# AUTOMATIC USAGE DISAMBIGUATION OF THE ENCLITIC DA IN TURKISH

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

ELIF EBRU ERSÖYLEYEN

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

AUGUST 2022

**AUTOMATIC USAGE DISAMBIGUATION OF THE ENCLITIC DA IN TURKISH**

submitted by **ELIF EBRU ERSÖYLEYEN** 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** _____

Dr. Ceyhan Temürcü
Head of Department, **Cognitive Science** _____

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

**Examining Committee Members:**

Prof. Dr. Gülsün Leyla Uzun
Linguistics, Ankara University _____

Prof. Dr. Deniz Zeyrek Bozşahin
Cognitive Science, Middle East Technical University _____

Assist. Prof. Dr. Umut Özge
Cognitive Science, Middle East Technical University _____

**Date: 16.08.2022**

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

**Name, Surname: Elif Ebru Ersöyleyen**

**Signature :**

# ABSTRACT


**AUTOMATIC USAGE DISAMBIGUATION OF THE ENCLITIC DA IN TURKISH**

Ersöyleyen, Elif Ebru

M.S., Department of Cognitive Science

Supervisor: Prof. Dr. Deniz Zeyrek Bozşahin

August 2022, 68 pages

Discourse is composed of several constituents that yield coherency in a structural form. One of the interesting aspects of discourse is discourse connectives and their contribution to discourse structure. They are lexico-syntactic elements that signal a semantic relation between two discourse units (clauses and sentences). Clitics are morphemes that are phonologically dependent on the lexical item to which they are attached, but have separate syntactic forms, and carry no meaning by themselves. They can function as a discourse connective in several languages; for example in Cuzco Quechua the clitic *pas* [1] and in Turkish, *dA* [2] can signal multiple senses, and have features that distinguish them from affixes and other words [3]. *dA* is essentially a focus-associated enclitic that also has discourse functions in Turkish, conveying contrast, addition, causal and condition senses. In other words, just like other linguistic expressions, *dA* is subject to ambiguity and creates a challenge in natural language automatization tasks. The aim of this study is two-fold: (a) to analyze the linguistic behavior of *dA*, annotating its discourse and non-discourse occurrences in corpora of written Turkish, (b) to develop machine learning models that distinguish its discourse usage from its non-discourse usage - i.e., its discourse connective vs. focus enclitic role. The thesis describes the annotation study and the machine learning models, which uses linguistic features. The results of our machine learning experiments show that we can disambiguate the discourse usage of *dA* with an F1-score of 0.83 in free texts.



Keywords: Clitic *dA*, Discourse Connective, Focus, Corpus Study, Natural Language Processing

# ÖZ

## TÜRKÇE'DEKİ ENKLİTİK DA'NIN SÖYLEM VE SÖYLEM DIŞI ROLÜNÜN OTOMATİK BELİRLENMESİ

Ersöyleyen, Elif Ebru

Yüksek Lisans, Bilişsel Bilimler Bölümü

Tez Yöneticisi: Prof. Dr. Deniz Zeyrek Bozşahin

Ağustos 2022, 68 sayfa

Söylem, yapısal bir biçimde tutarlılık sağlayan birçok dilsel bileşenden oluşur. Söylemin göze çarpan yönlerinden biri, bağlaçların katkısı ve söylem yapısındaki işlevleri ve de davranışlarıdır. İki söylem birimi (cümlecikler ve cümleler) arasında anlamsal bir ilişkiye işaret eden sözcüksel unsurlardır. Klitikler, sesbilimsel olarak sözcüksel öğeye bağlı, ancak ayrı sözdizimsel biçimleri olan ve kendi başlarına bir anlam taşımayan biçimbirimlerdir. Bu öğeler pek çok dilde söylem bağlacı işlevi görebilirler, örn. Cuzco Quechua klitik *pas* [1], Türkçe *dA*[2], çoklu anlamsal bağlantılar göstererek diğer kelimelerden ve eklerden ayırt edici işlevleriyle ayrılırlar [3]. *dA*, Türkçe'de söylem seviyesinde, örn. açıklama, zıtlık ve nedensellik belirtebilen çeşitli işlevlere sahip söylem dışı kullanımı da olan odak ilişkili enklitiktir. Başka bir deyişle, *dA* çoklu fonksiyon gösterebilmesi nedeniyle berimsel doğal dil analizlerinde doğru anlamını belirlemede güçlük yaratabilmektedir. Bu çalışmanın amacı ise iki aşamalıdır: (a) *dA*'nın yazılı Türkçe derlemindeki oluşumlarını açıklayarak dilsel davranışını analiz etmek, (b) söylem kullanımını söylem dışı kullanımından, yani, odak parçacığı rolünden, ayıran makine öğrenimi modelleri geliştirmek. Bu tez metin işaretleme çalışmasını ve dilbilimsel özellikleri kullanan makine öğrenimi modellerini açıklamaktadır. Makine öğrenmesi deneylerinden elde edilen sonuçlardan görülmektedir ki *F1-değeri* 0,83 oranıyla *dA*'nın söylem ve söylem dışı kullanımını ayrıştırabilmekteyiz.

Anahtar Kelimeler: Klitik *dA*, Söylem Bağlaçları, Odak, Derlem Çalışması, Doğal Dil İşleme

*To my beloved GSD boy, Dük,*
*may all the tennis balls of the world be yours...*

# ACKNOWLEDGMENTS

I would like to thank all my lecturers, whom I took lessons during my graduate education, for enabling me to determine my research interests thoroughly and improving my way of thinking in the field of Cognitive Science while teaching me graduate-level concepts.

I am beyond grateful and would like to offer my deepest gratitude to my supervisor Prof. Dr. Deniz Zeyrek Bozşahin for her contributions to my thesis. Through reflecting her rich seam of information on the field and bestowing her endless support and guidance, she provided me with many inventive thoughts and enlightened my knowledge on the topic of discourse.

I would love to express my gratitude and endless thanks to Fırat Öter for his creative ideas, continuous assistance, and for always initiating intriguing questions that led me to deepest thoughts on the topic.

I am truly thankful to Prof. Dr. Gülsün Leyla Uzun and Assist. Prof. Dr. Umut Özge for their comments and suggestions on the thesis.

I cannot thank my family enough for everything they have done for allowing me to gain all the experience throughout my life.

I want to thank Assoc. Prof. Dr. Aziz F. Zambak and Assoc. Prof. Dr. Samet Bağçe for encouraging and recommending me to apply for doing masters in Cognitive Science, which completely enriched my perspective at all degrees.

From the bottom of my heart, thank you for being in my corner :f.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ADD | Additive |
| ADJ | Adjective |
| ARG1 | Argument 1 |
| ARG2 | Argument 2 |
| C | Connective |
| CDTB | Chinese Discourse Tree Bank |
| CoNLL | Computational Natural Language Learning (Conference) |
| CRF | Conditional Random Field |
| CSV | Comma-Separated Values |
| DC | Discourse Connective |
| DISRPT | Discourse Relation Parsing and Treebanking |
| FDTB | French Discourse Tree Bank |
| JSON | JavaScript Object Notation |
| LSTM | Long Short-Term Memory |
| NDC | Non-Discourse Connective |
| NLG | Natural Language Generation |
| NLP | Natural Language Processing |
| NP | Noun Phrase |
| PDTB | Penn Discourse TreeBank |
| POS | Part of Speech |
| PREV | Previous |
| PROPN | Proper Names |

| | |
|---|---|
| RNN | Recurrent Neural Network |
| RST | Rhetorical Structure Theory |
| S | Sentence |
| SOV | Subject-Object-Verb |
| SVM | Support Vector Machine |
| STTS | Stuttgart-Tubingen Tagset |
| TCL | Turkish Connective Lexicon |
| TCD | Turkish Connective Disambiguator |
| TDB | Turkish Discourse Bank |
| TED-MDB | TED-Multilingual Discourse Bank |
| T-TED-MDB | Turkish section of TED-Multilingual Discourse Bank |
| VP | Verb Phrase |

# CHAPTER 1

# INTRODUCTION

Discourse connectives, e.g. subordinating and coordinating conjunctions *and*, *but*, *because*, and adverbs, e.g., *however*, are lexico-grammatical items that can indicate a semantic relationship between two discourse units such as verb phrases, clauses or sentences. Discourse connectives are one of the fundamental constituents in forming the discourse structure [6]. Thus, they have occurrences in two different linguistic levels: they can be used as conjunctions or adverbs, or as discourse connectives. Discourse connectives should be differentiated from *discourse particles*, short words that are responsible for maintaining the flow in discourse structure, e.g. *well* in English, *şey, yani, işte* in Turkish as well as clitics such as *dA*.

1. He bought a pair of glasses and a scarf. (non-discourse connective)

2. *He bought a pair of glasses* and **[he] left the store**. (discourse connective conveying additive sense)

For example, *and* is a discourse connective with an extensive use in English. While *and* has non-discourse usage in 1 by coordinating two nouns, in 2, it has an occurrence in discourse level, and conveys an additive sense that holds between two clauses.

Clitic is a term that is used for a morpheme, phonologically dependent on the bounded word, but has a separate syntactic form, and carries no meaning by itself. Turkish has several clitics functioning as a marker of yes/no questions (e.g., *mI*), copula (e.g., (*y*)-*sA*) or a discourse connective (*dA*) or subordinator (e.g., ki) [7]. The enclitic *dA*[1] can have various functions owing to its diverse positioning and semantics. [2] Since *dA* attaches to the word immediately preceding it, the preceding word has a significant role when determining the role of *dA*. In other words, the function of *dA* can be determined by the preceding word, thus, its location in the sentence. The following sentences are from [7] and [9] illustrating various clitics in Turkish.

3. Gitse miydik?
   Should we have gone? (the marker of yes/no questions)

[1] *dA*: Upper case letter "A" is used for representing the alteration of the vowel ("-e", "-a") with respect to the last syllable of the preceding word [8]

[2] see Section 3.

4. **Biraz erken gel**-se-**n** *iyi ol-ur*. (a copular clitic indicating conditional sense)
   *It would be good* if **you came a bit early**.

5. *Sana bunu söylüyorum* ki **sonradan şaşırmayasın**. (a subordinator connecting two clauses)
   *I'm telling you this* so that **you won't be surprised later**.

6. **Ayşe'ye temizlik yapmasını söyledim** de, *yapacak mı bilmiyorum*. (discourse connective conveying contrastive sense)
   **I've told Ayşe to do [some] cleaning**, but *I don't know if she will do it*.

7. *Seni seviyorum dedi*, **ben** de **inandım**. (discourse connective conveying additive sense)
   *S/he said "I love you"*, and **I believed him/her**.

Anderson [10] has shown that in languages such as Finnish and Northern Vogul, clitics can function as discourse connectives. The examples below are retrieved from his work [10].

8. Kalle on-kin ostanut auton (Finnish)
   Karl is-also bought car
   Karl also bought a car

9. χum jot-ke åleγėm ńaurem χani (Northern Vogul)
   man with-if I live child clings
   If I live with a man, the child belongs to me

This thesis is based on the understanding that Turkish is yet another language where clitics can serve as discourse connectives, as shown by the discourse role of *dA* in sentences 6, 7, 11. Most clitics do not have a single function, and their role in texts can differ (compare sentences 10 - 11). Hence, we need a comprehensive analysis on the linguistic behavior of these lexical items so that we can build a sound model regarding their multifunctionality. Given the lack of a comprehensive analysis of clitics in Turkish and their ambiguity, a computational approach to clitics becomes a complex task. This thesis aims to move towards filling this gap by focusing on *dA*.

10. Düşün bir iş bul artık. İlk parayla bir çeki kömür alacağım. Sana da lastik çizme.[11] (non-discourse connective)
    Just think, we've got a job now. With the first money (that comes in), I'll buy a load of coal. And a pair of wellies for you.

11. **Markete gittim** de *etrafta kimseyi görmedim*. (discourse connective)
    **I went to the market** but *[I] did not see anybody around*.

As already explained, discourse connectives are susceptible to usage ambiguity [6]. That is, given that discourse connectives are lexico-grammatical elements, they not only have a discourse connective (DC) role but also a grammatical role. For example, in English, *once* can function either as a temporal DC or a sentential adverb meaning *formerly* [12]. Human

language users can easily distinguish between these two functions but this ambiguity is a challenge for automatic discourse processing. Previous usage disambiguation tasks over connectives have been done for many languages. For English, well-known works involve Pitler and Nenkova [12] and Lin et. al. [13]. In other languages, similar tasks have been done by Shih and Chen [14] in Chinese, by Laali and Kosseim [15] in French, and by Dipper and Stede [16] in German. In Turkish, the only work we are aware of is the PhD thesis by Başıbüyük [4]. To facilitate Natural Language Processing (NLP) tasks such as text summarization, automatic translation, knowledge extraction, etc., usage ambiguity tasks have to be conveyed over clitics, just like other types of discourse connectives.

The goal of this study is to analyze the linguistic behavior of *dA* through annotated corpora[3] of written Turkish, and propose machine learning models that distinguish the discourse usage of *dA* from its non-discourse, normal clitic usage. By describing and disambiguating the discourse role of this clitic, the thesis aims to contribute to understanding clitics in Turkish as well as the computational analysis of Turkish discourse.

The outline of the rest of this thesis is as follows: Chapter 2 presents a review of the literature, explaining, so far, what has been done on the related topic, focusing on Turkish NLP, usage disambiguity of DCs, clitics and their function in other languages.

Chapter 3 will be focusing on the linguistic behavior of *dA* and a comprehensive examination of its functions by examples from the existing literature and our data. In this way, its usage ambiguity will be demonstrated. Then, a brief discussion of the syntax of *dA* will be introduced, describing where *dA* is placed and which functions it can bear regarding its position in a sentence.

In chapter 4, the annotation study and the process of data preparation will be described in detail. After introducing the characteristics of the data, the difficulties and problems confronted during various stages of annotation and data preparation will be discussed together with how we overcame the difficulties and how the annotation process is carried out. Computational methods for linguistic processing of the data will also be presented in this section.

In chapter 5, the experiments that are conducted for disambiguating the discourse usage of *dA* will be explained. We will describe our experimental setup by introducing the feature set and the learning algorithms we used.

The results of the experiments will be presented in Chapter 6. Tables and measurement values (e.g. F-scores) will be presented with a commentary on the results. An overall evaluation of the study will also be done in this part.

Finally, a general discussion on the topic will be provided in chapter 7. We will summarize the contribution of the thesis to the field, and what can be done in future works as a follow-up of this study will be discussed.

[3] Turkish Discourse Bank 1.1, TS Corpus v2.

# CHAPTER 2

# LITERATURE REVIEW

In this chapter, the studies on which the thesis was built will be mentioned, and it is briefly discussed what have been done related to clitic *dA* from linguistics and computational perspectives.

## 2.1 Clitics

There are many studies on clitics. In the field of linguistics, researchers have examined the behavior of clitics in general [17] [18], and from particular theoretical perspectives with examples from various languages [10] and [19].

It is known that clitics are difficult to distinguish from other grammatical elements. Due to their morphological and syntactic differences [17] and due to their exclusive behavior, an unconventional classification is necessary. Although clitics do not fall under a particular syntactic class, e.g. inflectional or derivational affixes, Zwicky states they are close to inflectional affixes in several ways [17].

Inflectional suffixes are attached to the syntactic form of their host without affecting the type of the word they are bound to. For example, the plural suffix in Turkish is an inflectional suffix.

