predicting the sense of explicit discourse connectives is performed by various researchers with high accuracy.

Working on PDTB, Pitler et al. (2008) achieved 93.09% classification accuracy in four-way classification on only explicit connectives. This classification is only for class level classification; between temporal, comparison, contingency and expansion sense labels. Four-way classification means all annotations of each of four class are provided as test and training data and the output of the classification is expected to be one of the four possibilities for every connective. What Pitler et al. suggests is that the connective itself is a good enough feature to disambiguate the sense in explicit relations. In the same study, Pitler et al. (2008) made a binary classification on explicit connectives, as well. This experiment is also conducted with only one feature, which is the explicit discourse connective. The accuracy is 95.4% for Temporal, 97.23% for Comparison, 93.99% for Contingency and 97.61% for Expansion. There are some other interesting findings from them. They found that temporal and comparison relations are overall mostly unambiguous; and that these relations tend to be explicit more often than contingency and expansion relations.

In Pitler & Nenkova (2009), a prediction system to label connectives as discourse or non-discourse connectives and then to label the discourse connectives' sense was developed by using syntactic features. Incorporating the syntactic features with the connective words, the prediction accuracy goes up to 94.15%. Pitler & Nenkova state that they seem to be approaching a performance ceiling as the inter-annotator agreement between real annotators is also not more than 94% in class level annotation. The syntactic features used by Pitler & Nenkova in addition to the connective word itself are:

- **Self category:** This is the highest node in the tree of the connective only. For single word connectives, it's Part of Speech (PoS) of that word but for multi-word connectives, it is not.
- **Parent category:** It is the category of the immediate parent of the Self category. Pitler & Nenkova claim this feature to be especially useful in distinguishing discourse connectives from non-discourse ones as for a discourse connective Parent category can hardly be a Nominal Phrase.
- **Left sibling category:** It is the syntactic category which is immediately on the left of the Self category. This will be NONE if there is no Left sibling.
- **Right sibling category:** In a similar way to the Left sibling, Right sibling category is the one immediately on the right of the Self category and will be NONE if it is not existent. This category is particularly important as English is a head-initial language, which means the dependents of a category will appear as right siblings. Pitler & Nenkova also used two additional features inside this Right sibling category. These are *Right Sibling Contains a VP* and *Right Sibling Contains a Trace* features to distinguish single word clause from a full embedded clause.

Pitler & Nenkova also reports that most of the errors is related to the Temporal class as it is the least frequent class in the training data. And the most frequently encountered error is the Contingency class being predicted as Temporal, which makes up 29% of all the errors.

In addition to these studies, Prasad et al. (2011) conducted an experiment on biomedical corpus by claiming that a biomedical discourse corpus would be beneficial. The experiment carried out on class level sense relations resulted in 90.9% accuracy by relying on only one feature, which is the explicit discourse connective for each relation.

2.3.2 Implicit Classifiers

Similar to explicit sense prediction, implicit sense prediction has also been center of attention in research and it is the most challenging component of discourse parsing (Lin et al., 2014) as there is no connective present and the prediction of the relation is made based on the interpretation of the sentences by the real annotators. Also, it is known that inter-annotator agreement on implicit sense relation can be lower than explicit sense relations (Zeyrek & Kurfalı, 2017).

Marcu & Echihabi (2002) worked on automatic prediction of implicit discourse connectives on data generated from explicit relations, i.e. artificial implicit relations. The data is, in fact, composed of explicit discourse relations. But, the researchers remove the explicit relation cue phrase, which is the discourse connective, from the relation and labels it. As an example, the label CONTRAST is assigned to the relation if the removed connective is *but*. The main feature for this study is word pairs. The data is composed of 41.147.805 unannotated English sentences and BLIPP, a corpus containing 1.796.386 English sentences which are automatically parsed. Marcu & Echihabi (2002) reports that for some relations an accuracy of 93% was captured.

Lin et al. (2009) reported an accuracy of 40.2% reached in implicit relation prediction which is reportedly 14.1% over the majority baseline. The prediction is conducted on data with natively implicit relations which differs from the data used by Marcu & Echihabi (2002). The features employed are put together under four categories. These are *contextual features* referring to the presence of dependencies between discourse relations, *constituent parse features*, *dependency parse features* and *lexical features* which include word pairs.

Pitler et al. (2009) conducted "the first study which reports results on classifying naturally occurring implicit relations in text and uses the natural distribution of the various senses." They worked on implicit sense relations on newspaper text from PDTB. The overall results proved an increase for all class level sense relations. The feature set used for this study includes:

- *polarity tags*, which refer to the sentiments of the words in the arguments. Polarity tags of the words are obtained from Multi-perspective Question Answering Opinion Corpus Wilson et al. (2005);
- *inquirer tags*, which are used to hold the semantic categories of the words according to the General Inquirer lexicon by Stone et al. (1966). Some categories include Understatement versus Overstatement, Rise versus Fall, Pleasure versus Pain;
- Levin verb classes according to Levin (1993). This feature also includes length of verb phrases, i.e. number of the words in a verb phrase;
- *first-last-first3*, referring to the first and last words of the arguments in addition to the combination of first words of each argument and combination of the last words of the arguments;
- *modality*, which indicates the modal words usages such as *can* and *may*, which are often used in conditional statements and which signal a Contingency relation.

As for Turkish, Kurfalı (2016) conducted experiments on Turkish to automatically predict the

sense of implicit relations on TDB 1.1 data. Following the work of (Pitler et al., 2009), (Kurfalı, 2016) developed a supervised model with linguistic features, including the polarity and tense of the arguments of the discourse relations, in order to assign a correct Top-level sense for Turkish Implicit discourse relations. Moreover, (Kurfalı, 2016) created pseudo implicit discourse relations, which are explicit discourse relations from which the overt connective is stripped. According to the results, adding these pseudo implicit discourse relations do not improve the performance systematically, which suggests that the explicit and implicit discourse relations are different in terms of discourse relations they convey.

2.4 Centering Theory

Centering Theory is a theoretical framework to examine the use of various referring expressions and their interaction with mechanisms used to maintain discourse coherence (Grosz et al., 1983). In order to do so, the notion of center is introduced. Each sentence, S, has a backward-looking center, Cb(S) and a set of forward-looking centers Cf(S). Cb(S) is a link to the preceding sentence while Cf(S) is a set of all utterances which are also backward-looking center candidates of the following sentence. The entities in the Cf(S) set are ordered depending on the entity's salience. Salience of an entity is about the degree of activation of that entity in the stock of shared knowledge between the conversants and it can be can be contributed by factors such as 'subjecthood' and 'pronominalization' (M. Walker et al., 1994). That is to say, the least salient item has the highest probability to be the backward-looking center of the next sentence and it is the first item in the Cf(S) set. In other words, the more prominent an item of $Cf(U_n)$ is, the more likely it will be $Cb(U_{n+1})$ (Grosz et al., 1995).

CT has constraints on which NP realizes as the Cb(U). The rule from Grosz, et al. specifies the basic constraint:

- If the Cb of the current utterance is the same as the Cb of the previous utterance, a pronoun should be used (1983):

  (25) a. How is Rosa?
  b. Did anyone see her yesterday?
  c. Max saw her. [Cb(c) = Rosa]

A follow-up rule is formulated in Grosz et al. (1995):

- If any element of $Cf(U_n)$ is realized by a pronoun in U_{n+1}, then the Cb(Un+1) must be realized by a pronoun also. This is exemplified by violating the rule:

  (26) a. He has been acting quite odd. [Cb = John = referent ("he")]
  b. He called up Mike yesterday. [Cb = John = referent ("he")]
  c. John wanted to meet him urgently. [Cb = John; referent ("him") = Mike]

  The violation is that even though the center is John, it is not realized with a pronoun but there is another entity which is realized as pronoun while it is not the center.

Grosz et al. (1995) states that this is validated by both psychological and cross-linguistic research and the correspondent of pronouns is zero-pronouns in some languages. Turkish is an example of such languages. Turan (1996) suggested that null pronouns are more likely to be the centers:

(27) a. Annem para verdi.
My mother gave money.
b. Adam ∅ almadı.
The man didn't take it.
c. *O/∅ cok az bir paraydı.
It was very little money.

As a bare NP, *para* can be accessed with a null pronoun rather than an overt pronoun. The attempt of accessing it with an overt pronoun makes it infelicitous.

A discourse segment's coherence is affected by the amount of the change made. Grosz & Sidner (1986) defined 'transitions' with some rules to define the amount of change made. These are *Center Continuation*, *Center Retaining* and *Center Shifting*.

- Center Continuation: $Cb(U_{n+1}) = Cb(U_n)$ which is also the most highly ranked element of $Cf(U_{n+1})$. This means that the center of a sentence is also the center of the next sentence and it is the most prominent element of that sentence.

  (28) a. John's been having a lot of trouble arranging his vacation.
  Cb:- Cf: [John, vacation]
  b. He called up Mike yesterday to work out a plan.
  Cb: [John] Cf: [John, Mike, plan] CONTINUE

- Center Retaining: $Cb(U_{n+1}) = Cb(U_n)$ but the entity is not the most highly ranked element of $Cf(U_{n+1})$ which means that the entity will probably not be the center of $Cf(U_{n+2})$.

  (29) a. John's been having a lot of trouble arranging his vacation.
  Cb:- Cf: [John, vacation]
  b. He called up Mike yesterday to work out a plan.
  Cb: [John] Cf: [John, Mike, plan] CONTINUE
  c. Mike gave him some good advice.
  Cb: [John] Cf: [Mike, John, advice] RETAIN

- Center Shifting: $Cb(U_{n+1}) \neq Cb(U_n)$ where the center of the sentence U_n is different from the center of U_{n+1}.

  (30) a. John's been having a lot of trouble arranging his vacation.
  Cb:- Cf: [John, vacation]
  b. He called up Mike yesterday to work out a plan.
  Cb: [John] Cf: [John, Mike, plan] CONTINUE
  c. Mike gave him some good advice.
  Cb: [John] Cf: [Mike, John, advice] RETAIN

d. He told John to talk to his boss.
Cb: [Mike] Cf: [Mike, John, boss] SHIFT

Brennan et al. (1987) extended Center Shifting in the formulation to two Center Shifting transitions because the formulation in Grosz & Sidner (1986) fails to resolve pronouns in certain examples. Brennan et al. (1987) gives this example:

(31) a. Brennan drives an Alfa Romeo.
Cb: [Brennan] Cf: [Brennan, Alfa Romeo]

b. She drives too fast.
Cb: [Brennan] Cf: [Brennan] CONTINUE

c. Friedman races her on weekends.
Cb: [Brennan] Cf: [Friedman, Brennan, weekend] RETAIN

d. She often beats her.
Cb: [Friedman] Cf: [Friedman, Brennan] SHIFT

They argue that "there seems to be more and less coherent ways to shift" suggesting the division of *Center Shifting* into two regarding the preferred center of the next sentence. This way, it is taken into account if the center of the next sentence is, at the same time, the preferred center, $Cp(U_{n+1})$, or not. Also, it is stated that an algorithm where Retain signals an intention to shift is more successful again with the help of the addition of Cp. So, the extended transitions are formulated as shown in Figure 2.3.

	$Cb(U_{n+1}) = Cb(U_n)$	$Cb(U_{n+1}) \neq Cb(U_n)$
$Cb(U_{n+1}) = Cp(U_{n+1})$	Continue	Smooth-Shift
$Cb(U_{n+1}) \neq Cp(U_{n+1})$	Retain	Rough-Shift

Figure 2.3: Extended Centering Theory transitions (Brennan et al., 1987)

As a theory to understand the means of coherence, Centering Theory may be quite useful in automatically predicting senses. The center transitions can be related to some specific senses. So, this part of the study will attempt to find out how Centering Theory help to improve the accuracy of automatic prediction of senses.

Though CT is believed to be of great help in automatically predicting the sense of a relation, this is only limited to inter-sentential relations where the arguments of the relation are in the different sentences. The claim that CT may not be helpful in intra-sentential relations became prominent after the suggestions of Strube (1998) and Miltsakaki (2002). Strube (1998) proposes a model, S-list, which is a model of CT with further modifications. S-list is a *salience list* and it covers the function of *backward-looking center* in CT. S-list is used to describe

the hearer's attentional state at any given point and it holds some discourse entities realized up to the given point. The entities in the list are ranked depending on their information status. Strube (1998) suggests that his model's predictive power is better than CT in anaphora resolution.

The following example from Miltsakaki (2002) suggests that CT is not capable of handling some intra-sentential relations:

(32) a. Dodge was robbed by an ex-convict.
b. The ex-convict tied him up
c. because he wasn't cooperating.
d. Then he took all the money and ran.

Miltsakaki (2002) claims that the algorithm proposed by Brennan et al. (1987) fails to resolve the *he* in example 32d to '*ex-convinct*' as anaphora resolution in CT tends to go for Continue transition; here it results in resolving *he* to '*Dodge*'.

In addition to these studies, Centering Theory is made use of by Prasad et al. (2010) while the researchers presented an approach to automatically identify the span of the first argument in a discourse relation. The first argument, which is not bound to the connective, represent a more challenging problem than the second argument. For this study, a sentence-based approach is employed distinguishing intra-sentential relations from inter-sentential ones. The researchers state that constraints of Centering Theory on anaphoric expressions partially inspired their coreference evaluations rules. The results indicate a 3% increase in performance to identify the span of first argument when compared to their baseline.

In this chapter, some of the existing discourse resources, namely TDB and PDTB is explained in terms of the annotation scheme they follow. Then, the literature about discourse parsing in general as well as the developed explicit and implicit discourse classifier are provided. Finally, Centering Theory and its notions are provided as transitions of the Centering Theory constitutes one of the unique aspects of the current study. In the next chapter, Chapter 3, data and the feature set is elaborated.

CHAPTER 3

DATA AND FEATURE SET

As an attempt to build an automatic system predicting the sense of explicit connectives, this study is employing gold annotated data and additional features. In this section, first the data composing the training and test data of this study is explained by mentioning how data is collected, by whom and how it is annotated and how inter-annotator agreement is calculated. After that, this section gives detailed explanation on which additional features are used and how they are employed. Finally, in this section, limitations of the study and modifications made to the data are explained.

3.1 Turkish Discourse Bank 1.1

The data employed for this study is a subcorpus of TDB, which is referred as TDB 1.1. During the creation of TDB 1.1, the genre distribution of the original TDB is maintained. The distribution of files in TDB 1.1 regarding their genre is as in table 3.1 Zeyrek & Kurfalı (2017). Also, CLASS level sense distribution can be seen from Table 3.2. For the distribution of individual explicit connectives, please see Appendix A, Data Distribution.

Table 3.1: Genre distribution in TDB 1.1 (Zeyrek & Kurfalı, 2017)

Genre	# of documents
Fiction (Novel; short story)	7 (35%)
News (Essay)	6 (30%)
Research monograph	2 (10%)
(Magazine) Article	2 (10%)
Memoir	2 (10%)
Interview	1 (5%)
SUM	**20 (100%)**

Table 3.2: CLASS level sense distribution

Temporal	Comparison	Contingency	Expansion	**TOTAL**
51	164	67	179	**461**

The data was annotated manually on the annotation tool developed for TDB (DATT) by Aktaş

et al. (2010).[1] Connectives (if any), both arguments, the sense of the relation, shared elements and spans are all kept in XML format. A sample output is given in Appendix B. Prior to the annotation process, the team first analyzed PDTB Annotation Manual 2.0 by Prasad et al. (2007) and analyzed TDB sentences to adapt the PDTB principles to TDB.

The data is divided into two halves and each half is annotated by two annotators who were graduate students in Cognitive Science Department of Middle East Technical University. As the expected process of an any doubly annotation process, the annotators conducted the annotations without consulting with their pairs. After the annotations were completed, inter-annotator agreement is calculated based on the exact match criterion; that is to say when the annotations of two annotators are exactly matching, the annotation is assigned 1; otherwise 0 is assigned to the annotation (Miltsakaki et al., 2004). Inter-annotator agreement results are given in the table 3.3 (Zeyrek & Kurfalı, 2017):

Table 3.3: Inter-annotator agreement results in TDB 1.1 (Zeyrek & Kurfalı, 2017)

Sense	Explicit	Implicit	AltLex
Level-1 (class)	88.40%	85.70%	93.90%
Level-2 (type)	79.80%	78.80%	79.50%
Level-3 (subtype)	75.90%	73.10%	73.40%

Regular meetings were held to discuss the disagreements of the annotators. In these meetings, a gold standard annotation was produced. If any new decision was made and the guidelines were affected from this change, a new annotation cycle to find and correct the inconsistencies in the old data started.

3.1.1 Sense Annotation

TDB 1.1 assigns sense to discourse relations using PDTB's hierarchical sense tags. Below, examples for class level senses can be found. The sense hierarchy can be seen in figure 3.1.

- Temporal: This tag is used when the arguments of the relation are related temporally.

  … **kadın terasa çıkmadan** önce *kaçıyordu*.

  Before **the woman come up to the terrace**, *he escaped*.

  ("TEMPORAL: Asynchronous: precedence", fileNO: 00001131 in TDB)

- Contingency: When there is a causal or conditional relation between the arguments, the tag Contingency is used.

  Bir çözüm bulmalıydı. Yoksa **delirecekti**.

  He had to find a solution. Otherwise, **he would go mad**.

  ("CONTINGENCY: Condition", fileNO: 00001231 in TDB)

- Comparison: The class level tag Comparison is used when the differences between the situations of the arguments are highlighted.

[1] This annotation process was a part of the project numbered BAP-07-04-2015-004 supported by METU, Informatics Institute.

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

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

("COMPARISON: Contrast: opposition", fileNO: 00001131 in TDB)

- Expansion: When the relation of an argument is expanded, the class level tag Expansion is used.

  Kızınca bir çocuk kadar bile olamazdım. **Bir tenekeye tekme atamazdım** mesela.

  I couldn't even be childish when I get angry. **I couldn't kick a tin**, for example.