1. çocuk -lar -ın -a (Göksel and Kerslake [7])
   child -PL -2SG.POSS -DAT
   NUMBER (-lar) POSSESSION (-ın) CASE (-a)
   to your children

Derivational suffixes, on the other hand, may alter the category of the word they belong to, for example, they derive a noun out of a verb and they become part of the new word that is formed. Some examples of Turkish derivational suffixes are as follows:

*gel*-ir (income) – derivation of a noun from a verb
*kız*-ış- (excalate) – derivation of a verb from a verb
*yan*-aş- (approach) – derivation of a noun from a verb [7]
*göz*-lük (glasses) – derivation of a noun from a noun

The difference between clitics and suffixes is that unlike suffixes, they are not included in the form of the word they are attached to. According to their type, i.e. pre-, en-, pro- and post- [17], they are positioned before or after their host. Although they are syntactically separate, they are phonetically sensitive to voice changes in their host. In other words, the harmony of their own sound changes according to the vowel of the word they are attached to. As a result, clitics are separated from suffixes in syntax, especially derivationals because they do not make any radical changes of their host's word form but usually influence the degree of semantics. However, they are phonetically similar to suffixes in that they are sensitive to their host's vocalizations.

## 2.2 Additive Markers and Clitics in Other Languages

Anderson [10] pointed out that there are two types of clitics: those functioning as grammatical markers and those that indicate semantic relations. The connective function of clitics seems to be neglected due to their definition as *particles*.

Examining the cross-linguistic variances of additive particles and clitics, Forker [2] shows that additive markers in many languages work similar to the additive function of Turkish *dA*. Faller [1] examines the functions of Cuzco Quechua *pas*, which resembles the semantics of the additive *dA*. In another study showing the similarity between *dA* and the suffix *-nde* in Cypriot Greek, Pavlou and Panagiotidis [20] state that the clitic *-nde* was borrowed from *dA* in Turkish and that they have similar semantic properties.

### 2.2.1 Multifunctionality of Additive Markers

Additive markers share common features in many languages. Forker states that markers with additive functions in many languages also have other functions. They can serve as an conjunctional adverb 'and then', convey concessive and contrastive senses and express topic switch [2]. These are very similar to the behaviours of *dA*. The examples, i.e. 2, 3, 4 and 5, below are retrieved from Forker [2] to show different functions of clitics in different languages.

2. Central Kurdish [21] (additive marker functions as an additive)
   gā-eke-ān-īš -im de-č-in
   (ox-DEF-PL-ADD-1SG.POSS IND-go.PRS-3PL)
   Not only I will not get the girl, I will also lose my oxen.

3. Sheko [22] (additive marker functions as a topic shifter)
   yordanos ʃik-n-s tə-k-ə, k'orint'os-k'əra ʃaad-n-s tə-k-ə
   Yordanos short-DEF-M COP-REAL-STI Qorinxos-ADD tall-DEF-M COP-REAL-STI
   Yordanos is short, Qorinxos is tall.

4. Udihe [23] (additive marker functions as an adversative (in contrastive context))
   min-du sata bie s'ei-de anči
   me-DAT sugar be.PRS.HAB salt-ADD no
   I have sugar, but no salt.

5. Tibetan [24] (additive marker functions as an adversative (in concessive context))
   khrom-la phyin-na-yang nyo-bya rgyab-ma-song
   bazar-LOC went-SUB-ADD shopping do-NEG-AUX
   Although he went to the bazar, he didn't do any shopping.

Forker also mentions that additive markers are focus-sensitive. This is similar to the approach of Göksel and Özsoy [8], who also mention the same for the clitic *dA*, i.e. it is a focus-sensitive clitic that also functions as an additive marker. They also state that the sense and focus scope of an additive marker such as *dA* depend on its position in the sentence, even, intonation in some languages.

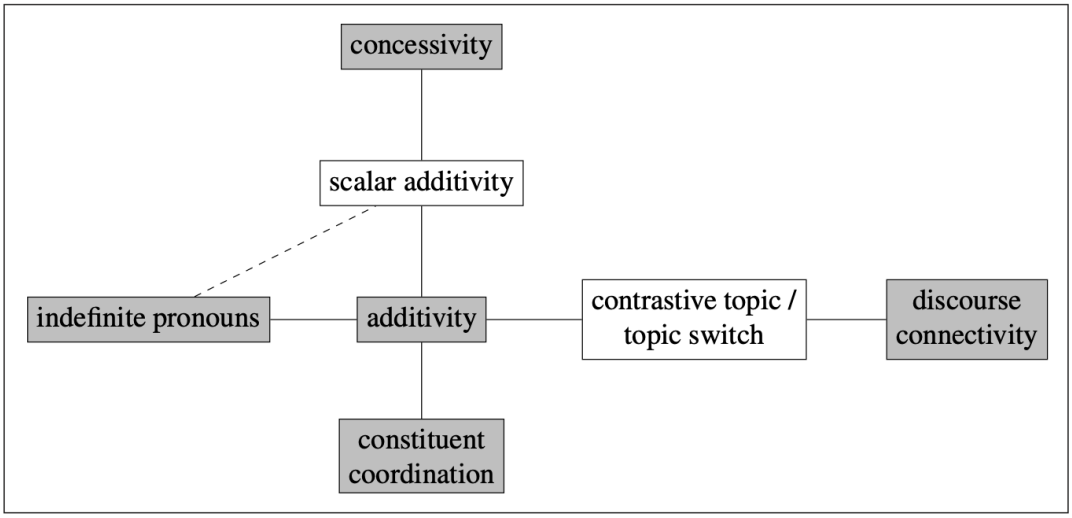


Figure 1: Semantic map of additive markers, grey boxes represent functions of the Cuzco Quecha clitic *pas* ([2]; as cited in [1]

Faller, similar to Forker, shows that additives have versatile uses and they function as discourse connectives in many languages [1], then, she demonstrates a semantic map, see Figure 1, of additive markers retrieved from [2], and defines the functions of the clitic Cuzco Quechua *pas*.

Our annotation study in Section 4 shows that similar to the clitics and additive markers mentioned by Forker and Faller, the Turkish clitic *dA* has the following functions; additivity, concessivity and contingency (condition or cause):

6. *Defterini masanın üstüne koydu*. **Kalemi** de **kenarına iliştirdi**. (additive)
   *S/he put the notebook on the table*, and **[s/he] placed the pen near**.

7. **Geldin** de *bir haber vermedin*. (concessive)
   **You came** yet *[you] did not let me know*.

8. **Anlatayım** da *konuyu öğrenin*. (cause)
   **Let me explain**, so *you get the point*.

## 2.3 Discourse Connectives in Turkish

Since discourse connectives are important components for discourse processing tasks,[1] it has been an essential pursuit to examine the discourse relations in the texts from a computational perspective. There are two recognized and fundamental frameworks that enable computational analysis of discourse relations: The Penn Discourse TreeBank (PDTB) framework [25] and Rhetorical Structure Theory (RST) [26]. The PDTB has a connective-based approach that focuses on revealing the predicate-argument structure of discourse connectives; that is, discourse connectives are considered discourse-level predicates and the text spans linked by discourse connectives are taken as arguments. These are annotated together with the sense of the discourse relation. RST, on the other hand, presents a hierarchical model of discourse structure by introducing the following terms: *nucleus* and *satellite*, which can be inferred as children (arguments) of a node (discourse) [2]. In other words, for instance, if an Elaboration relation (alternatively known as Expansion/Additive relations in different frameworks) holds between two arguments, the nucleus is the argument that gives basic information about the discourse relation, and satellite provides the additional information, presumably, in the adjacent sentence.

*Argument* is the term used to refer to the constitutive units of a discourse connective, which are usually realized as clauses or sentences. Turkish Discourse Bank (TDB, the first discourse corpus of Turkish) also considers nominalizations as potential arguments to discourse connectives, as in example 9 below.

9. Yaklaşan enflasyonun *çözümlenmesi* ya da **çözümlenmemesi** finansal tebdirlere dayanır.
   *Achieving* [to resolve] or **failing to resolve** the impending rise of the inflation depends on financial measures taken.

There are various discourse annotation tasks that have been done for several languages in the light of these two approaches to discourse structure. Arabic [27], Chinese [28], Hindi [29], English [25] and Turkish [30] were analyzed and annotated from the perspective of the PDTB framework, German [31] discourse structure was examined using the tree-like structure of RST, by a combined approach to different frameworks (RST and PDTB). Research was also conducted on German [32] and multiple languages that include German [3] [33] revealing the discourse structure of these languages. Parallel to the growing literature on discourse structure research in other languages, in the past decade, the analysis of different discourse connectives in Turkish, e.g. "ve" (and), "ama" (but), "çünkü" (because) has been done over written corpora by Zeyrek and Webber [34], *ama* and *fakat* in Zeyrek [35]. Demirşahin and Zeyrek [36] carried out a pilot study where explicit discourse connectives are annotated in spoken Turkish.

Zeyrek and Webber [34] defined *dA* as also a discourse connective, translating it as *and*, *but* in their preliminary list of connectives (see their examples 10, 11 below).

[1] We use the term discourse processing to mean NLP at the discourse level.
[2] The terms *children* and *node* are borrowed from Başıbüyük [4].
[3] Danish, English, German, Italian, and Spanish.

10. **Konuşmayı unuttum diyorum** da *gülüyorlar bana*.
**I said I've forgotton to talk** and *they laughed at me*.

11. **Belki bir çocuğumuz olsa onunla oyalanırdım** da *Allah kısmet etmedi*.
**If we had a child I would keep myself busy with her/him** but *God did not predestine it*.

Zeyrek and her team have been working on discourse structure. They have created different versions of Turkish Discourse Bank in the last decade and focused on different kinds of discourse connectives. For example, a set of explicit inter-sentential versus intra-sentential discourse connectives were revealed for the first time by Zeyrek and Kurfalı [37]. They have captured the relations specified by the intra-sentential discourse connectives in Turkish. There is also research that aims to create a lexicon of Turkish discourse connectives (Zeyrek and Başıbüyük [38]). One of the main contributions of this study is the creation of syntactic categories of discourse connectives together with the meanings they convey.

Furthermore, a multilingual discourse corpus where Turkish is included has been created by Zeyrek et. al [39]. With this study, a comprehensively annotated corpus was created in 6 languages. This has become one of the pioneering studies in which Turkish is compared to other languages. Nevertheless, *dA* was not annotated systematically in any of these resources.

This short review shows that even though Turkish discourse connectives have been studied extensively and in different corpora, the clitic *dA* has not been examined in any of these works in detail.

## 2.4 Usage Ambiguity of Discourse Connectives

### 2.4.1 Usage Ambiguity of Discourse Connectives in Different Languages

Usage ambiguity, that is, whether a word such as a conjunction or adverb functions as a discourse connective is an issue that has been observed in the literature [6], and, so far, various studies have been conducted on automatic detection of discourse connectives in order to solve the problem of usage ambiguity. Discourse connectives, such as *and*, *once* and *since* in English can be shown as examples of this situation. For example, *since* can appear both as a temporal preposition (12) and a discourse connective (13 and 14) that conveys causal and temporal senses. Automatic detection of the functions of discourse connectives, which have a significant role in discourse structure, is important for the realization of tasks such as text summarization and natural language generation (NLG), which are aimed to be done for natural language.

12. I have been waiting for this event since 2011. (temporal preposition)

13. *I have been waiting for this event* since **I was a child**. (temporal discourse connective)

14. Since **it was getting late**, *I had to go*. (causal discourse connective)

Recently, for the usage disambiguation of discourse connectives, studies have been carried out in many languages. For English, Pitler and Nenkova [12] extracted syntactic features from derivation trees for the disambiguation task, using the PDTB framework. They stated that the syntax was not used for usage disambiguation before and they saw it as a distinguishing feature in detecting discourse connectives. They created 4 categories regarding the syntactic features of discourse connectives:

- **Self Category** (First node in the tree denotes the syntactic features, e.g. POS tag of the one word connectives)
- **Parent Category** (Direct mother node of the Self Category)
- **Left Sibling Category** (Category containing syntactic features immediately to the left of Self category)
- **Right Sibling Category** (Category containing syntactic features immediately to the right of Self category)

Emphasizing that English is a right-branching language, they stated that the right sibling category is more important than other categories when detecting the discourse connectives, because it represents the features of words that are directly dependent on the discourse connective to be detected.

In their experiments, they used "connective only", "syntax only", "both connective and syntax" and "combination of connective and syntax features (using pairwise interaction of the connective and the individual syntactic feature)" as the feature space. With the F1-score of 94.19%, they achieved the highest result when pairwise interaction was included in the feature space.

Another important work on usage disambiguation of English discourse connectives is Lin et. al.'s [13] study, in which they developed an end-to-end discourse parser model. In order to understand the semantic relations between text units, discourse parsing is used as a natural language automatization task for segmenting sentences, where discourse relations, the argument spans and the discourse connectives are parsed. Similar to Pitler and Nenkova [12], Lin et. al. [13] used the PDTB framework for their model. In addition to knowing that connectives can be detected with their syntactic features, they added the lexico-syntactic features of the connectives, that is, the connective's direct relationship with previous and next words [4], to the feature space. It has been observed in the study that using lexico-syntactic features increases the success rate of the connective classifier of their discourse parser.

Laali et. al. [40] presented a shallow discourse parsing model for English. They used the C4.5 decision tree binary classifier [41], which is a standard machine learning algorithm, for identifying explicit discourse connectives. As a feature set, they used a total of 10 features inspired by the work of Pitler and Nenkova. Their features are as follows in their terms:

[4] C POS, prev + C, prev POS, prev POS + C POS, C + next, next POS, and C POS + next POS, path from C to the root and compressed path from C to the root (if two POS tags are the same it is compressed as one, e.g. -VP-VP- is combined into -VP-) [13].

- 1. The discourse connective text in lowercase.
- 2. The categorization of the case of the connective: all lowercase or initial uppercase.
- 3. The highest node (called the SelfCat node) in the parse tree that covers the connective words but nothing more.
- 4-6. The parent, the left sibling and the right sibling of the SelfCat.
- 7-10. The left and the right word of discourse connective and their parts of speech.

Even though they could not get promising results in full parsing (including detection of argument spans), the models trained on blind data set for explicit discourse connective identification reached the highest F1-score of 0.9020.

The usage disambiguation task was also attempted for languages other than English. Dipper and Stede [16] annotated the 9 ambiguous connective in German as 'DC', and for disambiguating these connectives, they conducted an experiment using the POS tagger (Brill [42] tagger trained on STTS [43][5]) which learns symbolic rewriting rules [6]. Then they disambiguated these connectives in different scenarios.

These scenarios involved: (1) only positive instances where discourse connectives are used and marked as discourse connective, (2) positive instances where STTS-tag is recorded, (3) positive and negative instances where discourse connective and non-discourse connective uses are marked, (4) positive and negative instances where STTS-tag is recorded. As a result, even though using only POS tags that are attached to potential discourse connectives to disambiguate their usage did not yield robust results, they managed to achieve better results both by re-training the data multiple times[7] and adding a basic mapping method as "(DC candidate, STTS tag) to the set of {connective, non-connective}", as STTS tags are good for extracting a decent amount of information of discourse connectives [16].

Shih and Chen [14] presented a study aiming to disambiguate the discourse usage of connectives in Chinese. They used a binary classifier to detect connectives by string matching with connective strings extracted from the connective lexicon of Chinese Discourse Tree Bank (CDTB) [44]. As a feature set, they used multiple ranges: (1) the feature space used by Pitler and Nenkova (P&N), (2) connective itself as a string (CONNECTIVE), (3) the POS tags of potential connectives and all components that are constituents of the connective (POS), (4) the number of components that are constituents of the connective (NUM), (5) vectorial representations of the potential connective and all components that are constituents of the connective (VECTOR). Their model is trained using 10-fold cross-validation, first, using solely individual features, then, in combination of all features with vectors that are created by skip-gram model, yielding a set of words closest to the given word. They achieved the highest F1-score of 0.7506 for usage disambiguation task, when all features are combined with skip-gram.