  ("EXPANSION: Instantiation", fileNO: 00001131 in TDB)

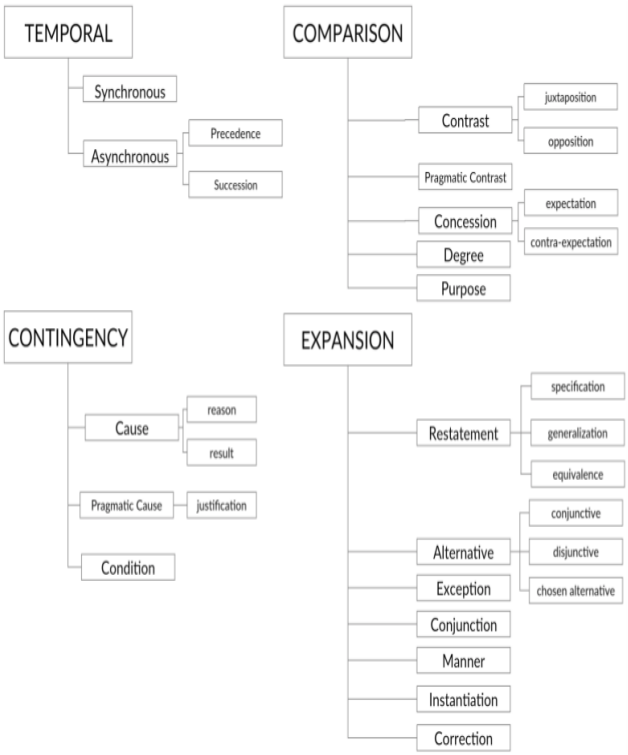

Figure 3.1: Hierarchy of Sense tags in TDB 1.1 (Kurfalı, 2016)

TDB, in a similar way to PDTB, works on ambiguous connectives. In TDB, such connectives are annotated for their all possible senses. For example, "ancak" in Turkish is polysemic in various ways. It is observed to be Comparison, Contingency and Expansion. It is still polysemic in type classification. In Comparison, it is observed that it can be Contrast or Concession. So, "ancak" is an ambiguous discourse connective in Turkish.

3.2 Feature Set

This section gives detailed explanation and examples on the features employed for this study. As as already been mentioned, this study makes use of two features: Centering Theory and the semantic information conveyed by morphosyntactic characteristics of discourse units. These features are added directly to the output of the TDB 1.1 data manually. A sample data after Centering Theory and morphosyntactic feature additions is provided in Appendix C.

As the arguments of a relation can be composed of more than one sentence or clause, a decision on which sentence to annotate was made here. Centering Theory analyzes centers of the sentences that are immediately following each other. Thus, it is decided that if an argument contains more than one sentence, the last sentence of the argument which is earlier in the text and the first sentence of the argument which is later in the text are taken into account. For instance, the first argument of the example 33 consists of two clauses in it. So, only the clause ending the argument is taken into consideration.

(33) *Ben İngilizce bilmiyordum, İngilizce ismini şu anda hatırlamıyorum*, ama **Nurhan tayf analizlerinden söz eden İngilizce bir fizik kitabı edinmişti**.
I didn't know English, I can't remember its English name now but **Nurhan acquired an English physics book mentioning about the spectrum analysis**.
("COMPARISON: Pragmatic Contrast", fileNO: 00050220 in TDB)

This decision is kept consistent for all the feature annotations, which means that the same decision is applied when the data is annotated morphosyntactically.

3.2.1 Centering Theory

The transition types (Brennan et al., 1987) for all explicit connectives (i.e. *continue*, *retain*, *smooth-shift* and *rough-shift*) are doubly-annotated in the same way as TDB annotations. Both annotators are graduate students at Cognitive Science Department of Middle East Technical University. Both annotators are self-trained on Centering Theory using existing resources (Turan, 1996; Grosz et al., 1983; Grosz & Sidner, 1986; Grosz et al., 1995; Brennan et al., 1987) to learn about the transitions. The annotators held regular meetings to do test annotations together, and then they annotated some toy data without consulting each other. When the inter-annotator agreement between the annotators in this toy data reached 80%, they started to annotate the data to be used for this study independently. Both annotators annotated the whole data. The result of inter-annotator agreement, in an exact-match style, is 85.6%. Exact-match style requires giving points to the annotations. When annotations of both annotators are the same for one specific case, the annotation is assigned 1; otherwise it is assigned 0 (Miltsakaki et al., 2004).

The following examples illustrate the Centering transitions added to the data:

- Continue: $Cb(U_{n+1}) = Cb(U_n)$ which is also the most highly ranked element of $Cf(U_{n+1})$. This happens when the center is preserved.

  (34) *Elleri titreyerek cebindeki çakmağı çıkardı* ve **zorlukla yere eğilerek mumu aramaya başladı**.

He took the lighter off from his pocket and **looked for the candle with great difficulty on the floor**.
("TEMPORAL: Asynchronous: precedence", fileNo: 00001231 in TDB)

- Retain: $Cb(U_{n+1})$ f= $Cb(U_n)$ but the entity is not the most highly ranked element of $Cf(U_{n+1})$ which is the case when the center is retained.

  (35) *Nakışçı çok yaşlı bir adamdı* ve **kendisine kızı yardımcı oluyordu**.
  The embroiderer was a very old man and **her daughter was helping him**.
  ("EXPANSION: Conjunction", fileNo: 00001131 in TDB)

- Smooth-Shift: $Cb(U_{n+1}) \neq Cb(U_n)$ but $Cb(U_{n+1}) = Cp(U_{n+1})$ which means that the center of the previous sentence is no more the center for the current sentence but the preferred center of the current sentence is likely to be the center for the next one.

  (36) *Evler kerpiçten*, ama **tapınağın temelinde taş var**.
  The houses are made of mudbrick but **there is stone in the base of the temple**.
  ("COMPARISON: Contrast: juxtaposition", fileNo: 00013112 in TDB)

- Rough-Shift: $Cb(U_{n+1}) \neq Cb(U_n)$ but $Cb(U_{n+1}) \neq Cp(U_{n+1})$ which means that the center of the previous sentence is not the center of the current sentence and the preferred center of this current sentence is not the highest ranked element in the preferred entities set.

  (37) *Vazgeçmek kolaydı, ertelemek de*. Ama **tırmanmaya başlandı mı bitirilmeli**!
  It was easy to give up, to delay, as well. However, **once one start to climb, it should be completed**.
  ("COMPARISON: Pragmatic Contrast", fileNo: 00001131 in TDB)

Transitions from Centering Theory are manually added as features to the data. We expected the Center transitions to correlate with some CLASS level sense tags. The Expansion tag is used when the discourse is expanded and its narrative or exposition is moved forward (Prasad et al., 2007). In a similar way, the Continue transition is observed when the center of the earlier sentence is preserved. Thus, we expect a correlation between the Continue transition and the Expansion class.

The Center Transitions are expected to have an effect on the Comparison class, as well. The Comparison tag applies when the prominent differences between the arguments are highlighted. This is expected to correlate with the Shifting transition from Centering Theory as Shifting happens when the center of the preceding sentence is different from the current sentence, which signals a difference between the situations arguments.

3.2.2 Morphosyntactic Features

In Turkish, numerous forms of semantic information related to discourse is conveyed by morphosyntactic features. Thus, in addition to Centering Theory transitions, morphosyntactic features are used in this study. They are named Predicate, Tense, Aspect, Modality, Person, Polarity and Suspended Affixation. To decide the features and their values, Göksel & Kerslake (2005) was followed. The features are explained in detail below.

- Predicate: This feature is used to see if the predicate of the argument is Verbal or Nominal. There is a predicate type feature for each argument in a relation; thus, the features Pred1 and Pred2 are used for the predicate of the first argument and for the second one respectively. This feature can be "Verbal" or "Nominal". In addition to these two features, there is also a feature called SamePred, which is True when both predicates are the same type and False when they are different.

  The idea of assigning predicate type as a feature for this study is that the sameness or the distinctness of the predicates can help distinguishing Expansion and Comparison classes, respectively. Also, we coded this difference since we expected verbal predicates to overweight nominal predicates in the Temporal class.

  (38) *Bu geç vakitte bile ortalıkta dolaşanlar vardı*, ama **Halil'le ilgilenmiyorlardı**.
  There were people around even at this late hour, but **they didn't care about Halil**.
  ("COMPARISON: Concession", fileNo: 00001131 in TDB)
  Pred1: "Nominal" annotated for '*vardı*', Pred2: "Verbal" annotated for '*ilgilenmiyorlardı*'

  (39) *Yapılarını kerpiçten yapıyorlar*, sonra **taşı kullanmayı öğreniyorlar**.
  They are building their constructions with mudbrick, then **they learn how to use stone**.
  ("TEMPORAL: Asynchronous: precedence", fileNo: 00013112 in TDB)
  Pred1: "Verbal" annotated for '*yapıyorlar*', Pred2: "Verbal" annotated for '*öğreniyorlar*'

- Tense: This is the tense of the verb in each argument. Again, this has two variants: Tense1 and Tense2. This feature can get only one of the following values: "Present", "Past" and "Future". Additionally, there is a binary feature called SameTense, which can be True or False depending on the equality of both tenses.

  The Tense feature is expected to have a positive effect on the Temporal class as the Temporal tag is assigned to a relation when the arguments of the relation are in a Temporal relation. We especially expect to see an increase in prediction score of the Temporal class with the help of SameTense feature with the True value for the Temporal class. As a negative effect, this feature may yield to misclassifying other classes as the Temporal class depending on the similarity of the tenses of the arguments.