[5] A POS tagger for German.

[6] For example, "S -> NP VP" means that S can take NP if followed by VP, thus, S can be substituted by NP VP.

[7] Dipper and Stede [16] describes it as "a tagger trained on the STTS tagset can be trained further on a modified tagset."

Laali and Kosseim [15], investigated how applicable the features used for the English usage disambiguation task of connectives is to French. In the French Discourse Tree Bank (FDTB) [45], created with the discourse structure style of the PDTB, they showed with 94.2% accuracy that the syntactic and lexical features used for English also work in French. They expanded the feature set (categories + case sensitive connective itself) proposed by Pitler and Nenkova [12] by modifying case-sensitive connective itself (*Conn*) and adding POS feature of connective (*Pos*), which gives the position of the discourse connective in a sentence in order to detect if the connective is placed at the beginning or at the end. As a result, they have shown that using syntactic and lexical features grants high results for the usage disambiguation of connectives for languages other than English.

Xue et. al. [46] carried out a comprehensive shared task for discourse parsing that also covers explicit connective detection. They used conditional random fields (CRF) for the connective disambiguation task with features as separate "discourse connective head" and "its syntactic head", and "non-head constituents" which grants limited semantic inference. They demonstrate the discourse connective heads as follows: For an expression such as "At least not when", "when" is the discourse connective head [46]. For performing explicit discourse connective detection, only the correct detection of the head of the discourse connective was necessary. The teams from 16 countries from 3 continents set up their systems to be run on the test set, but the results for the parsing task were low and the F1-score could not exceed 29.69%. However, they have made some progress considering the F1-score 19.98% shown by the existing end-to-end parsers [47].

As the second step of building a discourse parser, Xue et. al. [48] presents the 2nd edition shared task study on Multilingual Shallow Discourse Parsing. This study also aims to detect explicit discourse connectives in Chinese as well as English. They described two methods for the disambiguation task: (1) train the learning model individually for each connective and (2) train the learning model for all connectives. They claimed that connective detection could be handled as a token-level segmentation task and solved by algorithms that perform sequential labeling on data sets, such as CRF.

Another work on connective detection is the multilingual discourse work that was conducted by Zeldes et. al. [49]. In this study, which includes 15 different data sets (the DISRPT workshop) in 10 languages [8], 4 different recurrent neural networks (RNN) systems, with word embeddings or decision trees, where linguistic features are gathered, or both, were used for connective detection, which was planned to be performed in 3 languages (English, Mandarin and Turkish). They claimed that syntax is not a very important factor in connective detection because the results are close to each other when compared to tree-banked and plain sentences. Moreover, they observed that accuracy is directly proportional to the data set, probably because the disambiguating factors of the connective are affirmed more extensively as the data set expands [49].

The identification of discourse connectives for biomedical texts has also been developed [50]. They performed the task using the connective features (Words -previous+current+next-, POS of words, Combination of POS and Chunk Combination of words) in the PDTB style annotated data. They also used connective itself for the sense relation identification, but it was

[8] English, German, Basque, French, Dutch, Portuguese, Russian, Spanish, Mandarin and Turkish

excluded for usage disambiguation task. Using 10 fold cross-validation on conditional random field, they achieved the success rate of 90.53 in F1-score measurement.

### 2.4.2 Usage Disambiguation of Turkish Discourse Connectives

A comprehensive disambiguation task for Turkish connectives was done by Başıbüyük [4]. In this study, both usage and sense disambiguation tasks of Turkish discourse connectives were carried out in TDB 1.0 [51], TDB 1.1 [52] and Turkish section of TED-Multilingual Discourse Bank (T-TED-MDB) [39] data sources. After creating a Turkish Connective Lexicon (TCL) that includes discourse connectives with their syntactic and semantic features, they developed a rule-based Turkish Connective Disambiguator (TCD) for the usage and sense disambiguation of the connectives extracted from TCL.

Discourse parsing task is planned in 3 steps, as usage disambiguation, argument identification and sense identification. The features based on 5 different classification rules and they were decided to be used after a comprehensive literature review on the disambiguation of connectives in other languages. The model performed the disambiguation task of single and phrasal connectives by combining features similar to the existing sources ([12], [13], [40], [50]). Later on, Başıbüyük defines another set of features based on 4 different classification rules for disambiguating the suffixal connectives. These features are listed below. While the model yielded an and F1-score of 84% for single/phrasal connectives, it managed to reach 80% for the usage disambiguation of suffixal connectives.

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

## 2.5 Clitic *dA*

### 2.5.1 Linguistic Studies

There is plenty of research and analysis on the behavior of *dA* from a linguistic perspective, describing its role in syntax such as its modifier role. Various studies (Ergin [53], Göksel and Kerslake [7], [9], Kornfilt [54], Lewis [55]) have examined the linguistic characteristics of *dA*. Other works that deal with the connective function of *dA* comprehensively are the studies done by Kerslake [11], Göksel and Kerslake [7] and Göksel and Özsoy [8].

Kerslake discussed in length the discourse connective functions of *dA*. In particular, she put the continuative *dA* under the category of additives due to their close meaning. She also showed that *dA* is a subordinate conjunction, and mentioned the senses of *dA* in sentences, such as contrastive, concessive and causal. Along with the research on *dA*, she also presented a comprehensive table of conjunctions in Turkish. The following examples are taken from [11], unless otherwise stated.

15. Ne iyi ettin de geldin. (subordination)
    How good that you have come. (Lit. "How well you have done that...")

16. **Neden ölümünden otuz yıl sonra bile O'nu bu kadar seviyorsunuz** da *O'nun dediklerine kulak asmıyorsunuz*? (concessive)
    **Why is it that you love him so much even thirty years after his death** but *you don't pay any attention to the thing he said*?

17. *Radyoda Türk dili konulu konuşmalar yapıyordum*, **dönem dönem** de **ozanlardan örnekler getiriyordum**. (additive)
    *I was doing [some] talks on the Turkish language on the radio*, and **I was giving examples from poets, period by period**.

18. *TRT yetkilileri [. . . ] bana başka bir ozandan örnek getirmemi bile önerdiler*, **ben** de **konuşmalarımı yarıda kestim**. (causal)
    *The TRT staff [. . . ] even suggested that I give examples from another poet*, so **I cut short my [series of] talks**.

She noted that the additive *dA* can also occur at the end of the sentence, and its host is the word after the predicate.

19. Osmanlı Bankasının orada ineceğim ben de. (Additive)
    I'm going to get out somewhere near the Ottoman Bank.

*dA* has been known as a focus particle in Turkish. Contrarily, Göksel and Özsoy [8] discussed that categorizing *dA* as a focus particle is actually false. Instead, they argued that *dA* is a focus-associated clitic and does not always change the focus of the sentence. Additionally, they argued that clitic *dA* gives us information about the truth conditions of the arguments it connected to.

### 2.5.2 Computational Approaches

To the best of our knowledge, the existing computational works on *dA* concern its spelling rather than syntactic or discourse functions. For example, Arıkan et. al. [56] aim to automatically detect the spelling mistakes of this clitic, which is mostly misspelled due to its confusion with the spelling of the locative suffix *-de*/*-da*. Stating that the misspelling will also change the meaning of the sentence, Arıkan et. al. [56] demonstrated its two usages. When "Araba da gördüm." (I also saw a car) is misspelled as "Arabada gördüm." (I saw [it] in the car),

the semantics of the sentence is completely changed. Thus, incorrect spelling of *dA* creates semantic ambiguities.

They argued that morphological analysis would be insufficient in spelling correction studies, since *dA* can form a meaningful sentence both in concatenated and separated spelling. Therefore, for their problem set, they used word embedding vectors, the method of assigning numerical values to the sentence created according to the word flow in the sentence, for representing and calculating the semantics of the sentences. They anticipated that this method would catch the syntactic and semantic relationships.

In a synthetically generated data containing 75 million Turkish sentences, they have presented a neural sequence tagging model that corrects the spelling mistakes occurring in the usage of the clitic *dA* and the locative suffix *-dA*. They extracted sentences from several websites, then replaced the correctly spelled *dA* with the wrong ones, and trained the model on their data set. Their model runs a CRF (conditional random field) algorithm, which predicts sequences by collecting previous label information and enables the model to make better predictions using multi-layered bidirectional long short-term memory (LSTM). With hyperparameter tuning, they run their model in 10 epochs by aiming to complete the training with the highest F1-score, and they achieved their goal with an F1-score of 0.86.

The main objectives of the study are (a) to take a step into creating a comprehensive spell checking tool model for Turkish and (b) to present data that would be challenging to differentiate the use of clitic and suffix *dA*. As a result, a promising solution to the misspelling of *dA* has been provided by their work.

In another NLP research in Turkish, Aktaş et. al. [57] briefly mentioned the argument span of *dA* and the discourse connective *madem*, without further scrutiny. Their work shows that *dA* can occur as a modifier of a discourse connective which it is not syntactically attached to.

20. [..] madem **yanlış bir yerde olduğumuzu düşünüyoruz da doğru denen yere asla varamayacağımızı biliyoruz** , *senin gibi biri nasıl böyle bir soru sorar* , [..] (Aktaş et. al. [57])
    [..] madem **yanlış bir yerde olduğumuzu düşünüyoruz** da *doğru denen yere asla varamayacağımızı biliyoruz* , senin gibi biri nasıl böyle bir soru sorar , [..]
    [..] if we think that we are in a wrong place, and we know that we will never never reach the right place; how come a person like you ask such a question? [..]

In this chapter, we mentioned the clitics and additive markers in other languages that share similar functions with clitic *dA*, and linguistic studies that are conducted on *dA*. Further, we introduced how the usage disambiguation task of other discourse connectives, both in other languages and Turkish, has been done so far. In the next chapter, we will discuss functions of *dA*, and explain the senses that can be conveyed through *dA*.

# CHAPTER 3

# FUNCTIONS OF *DA*

In this chapter we will discuss *dA*'s functions and position in a sentence and the meanings it carries.

## 3.1 *dA* as a Discourse Connective

For an item to be a discourse connective, it basically needs to connect two separate arguments, which are taken as verb phrases (both finite and non-finite), clauses or sentences in Turkish. So we can conclude that if a relation holds between two clauses, sentences or verb phrases combined with *dA*, it is potentially a discourse connective. Our preliminary analysis has shown that the following discourse relations may hold between two arguments of *dA*:

- Causal
- Conditional
- Concessive
- Contrastive
- Additive (conjunction and continuation in different frameworks)

In the rest of this chapter, we will describe the senses conveyed by *dA*, using the PDTB's sense tags. In the PDTB framework, the two arguments of a discourse relations are referred to as ARG1 and ARG2. Each discourse connective is assigned a sense from a hierarchically organized tagset 1. The tagset has four Level-1 senses, i.e. Expansion, Comparison, Contingency and Temporal. Each of these senses is further specified at Level-2. A third level specifies the contribution of the arguments [25].

### 3.1.1 Contingency Relations

In its discourse connective function, the clitic *dA* (underlined below) can convey Contingency relations, which include causal and conditional senses. In the examples 1, 2 and 3 ARG2 is the reason for ARG1. Following the PDTB, we consider the text piece that hosts *dA* as ARG2.

Table 1: PDTB 3.0 Sense Hierarchy

| **Level-1** | **Level-2** | **Level-3** |
|---|---|---|
| TEMPORAL | Synchronous | - |
| | Asynchronous | Precedence |
| | | Succession |
| CONTINGENCY | Cause | Reason |
| | | Result |
| | | NegResult |
| | Cause+Belief | Reason+Belief |
| | | Result+Belief |
| | Cause+SpeechAct | Reason+SpeechAct |
| | | Result+SpeechAct |
| | Condition | Arg1-As-Cond |
| | | Arg2-As-Cond |
| | Condition+SpeechAct | - |
| | Negative-Condition | Arg1-As-NegCond |
| | | Arg2-As-NegCond |
| | Purpose | Arg1-As-Goal |
| | | Arg1-As-Goal |
| COMPARISON | Concession | Arg1-As-Denier |
| | | Arg2-As-Denier |
| | Concession+SpeechAct | Arg2-As-Denier+SpeechAct |
| | Contrast | - |
| | Similarity | - |
| EXPANSION | Conjunction | - |
| | Disjunction | - |
| | Equivalence | - |
| | Exception | Arg1-As-Excpt |
| | | Arg2-As-Excpt |
| | Instantiation | Arg1-As-Instance |
| | | Arg2-As-Instance |
| | Level-Of-Detail | Arg1-As-Detail |
| | | Arg2-As-Detail |
| | Manner | Arg1-As-Manner |
| | | Arg2-As-Manner |
| | Substitution | Arg1-As-Subst |
| | | Arg2-As-Subst |

The other text piece is ARG1. In the examples, we show ARG1 in italic fonts, ARG2 in bold fonts.

1. **İlişkileri bana söylediğinin ötesindeydi** de *o gece Eda'nın kalkıp benimle gelmesine bozulmuştu*. (TDB 1.1, file 00005121) (Contingency:Cause:Reason)

Because **their relationship was beyond what he told me** *he was upset that Eda came with me that night*.

2. *Seni arayamadım*, **işim çıktı** da. (Contingency:Cause:Reason)
   *I could not call you*, since **something had come up**.

3. *Kapıyı açar mısınız*, **marketin orada ineceğim** de (ben)? (Contingency:Cause:Reason)
   *Could you open the doors*, (because) **I will get off at the market**?

4. **Elini ver** de *falına bakayım*. (TDB 1.1, file 00005121) (Contingency:Condition:Arg2-as-cond)
   If **you allow me your hand**, *I will see your fortune*.

So, examples 1, 2 and 3 demonstrate the cases where *dA* indicates a cause-effect relation with different ordering of arguments.

Another sense that *dA* can carry is conditional. When the occurrence of one argument depends on the other, that is, one is the condition of an unrealized event described in the other, the relationship between these arguments is called conditional. In this function, the sentence usually has a clause that can only happen if the specified condition is met. Thus, in example 4 'giving the hand' is the condition of seeing the subject's fortune.

### 3.1.2 Comparison Relations

*dA* can also indicate concessive and contrastive meanings, which are Level-2 senses of Comparison in the PDTB framework. The concessive sense can be understood as denial of expectation, in other words, an event or act that is expressed in a sentence and expected to happen, but the other sentence gives information from which we can infer that it did not occur. Contrastive relations, on the other hand, occur when an argument contains a proposition that is opposite to the information specified in the other.

*dA* conveys the concessive sense in 5, as 'giving the card' is an expected behavior from the hearer, however, the hearer, instead, writes down the name on paper and acts as if s/he is flaunting it, which is thought to be improbable from the perspective of the speaker. In 6, *dA* conveys that although an email is sent, the speaker does not think that it will be enough as an answer. Thus, it can be deduced that ARG1 expresses contrastive information with respect to the content of ARG2.

5. **Niye kartını vermiyorsun** da, *hava yapar gibi kâğıda adını yazıp veriyorsun*? (TDB 1.1, file 00050220) (Comparison:Concession:Arg1-as-denier)
   **Why don't you give your card**, but *write your name on the paper as if you are flaunting it*?

6. **Elektronik posta attım** da *cevabın yeterli olacağını sanmıyorum*. (Comparison:Contrast)
   **I sent an email** but [I] *do not think that the reply would be efficient*.

### 3.1.3 Expansion Relations

The most common sense of *dA* is additive in Turkish. It can function similarly to *ve* (and) in English [54]. Given that the term *additive* is too broad, we will propose a further distinction within the PDTB's Expansion relations, distinguishing *conjunction* from *continuation*. In the PDTB framework, *conjunction* is described as follows: "Arg2 provides additional, discourse new, information that is not related to Arg1 in any of the ways described for other types of EXPANSION." [25], p. 37. On the other hand, Asher [58] explains *continuation* as follows: "Intuitively, a constituent $\alpha$ continues a constituent $\beta$ just in case $\alpha$ and $\beta$ have the same topic."

Based on these descriptions, in our analysis, if ARG2 elaborates on the content of ARG1 and provides non-identical information that is related to the content of ARG1 (or in the reverse order), we will refer to it as a conjunction relation. On the other hand, we will use continuation as a discourse relation where both arguments are on the same topic. In 7, *dA* has a conjunction function, connecting ARG1 and ARG2, whereas in 8, *dA* shows a continuation relation in which ARG1 and ARG2 are about the same topic. In the latter, it is likely that also the two arguments share common words, e.g. *delik-deliği* (hole).

7. **Gel** de, *bir çay iç*. (TDB 1.1, file 00003121) (Expansion:Conjunction)
   **Come**, and *have some tea*.

8. *Halil gecelerce uğraşarak belirlediği yere bir delik daha açtı*. **Deliği açtığı yerde** de **eski takunyalar yığılıydı**. (TDB 1.1, file 00001131) (Expansion:Level-of-detail:Arg2-as-detail)
   *Halil worked for nights and dug another hole in the place he had marked.* And **old clogs were piled up where he made the hole**.

Many continuity relations can be identified by observing temporal adverbs, subordinators (e.g. *after*, *while* etc.) and grammatical dependencies (e.g. past tense verb followed by past tense auxiliary: "I **had** a great night, I **was** killing time with a new released computer game.", on the one hand, and lexical relations and referential links, that is, "coherence strands" [59] on the other. In other words, arguments involving lexical relations (e.g. synonyms, antonyms, repetition) and syntactic parallelism (e.g. consecutive sentences in the past tense or consecutive sentences in passive voice) are the precursors of continuity. Hence, syntactic and lexical dependencies are the prominent features in determining continuation, and one should not rule out these features when determining the continuation sense.

In that regard, continuative *dA* can show some observable phenomena in the discourse. Thus, its continuation sense can be detected by observing certain features in discourse, summarized below:

- ***dA* occurs within or at the end of ARG2:** e.g. 'Sen çağırdın. Ben de geldim.' (Then/so, I came.) and 'Tabakları yıkamışsın. Bu kupayı da yıka!' (Also wash that mug!)
- ***dA* occurs in syntactically parallel sentences:** e.g. 'Eski kitaplar toplandı, kitaplığa da yerleştirildi.' (Old books were piled up and [were] placed in the bookshelves.)

- ***dA* occurs in sentences containing lexical links:** e.g. 'Karnım açtı, ben de yemek siparişş ettim.' (I felt hungry, then/so I ordered a meal.)

#### 3.1.3.1 *dA...dA*

*dA* can be placed in a sequential order, indicating the additive sense. This composition, unlike the enumeration function, always needs to take two arguments to be considered at discourse level. The syntax of *dA...dA* can be represented as S(NP VP) *dA* S(NP VP).

9. *Geldi* de **uğradı** da. (Expansion:Conjunction)
   *S/he came*, (moreover) **s/he stopped by**.

## 3.2 Other Functions

So far, we illustrated the discourse functions of *dA* that are considered in the current work. In this section, we will give detailed descriptions of other uses of *dA*. Although some of these functions are associated with discourse (e.g. the topic shifting role of *dA*), we have left them out of our scope leaving their analysis for further research.

### 3.2.1 Enumeration, Topic Shifting and Modifier

There are other uses of *dA* in Turkish in which it does not indicate relations between clauses, but enhance the meaning, shift the subject, or coordinate nouns. These functions are examined under the categories: Enumeration, topic shifting and modifier.

The enumerative function of *dA* was discussed by Göksel and Kerslake [7], Ergin [53], Lewis [55] and Kornfilt [60]. In this function, *dA* coordinates items that are not clauses, e.g. nouns, adjectives.

10. Ayşe de Semra da... [7] (Enumeration)
    Both Ayşe and Semra...

For example, in 10 *dA* is used in a discourse-initial sentence with no sentence that precedes it. However, its syntax should not be confused with sentences in which *dA* connects predicates in an additive sense[1].

11. Bu kitabı da Ayşe aldı. (Topic shifting)
    As for this book, Ayşe bought it.

[1] e.g. "*Gittik* **de** *kaldık* **da**.", "*We went* (and) **also** *we stayed*."

In 11 *dA* indicates a change in subject, that is, shifts the topic so that it does not show any semantic relation within the preceding text. In 11, it can be seen that the syntax is similar to *dA*'s behavior in an additive discourse connective sense; however, whether or not there is a difference can be decided by looking at the antecedent sentence. For example, if the previous sentence was 'Ahmet bought the flowers', it would be marked as a discourse connective in the additive sense, whereas if it was, let us say, 'I was late to the party', it would signify a topic shift, and 11 acts in that sense as there is no related previous context, hence not a discourse connective in our approach.

Although Göksel and Kerslake examine the continuative and topic shifting functions of *dA* in a single category, topic shifting *dA* represents a different function than continuative *dA*. Contrary to the continuative, topic shifting signals a change in subject [7], which does not represent a semantic relation to the preceding discourse. They also stated that although it is common for the topic shifting *dA* to come after the first constituent in the sentence, there are other cases in which *dA* is placed in other positions when signaling a change in topic [7].

Another occurrence of *dA* is its use right after an adverb, and taking the modifier or enhancer role. The modifier function of *dA* was introduced by Zeyrek et. al. [51] and is defined as the most used modifier by making up 49.07% of all modifiers in the examined data, i.e. the Turkish Discourse Bank 1.0. It is demonstrated in 12 how these kinds of sentences were composed. Other adverbs can be; *bazen*, *belki*, *özellikle* and so on.

12. Gerçekten de babanın eli açıktı. (TDB 1.1, file 00050120) (Modifier)
    Indeed, your father was very generous.

Temporal adverbs can also be hosts of *dA*, as in 13. In the combination of the two, *dA* is considered as an enhancer, i.e. modifier of 'sonra', as enhancing the temporal sense conveyed by 'sonra'.

13. *Balığı kızarttım*. **Biraz** sonra da **yiyeceğim**. ([7]; p. 442)
    *I have fried the fish* and **will eat it in a few minutes**.

*dA* can be attached to other discourse connectives, *ve* (and), *ayrıca* (moreover) etc. It has two syntactic occurences with this use. One can be observed as 'ARG1 ve de ARG2' or 'ARG1. Ayrıca da ARG2', the other can be observed 'ARG1 ve ARG2 *dA* ARG2' or 'ARG1. Üstelik ARG2 *dA* ARG2', where, arguments linked by the main discourse connectives, i.e. *ve* (and), *ayrıca/üstelik* (besides/moreover). In this case, *dA* is the modifier of the discourse connective to which it is attached. Since one of the functions of *dA* is enhancing and emphasizing the meaning of its host, *dA* cannot be thought of separately from the main connective which conveys an Expansion relation when it is used right after a discourse connective. Therefore, in these sentences *dA* is inferred as a modifier of the main connective, and it takes the emphatic function when used with other explicit discourse connectives that indicates Expansion relation, thus left out of the scope of our analysis.

14. *80 yaşında Almanca öğrenmeye başladı* ve de **bundan çok memnun**. (*loc. cit.*) (Modifier of *ve*)
    *S/he has started learning German at 80*, and what's more **s/he's very happy about it**.

15. *Çok küstahsınız*, üstelik de **sarhoşsunuz**. [61] (Modifier of *üstelik*)
*You are vain*; moreover, **you are drunk**.

Furthermore, there are sentences in which *dA* is used for emphasizing a certain word in the sentence. These forms can be observed for verbs and nouns; *anlattı da anlattı* (s/he told and told), signifying that s/he talked constantly and probably without taking a break. Another use is for creating an emphasis on nouns; *masa da masaymış!*, underlining the strengths of features that the *table* has. These cases are not considered discourse-level uses.

## 3.3 Combination of *dA* with Other Clitics and Verbal Suffixes

*dA* can also be used with other clitics and suffixes in Turkish. It can occur with the conditional converbial suffixes, i.e., *-sA* and *-(y)sA* ([7]; pp. 431-434, [54]). If it is used with these affixes, it contributes concessive and alternative meanings, as can be seen in examples 16 and 17. However, these functions have been excluded from this study for now, since the analysis of its use requires a different research and goes beyond the subject of this study.

16. **Birkaç arkadaşımıza da aynı öneriyi sunduk**sa da, *kimse yanaşmadı*. (Concessive conditional)
Although, **we made the same suggestion to a few of our friends too**, *no one acceded*.

17. **Okula git**sen de **gitme**sen de *kahvaltını etmelisin*. (Alternative conditional)
Whether **you go to school or not**, *you should have your breakfast*.

### 3.3.1 *-mI* and *ki*

One of the lexical items that can be used together with *dA* is the question particle *mI*. Apart from being used as in 18, these two can also be used with *dA* on the right of *mI*. In these sentences, *dA* conveys the concessive sense. It is also possible to use *mI*, *ki* and *dA* together as in 19b. While it is likely to have these kinds of sentences in spoken language, again, 19b has a concessive meaning indicated by *ki* and *dA*. If we remove either one of them from the sentence, they will continue to carry concessive sense, but together they convey an enhanced meaning.

18. Kağıtları da okumuş mu? [7]
Has s/he read the papers too?

19. (a) **Sen üstüne düşeni yaptın** *mı* da *bana bunları söylüyorsun*? (Comparison:Concession:Arg2-as-denier)
Well, *you are telling me these* but **have you done your part**?

    (b) **Sen üstüne düşeni yaptın** *mı* ki de *bana bunları söylüyorsun*? (Comparison:Concession:Arg2-as-denier)
Well, *you are telling me these* but **have you done your part**?

### 3.3.2 *bile*

When clitic *bile* (even) is used with *dA*, it creates sentences that are more emphatic, and has an additive sense carried by *dA*. As seen in 20, *dA* is placed on the right of *bile* and shows an additive (conjunction) relation.

20. *Hiç aramadı*, **bir gün** bile de **nasılsın diye sormadı**. (Expansion:Conjunction)
*He never called*, (and) also **did not** even **ask for a day how I was**.

## 3.4 Grammaticalization of *dA*

Although *dA* is not referred to as a copular clitic, it can be inserted between the verb and its suffix, but it has isolated syntax. In this case, *dA* is written separately from both stems, and its syntax can be expressed as: *verb stem-dA-verbal suffix*. *dA* has an occurrence with these second verb stems: *A-bil-, A-dur-* [62] as cited in [20], *A-gel-*, *A-yaz-*, *A-ver-* and *A-kal-*. When used with either one of them, *dA* conveys an additive sense, therefore, it functions as a discourse connective. Example sentence structures in which *dA* is used with verbal suffixes and conveys additive sense can be seen below.

21. *Ben seni kapıda beklerim*. **Seni köşede bekle-ye** de **bil-irim**.
*I will wait for you at the doorway*. **I can** also **wait for you in the corner**.

22. *Kimse yaptıklarına sesini çıkarmamış*. **Bu yıllarca böyle sür-e** de **gelmiş**.
*No one raised a voice against what she did*. And, **it has been that way for many years**.

In the annotation process, which will be described in detail in Chapter 4, we took these cases as discourse connective *dA*.

## 3.5 Focusing and Non-focusing *dA*

When *dA* functions as a focus particle, it can either signify a discourse relation or has non-discourse use within its text span. But the non-focusing *dA* can also occur in discourse connective and non-discourse connective uses. Below are two discourse blocks, namely examples 23(a-b) and 24 (a-b), illustrating both the focusing and the non-focusing *dA* and their discourse or non-discourse connective readings (focus is shown by capital letters).

23. (a) *O gün kimse derse gitmemiş*. **BEN** de **derse gitmedim**. (focusing *dA* indicates additive relation)
*No one attended the class that day*. Also, **I did not attend the class**.

    (b) Ahmet bu arada SINAVA da hazırlanacaktı. [8] (focusing *dA* as non-discourse connective)
In the meantime, Ahmet was supposed to get prepared for the EXAM.

24. (a) *Ödevlerimi yapmadım*. **BEN derse** de **gitmedim**. (non-focusing *dA* indicates additive relation)
*I did not do my homework*. **I** also **did not attend the class**.

(b) Zeynep de bu deneyi BİTİRMEK istiyor. [8] (non-focusing *dA* as non-discourse connective)
As for Zeynep, (she) wants to FINISH this experiment.

## 3.6 Syntactic configurations where *dA* appears in a discourse and non-discourse sense

Since *dA* is a clitic, its position can change within the clause. Although clitics cannot be moved independently of their host, they can be moved freely in a sentence with their host [17]. Therefore, it is important to determine the variability in their word order before attempting a computational analysis. Here, we will briefly discuss the position of *dA* in the clause; more specifically, which relations it can convey in specific syntactic configurations.

We have observed that when expressing Contingency and Comparison relations, *dA* tends to appear in clauses with a specific syntactic configuration [2]. The syntactic configurations in which *dA* occurs and the relations that can be expressed through these configurations can be listed as follows.

- ***dA* conveying Contingency and Comparison senses:** "S(NP VP) dA S(NP VP)"
- ***dA* conveying Causal and Additive senses:** "S, S(NP *dA* VP)" and "S, S(NP VP) *dA*"
- ***dA* conveying the Additive sense and its Non-Discourse usage:** "S, S(NP *dA* S | VP) [3]"
- ***dA* conveying Constrastive and Additive senses *dA*:** "S(VP) *dA* S(VP)"

***dA* conveying Contingency and Comparison senses:**

25. S(NP VP) dA S(NP VP)

(a) **Aceleyle çağırdı** da *geldim*. (Contingency:Cause:Reason)
*I came*, because **she called in haste**.

(b) **Aceleyle çağırdı** da *geldiğimde kimse yoktu*. (Comparison:Contrast)
**She called in haste** but *no one was around when I arrived*.

Even though sentences in example 25 share the same word sequence, the sense conveyed by the two are completely different from each other. While 25a has a causal sense, 25b has a contrastive sense.

[2] See Section 3.1.1 and 3.1.2

[3] The symbol "|" between S and VP means that there can be a sentence (S) that can only be composed by a verb phrase (VP).

***dA* conveying Causal and Additive senses:**

26. S, S(NP *dA* VP)

    (a) *Aceleyle çağırdı*, **ben** de **geldim**. (Contingency:Cause:Result)
    *She called in haste*, so **I came**.

    (b) *Aceleyle çağırdığı kişiler geldi*, **ben** de **geldim**. (Expansion:Conjunction)
    *The people, she called in haste, came*, **I**, also, **came**.

27. S, S(NP VP) *dA*

    (a) *Kapıyı açar mısın*? (ben) **geldim** de. (Contingency:Cause:Reason)
    *Could you open the door*? (Because) **I came**.

    (b) *Ben geldim*, (ben) **eve uğradım** da. (Expansion:Conjunction)
    *I came*, and (I) **also stopped by the house**.

The semantic relation of 26a is cause-result, that is, the reason why I came is that I have been called hastily, whereas in 26b the additive sense is conveyed through *dA*, as can also be clued by the verb *gel-* which is used in both sentences, it can be inferred that some people other than the subject, i.e. "I", also came.

The syntax of *dA* in 26 is NP *dA* VP, while it is NP VP *dA* in 27a, 27b. Examining the examples in 27 in detail: a Causal relation is inferred from 27a, while *dA* has an additive function in 27b.