  (40) *... biyolojik strese girer* ve **ölürler**.
  ... they undergo biological stress and **they die**.
  ("TEMPORAL: Asynchronous: precedence", fileNo: 00011112 in TDB)
  Tense1: "Present" annotated for '*girer*', Tense2: "Present" annotated for '*ölürler*'

- Aspect: Aspect is about the completeness and recurrence of a situation. A situation can be completed, ongoing or it can have a recurring pattern. There is one aspect feature for each argument: Aspect1 and Aspect2 and these features' possible values are "Perfective", "Progressive" and "Habitual". This feature is added to the feature set whenever an aspective feature is available. Also, there is a feature called SameAspect, which is assigned True when both aspects are the same and False when they are different.

We expect this feature to have significance in the Comparison, Expansion, Contingency classes. Especially, the SameAspect feature is expected to have a positive effect on Expansion when its value is True, which means that we expect an increase in the classification score when SameAspect is True. When SameAspect is False, which means the aspects of the main verbs of the arguments are different from each other, we expect a raise in the classification score of Comparison as the Comparison class means a difference in the situations of the arguments in a relation. We expect to observe an increase in prediction score of the Contingency tag when the arguments have verbs with the specified aspect features. The reason for this is that we expect such arguments to be in causal relations between each other, which is a signal for the Contingency tag.

(41) *Kaptandı*, ama **yüzme bilmezdi amcam**.
He was a captain but **my uncle didn't know how to swim**.
("COMPARISON: Concession", fileNo: 00003121 in TDB)
Aspect2 : "Habitual" annotated for '*bilmezdi*'

(42) *Kapım çalındığında karşımda duran yüzü hatırlamaya çalışıyordum*, ama **öyle zorlanıyordum** ki, eski dostum adını ve nerede tanıştığımızı söylemek zorunda kalıyordu.
When my door is knocked, I was trying to recognize the face in front of me but **I was having such a hard time** that my old friend felt obliged to his name and where we met.
("COMPARISON: Contrast: juxtaposition", fileNo: 00001131 in TDB)
Aspect1: "Progressive" annotated for '*çalışıyordum*', Aspect2: "Progressive" annotated for '*zorlanıyordum*', SameAspect: "True"

(43) *Zincirleri çözülmemişti*, ama **her an koparabilirlerdi**.
Their chains were not yet loosed but **they could be broken any moment**.
("COMPARISON: Contrast: opposition", fileNo: 00001131 in TDB)
Aspect1: "Perfective" annotated for '*çözülmemişti*'

- Modality: Modality is a linguistic feature about the attitude of the verb regarding possibility and necessity. It can also be about the desire of a speaker. There is one modality feature corresponding to each argument: Modality1 and Modality2. The available values for modality feature are "Ablitative" and "Necessity". The additional feature is SameModality and gets the value of True when both modalities are the same and False when they are different.

  In a similar way to the Aspect feature, we expect the Modality feature to be affecting the Expansion and Comparison classes positively. The sameness of Modality feature between the arguments is expected to increase the prediction score of the Expansion class while the distinctness is expected to have a positive impact on the Comparison class.

  (44) *İnsanların onu bulamayacağı bir yere gitmeliydi*, ama **Mihriban'ı da bırakamazdı**.
  He had to go to a place where no one could find him but **he couldn't leave Mihriban, as well**.
  ("COMPARISON: Pragmatic Contrast", fileNo: 00001231 in TDB)
  Modality1: "Necessity" annotated for '*gitmeliydi*', Modality2: "Ablitative" annotated for '*bırakamazdı*', SameModality: "False"

- Person: Person is the feature used to understand the person agreement on the verb in each argument. There are Person1 and Person2. These features can be assigned first-person singular “1sg”, second-person singular “2sg”, third-person singular “3sg”, first-person plural “1pl”, second-person plural “2pl” or third-person plural “3pl”. There is also SamePerson feature, which is not automatically filled depending on Person1 and Person2 but manually annotated. The reason behind this is that each person’s being 3rd singular, for example, would not necessarily mean that they are the same subjects.

  We expect that the Person feature, when specifically both arguments’ subjects are the same, to affect the Expansion class in positive way as it is observed that in the Expansion class the subjects of the arguments are frequently the same.

  (45) *Mimarlık açısından çok önemli*, çünkü **bir yapı malzemesini başka bir malzemeyle beraber kullanmayı, ilk defa burada görüyoruz**.
  This is very important regarding the architecture because **we, for the first time, see here that one construction material is used with another one**.
  ("CONTINGENCY: Cause: reason", fileNo: 00001231 in TDB)
  Person1: "3sing" annotated for ’*önemli*’, Person2: "1plu" annotated for ’*görüyoruz*’, SameSubject: ”False”

- Polarity: Each argument is given a feature for polarity: Polarity1 and Polarity2. The values are "Positive" and "Negative". There is also a feature called SamePolarity, which gets values of True or False depending on the equality of the polarity of each argument.

  Regarding polarity, the sameness of polarity of the arguments can affect the Expansion class positively while the distinctness is expected to be effective to correctly classify the Comparison class. This is because the Expansion class signals the sameness while the Comparison class signals a difference in the discourse.

  (46) *Paniğe kapıldı*. Aslında **böyle şeyler onu asla korkutmazdı**.
  He panicked. Actually, **such things never frightened him**.
  ("EXPANSION: Exception", fileNo: 00001231 in TDB)
  Polarity1: "Positive" annotated for ’*kapıldı*’, Polarity2: "Negative" annotated for ’*korkutmazdı*’, SamePolarity: ”False”

- Suspended Affixation: Suspended affixation describes situations when the scope of an affix covers two or more words (Lewis, 1967). Kabak (2007) shows that the suspension of affixes in Turkish happens if the conjunct is a morphological word, which is a word form that can occur in isolation in a text. This feature conveys a purely morphological information unlike other features employed in this study, which convey semantic information through morphosyntax.

  The absence or existence of suspended affixation in the data is indicated with the feature SuspendedAffixation, which becomes "True" when there is a suspended affixation and "False" otherwise. This characteristic feature of Turkish is taken into account when encountered in the predicates of the arguments which signals a shared tense or person agreement between the predicates of the arguments.

  We expect that suspended affixation is found most frequently in the Expansion class. This is because, as highlighted by Kabak (2007) as well, we find suspended affixation mostly in coordinate constructions and coordinate constructions are signals of the

Expansion class. This syntactic phenomena indicates that the discourse is expanded. When the tense is elided, the person agreement on the predicate is also elided, which means that there is little change in the discourse. And we expect that this is the characteristic feature of the Expansion class.

(47) Ante *bu genç ressamın her sırrını biliyor* ve **onu çok iyi anlıyordu**.
Ante *knew every secret of this young painter* and **understood him very well**.
(fileNo: 00001131 in TDB)
SuspendedAffixation: ”True” annotated for ’*biliyor(du)*’

With these features employed, this study will attempt to find out if these morphosyntactic features can help improving the accuracy of sense prediction of explicit discourse connectives. These features are all tested alone and in combination with others to see if any of them is more effective when with another feature.

This study is limited with relations the relations whose arguments contain tensed clauses and all relations composed of converbs are left out of the scope. The reason is that we can’t apply the same feature set to both kinds of the relations. In relations with converbs, Arg2, which is syntactically bound to the connective, is never a tensed Predicate so Tense, Person, Predicate etc. features as explained in 3.2.2 would not be applied to those.

The data is first annotated for its tensed clauses. All relations are analyzed and filtered depending on their syntactic structure, i.e. they are added a label called TensedClause. This label’s value is "True" if both arguments of a relation contain a tensed clause, and it is assigned "False" if otherwise. Example 48 is annotated as TensedClause="True".

(48) *Bu geç vakitte bile ortalıkta dolaşanlar vardı*, ama **Halil’le ilgilenmiyorlardı**.
There were people around even at this late hour, but **they didn’t care about Halil**.
(TensedClause="True", fileNo: 00001131 in TDB)

This chapter provided information on data and how data is annotated with the feature set which this study employs. In the next chapter, Chapter 4, methodology and experimental design is elaborated.

CHAPTER 4

METHODOLOGY

In this chapter, the classifier used in experiments, the experimental design and how the data is processed for this study are described.

4.1 Maximum Entropy

A classifier, as its name suggests, is a machine learning tool to take data and classify parts of it into classes. The classes that the data is tried to be fit into are limited and known and as the instances given to be trained are labeled, this kind of models are called supervised (Kotsiantis et al., 2007). For this study, a supervised machine learning model, Maximum Entropy (MaxEnt), will be used for multiclass classification. MaxEnt is a probabilistic classifier that depends on the features to classify the given data. The difference between MaxEnt and other classifiers is that MaxEnt does not assume that the features in the given set are conditionally independent of each other. It enables the features to be implemented freely. MaxEnt goes for minimum assumptions, which allows the classification without knowledge of the prior distribution. That is to say maximum entropy is a method learning a distribution with the vector of given features over the labels. It is calculated as follows:

$$P(c|d) = \frac{exp[\Sigma_i \lambda_i f_i(c,d)]}{\Sigma_{c'}[exp[\Sigma_i \lambda_i f_i(c',d)} \tag{4.1}$$

which contains λ, the weight vector; c referring to the CLASS level sense tag and d meaning the discourse relation.