***dA* conveying Additive sense versus its non-discourse usage:**

28. S, S(NP *dA* S | VP)

    (a) *Ben sinemaya gitmek istiyorum*. **Ayşe** de **sinemaya gitmek istiyor**. (Expansion:Conjunction)
    *I want to go to the cinema*. **Ayşe**, also, **wants to go to the cinema**.

    (b) Ben sinemaya gitmek istiyorum. Ayşe de tiyatroya gitmek istiyor. (Topic Shifting)
    I want to go to the cinema. As for Ayşe, (she) wants to go to the theater.

The discourse (Expansion) and the non-discourse uses of *dA* show syntactic similarity. In particular, since *dA* is usually positioned right after the first constituent of the sentence when it functions as an additive marker, and it usually has the same positioning in the topic shifting use, it becomes difficult to distinguish between these two different levels. In this case, clauses, that is preceded by the sentence containing *dA*, has a great importance in terms of semantic inference. In 28a, ARG2 allows us to infer that someone other than "Ayşe" wants to go to the "cinema". While, in 28b, different activities of two distinct subjects (Ayşe and I) are mentioned. In other words, the intentions of two subjects are expressed by using *dA* (*as for* in English).

***dA* conveying Constrastive and Additive senses:**

29. S(VP) *dA* S(VP)

    (a) **Yetiyor** da *artıyor*. (Expansion:Conjunction)
    **There is enough** and *to spare*.

    (b) **Geldi** de *uğramadı*. (Comparison:Contrast)
    *He did not stop by* although **he came**.

As already noted, in Turkish, verb phrases can be taken as discourse-level arguments. When *dA* is located between two verb phrases, it can indicate Comparison or Expansion relations. In 29a it can be inferred that something is more than enough, that is, it suffices, moreover, it can be spared, whereas, in 29b it is expressed that the subject has come, however, he did not stop by the expected location.

In conclusion, even though the syntax of *dA* may be helpful in some cases for its usage disambiguation, it does not appear to root out the problem, since its syntax as a discourse connective and a non-discourse connective can be very similar. This issue may also be due to the property of clitics that can be moved freely within a sentence with their host [17]. In summary, the flexible word order of Turkish and the freedom of movement of *dA* with its host can create syntactic ambiguities, and these ambiguous syntax types may appear as a problem when we attempt to disambiguate the usage and the sense of *dA* by using solely linguistic features. The sense disambiguation of *dA* is out of our scope. Nevertheless, our observations on the syntactic configurations of *dA* according to its sense are hoped to contribute to future studies.

To sum up, as an initial step to the annotation and usage disambiguation tasks, this chapter is devoted to explaining the discourse and non-discourse use of *dA* and the syntax of *dA* in several sentence structures conveying different senses. In the following chapter, we will explain the data and methods we use for performing the discourse usage disambiguation of *dA*.

# CHAPTER 4

# DATA AND METHOD

This chapter will present the steps prior to the machine learning method used to classify the discourse vs. non-discourse usage of *dA*. First of all, the data will be introduced. Then, the annotation method, and the inter-annotator agreement results will be described. Finally, the data preprocessing stage for the machine learning algorithm will be explained.

## 4.1 Data

Here, our aim was to create a sufficiently large data for machine learning (ML) algorithms to learn from. We used 407 text pieces from TDB 1.1 [52] and 407 from the TS Corpus v.2 (see below), amounting to a total of 814 text pieces containing the use of *dA*. The texts were converted to a spreadsheet, and (a) annotated by two independent annotators for the discourse/non-discourse usage of *dA*, and (b) when *dA* is used in a discourse role, its senses were annotated as described below.

### 4.1.1 Turkish Discourse Bank 1.1

In order to create our training data, we started with TDB 1.1, which is an annotated corpus of implicit and explicit discourse connectives, their binary arguments and senses but *dA* has not been annotated systematically on this data set. Thus, before undertaking a computational analysis, we wanted to annotate the occurrences of *dA* in the data as the first step of the analyses conducted in the thesis. TDB 1.1 is a multi-genre, written corpus of modern Turkish and it is a 40.000-word subcorpus of METU Turkish Corpus or MTC [63] consisting of 20 texts containing novels, research surveys, articles, interviews, memoirs and news articles.

However, an analysis showed that the samples of *dA* extracted from TDB 1.1 does not have an equal distribution of its discourse and non-discourse occurrences.

### 4.1.2 The TS Corpus

Since the data from TDB 1.1 was unevenly distributed, an expansion of the data was necessary, also, increasing the number of samples for training would also increase the diversity in

sentence structures. Thus, we decided to extend our data by adding sentences from the TS Corpus v2 [1] [64].

TS Corpus is a Turkish corpus composed of over 491M units, where all units are marked on the basis of word type (POS tag), morphological structure tag (Morphological Tagging) and root word (Lemma). It is created with the IMS Open Corpus Workbench [2] (CWB) /Corpus Query Processor [3] (CQP) [64]. TS Corpus was introduced by Sezer & Sezer [64], presenting the project as an online publicly available dictionary of TS DIY (TS do-it-yourself) Corpus [65]. In order to create the TS Corpus, they used the BOUN Web Corpus [66], containing data from news and other internet websites.

The users can reach the corpora through CQPWeb, which is an internet interface. The published corpus dictionary offers several advanced functions in the user interface, e.g. keyword queries are enhanced by displaying collocations and tri-grams of the searched word [65]. As a result, we have extended the TDB data by 407 samples by using the internet interface of the dictionary. In both cases, the selection of the samples was manual, and care was taken to include sentences showing variety in syntax and semantics.

### 4.1.3 Data Preparation and Data Format

For annotating the data, excel is used. We examined texts containing the clitic *dA*, that is, its occurrence within a sentence or across sentences. Firstly, the texts containing *dA* are divided into three parts: (1) The part denoted as 'ANT', consists of the text span which occurs before *dA*. (2) The part denoted as 'dA' consists only *dA*. (3) The part denoted as 'POST', consists of the text span which occurs after *dA*. Once this segmentation has been done, all sentences were taken into a separate excel document for each annotator, discourse connective (DC) and non-discourse connective (NDC) options were assigned as annotation tags, see Figure 2.

The first set of annotations was created over 407 sentences from TDB 1.1. Then, the remaining 407 sentences from TS Corpus v2 were prepared in the same format.

## 4.2 Annotation of the data

All the *dA* samples were annotated by two independent annotators for discourse and non-discourse usage of *dA*. The annotators were native in Turkish and fluent in English. The first annotator is the author of the present thesis. The second annotator is a graduate student in Economics who has no background knowledge on linguistics. Therefore, first, a set of guidelines was created to describe the annotation task and the annotation method to the second annotator. In two sessions, which lasted approximately two hours, the guidelines were explained to the second annotator and a few examples that are not involved in the data, were annotated jointly. The first session was held for the discourse usage annotation task and the second session for

[1] http://tscorpus.com

[2] CWB is a software that is created by combining several open source tools and can process language data up to 2 billion words, including their linguistic labels.

[3] CWB software uses CQP (Corpus Query Processor) as a core.

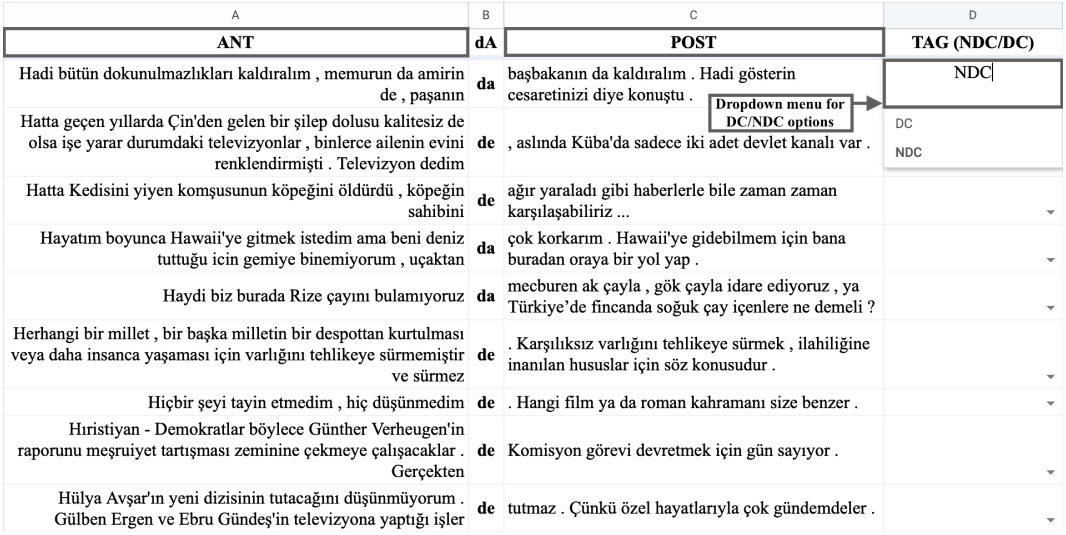


Figure 2: Excel Format for the Annotation Task

the sense annotation task. For sense-level annotations, the PDTB 3.0 sense hierarchy was explained to the second annotator by using the PDTB 3.0 Annotation Manual [67]. Then, the annotators were given the texts to be annotated and were expected to annotate the data by the given tag sets, i.e. the DC, NDC annotations were done independently, and sense-level annotations were requested from the second annotator for the samples that annotator 1 tagged as DC. Sense-level annotations were performed as will be described below.

### 4.2.1 Discourse Usage Annotation (Discourse Relation Spotting)

As it has been already explained in Section 3.1, discourse relations emerge as a result of two text spans creating coherence. Discourse connectives play an important role in that regard. They are clear signals that show how the arguments are linked by a semantic relationship. In accordance with our definition, at this stage, where *dA* fits this definition, the function of it is marked as discourse connective (DC). If no discourse relation is detected, *dA* is tagged as an non-discourse connective (NDC).

1. *Hiç üzülmedi*, **üzmedi** de. (*dA* tagged as DC, Expansion:Conjunction)
   *She was never hurt*, (nor) **did she hurt others**.

2. Koştu da koştu. (*dA* tagged as NDC)
   She ran (and ran).

### 4.2.2 Sense Annotation

In annotating discourse and non-discourse use of *dA*, the senses that are conveyed by *dA* have to be inferred from the texts.

An extensive linguistic review was carried out in the literature to decide the list of sense tags of *dA*, as already mentioned in Section 3.1. As a result, discourse (DC) relations conveyed by *dA* were identified by using the PDTB 3.0 sense hierarchy. The results of the sense annotation task are presented in Table 3.

### 4.3 Annotation Results and Evaluation

Discourse relation spotting was done for all samples by both annotators. Taking the sense hierarchy of the PDTB 3.0 as the basis, the samples that were annotated as DC by annotator 1 were annotated for their sense-level relations by annotator 2 considering the previous and right contexts of *dA*. The annotators did not have to identify the exact spans of the texts before and after *dA* because the data already contained either a complete sentence where *dA* was used intra-sententially, or a pair of sentences where *dA* was used inter-sententially.

Of the 814 samples, as can be observed in Table 2, 336 samples were annotated as DC, 349 samples were decided to be tagged as NDC by both annotators. There is a disagreement for 76 samples which were tagged as DC by annotator 1, NDC by annotator 2; there is also disagreement for 53 cases which were annotated as NDC by annotator 1, DC by annotator 2.

Table 2: Inter-rater Results for Discourse Relation Spotting for Clitic *dA*

| **ANN I** \ **ANN II** | **DC** | **NDC** |
|---|---|---|
| **DC** | 336 | 76 |
| **NDC** | 53 | 349 |

Table 3 shows the sense-level annotation results for the samples that are annotated as discourse connective by the first annotator.

Table 3: Sense Annotation Results for Clitic *dA* Created by Annotator I and II

| **Relations** \ **Annotators** | **ANN I** | **ANN II** |
|---|---|---|
| **Expansion.Conjunction** | 166 | 179 |
| **Expansion.Level-of-detail.Arg2-as-detail** | 67 | 61 |
| **Contingency.Condition.Arg2-as-cond** | 63 | 61 |
| **Comparison.Contrast** | 45 | 42 |
| **Contingency.Cause.Reason** | 35 | 35 |
| **Comparison.Concessive.Arg1-as-denier** | 26 | 21 |
| **Comparison.Concessive.Arg2-as-denier** | 7 | 11 |
| **Contingency.Cause.Result** | 3 | 2 |

Examples for each sense category are given below.

3. *Dediklerini yapmadım*, **yapmayacağım** da. (Expansion:Conjunction)
   *I did not do what they said*, and **I won't**.

4. *Bir eliyle bardağa uzandı*, **öbür elini** de **duvara yasladı**. (Expansion:Level-of-detail:-Arg2-as-detail)
   *He reached for the glass with one hand* and **leaned against the wall with the other**.

5. **Bir örnek ver** de *görelim*. (Contingency:Condition:Arg2-as-cond)
   (If) **you give an example**, *we'll see*.

6. **Yazıyı okudum** da *anlamadım*. (Comparison:Contrast)
   Although **I read the text**, *I did not understand*.

7. **Tanıştığımız kafeden bahsetti** de *aklıma geldin*. (Contingency:Cause:Reason)
   *I thought of you*, because **she mentioned the cafe we met**.

8. *Kalemimi istedi*, **ben** de **verdim**. (Contingency:Cause:Result)
   *He asked for my pen*, so **I handed it**.

9. **Bu kadar yolu geldin** de *bana hediyemi getirmedin*. (Comparison:Concession:Arg1-as-denier)
   Although **you came the all this way**, *you didn't bring me my present*.

10. **Beni merak etmiyorsun** da *hala benden iyilik bekliyorsun*. (Comparison:Concession:-Arg2-as-denier)
    Although **you do not worry about me**, *you are still expecting a favor*.

### 4.3.1 Inter-annotator Agreement (IAA)

To assess the stability of the discourse usage annotations, we calculated IAA by the standard metrics provided below. These metrics can be used for various purposes. In this section, we describe how we used these formulae for calculating IAA, and in Chapter 6, we make use of these formulae for measuring the success of our classification models.

Table 4: The Standard Parameters We Used to measure IAA and Classification Performance

| | |
|---|---|
| **true positive (tp)** | The number of samples that are correctly predicted as positive. |
| **true negative (tn)** | The number of samples that are correctly predicted as negative. |
| **false positive (fp)** | The number of samples that are falsely predicted as positive. |
| **false negative (fn)** | The number of samples that are falsely predicted as negative. |

**Accuracy**

Accuracy is the ratio of the amount of correct classifications to the total amount of classifications. Its formula is as follows:

$$\textbf{ACC} = \frac{tp + tn}{tp + tn + fp + fn}$$

**Precision**

Precision, namely the positive predicted value (PPV), is the proportion of the correctly positive predicted samples to all samples that are predicted as positive.

$$\textbf{PPV} = \frac{tp}{tp + fp}$$

**Recall**

Also known as true positive rate (TPR) or sensitivity, recall is the percentage of the positive predicted samples. In other words, it is the ratio of positive predicted samples to the sum of positive predicted samples and negative predicted samples that should have been predicted as positive.

$$\textbf{TPR} = \frac{tp}{tp + fn}$$

**F1-score**

F1-score is a measurement of the average rate, sometimes called harmonic mean, of precision and recall values.

$$\textbf{F1} = \frac{2}{recall^{-1} + precision^{-1}} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

#### 4.3.1.1 Cohen's Kappa

Cohen's Kappa [68] is a statistical method for measuring the reliability among raters in order to assess the quality of the rating task. For annotation tasks with multiple tags, The equation of Cohen's Kappa is given below, where $P_o$ denotes the observed agreement and $P_e$ designates the probability of agreement between raters by chance. The Cohen's Kappa coefficient formula is as follows :

$$\kappa = \frac{Po - Pe}{1 - Pe}$$

Regarding representative letters in Table 5: $P_o$ is calculated as:

$$\frac{x + l + z}{S}$$

$P_e$ is calculated as:

$$\frac{a \times d}{S \times S} + \frac{b \times e}{S \times S} + \frac{c \times f}{S \times S}$$

where *a* is equal to the sum of *x, k, t*; *b* is equal to the sum of *w, l, v*; *c* is equal to the sum of *y, n, z* and *d* is equal to the sum of *x, w, y*; *e* is equal to the sum of *k, l, n*; *f* is equal to the sum of *t, v, z* and *S* is the total amount of samples.

Table 5: A Representative Table Showing Agreement and Disagreement Figures Among Annotators for Multiclass Label Annotation

| Rater I \ Rater II | **A** | **B** | **C** | **Total** |
|---|---|---|---|---|
| **A** | x | w | y | d |
| **B** | k | l | n | e |
| **C** | t | v | z | f |
| **Total** | a | b | c | **S** |

#### 4.3.1.2 Inter-annotator Agreement of Discourse Usage Annotation

We assessed IAA by adapting a method used in Zeyrek et. al. [39], which takes one set of annotations as the correct annotations. We considered annotator 1's work as correct because as the first step of our annotation task, some samples annotated by her were thoroughly examined and corrected by the supervisor of the current thesis and a graduate student in Cognitive Science, where needed. Annotator 1 performed the rest of the annotations by taking into account their comments.

Thus, our formulae for inter-rater reliability measurements are based on the following parameters:

- TP: Ann1's DC annotated as DC by Ann2
- TN: Ann1's NDC annotated as NDC by Ann2
- FP: Ann1's NDC annotated as DC by Ann2
- FN: Ann1's DC annotated as NDC by Ann2

Our formulae were as follows:

**Precision:**

$$\frac{Ann1's\ DC\ annotated\ as\ DC\ by\ Ann2}{(Ann1's\ DC\ annotated\ as\ DC\ by\ Ann2) + (Ann1's\ NDC\ annotated\ as\ DC\ by\ Ann2)}$$

**Recall:**

$$\frac{Ann1's\ DC\ annotated\ as\ DC\ by\ Ann2}{(Ann1's\ DC\ annotated\ as\ DC\ by\ Ann2) + (Ann1's\ DC\ annotated\ as\ NDC\ by\ Ann2)}$$

Calculating the IAA for discourse relation spotting using the numbers given in Table 2 by these formulae, we obtained the following values:

$$Precision = \frac{336}{336 + 53} = 0.86$$

$$Recall = \frac{336}{336 + 76} = 0.81$$

$$F1 = \frac{2 \times 0.86 \times 0.81}{0.86 + 0.81} = 0.83$$

Calculating agreement for annotation tasks is necessary in order to understand the quality of samples contained in our data. The perfect F1 measure is 1.0. Our findings show that we reached an F1 score of 0.83. Thus, the agreement results showed that the annotation task has been successful at an acceptable rate, and the guidelines we developed for the discourse relation spotting were successful, meaning that our training data consists of samples, on which two human annotators showed sufficient agreement on their discourse usage and senses. As a result, we obtained quite a stable set of annotated data for predicting the discourse usage of *dA* to be trained in ML algorithms (described in detail in Section 5.2).

Table 6: Inter-rater Reliability Measurement Results for Discourse Relation Spotting of *dA*

| Precision | Recall | F-score |
|---|---|---|
| 0.86 | 0.81 | 0.83 |

#### 4.3.1.3 Inter-annotator Agreement of Sense Annotation

For the second part, the sense annotation of the samples that were tagged as discourse connective (DC) by annotator 1 was independently annotated for their senses by both annotators. We calculated the agreement between two annotators only for the third-level senses in the PDTB 3.0 hierarchy [4], which is the most fine-grained semantic level. Cohen's Kappa coefficient [68] was used to calculate inter-rater reliability.

Table 7: Agreement and Disagreement Figures Between Two Annotators for Sense-level Annotation

| ANN I \ ANN II | Conjunction | Arg2-as-detail | Arg2-as-cond. | Contrast | Reason | Arg1-as-denier | Arg2-as-denier | Result | **Total** |
|---|---|---|---|---|---|---|---|---|---|
| Conjunction | 163 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | **166** |
| Arg2-as-detail | 12 | 54 | 0 | 0 | 0 | 0 | 0 | 1 | **67** |
| Arg2-as-cond. | 1 | 0 | 60 | 0 | 2 | 0 | 0 | 0 | **63** |
| Contrast | 0 | 4 | 0 | 40 | 1 | 0 | 0 | 0 | **45** |
| Reason | 2 | 0 | 0 | 0 | 32 | 0 | 1 | 0 | **35** |
| Arg1-as-denier | 0 | 0 | 0 | 2 | 0 | 21 | 3 | 0 | **26** |
| Arg2-as-denier | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | **7** |
| Result | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | **3** |
| **Total** | **179** | **61** | **61** | **42** | **35** | **21** | **11** | **2** | **412** |

Considering the coefficient formulae presented in Section 4.3.1.1, and the results given in Table 7 the steps for calculating Cohen's Kappa score are as follows:

$P_o$ is:

$$\frac{163 + 54 + 60 + 40 + 32 + 21 + 7 + 1}{412} = 0.91$$

[4] Except for Expansion.Conjunction and Comparison.Contrast as they do not cover a third level sense, we took their second level sense in consideration.

So, $P_o$ is calculated as 0.91.

$P_e$ is:

$$\frac{179 \times 166}{412 \times 412} + \frac{61 \times 67}{412 \times 412} + \frac{61 \times 63}{412 \times 412} + \frac{42 \times 45}{412 \times 412} + \frac{35 \times 35}{412 \times 412} + \frac{21 \times 26}{412 \times 412} + \frac{11 \times 7}{412 \times 412} + \frac{2 \times 3}{412 \times 412} = 0.279$$

and $P_e$ is calculated as 0.279. According to the coefficient formula:

$$\kappa = \frac{0.91 - 0.279}{1 - 0.279} = 0.87$$

Cohen's Kappa values $0.81 \leq \kappa \leq 1.00$ are usually interpreted as almost perfect agreement, $0.61 \leq \kappa \leq 0.80$ as substantial agreement, $0.41 \leq \kappa \leq 0.60$ as moderate agreement, $0.21 \leq \kappa \leq 0.40$ as fair agreement, $0.01 \leq \kappa \leq 0.20$ as slight agreement and $0 \leq \kappa$ means there is no consensus between raters. For discourse annotation, Spooren and Degand [69] suggest that a score of 0.70 and above is acceptable, and, as can be seen in Table 8 our Kappa score is above this standard.

We also measured simple ratio agreement, which is calculated by dividing accepted samples (the sum of samples having the same annotations by both annotators) into all annotated samples involved in the annotation task. This ratio gives the proportion of consensus among the annotators without taking into account the agreement by chance. Cohen's Kappa, on the other hand, takes this ratio into account by adding the probability of chance agreement in calculation process by reckoning expected agreement ($P_e$), and yields us the percentage of accepted samples that could have been made by chance. As can be seen from Table 8, the agreement rate decreases when the probability of agreement by chance is included in the calculation.

Table 8: Inter-rater Reliability Measurement Results for Third-level Senses

| Simple Ratio Agreement | Cohen's Kappa |
|---|---|
| 0.91 | 0.87 |

## 4.4 Linguistic Processing

In order to obtain features that will be used for training the ML algorithms, we need to extract properties (e.g. POS tags) of the raw language data through computational steps by performing lemmatization and tagging. Therefore, we create a feature set that a machine learning algorithm can learn from.

In our work, all linguistic processing were done in Python [70] programming language. Punctuation marks, except sentence-final marks such as full stops, question marks, exclamation marks etc., were eliminated and raw sentences were clustered together with their annotations as separate values. Then, annotated sentences were collected in a separate place as

raw sentences, with *anterior-dA-posterior* headings and their annotated tags, i.e. DC, NDC, Expansion, Contingency, Comparison (second and third level of these top-level senses).

In order to create a hierarchical structure, data was configured as a dictionary that contains sub-dictionaries as keys and values. As a demonstration of a text sample, the data is formed as:

> "**data-points**: {0: {ANT: {Onlar geldi. Ben}, dA: {de}, POST: {geldim.}, TAG: {DC}, SENSE: {ADDITIVE}, Level-1: {Expansion}, Level-2: {Conjunction}, Level-3: {-}, 1: ..., 2: ..., ..., *nth sentence*}"

The data was structured, and the samples were prepared for part-of-speech-tagging and lemmatization, which were needed for extraction of features which will be used by ML models for usage disambiguation of *dA*. The raw data was processed through UDPipe 2.0 pipeline [71], [72], [73] and we obtained tagged and lemmatized samples.

### 4.4.1 Structuring the Samples

After the data were processed in this way, the position information of the words relatively to *dA* was extracted. It was predicted that the discourse function of *dA* can be determined by the distance of words from *dA*, that is, by syntactic sequence. The position information of lemmas was also stored in dictionary form. For example, if the sentence is: "Onlar geldi. Ben de geldim.", then the positions are as follows:

> -4: {o}, -3: {gel}, -2: {.}, -1: {ben}, 0: {de}, 1: {gel}, 2: {.}

In the next step, lemmas and part of speech (POS) tags for each position were obtained. Then, these parameters (position, word root (lemma), POS tag) were combined in a hierarchical data structure, so the position of a word would be the key, and the lemma and the POS tag information would be the values, as can be seen below.

> -4: {o, PRON}, -3: {gel, VERB}, -2: {., PUNCT}, -1: {ben, PRON}, 0: {de, CCONJ}, 1: {gel, VERB}, 2: {., PUNCT}

With these calculated positions, lemmas and tags, a discrete window size was defined according to the observed length of the samples that contains *dA*, and the narrowness of this window is set as in range (-3, +3). Then, all the samples that include words from window size -3 to +3 were collected. These samples are set aside for the training with their POS tags and lemmas. As a result, 757 samples were obtained after filtering through the specified window size, i.e. (-3, +3). The data decreased by 57 samples, because these samples did not contain enough words to fill the (-3, +3) range. [5]

[5] The eliminated sentences included 56 samples annotated as DC and the top-level sense tagged as Expansion and 1 sample was annotated as NDC.

### 4.4.2 POS Tags & Lemmas

The discourse use of *dA* is usually signalled by its occurrence after the main verb of a clause unless it functions as the additive connective, though there are some cases in which additive *dA* is placed after the main verb of the clause. Thus, we anticipate that the POS tag frequencies of the words around *dA* may provide information on how it behaves. The following instances, mentioned in Chapter 3 in examples 2, 7 and 8, are the instances where *dA* is annotated as a DC and occurs before the main verb (shown again below in examples 11,12). In example (13), *dA* is annotated as a DC, but does not occur before the verb (this is when it functions as an additive particle).

11. *Seni arayamadım*, **işim çıktı** da. (Contingency:Cause:Reason)
    *I could not call you*, since **something had come up**.

12. **Gel** de, *bir çay iç*. (TDB 1.1, file 00003121) (Expansion:Conjunction)
    **Come**, and *have some tea*.

13. *Halil gecelerce uğraşarak belirlediği yere bir delik daha açtı*. **Deliği açtığı yerde** de **eski takunyalar yığılıydı**. (TDB 1.1, file 00001131) (Expansion:Level-of-detail:Arg2-as-detail)
    And *Halil worked for nights and dug another hole in the place he had marked*. **Old clogs were piled up where he made the hole**.

We also decided to extract word roots (lemmas) from the samples, as they can inform us about the semantic relations that *dA* conveys. Examples of the samples with extracted POS tags and lemmas can be seen below.

Table 9: An Example of Causal and Additive *dA* Occurring Right After the Verb, and Additive *dA* Right After the Noun

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| DC-Raw | arayamadım | işim | **çıktı** | da | . | Sen | aramayınca |
| POS Tag | VERB | NOUN | **VERB** | CCONJ | PUNCT | PRON | VERB |
| Lemma | ara | iş | **çık** | da | . | sen | ara |
| | | | | | | | |
| DC-Raw | oradan | sesleniyor | **Gel** | de | bir | çay | iç |
| POS Tag | PRON | VERB | **VERB** | CCONJ | NUM | NOUN | VERB |
| Lemma | ora | seslen | **gel** | de | bir | çay | iç |
| | | | | | | | |
| DC-Raw | Deliği | açtığı | **yerde** | de | eski | takunyalar | yığılıydı |
| POS Tag | ADJ | VERB | **NOUN** | CCONJ | ADJ | NOUN | VERB |
| Lemma | delik | aç | **yer** | de | eski | takunya | yığ |

Including lemmas in the training data is also helpful for detecting the non-discourse use of *dA*. As *dA* can function as a modifier of certain adverbs (see Section 3.2.1), the non-discourse usage of *dA* can be captured from the lemma of the previous word, e.g. when found in patterns such as "... Özellikle de..." (Especially/In particular in English), we can safely assume that *dA* has a non-discourse role.

Table 10: An Example of *dA* Occurring as a Modifier *dA* due to Its Position Right After an Adverb

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| NDC-Raw | baktı | . | **Özellikle** | de | seni | sordu | . |
| POS Tag | VERB | PUNCT | **ADV** | CCONJ | PRON | VERB | PUNCT |
| Lemma | bak | . | **özellikle** | de | sen | sor | . |

### 4.4.3 Proper Names

As another method to catch the distinguishing linguistic properties of additive *dA*, it was also checked whether the word in the (-3, +3) range was a proper name or not. This is because it is very common to place additive *dA* in sentence initial position, and in such sentences, proper names are likely to be the host of *dA*. In that regard, it means that additive *dA* has more flexible word order compared to the other discourse connective functions it conveys, and we need a feature to distinguish its syntax when it functions as an additive. Hence, we wanted to include the property of being a proper name into our feature space in order not to miss the syntax-semantics relationships between lemmas. As a demonstration, after text processing, the sample 14 will be displayed as seen in Table 11.

14. *Halil'in geldiğini fark etmediler*. **Halil** de **içtiği içkilerin etkisiyle kadınlara bir şaka yapmaya karar verdi**. (Expansion:Conjunction) [6]
*They did not notice that Halil came*, and **Halil decided to play a joke on the ladies by the effect of the drinks he drank**.

Table 11: An Example of Additive *dA* Occurring Right After a Proper Name

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| DC-Raw | etmediler | . | **Halil** | de | içtiği | içkilerin | etkisiyle |
| POS Tag | VERB | PUNCT | **PROPN** | CCONJ | VERB | NOUN | NOUN |
| Lemma | et | . | Halil | de | iç | içki | etki |
| Isproper | FALSE | FALSE | **TRUE** | FALSE | FALSE | FALSE | FALSE |

Finally, for the annotation and for mapping the annotated samples with their linguistically processed versions described in Section 4.4, each sample was configured to include their raw texts in three parts (*ANT-dA-POST*), as mentioned in Section 4.1.3. Regarding this mapping, each data point also contains raw samples as text as well as tags regarding *dA*'s discourse usage and senses. Rich texts (words in samples that have been processed through lemmatization and tagging), and position information of all sample pieces (words) including lemmas

[6] There may be an ambiguity for some sentences in which *dA* conveys multiple senses. In the example 14, there are more than one senses conveyed by *dA*: (1) Conjunction: Where ARG2 gives new information about ARG1, (2) Purpose: Where ARG2 describes the action to be accomplished by Halil in order to be noticed by the ladies, (3) Cause: Where ARG2 is the result of ARG1, that is, Halil wanted to be noticed by the ladies, so he decided to play a joke on them. However, we annotated all samples just for one sense relation, which is primarily inferred by the reader, even though it is possible to interpret multiple senses by the same reader. In further studies, two/three-fold annotations can be made for ambiguous samples. In particular, it is important to annotate multiple senses for performing the sense disambiguation of *dA*.

and POS tags was also included for all data points. Samples of the mapped annotations and parsed data can be seen in Appendix C.