The results of the classification are shown by giving Accuracy, Recall, Precision and F1 scores for each class. Also, overall calculated micro- and macro averaged F1 scores are provided. A brief explanation is given here to get a better understanding of how these numbers are calculated. Before understanding the calculations, the units of these calculations will be mentioned. These are TP, TN, FP and FN where:

- "TP" stands for True Positives, which are correctly predicted positive values.
- "TN" stands for True Negatives, which are correctly predicted negative values. This happens when the actual value of the class and the predicted value are both "no".
- "FP" stands for False Positives, which occurs when the actual value is "no" but the predicted one is "yes".

- "FN" stands for False Negatives, which happens when the actual value is "yes" but the system cannot predict it.

Accuracy is simply a ratio of correctly predicted observation to the total observations. It is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations and calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4.3}$$

Recall is the ratio of correctly predicted positive observations to the all observations in actual class and calculated as follows:

$$Recall = \frac{TP}{TP + FN} \tag{4.4}$$

F1 is the weighted average of Precision and Recall. Then, the overall averages are divided into two: micro- and macro-averaged F1 scores. For micro-averaged F1 score, each test item counts equally while for macro-averaged F1 score, each class counts equally.

There are various classification methods depending on the number of the classes available in the training and test data. These are *one-class*, *binary* and *multi-class* classification methods.

One-class classification is used when there are no negative examples (Khan & Madden, 2009). As an example we can imagine a situation where one tries to recognize A instances and distinguish them from $\hat{A}$ (i.e. not A) instances. One-class classification is used when the training data does not include any instances of $\hat{A}$. The system will only know what it is looking for but will not have any prior knowledge of what to exclude as it is trained solely on the positive instances. To the best of our knowledge, this method has not been employed for an automatic prediction system of discourse connectives.

Binary classification is used when the data has both positive and negative examples. This method is employed by many researchers including but not limited to Pitler et al. (2008), Pitler et al. (2009), Zhou et al. (2010), Patterson & Kehler (2013) and Kurfalı (2016). While doing a binary classification on automatically prediction of class level sense tags, the classifier is run four times as there are four classes, namely Temporal, Contingency, Comparison and Expansion. Each time the classifier runs, it trains on the data where the positive instances are labelled, for example, as *Temporal* and negative instances are labelled as *non-Temporal* for the Temporal class. As Pitler et al. (2009) also mentioned, binary classification is different from annotation process conducted by human annotators. That is because human annotators, while annotating a piece of text, analyze the sentences and choose between all the possibilities on how to label a relation. What is more similar to the realistic approach to the data here is a multiclass classifier.

Multiclass classifiers train and test on data where there are more than two classes. The classifier decides on what class an instance belongs to. This method is more similar to what human annotators do while annotating a text for discourse connectives. This is employed by Pitler et al. (2008) and from the results of binary and four-way classifications, it can be inferred that the predictive power of the system drops when switched to four-way classifier.

4.2 Experimental Design

This study makes use of relations with tensed clauses as arguments only. After filtering the data to include only such relations and exclude the ones containing converb as their connective, each relation, gold annotation of Centering Theory and singly annotated morphosyntactic characteristics of the arguments are registered to the XML-formatted output files of DATT. Later on, as this study does not have any feature based on the surface form of the arguments, only connective and annotated sense tags are extracted to create training and test data. In order to do that, a script is written. Extensive care was taken to create the training and test data in a balanced way. For every sense annotation, the threshold of 5 is applied, which means that there must be at least 5 instances of a specific connective and sense combination for those instances to be in the data for this study. The training data is composed of 80% of such instances and the remaining 20% is chosen for the test data. That is to say for every instance in the test data, there must be four of the same instance in the training data.

In order to assess the effect of individual features, training and test files involving only one feature is also prepared. To be able to do a MaxEnt classification, Stanford Classifier is used (Manning & Klein, 2003). Thus, the method of multiclass classification is employed as opposed to one-class and binary classification methods.

This chapter provided information on MaxEnt classifier, how its results are calculated and the experimental design of this study. The next chapter, Chapter 5, presents results of the classification experiments.

CHAPTER 5

RESULTS

In this chapter, results of the experiments conducted to automatically predict the sense tags of explicit discourse connectives are provided. The chapter is divided into nine sections each of which stands for a feature or a combination of features that were tested. While the first section gives the prediction score of the classifier obtained from the classifications based solely on the explicit discourse connective, the rest of the chapter includes the prediction scores based on Centering Theory transitions feature, Tense-Aspect-Modality feature, predicate type feature, person agreement feature, polarity feature, suspended affixation feature, in this order. After giving results on these features individually, the prediction rates of the classification based on all morphosyntactic features combined are given. Finally, the results of the classification based on all features are provided.

5.1 Baseline scores based on connective only

The prediction rate of the classifier based solely on the explicit discourse connective is presented in Table 5.1. That is to say for each class there was only one feature available, which was the the explicit discourse connective. Micro- and macro-averaged F1 scores are given in Table 5.2. The highest prediction rate belongs to the Contingency class where the lowest belongs to the Expansion class.

These scores are taken as the baseline for this study and the results of any feature or any combination of features are compared to this baseline.

Table 5.1: Prediction rate calculated with only connective feature

	Accuracy	Precision	Recall	F1
Temporal	0.958	1.000	0.636	0.778
Comparison	0.695	0.525	1.000	0.688
Expansion	0.695	0.714	0.286	0.408
Contingency	0.958	1.000	0.765	0.867

Table 5.2: Micro- and macro-averaged F1 scores of the prediction with only connective feature

Accuracy/micro-averaged F1	65.2%
Macro-averaged F1	68.5%

5.2 Transitions of Centering Theory

When transitions of Centering Theory are added to the feature set in addition to the explicit discourse connectives, there is an increase of prediction score in two classes, i.e. Comparison and Expansion. This can be seen in Table 5.3. The F1 scores increased $\sim 6\%$ for the Comparison class and $\sim 22\%$ for the Expansion class. The difference is resulting from the fact that number of both TNs and TPs for Comparison and Expansion classes are higher with this feature. That is to say some instances which were mis-classified between these two classes are now correctly classified. This feature seems not to be affecting Temporal and Contingency classes.

Overall results are also affected from these changes: micro-averaged F1 score increased by 7.4% to 72.6% and macro-averaged F1 score increased by $\sim 6.9\%$ to 75.4%. The increases are shown with bold font in the tables.

Table 5.3: Prediction rate calculated with connective and transition feature

	Accuracy	Precision	Recall	F1
Temporal	0.958	1.000	0.636	0.778
Comparison	0.800	0.651	0.875	**0.747**
Expansion	0.737	0.656	0.600	**0.627**
Contingency	0.958	1.000	0.765	0.867

Table 5.4: Micro- and macro-averaged F1 scores of the prediction with connective and transition features

Accuracy/micro-averaged F1	**72.6%**
Macro-averaged F1	**75.4%**

5.3 Tense, Aspect and Modality features

Tense, aspect and modality features are tested together as a feature set. This feature set is the only feature affecting the Temporal class in a positive way. F1 score of the Temporal class increased by 0.5%. This results from the fact that this feature set helps the classifier to correctly predict more Temporal class items correctly, which means number of TPs in the Temporal class is higher with this feature. Similar to the Temporal class, F1 scores of Comparison and Expansion classes also increase; by 3% and 15% respectively. However, this feature set has a negative effect on the prediction of the Contingency class as F1 score of the Contingency class decreases by $\sim 2\%$. These scores are provided in Table 5.5.

Also, the overall increase of $\sim 4\%$ in micro- and macro-averaged F1 scores can be seen in Table 5.6. The increases are shown with bold font in the tables.

Table 5.5: Prediction rate calculated with connective and TAM feature

	Accuracy	Precision	Recall	F1
Temporal	0.947	0.818	0.750	**0.783**
Comparison	0.766	0.609	0.875	**0.718**
Expansion	0.713	0.654	0.486	**0.557**
Contingency	0.957	1.000	0.733	0.846

Table 5.6: Micro- and macro-averaged F1 scores of the prediction with connective and TAM features

Accuracy/micro-averaged F1	**69.1%**
Macro-averaged F1	**72.6%**

5.4 Predicate Type feature

The Predicate type feature increases F1 scores of Comparison and Expansion classes and decreases the F1 score of the Temporal class. However, it does not affect the Contingency class. The increase for Comparison and Expansion classes are $\sim 3\%$ and $\sim 17\%$ respectively. The decrease in the Temporal class is $\sim 7\%$ and this results from mis-classification of one instance of explicit discourse connective "önce". When this is classified wrong, the number of TPs decrease while the number of FNs increase. The detailed results of this run can be seen in Table 5.7.

Also, the increase of $\sim 4\%$ in overall micro- and macro-averaged F1 scores are provided in Table 5.8. The increases are shown with bold font in the tables.

Table 5.7: Prediction rate calculated with connective and predicate type feature

	Accuracy	Precision	Recall	F1
Temporal	0.947	1.000	0.545	0.706
Comparison	0.747	0.577	0.938	**0.714**
Expansion	0.737	0.708	0.486	**0.576**
Contingency	0.958	1.000	0.765	0.867

Table 5.8: Micro- and macro-averaged F1 scores of the prediction with connective and predicate type features

Accuracy/micro-averaged F1	**69.4%**
Macro-averaged F1	**71.5%**

5.5 Person Agreement feature

The person agreement feature affects the Expansion class positively. F1 score of the Expansion class increased by $\sim 11\%$. Other three classes are affected negatively from this feature. While F1 scores of Comparison and Contingency classes decrease slightly, by 0.5% and 2.8% respectively, there is a sudden drop in the Temporal class. This drop is due to the decrease of TPs. No item in the test set could be identified as a member of the Temporal class with this feature. The detailed results can be seen in Table 5.9. The increases are shown with bold font in the tables.

Micro- and macro-averaged F1 scores are present in Table 5.10.

Table 5.9: Prediction rate calculated with connective and person agreement feature

	Accuracy	Precision	Recall	F1
Temporal	0.884	1.000	0.000	0.000
Comparison	0.726	0.560	0.875	0.683
Expansion	0.663	0.548	0.486	**0.515**
Contingency	0.947	0.929	0.765	0.839

Table 5.10: Micro- and macro-averaged F1 scores of the prediction with connective and person agreement features

Accuracy/micro-averaged F1	61.0%
Macro-averaged F1	50.9%

5.6 Polarity feature

Polarity tags of the predicates of the arguments in a discourse relation is observed to be affecting all classes. While F1 scores of Comparison and Expansion classes increase, there is a decrease in F1 scores of the Temporal and Contingency classes. The increase in the Comparison and Expansion classes are 0.6% and $\sim 17\%$, respectively; and, the decrease in Temporal and Contingency classes are $\sim 7\%$ and 5.5% respectively. The detailed results are provided in Table 5.11.

Also, slightly increased micro- and macro-averaged F1 scores can be seen in Table 5.12. The increases are shown with bold font in the tables.

Table 5.11: Prediction rate calculated with connective and polarity feature

	Accuracy	Precision	Recall	F1
Temporal	0.947	1.000	0.545	0.706
Comparison	0.768	0.625	0.781	**0.694**
Expansion	0.695	0.588	0.571	**0.580**
Contingency	0.937	0.867	0.765	0.812

Table 5.12: Micro- and macro-averaged F1 scores of the prediction with connective and polarity features

Accuracy/micro-averaged F1	**67.3%**
Macro-averaged F1	**69.8%**

5.7 Suspended Affixation feature

The suspended affixation feature slightly affects the Comparison and Expansion classes positively while it does not affect Temporal and Contingency classes. F1 scores of the Comparison and Expansion classes increase by 2.3% and 1.5%, respectively. The detailed results of this run are provided in table 5.13.

Also, micro- and macro-averaged F1 scores are present in Table 5.14. The increases are shown with bold font in the tables.

Table 5.13: Prediction rate calculated with connective and suspended affixation feature

	Accuracy	Precision	Recall	F1
Temporal	0.958	1.000	0.636	0.778
Comparison	0.726	0.552	1.000	**0.711**
Expansion	0.684	0.647	0.314	**0.423**
Contingency	0.958	1.000	0.765	0.867

Table 5.14: Micro- and macro-averaged F1 scores of the prediction with connective and suspended affixation features

Accuracy/micro-averaged F1	**66.3%**
Macro-averaged F1	**69.4%**

5.8 All Morphosyntactic features combined

When all morphosyntactic features are combined as a feature set and the system is trained on this, there are improvements in Comparison and Expansion classes. The increases in F1 scores of Comparison and Expansion are $\sim 4\%$ and $\sim 12\%$, respectively. While the F1 score of the Temporal class decreases by 24%, the F1 score of the Contingency class is not affected. The reason behind the drop in the Temporal class is that the items belonging to the other classes are mis-classified as Temporal, which means that there is an increase in FPs while the numbers of TPs and FNs are the same with the ones obtained from the experiment when the classifier is trained only on explicit discourse connectives. The detailed scores are provided in Table 5.15.

Micro-averaged F1 score slightly increases by 1.1% and macro-averaged F1 score slightly decreases by 2.1%, as can be seen in table 5.16. The increases are shown with bold font in the tables.

Table 5.15: Prediction rate calculated with connective and all morphosyntactic feature

	Accuracy	Precision	Recall	F1
Temporal	0.874	0.467	0.636	0.538
Comparison	0.779	0.622	0.875	**0.727**
Expansion	0.716	0.682	0.429	**0.526**
Contingency	0.958	1.000	0.765	0.867

Table 5.16: Micro- and macro-averaged F1 scores of the prediction with connective and all morphosyntactic features

Accuracy/micro-averaged F1	**66.3%**
Macro-averaged F1	66.4%

5.9 All features combined

When both the morphosyntactic features and the transition features of Centering Theory are combined with the connective feature, the classification results for Comparison and Expansion classes display an increase. While F1 score of the Comparison class increases by $\sim 3.5\%$, F1 score of the Expansion class increases by $\sim 15\%$. F1 scores of Temporal and Contingency classes drop by 19.5% and $\sim 8\%$, respectively. The drop in these classes result from the change in the number of FPs of these classes, which means that more items are mis-classified as Temporal or Contingency in this run. The detailed results are provided in Table 5.17. Micro-averaged F1 score of this run is 66.3%, which is 1.1% higher than the baseline. However, macro-averaged F1 score drops by 2.4% to 66.2%. Micro- and macro-averaged F1 scores are present in Table 5.18. The increases are shown with bold font in the tables.

Table 5.17: Prediction rate calculated with connective and all features combined

	Accuracy	Precision	Recall	F1
Temporal	0.895	0.538	0.636	0.583
Comparison	0.789	0.650	0.812	**0.722**
Expansion	0.716	0.654	0.486	**0.557**
Contingency	0.926	0.812	0.765	0.788

Table 5.18: Micro- and macro-averaged F1 scores of the prediction with connective and all features combined

Accuracy/micro-averaged F1	**66.3%**
Macro-averaged F1	66.2%

This chapter provided results of the experiments conducted based on several features. The results are given in tables with their Accuracy, Precision, Recall and F1 scores. The highest micro- and macro-averaged F1 scores are obtained with the feature of transitions of Centering Theory. In the next chapter, Chapter 6, each class level sense tag is elaborated on which feature or features contribute them the most with examples.

CHAPTER 6

DISCUSSION

In this chapter, the analysis of the results obtained from the classification and reported in Chapter 5 is provided. The results are interpreted for each CLASS level. Then, a brief and more general analysis is provided. For each sense, the feature set which yields the most increase over the baseline is discussed with examples.

6.1 Temporal

The Temporal class has the least number of instances among the four classes, which makes it difficult to learn. It has 50 examples overall; 40 in training and 10 in test data in this study. The only feature set affecting the learning of the Temporal class positively is the TAM feature set as expected. This feature set includes tense, modality and aspect features of the arguments and if any of these features are the same between the two arguments, there is another feature added to the set signing the sameness. Otherwise, a feature showing the difference is added to the set. The reason behind this increase can be explained with the plain fact that both the Temporal class and TAM feature set are time-related.

For the Temporal class, the most weighted feature in this set seems to be "same_asp", which stands for the sameness of the aspective feature of the two arguments. An example of this feature in a Temporal sense relation is given in example 49.

(49) *"Benim amcam kaptandı," diyor durup dururken* ve **Nesli'nin araştıran şımarık gözlerine bakıyor**.
"My uncle was a captain", he was saying suddenly and **was looking at Nesli's investigating spoilt eyes**.
("TEMPORAL: Asynchronous: precedence", fileNo: 00003121 in TDB)

In addition to the "same_asp" feature, the arguments' being in future tense seems to be an important feature for the Temporal class. Example 50 is provided to demonstrate this feature.

(50) *Kapalı, ağır, dokunsan ağlayacak bir hava olacak* ve **ben bir başıma basacağım pedallara**.
There will be an overcast, heavy weather which would cry once you touch it and **I will pedal all alone**.
("TEMPORAL: Synchronous", fileNo: 00003121 in TDB)

In the case of polarity feature, one instance of *önce* 'before' is misclassified as the Comparison class. The features available for this instance are "Positive_Arg1", "Negative_Arg2" and "diff_pol". Combinations of these features are observed to be more frequently found for the Comparison class. The misclassified instance is given in 51.

(51) **Nevtan** önce **bu çizimlerin ne olduğunu anlayamamıştı**. *Biraz düşündükten ve hayal gücünü zorladıktan sonra vazgeçti, uzun bir uykuya daldı.*
First, **Nevtan couldn't understand of these drawing**. *After some thinking and forcing her imagination, she gave up, fell in a long sleep.*
("TEMPORAL: Asynchronous: succession", fileNo: 00001231 in TDB)

Also, although it was expected that predicate type feature could have some positive effect on classifying the Temporal class, it has negative effect. The reason behind this is again an instance of "önce" which is misclassified as Comparison. The distinct feature about this specific example is that it has different predicate types. It seems from the classification that the Temporal class has a trend to have same predicates (i.e. the SamePred feature with True value). This specific instance of "önce" is outside of the trend; thus, it is misclassified.

As one last note about the Temporal class, the F1 score drops when all morphological features and all features are taken into consideration when trying to classify the Temporal class. The reason behind this drop is the fact that the items belonging to the other classes are misclassified as Temporal, which means that there is an increase in False Positives

6.2 Comparison

For the Comparison class, the feature helping the classification the most is the transitions feature from Centering Theory. This was expected as both Comparison and Center Shifting are a sign of the difference between the arguments.

The most obvious weight is "SShift" standing for Smooth-Shifting in Centering Theory. One example is given in 52.