Before concluding this chapter, we should note why we have developed a new feature set for the disambiguation of *dA* rather than using the feature set developed by Başıbüyük [4]. The features used by her study (provided in Table 12 below) provided successful results reaching the F1 score of 0.84 in the disambiguation of single/phrasal connectives. We did not use the exact same features in our feature set for the usage disambiguation of *dA*, mainly because the syntax of *dA* is not compatible with that of other discourse connectives considered in that work, as mentioned in Chapter 3.

Table 12: Features used by Başıbüyük [4] for the usage disambiguation of single/phrasal connectives, and features used by the present thesis for the usage disambiguation of *dA*

| **Features used by Başıbüyük [4]** | **Explanation** |
|---|---|
| Previous, current, next words | The word before and after the connective, and the connective itself |
| POS tag of previous, current and next words | The POS tags words before and after the connective, and the connective itself |
| Mostly DC, mostly NDC | If the connective's role is mostly DC or NDC [7] |
| Sentence-initial, sentence-final | If the connective is placed sentence-initial position or sentence-final position |
| LeftVerb | If the POS tag of one of the three words before the connective is VERB |
| RightVerb | If the POS tag of one of the words after the connective is VERB |
| **Features used by the present thesis** | |
| POS of words | The POS tag of three words right before and right after *dA* |
| Lemma of words | The lemma of three words right before and right after *dA* |
| Isproper | If the POS tag of one of the three words before *dA* is proper name |

Although both studies used the linguistic features of words around the connective, there are differences in the features sets they exploited. For example, Başıbüyük used the POS tag feature of the word right before and right after the connective, and also the connective itself. We used the POS tag of the three words before and after *dA*, as we aimed to detect other POS tags as well, e.g. nouns. Also, contrary to many other discourse connectives, such as *çünkü* (because/since), *ancak* (however/yet), a sentence cannot begin with *dA* in Turkish, so Başıbüyük's sentence-initial feature could not be added to our feature set.

In summary, in this chapter, we introduced our data and explained its annotation process. Then we assessed the reliability of our discourse relation spotting annotation and sense annotation, showing that annotator 1's annotations can be reliably used in a machine learning algorithm. We also explained how the training data is processed for the experimental set up. Briefly, we showed that we extract the position, lemma and POS tag information from the annotated samples only if a sample contains words in the (-3,+3) range, and we checked whether a word is proper name or not through an if loop run in excel. Then, we combine all the features in comma separated value (CSV) format. In the next chapter, we will explain how we used this data in machine learning experiments.

[7] Based on predictions made by their TCD program.

# CHAPTER 5

# MACHINE LEARNING EXPERIMENTS

This chapter is devoted to the description of the machine learning experiments performed for disambiguating the discourse usage of *dA*. We will describe the algorithms we use; then, we introduce our feature set and the experimental procedure.

## 5.1 Methodology

The methodology of the present work involves several steps: As already explained in Chapter 3, first, in a corpus of samples extracted from two Turkish corpora (i.e. TDB 1.1 and the TS Corpus v2), we analyzed the samples (clauses within sentences or a pair of sentences) that include the clitic *dA*. Then, based on the linguistic analysis that we presented, we performed an annotation task over these texts marking the discourse usage of the clitic *dA* and its senses. Finally, we adapted the methods used in the usage disambiguation tasks of discourse connectives in other languages to our problem set, adapting the task to the linguistic characteristics of *dA*. The steps of the current work can be summarized as follows:

- Analyzing the TDB 1.1 and the TS Corpus v2 for collecting samples with the clitic *dA*.
- Analyzing linguistic features of clitics in other languages along with the clitic *dA* in Turkish.
- An extensive literature review on the functions of clitic *dA* and usage disambiguation of discourse connectives in several languages.
- Annotating the *dA* samples.
- Testing several machine learning models for classifying the discourse vs non-discourse usage of *dA* over the annotated samples.
- Reporting the performance of the models.

Machine learning is a field of study that practices understanding and building of artificial intelligence (i.e. creating algorithms) which can perform a specific set of tasks by minimizing human effort and the time spent. ML techniques are commonly used in automating tasks in many scientific fields, as also widely used for performing several NLP tasks. When the

features of a language data are presented to the ML algorithm, it can perform classification tasks (binary or multiclass) by supervised learning, and classify the data according to the target set (DC and NDC in our case). Below we will explain three ML models we use for the disambiguation task. These models were developed using Python and sklearn [74], a public machine learning library and softwares. The feature set for our classification models will be explained later in Section 5.3.

## 5.2 Classifiers

Since the DC and NDC labels are target values to be classified, binary classification algorithms were explored. We performed training on a data set of 381 DC and 376 NDC samples (see Section 4.4.1) with three ML algorithms: (1) Logistic Regression, (2) Support Vector Machines and (3) Random Forest. Below we describe these models and their classification process in detail.

### 5.2.1 Logistic Regression

Logistic regression is a binary classification model that calculates and predicts the probability of a value from two classes by the linear integration of independent variables, that is, features. It uses *logistic function-f(x)* to estimate the probability.

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

*Where*:

- $L$ = the maximum value of the curve
- $k$ = the steepness of the curve
- $e$ = the value of x of the sigmoid point
- $x$ = a real number

We use LIBLINEAR [75] as a solver, as it is suitable and yields faster results for small data sets, and tuned max_iter=1000.

### 5.2.2 Support Vector Machine (SVM)

SVM, presented by Cortes and Vapnik [76] and developed by Boser et. al. [77], is a statistical learning method that performs classification from two opposite maximized distance classes (by locating two support vectors), and assigns a vector to each data point in feature space, and calculates the optimal hyperplane for current data point by deciding which one of the

classes the data point is closest to. It is one of the popular learning methods used to solve classification problems in NLP tasks, e.g. [78].

"Linear separability" is an essential concept for SVM. If a line can be drawn between two classes, and if all data points belonging to each class can be separated by that line, the data is called linearly separable. The equation for this line is: "y = a . x + b". If the data is not linearly separable, the SVM algorithm moves the data point to a higher dimension until it achieves linear separability. There can be several hyperplanes that separate the data linearly, however, SVM finds the optimal one. The equation of the hyperplane is:

$$w \times x + b = 0$$

where the variables *w* is the weight vector and *b* is the bias. Both are calculated during the classification process, and *x* is the input vector representing the data point. Once SVM finds the hyperplane, it performs the prediction task between classes. If the value that is calculated for a data point is above or equal to the hyperplane, then it is classified as +1, that is the class 1, if it is below the hyperplane, it is classified as -1, that is the class 2.

$$f(x_i) = \begin{cases} +1 & if w \times x + b \geq 0 \\ -1 & if w \times x + b < 0 \end{cases}$$

SVM aims to find the optimal hyperplane farthest from both support vectors by maximizing the margin, so minimizing the weight vector (*w*). Thus, the equation of the distance between the support vectors (d) is: $\frac{2}{||w||}$.[1].

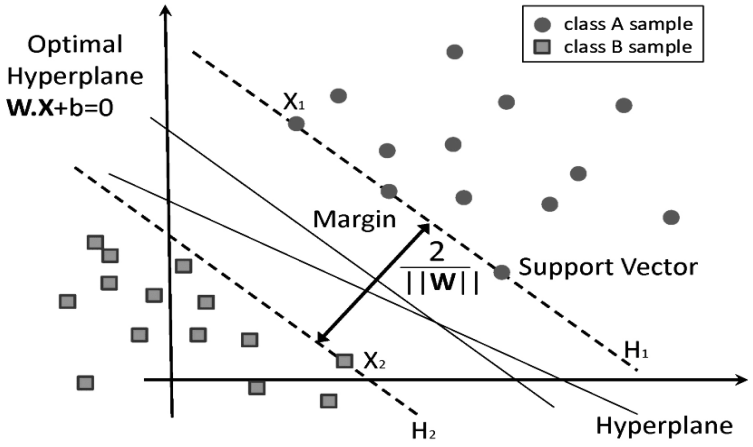


Figure 3: A visual demonstration of classifying data by using SVM [5].

[1] e.g. $||x||$ means that the euclidean distance of *x*.

### 5.2.3 Random Forest

Since ensemble learning methods showed higher results compared to other learning algorithms for specific problems [79], we also trained ensemble learning algorithms to test whether they yield better results or not. Thus, our third algorithm is the random forest classifier [80], an ensemble learning method which combines tree algorithms and processes the data set in a number of trees, then yields the best result possible with an advanced accuracy by preventing the overfitting problem.

By default, Random Forest uses the **Gini** index (or Gini impurity) to find the best split when branching from each node. The formula of Gini index is:

$$Gini(x) = 1 - \sum_{i=1}^{n} P(i|x)^2$$

where *n* denotes the number of classes, and *P*(*i*|*x*) denotes the probability of *i* given *x*. After branching each node with the information calculated by the *Gini* index, Random Forest finds the best class to match with the current data point.

## 5.3 Features for the Usage Disambiguation of the Clitic *dA*

As explained in earlier chapters, we surveyed the literature to find out how other researchers approached the usage ambiguity problem, and selected the features to be used in usage disambiguation of *dA*. Pitler and Nenkova [12] attempted to solve the connective disambiguation problem for English using syntactic properties, and they reached results showing as high as 0.94 accuracy, see Section 2.4.1. In that regard, their study gave us an insight that syntax has importance when detecting the discourse connectives.

Inspired by Lin et. al. [13], we decided to add POS tags of other words in the sentence, where *dA* is used. We extracted the POS tag of the 3 words before *dA* and the 3 words after *dA*. We also included the lemma (word root) of words in this range, (-3,+3).

As a final feature category, we assumed that proper names would be useful in distinguishing discourse connective-additive *dA*, when *dA* is hosted by a proper name, so we checked if the previous 3 words were proper names.

A total of 15 features were extracted from the data:

- **(1-6)** POS tags in (-3, +3) range
- **(7-12)** Word roots (lemma) in (-3, +3) range
- **(12-15)** Whether a word is proper name or not (expressed as “isproper”)

The supervised learning models were trained to predict whether *dA* indicates a discourse relation (DC or NDC). For each learning method, the training set was adjusted to 80% and

Table 13: Feature Set for Usage Disambiguation of *dA*

| **Feature** | **Range** | **Definition** |
|---|---|---|
| POS | (-3, +3) | The POS tags of 3 words before and 3 words after *dA* |
| LEMMA | (-3, +3) | The lemmas of 3 words before and 3 words after *dA* |
| ISPROPER | (-3, -1) | Whether the 3 words before *dA* is proper name or not |

the test was adjusted to 20%. We run logistic regression with LIBLINEAR solver and max_-iter=1000, and SVM algorithm without tuning, and random forest with n_estimators=100.

To summarize, in this chapter, we explained the experimental set up for the usage disambiguation task of *dA*. We presented the ML algorithms that were trained to perform the usage disambiguation of *dA* and defined our feature space. In the next chapter, we will present the performance rates of classification models by using the parameters given in section 4.3.1.

# CHAPTER 6

# RESULTS AND EVALUATION

In this chapter, the measurement parameters used for the evaluation of the ML results will be introduced, and the overall success rate (the learning performances) of our models, i.e. Logistic Regression, SVM, Random Forest, will be presented.

## 6.1 Results of the Usage Disambiguation of *dA*

The performance rate of the models have been calculated so as to give F1-score, accuracy, precision and recall by using the standard *classification report* and *confusion matrix* libraries in Python [1], where we took TP as correctly predicted as DC, FP as falsely predicted as DC, FN as falsely predicted as NDC, and TN as correctly predicted as NDC, as explained in Chapter 5.

We used features mentioned in Section 5.3 for automatically disambiguating the discourse usage of *dA*. We performed the training in three cycles, and wanted to observe if the models yield erratic or unreliable results. It has been observed that the results are consistent when the models are compared with each other. The models correctly disambiguated *dA* with an average F1-score of 0.82-0.83, as can be seen in Table 14 [2]. Classification results are also given in Table 33, 34 and 35 for all training cycles in Appendix D.

Table 14: Evaluation Results for Classification

| **Classifiers** / **Parameters** | **Logistic Regression** | **SVM** | **Random Forest** |
|---|---|---|---|
| **Accuracy** | 0.82 | **0.83** | **0.83** |
| **Precision** | 0.86 | **0.92** | 0.83 |
| **Recall** | 0.78 | 0.73 | **0.83** |
| **F1** | 0.82 | 0.82 | **0.83** |

[1] Although Isaksson et. al. [81]; as cited in [82] stated that cross validation yields unreliable results for small data sets (# samples <1000), we also performed 5 fold cross-validation on our model in order to observe results are unstable, and they were polarized as we run the model several times. So we thought that it would not reflect accurate results if we present the highest cross-validation results which reached F1-score average of 0.86. Nonetheless, we presented cross-validation results in Appendix B

[2] The highest scores achieved are written in **bold**.

While all three algorithms showed close results, Random Forest had the highest success rate of 0.83. In other models, it was observed that precision has slightly higher ratio than recall rate. This means that those models developed for the disambiguation task have a high positive predictive rate, in other words, they can classify positive instances with high accuracy, but the instances with true negative classifications are less accurate, so the rate of detecting DC as NDC is high.

## 6.2 Important Features for Classification

We trained each of our models to examine the predictive strength of each feature (and feature group) we used. Table 15 shows the POS tag is the most predictive feature among the other features in the disambiguation task. However, we achieved higher success rates with the combinations of POS + Lemma and POS + Isproper, though the former is the second best combination for the task at hand. As a result, it has been observed, the results incrementally got higher as we combined three feature set.

Table 15: Accuracy of the Individual Features Used in the Classification and the Best Combination

| **Features** | **# of Features** | **Logistic Regression** | **SVM** | **Random Forest** |
|---|---|---|---|---|
| POS + Lemma + Isproper | 15 | 0.82 | 0.83 | 0.83 |
| POS + Lemma | 12 | 0.78 | 0.76 | 0.76 |
| POS + Isproper | 9 | 0.74 | 0.74 | 0.76 |
| POS | 6 | 0.76 | 0.72 | 0.75 |
| Lemma + Isproper | 9 | 0.72 | 0.67 | 0.66 |
| Lemma | 6 | 0.70 | 0.67 | 0.63 |

When we calculated the feature weights of our models, we observed that features in -1 position 'VERB' POS tag is higher than other feature weights in different positions, see Table 36, 37, 38 in Appendix E. It was also observed that POS tag 'PUNCT' had high predictive weight as a feature at the +1 position, the signifying sentence/clause-final position of *dA*. Therefore, we conducted additional experiments on our data in order to understand if our models are biased for sentence/clause-final positions when predicted discourse connective *dA*, distribution of sentence/clause-final *dA* is given in Table 16 (see Example 1 and 8 for clause-final and clause-medial position of *dA*, respectively). We anticipate that training our data without position -1 POS tag and +1 POS tag might inform us whether there is a bias in our model. So, Table 17 shows the evaluation scores of training without '-1 POS' and '+1 POS'. From the results, we can infer that the models show an above average success when they do not use the specified positions and POS tags.