(52) *Onu da zorluyorlar.* Ama **gitmiyor**.
They force him, as well. But **he doesn't go**.
("COMPARISON: Concession", fileNo: 00003121 in TDB)

Different morphological features highlighting the differences between the arguments were expected to help to increase the overall score of the classification of the Comparison class. Following the transitions feature of Centering Theory, the biggest improvement is observed to happen when the classifier is trained with with all morphosyntactic features combined. The most affecting feature is "diff_tense" standing for the existence of a difference in tenses of the two arguments.

(53) *Şimdilik yerleşimde, 3. tabakaya ait, fazla geniş bir kazı alanı yok.* Ama **2. tabakaya ait çalışmalarımızda ilerledik**.

For now, there's no very large axcavation field belonging to the 3. layer in the settlement. But **we have made progress in our works in the 2. layer**.
("COMPARISON: Contrast", fileNo: 00013112 in TDB)

6.3 Expansion

The Expansion class is best learned by the classifier when the transitions feature of Centering Theory is provided. The learning improved by 21.9% when compared to the baseline score where the classifier was trained only on explicit discourse connectives. The most weighted feature seems to be "Cont", which stands for Center Continuation in Centering Theory where the center of the current and previous sentences are the same and the center of the current sentence is the most highly ranked element. The Continue transition is used when the center is preserved, which denotes that the discourse is moving forward in the way the Expansion tag specifies.

One such example from the data is given in 54.

(54) *Kızınca bir çocuk kadar bile olamazdım*. **Bir tenekeye tekme atamazdım** mesela.
I couldn't even be childish when I get angry. **I couldn't kick a tin**, for example.
("EXPANSION: Instantiation", fileNO: 00001131 in TDB)

The highest frequency of Center Continuation belongs to the Expansion class among four classes. The numbers of the transitions for the class Expansion found in the data are provided in table 6.1.

Table 6.1: Number of transitions of the instances in the Expansion class

Continuation	Retaining	Smooth-Shifting	Rough-Shifting	**TOTAL**
99	5	3	37	**144**

The Expansion class is observed to be affected positively with features same modality, same predicate and suspended affixation. These features were expected to have a positive effect on the Expansion class as these features help to expand the discourse from one argument to another. So, our expectations for the Expansion class are fulfilled.

6.4 Contingency

No feature set presented in this study increased the F1 score of the Contingency class. The weakest feature set attempted for the Contingency class is when all the fetures are combined. When compared to the score calculated as the baseline for this study (by providing only the connectives), there are more FPs in the run when all features are provided to the classifier. The worst performing feature set seems to be the combined set of the following features.

- RShift: This feature stands for Rough-Shifting in the center.

- Verbarg1 Verbarg2 same_pred: These features are used when both arguments' predicates are verb.

- Presentarg1 Presentarg2 same_tense: These features mean that both arguments' predicates are in present tense.

- 3singarg1 3pluarg2 diff_subj: These features signify that subjects of the arguments are different from each other. The first argument's subject agreement is 3rd person singular while the second one's is 3rd person plural; thus, diff_subj is assigned.

When together, these features seem to be existing not only for some items in the Contingency class but also for items belonging to the Expansion and Comparison classes. Such an example from the Contingency class is provided in example 55.

(55) *Recaizade Ekrem, kendisinden önceki Tanzimat romanlarında, romanın bütününü oluşturmuş bir kurguyu bir paragrafta özetledikten sonra, sanki bu kurgunun geleneksel trajik sonuyla yetinmeyerek romanını başka sonuçları incelemek üzere geliştirir.* Çünkü **Bihruz Bey'in yetim kalışı, tahsilini yarım bırakışı, ve sefahata dalarak servetini tüketişi, romanın daha ilk yirmi sayfasında özetlenir**.
Recaizade Ekrem, after summarizing a fiction creating a whole novel in one paragraph in the Tanzimat novels before him, develops his novel to analyze other results as if he couldn't content himself with this fiction's traditional tragic result. Because, **Mr. Bihruz's being left as an orphan, leaving his education in half and consuming his fortune on his pleasures were all summarized in the first twenty pages of the novel**.
("CONTINGENCY: Cause: reason", fileNO: 00027213 in TDB)

6.5 General Analysis

This study helped to improve the automatic prediction score for three class level sense tags, namely Temporal, Comparison and Expansion. In Table 6.2, overall scores based on features for each class is provided.

Table 6.2: F1 scores of predictions score with each feature for each class

	Temporal	**Comparison**	**Expansion**	**Contingency**
Connective only	0.778	0.688	0.408	0.867
Centering Theory	0.778	0.747	**0.627**	0.867
TAM	**0.783**	0.718	0.557	0.846
Predicate	0.706	0.714	0.576	0.867
Person	0.000	0.683	0.515	0.839
Polarity	0.706	0.694	0.580	0.812
Suspended Affixation	0.778	0.711	0.423	0.867
All morphosyntactic Features	0.538	0.727	0.526	0.867
All Features	0.583	**0.772**	0.557	0.788

As can be seen from the table and as explained in sections above for each class, Tense-Aspect-Modality feature gave the best results for the Temporal class, which is quite expected. For the Comparison class, the highest score is obtained when all features are combined. While every and each one of the feature set increased the prediction score for Expansion, the highest score is obtained from transitions of Centering Theory. All these correct classifications are compatible with our expectations.

Regarding misclassifications, Table 6.3 is provided for an overall analysis. This table gives the percentages of the misclassifications, which means that it provides the information on how much a specific class is misclassified as another class.

Table 6.3: Percentages of Class-specific misclassifications

		classified classes			
		Temporal	**Comparison**	**Expansion**	**Contingency**
gold classes	**Temporal**	-	22.8	77.2	0
	Comparison	12.9	-	83.9	3.2
	Expansion	6.5	90.5	-	3
	Contingency	0	41.6	58.4	-

It was observed that the Temporal, Comparison and Contingency classes were most frequently misclassified as the Expansion class. While for Temporal and Comparison classes the percentage of this kind of misclassification over all misclassifications is high, 77.2% and 83.9% respectively, it is relatively lower for the Contingency class, 58%. The reason behind this misclassification trend can be interpreted as related to the feature set. The Expansion class has been the only class which is affected positively from every and each of the features used in this study. Also, the number of items in the Expansion class is the highest among four classes. This fact can also be affecting the classification.

On the other hand, items in the Expansion class are most frequently misclassified as Comparison. The percentage of such misclassifications over all misclassifications of the Expansion class is 90.5%. This percentage implies that the Expansion class is more easily distinguished from Temporal and Contingency classes than the Comparison class.

In this chapter, results provided in the previous chapter (i.e. Chapter 5) are discussed under five sections, four of which correspond to one class level sense and their examples. The remaining section is provided to give a brief overall analysis. The following chapter, Chapter 7, summarizes and concludes the thesis by giving overall discussion, limitations and suggestions for future work.

CHAPTER 7

CONCLUSION

This thesis attempted to build an automatic prediction system to predict the sense tags of the explicit discourse connectives found in manually annotated TDB 1.1 corpus. This process is the first step of the automatic discourse parsing pipeline. In this study, only explicit discourse connectives are analyzed and only the relations with arguments composed of tensed clauses are taken into consideration. To our best knowledge, this study is the first one to attempt to automatically predict the sense tags of explicit discourse connectives in Turkish.

In this study, a maximum entropy classifier is employed. This classifier is run on training and test data, which were balanced in the items they contained. The data was annotated with connectives, transitions of Centering Theory and morphosyntactic features such as predicate type and tense of the arguments in a discourse relation.

The current study achieved to increase the prediction score for three class level senses; namely Temporal, Comparison and Expansion. For the Temporal class, a slight increase is observed to have happened when Tense, Aspect and Modality features are provided as a combined feature set.

In the Comparison class, the most effective feature comes from the transitions of Centering Theory. Smooth-Shifting seems to be a strong indication of the Comparison class level sense.

Finally, with the experiments in this study, the best achievement is obtained for the Expansion class. The baseline score for the Expansion score is 40.8% when the classifier is trained only with the connective itself for each relation. This score increased by 21.9% to 62.7% with the transitions features of Centering Theory. The strongest feature seems to be the Center Continuation feature, which is found frequently in the environment of the Expansion class. The Expansion class is signaling the further expansion of the discourse. Thus, by this nature of the Expansion class, it was expected to contain Center Continuation. That is because Center Continuation signals the maintenance and expansion of the center, as well.

In spite of the expectations, no features or feature sets in this study could achieve an increase in the prediction score of the Contingency class. When the classifier is run with only the explicit discourse connective feature, a prediction score of 86% is obtained, which means that only the connective feature is a good indicator for the sense tag of the connective.

This study is conducted on TDB 1.1 annotations. Similar studies that are carried out for PDTB have the opportunity to work on greater amount of annotations. Depending on the amount of the data, the results may give more insights about the discourse relations. Also, this study is limited with only explicit discourse relations. The feature set used for this study may con-

tribute to prediction of implicit discourse connectives, as well. Finally, this study attempted to predict class level sense tags. Some features or feature combinations may be confusing when put together for class level sense tag and may be associated more meaningfully for type or sub-type level sense tags, which are more informative in terms of the sense of discourse relations.

Bibliography

Aktaş, B., Bozsahin, C., & Zeyrek, D. (2010). Discourse relation configurations in turkish and an annotation environment. In *Proceedings of the fourth linguistic annotation workshop* (pp. 202–206).

Asher, N. (2012). *Reference to abstract objects in discourse* (Vol. 50). Springer Science & Business Media.

Brennan, S. E., Friedman, M. W., & Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of the 25th annual meeting on association for computational linguistics* (pp. 155–162).

Demirşahin, I., Sevdik-Çallı, A., Balaban, H. Ö., Çakıcı, R., & Zeyrek, D. (2012). Turkish discourse bank: Ongoing developments. In *Proc. lrec 2012. the first turkic languages workshop.*

Forbes-Riley, K., Webber, B., & Joshi, A. (2005). Computing discourse semantics: The predicate-argument semantics of discourse connectives in d-ltag. *Journal of Semantics*, *23*(1), 55–106.

Göksel, A., & Kerslake, C. (2005). *Turkish: A comprehensive grammar.* Psychology Press.

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive science*, *17*(3), 311–347.

Graesser, A. C., Millis, K., & Graesser, A. (2011). Discourse and cognition. *Discourse Studies: A Multidisciplinary Introduction, London: Sage*, 126–142.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st annual meeting on association for computational linguistics* (pp. 44–50).

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational linguistics*, *12*(3), 175–204.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, *21*(2), 203–225.

Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in english.* Routledge.

Hobbs, J. R. (1978). *Why is discourse coherent* (Tech. Rep.). SRI INTERNATIONAL MENLO PARK CA.

Kabak, B. (2007). *Turkish suspended affixation.* Walter de Gruyter.

Khan, S. S., & Madden, M. G. (2009). A survey of recent trends in one class classification. In *Irish conference on artificial intelligence and cognitive science* (pp. 188–197).

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). *Supervised machine learning: A review of classification techniques.*

Kurfalı, M. (2016). *Automatic sense prediction of implicit discourse relations in turkish* (Unpublished doctoral dissertation). MIDDLE EAST TECHNICAL UNIVERSITY.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Lewis, G. (1967). *Turkish language.* Oxford: Oxford University Press.

Lin, Z., Kan, M.-Y., & Ng, H. T. (2009). Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1-volume 1* (pp. 343–351).

Lin, Z., Ng, H. T., & Kan, M.-Y. (2014). A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, *20*(2), 151–184.

Manning, C., & Klein, D. (2003). Optimization, maxent models, and conditional estimation without magic. In *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology: Tutorials-volume 5* (pp. 8–8).

Marcu, D. (1997). The rhetorical parsing of natural language texts. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics* (pp. 96–103).

Marcu, D., & Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 368–375).

Miltsakaki, E. (2002). Toward an aposynthesis of topic continuity and intrasentential anaphora. *Computational Linguistics*, *28*(3), 319–355.

Miltsakaki, E., Prasad, R., Joshi, A. K., & Webber, B. L. (2004). The penn discourse treebank. In *Lrec.*

Patterson, G., & Kehler, A. (2013). Predicting the presence of discourse connectives. In *Emnlp* (pp. 914–923).

Pitler, E., Louis, A., & Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the joint conference of the 47th annual meeting of the acl and the 4th international joint conference on natural language processing of the afnlp: Volume 2-volume 2* (pp. 683–691).

Pitler, E., & Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the acl-ijcnlp 2009 conference short papers* (pp. 13–16).

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

Prasad, R., Joshi, A. K., & Webber, B. L. (2010). Exploiting scope for shallow discourse parsing. In *Lrec.*

Prasad, R., McRoy, S., Frid, N., Joshi, A., & Yu, H. (2011). The biomedical discourse relation bank. *BMC bioinformatics*, *12*(1), 188.

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

Ramesh, B. P., & Yu, H. (2010). Identifying discourse connectives in biomedical text. In *Amia annual symposium proceedings* (Vol. 2010, p. 657).

Robaldo, A. L. E. M. L., Prasad, A. J. R., Dinesh, N., & Webber, B. (2008). The penn discourse treebank 2.0. In *Proceedings of the sixth international conference on language resources and evaluation (lrec'08), marrakech, morocco, may. european language resources association (elra). http://www. lrec-conf. org/proceedings/lrec2008.*

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

Say, B., Zeyrek, D., Oflazer, K., & Özge, U. (2002). Development of a corpus and a treebank for present-day written turkish. In *Proceedings of the eleventh international conference of turkish linguistics* (pp. 183–192).

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Strube, M. (1998). Never look back: An alternative to centering. In *Proceedings of the 17th international conference on computational linguistics-volume 2* (pp. 1251–1257).

Turan, Ü. D. (1996). Null vs. overt subjects in turkish discourse: A centering analysis. *IRCS Technical Reports Series*, 95.

Walker, M., Cote, S., & Iida, M. (1994). Japanese discourse and the process of centering. *Computational linguistics*, *20*(2), 193–232.

Walker, M. A., Joshi, A. K., & Prince, E. F. (1998). *Centering theory in discourse*. Oxford University Press.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347–354).

Zeyrek, D., & Kurfalı, M. (2017). Tdb 1.1: Extensions on turkish discourse bank. *LAW XI 2017*, 76.

Zeyrek, D., & Webber, B. L. (2008). A discourse resource for turkish: Annotating discourse connectives in the metu corpus. In *Ijcnlp* (pp. 65–72).

Zhou, Z.-M., Xu, Y., Niu, Z.-Y., Lan, M., Su, J., & Tan, C. L. (2010). Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 1507–1514).

Appendix A

DATA DISTRUBITON

Connective	Temporal	Comparison	Contingency	Expansion	**TOTAL**
ve	12	0	15	124	**151**
ama	0	109	0	7	**116**
çünkü	0	0	36	0	**36**
sonra	25	0	0	0	**25**
ancak	0	19	1	0	**20**
ayrıca	0	0	0	19	**19**
oysa	0	14	0	0	**14**
fakat	0	10	0	0	**10**
önce	9	0	0	0	**9**
aslında	0	2	0	7	**9**
böylece	0	0	8	0	**8**
dolayısıyla	0	0	5	0	**5**
ya da	0	0	0	4	**4**
hem	0	0	0	4	**4**
ardından	3	0	0	0	**3**
halbuki	0	3	0	0	**3**
bir yandan	2	0	0	1	**3**
örneğin	0	0	0	3	**3**
dahası	0	0	0	3	**3**
yoksa	0	0	2	1	**3**
mesela	0	0	0	2	**2**
gene de	0	2	0	0	**2**
ne	0	0	0	2	**2**
adeta	0	0	0	1	**1**
yine de	0	1	0	0	**1**
ne ki	0	1	0	0	**1**
ne var ki	0	1	0	0	**1**
yalnız	0	1	0	0	**1**
iken	0	1	0	0	**1**
veya	0	0	0	1	**1**
TOTAL	**51**	**164**	**67**	**179**	**461**

Table A.1: Data Distrubition of TDB 1.1 Regarding CLASS level sense tags

Appendix B

SAMPLE DATT OUTPUT

```
<Relation genre="novel" note="precedence" sense="Temporal: Asynchronous: precedence"
type="EXPLICIT">
        <Conn>
            <Span>
                <Text>ve</Text>
                <BeginOffset>10595</BeginOffset>
                <EndOffset>10597</EndOffset>
            </Span>
        </Conn>
        <Mod/>
        <Arg1>
            <Span>
                <Text>Elleri titreyerek cebindeki çakmağı çıkardı</Text>
                <BeginOffset>10551</BeginOffset>
                <EndOffset>10594</EndOffset>
            </Span>
        </Arg1>
        <Arg2>
            <Span>
                <Text>zorlukla yere eğilerek mumu aramaya başladı</Text>
                <BeginOffset>10598</BeginOffset>
                <EndOffset>10641</EndOffset>
            </Span>
        </Arg2>
        <Supp1/>
        <Supp2/>
        <Shared/>
        <Supp_Shared/>
    </Relation>
```

Figure B.1: Sample TDB Output from DATT annotation tool

Appendix C

SAMPLE XML FILE WITH FEATURES ANNOTATED

The XML file provided in Appendix B is enriched with the features explained in Section 3.2. The features are added to DATT produced XML files as attributes by one or two annotators where necessary.

```
<Relation genre="novel" note="precedence" sense="Temporal: Asynchronous: precedence"
type="EXPLICIT" TensedClause="True" trans="Cont" Pred1="Verb" Pred2="Verb"
Tense1="Past" Tense2="Past" Person1="3sing" Person2="3sing" SameSubject="True"
Polarity1="Positive" Polarity2="Positive" SuspendedAffixation="False">
        <Conn>
            <Span>
                <Text>ve</Text>
                <BeginOffset>10595</BeginOffset>
                <EndOffset>10597</EndOffset>
            </Span>
        </Conn>
        <Mod/>
        <Arg1>
            <Span>
                <Text>Elleri titreyerek cebindeki çakmağı çıkardı</Text>
                <BeginOffset>10551</BeginOffset>
                <EndOffset>10594</EndOffset>
            </Span>
        </Arg1>
        <Arg2>
            <Span>
                <Text>zorlukla yere eğilerek mumu aramaya başladı</Text>
                <BeginOffset>10598</BeginOffset>
                <EndOffset>10641</EndOffset>
            </Span>
        </Arg2>
        <Supp1/>
        <Supp2/>
        <Shared/>
        <Supp_Shared/>
    </Relation>
```

Figure C.1: Sample annotation output after the features for this study are annotated