Table 16: Distribution of clause-final and clause-medial *dA* in the data set

| # clause-final *dA* | # clause-medial *dA* |
|---|---|
| 203 | 611 |

Table 17: Evaluation Results for Classification Without Using Position -1 POS Tag and +1 POS Tag Features

| Classifiers / Parameters | Logistic Regression | SVM | Random Forest |
|---|---|---|---|
| **Accuracy** | 0.70 | 0.70 | 0.70 |
| **Precision** | 0.71 | 0.72 | 0.70 |
| **Recall** | 0.73 | 0.73 | 0.74 |
| **F1** | 0.72 | 0.72 | 0.72 |

In addition, we also used a rule-based model tags a sample as DC only if the word immediately before *dA* is VERB or word immediately after *dA* is PUNCT, otherwise it tags the sample as NDC. In this experiment, it was observed recall value dropped significantly, and has the ratio of 52%, that is, most of the samples labeled as DC were incorrect.

As a result, based on our findings, we can argue that the models do not make biased predictions. In other words, the models do not solely rely on the POS tags before or after *dA* (VERB and PUNCT, respectively); neither do they use the sentence/clause-final position of *dA* as the only feature.

## 6.3 Error Analysis

After calculating the success rates, we carried out an analysis to understand the possible causes of classification errors. From this analysis, it appears that the errors are likely to arise due to the incorrect POS tagging of words by the part-of-speech tagger (UD Pipe 2.0).

The first incorrect tagging is observed on the predicate *var* "there is/are". In our data, occasionally, *var* was tagged as VERB if it has a tense suffix (e.g. -mış, -dım). But the part-of-speech tagger tagged this word as adjective (ADJ) when a tense suffix is missing. For our problem, whether *dA* is preceded by a verb or not is an essential piece of information when detecting if *dA* is a discourse connective, due to its distinct syntax (described in Section 3.6) when functioning as a discourse connective. Thus, wrong part-of-speech tags cause the loss of information regarding certain linguistic properties and the classifier fails.

The other tagger error is the inaccurate tagging of the proper names. Proper names can be the host of *dA* and act as an important linguistic cue for *dA*'s non-discourse and discourse-additive function. In case proper names are tagged incorrectly, e.g. if they are assigned the NOUN tag, classifier performance tends to drop. For example, one of the errors encountered in the data is the abbreviated proper name TBMM (Türkiye Büyük Millet Meclisi 'Grand National assembly of Turkey'), tagged as NOUN. But TBMM is a proper name and can host *dA*, as shown in the example below.

1. TBMM de kararını verdi.
   (And) TBMM has made its decision.

To summarize, in this chapter, first, success rates of the classification models were presented. We showed the predictive role of different features and their combinations in the classification. Then, we examined the reasons why the developed models could not classify some of the samples correctly. In the following chapter, we will summarize the thesis and offer future directions for further research.

# CHAPTER 7

# CONCLUSION AND FUTURE WORK

This thesis is based on (1) an annotated data set consisting of 757 samples of the clitic *dA* as the training data, and (2) the identification of linguistic features specific to *dA* used in machine learning experiments. These features are identified by a thorough analysis of the samples. (3) The thesis finally reports the usage disambiguation task of *dA* done with the supervised machine learning methods. We experimented with three models that perform the usage disambiguation task of *dA* and achieved promising results.

One of the contributions of this thesis has been the creation of a data set where discourse and non-discourse usage of *dA* are reliably annotated. We created this data set by analyzing the differences between the discourse usage of *dA* and its other, non-discourse uses, i.e. its enclitic role. The data set was compiled from two different data sources (TDB 1.1, TS Corpus v2). In addition to the discourse/non-discourse usage annotation of the data, our data set contains annotations of *dA*'s discourse senses, where we used the PDTB 3.0 sense tagset.

The second contribution of the thesis has been the automatic disambiguation of *dA*'s discourse usage by training our annotated *dA* samples. Three models were developed from Logistic Regression, SVM and Random Forest classifiers. Using these models, a binary classification task was done for detecting whether *dA* is used as a discourse connective or a non-discourse connective. Linguistic features are used for the task, and it appears that it carries significant information when detecting the function of *dA*. Thus, from the computational perspective, even though *dA* is ambiguous between discourse and non-discourse uses, our results supported the idea that the syntax of *dA* provides us with credible information when determining the functions of *dA*, and the classification models make use of syntactic structure of *dA*, albeit to a limited extent, see Table 15. We can also argue that the (-3,+3) word window size defines the syntactic scope for *dA*, and that proper names are important features since they can occur as the host of *dA*.

As a result, by investigating and analyzing the state-of-the art usage disambiguation works in discourse connectives in other languages, we created a framework for a comprehensive analysis of *dA* in Turkish, and within this framework, we determined the discourse and non-discourse use of *dA*.

However, the thesis is not without its limitations. For example, our data size is still small, and it is still open to further extensions. It took time to expand the data set, as the annotation task is a long and demanding task, and requires experienced annotators. The annotation part of the thesis was carried out by two annotators. In the future, a third annotator can be involved in the

process. The data annotated by all three annotators can be adjudicated and the ML models can be re-run on this data for comparison purposes. Moreover, it is possible to get better results if the models are trained with the data set above 1000 samples, because in case of reaching over 1000 sample sentences, more linguistic features of a discourse connective can be attested as the samples in the data set are increased [49]. Also, in case of reaching over 1000 samples, the use of Recurrent Neural Networks (RNNs) [1] (e.g. LSTM) will be possible, in order to achieve higher results for such a complex problems in language.

The second limitation is that a model that disambiguates the sense of *dA* has not been built, because it requires an analysis beyond linguistic features. It is clear that for the sense disambiguation of *dA*, we certainly need further investigations, because even when *dA* expresses different senses, its position in the clause may not differ, as demonstrated in Section 3.6. Moreover, the long-distance relations that *dA* can convey, and its relation with other verbal suffixes, e.g. (*-y*)-*sA* (concessive conditional, alternative conditional), could not be studied as they require a different syntactic approach.

In conclusion, we showed that with the ML methods we propose, we can achieve success rates of up to F-score 0.83. We have shown, therefore, how a versatile linguistic element, a clitic, can be studied from a computational approach by making use of NLP techniques.

In the future, the success rate we have achieved can be increased by expanding both the data and the feature set. If we can represent the semantic values vectorially (i.e. by word embeddings), we are likely to increase the success rate of the usage disambiguation of *dA*. By doing that, we will have taken the first step of the automated sense disambiguation of *dA* as well. It is hoped that the models we presented to disambiguate the discourse usage of *dA* will enable the automated sense prediction of *dA* also contributing to the discourse parsing research of Turkish. By using similar techniques, the usage disambiguation of other clitics in Turkish, e.g. *ki*, can also be conducted.

In this thesis, we have not dealt with the potential role of focus in determining the DC or NDC role of *dA*. Hence, in further studies, *dA* should also be analyzed in syntactic, semantic and intonational layers in order to understand the relation between the prosodic features of *dA* and its discourse connective function.

[1] In which connections between nodes is provided by graphs, directly or indirectly, in temporal order using forward/backward connections to process long sequenced data.

# REFERENCES


[1] M. Faller, "The many functions of cuzco quechua =pas: implications for the semantic map of additivity," *Glossa: a journal of general linguistics*, vol. 5, pp. 1–36, 03 2020.

[2] D. Forker, "Toward a typology for additive markers," *Lingua*, vol. 180, 04 2016.

[3] M. Erdal, "Clitics in turkish," in *Studies on Turkish and Turkic Languages, Proceedings of the Ninth International Conference on Turkish Linguistics* (A. Göksel and C. Kerslake, eds.), pp. 41–48, Wiesbaden: Harrassowitz Verlag, 08 2000.

[4] K. Başıbüyük, *Automatic Disambiguation of Turkish Discourse Connectives Based on a Turkish Connective Lexicon*. PhD thesis, Middle East Technical University, 2021.

[5] E. García-Gonzalo, Z. Fernández-Muñiz, P. J. Garcia Nieto, A. Sánchez, and M. Menéndez, "Hard-rock stability analysis for span design in entry-type excavations with learning classifiers," *Materials*, vol. 9, p. 531, 06 2016.

[6] B. Webber, R. Prasad, and A. Lee, "Ambiguity in explicit discourse connectives," in *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, (Gothenburg, Sweden), pp. 134–141, Association for Computational Linguistics, May 2019.

[7] C. Kerslake and A. Göksel, *Turkish: A Comprehensive Grammar*. Routledge, 2005.

[8] A. Göksel and S. A. Özsoy, "*dA*: a focus/topic associated clitic in turkish," *Lingua*, vol. 113(11), pp. 1143–1167, 2003.

[9] A. Göksel and C. Kerslake, *Turkish: An Essential Grammar*. Essential grammar, Routledge, 2011.

[10] S. R. Anderson, *Aspects of the theory of clitics*. New York: Oxford University, 2005.

[11] C. Kerslake, "The role of connectives in discourse construction in turkish," in *Modern Studies in Turkish Linguistics; Proceedings of the 6th International Conference on Turkish Linguistics* (A. Konrot, ed.), (Anadolu University, Eskişehir), pp. 77–104, 1996.

[12] E. Pitler and A. Nenkova, "Using syntax to disambiguate explicit discourse connectives in text," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, (Suntec, Singapore), pp. 13–16, Association for Computational Linguistics, Aug. 2009.

[13] Z. Lin, H. T. Ng, and M.-Y. Kan, "A pdtb-styled end-to-end discourse parser," *Natural Language Engineering*, vol. 20, no. 2, pp. 151–184, 2010.

[14] Y.-S. Shih and H.-H. Chen, "Detection, disambiguation and argument identification of discourse connectives in Chinese discourse parsing," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, (Osaka, Japan), pp. 1891–1902, The COLING 2016 Organizing Committee, Dec. 2016.

[15] M. Laali and L. Kosseim, "Automatic disambiguation of french discourse connectives," *International Journal of Computational Linguistics and Applications*, vol. 7, no. 1, pp. 11–30, 2016.

[16] S. Dipper and M. Stede, "Disambiguating potential connectives," 01 2006.

[17] A. Zwicky, *On Clitics*. Bloomington: Indiana University Linguistics Club, 1977.

[18] A. Zwicky, "Clitics and particles," *Language*, vol. 61, pp. 283–305, 1985.

[19] B. Gerlach, *Clitics between Syntax and Lexicon*, vol. 51. John Benjamins Publishing Company, 2002.

[20] N. Pavlou and P. Panagiotidis, *The morphosyntax of -nde and post-verbal clitics in Cypriot Greek*, vol. 206 of *Linguistik Aktuell/Linguistics Today*. John Benjamins Publishing Company, 2013.

[21] E. Opengin, "Topicalisation in central kurdish," in *Talk at the workshop Information structure in spoken language corpora. Bielefeld*, 2013.

[22] A.-C. Hellenthal, *A grammar of Sheko*. Netherlands Graduate School of Linguistics, 2010.

[23] I. Nikolaeva and M. Tolskaya, *A Grammar of Udihe*. Mouton Grammar Library [MGL], De Gruyter, 2011.

[24] P. Denwood, *Tibetan*. London Oriental and African Language Library, John Benjamins Publishing Company, 1999.

[25] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The Penn Discourse TreeBank 2.0.," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, (Marrakech, Morocco), European Language Resources Association (ELRA), May 2008.

[26] W. Mann, C. Matthiessen, and S. Thompson, "Rhetorical structure theory and text analysis," *Discourse Description: Diverse Linguistic Analyses of a Fund Raising Text*, p. 66, 11 1989.

[27] A. Al-Saif and K. Markert, "The Leeds Arabic discourse treebank: Annotating discourse connectives for Arabic," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, (Valletta, Malta), European Language Resources Association (ELRA), May 2010.

[28] N. Xue, "Annotating discourse connectives in the Chinese treebank," in *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, (Ann Arbor, Michigan), pp. 84–91, Association for Computational Linguistics, June 2005.

[29] U. Oza, R. Prasad, S. Kolachina, D. Misra Sharma, and A. Joshi, "The Hindi discourse relation bank," in *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, (Suntec, Singapore), pp. 158–161, Association for Computational Linguistics, Aug. 2009.

[30] D. Zeyrek, D. Turan, C. Bozşahin, R. Çakıcı, A. B. Sevdik-Çallı, I. Demirsahin, B. Aktas, İ. Yalçınkaya, and H. Ögel, "Annotating subordinators in the Turkish discourse bank," in *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, (Suntec, Singapore), pp. 44–47, Association for Computational Linguistics, Aug. 2009.

[31] M. Stede, "Disambiguating rhetorical structure," *Research on Language and Computation*, vol. 6, pp. 311–332, 2008.

[32] M. Stede, "The Potsdam commentary corpus," in *Proceedings of the Workshop on Discourse Annotation*, (Barcelona, Spain), pp. 96–102, Association for Computational Linguistics, July 2004.

[33] M. Buch-Kromann and I. Korzen, "The unified annotation of syntax and discourse in the copenhagen dependency treebanks," in *Proceedings of the Fourth Linguistic Annotation Workshop*, (Uppsala, Sweden), pp. 127–131, Association for Computational Linguistics, July 2010.

[34] D. Zeyrek and B. Webber, "A discourse resource for Turkish: Annotating discourse connectives in the METU corpus," in *Proceedings of the 6th Workshop on Asian Language Resources*, 2008.

[35] D. Zeyrek, *On the distribution of the contrastive-concessive discourse connective ama 'but/yet' and 'fakat' in written Turkish*, pp. 251–274. John Benjamins, 01 2014.

[36] I. Demirsahin and D. Zeyrek, "Annotating discourse connectives in spoken turkish," in *LAW VIII - The 8th Linguistic Annotation Workshop*, pp. 105–109, 01 2014.

[37] D. Zeyrek and M. Kurfalı, "An assessment of explicit inter- and intra-sentential discourse connectives in turkish discourse bank," in *LREC*, 2018.

[38] D. Zeyrek and K. Başıbüyük, "TCL - a lexicon of Turkish discourse connectives," in *Proceedings of the First International Workshop on Designing Meaning Representations*, (Florence, Italy), pp. 73–81, Association for Computational Linguistics, Aug. 2019.

[39] D. Zeyrek, A. Mendes, Y. Grishina, M. Kurfalı, S. Gibbon, and M. Ogrodniczuk, "Ted multilingual discourse bank (ted-mdb): a parallel corpus annotated in the pdtb style," *Language Resources and Evaluation*, vol. 54, 04 2019.

[40] M. Laali, A. Cianflone, and L. Kosseim, "The CLaC discourse parser at CoNLL-2016," in *Proceedings of the CoNLL-16 shared task*, (Berlin, Germany), pp. 92–99, Association for Computational Linguistics, Aug. 2016.

[41] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[42] E. Brill, "A simple rule-based part of speech tagger," *Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy*, 05 2002.

[43] A. Schiller, S. Teufel, and C. Thielen, "Guidelines für das Tagging deutscher Textcorpora mit STTS," tech. rep., Technical Report, IMS-CL, University Stuttgart, 1995, 1995.

[44] Y. Li, W. Feng, J. Sun, F. Kong, and G. Zhou, "Building Chinese discourse corpus with connective-driven dependency tree structure," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 2105–2114, Association for Computational Linguistics, Oct. 2014.

[45] L. Danlos, D. Antolinos-Basso, C. Braud, and C. Roze, "Vers le FDTB : French discourse tree bank (towards the FDTB : French discourse tree bank) [in French]," in *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, (Grenoble, France), pp. 471–478, ATALA/AFCP, June 2012.

[46] N. Xue, H. T. Ng, S. Pradhan, R. Prasad, C. Bryant, and A. Rutherford, "The CoNLL-2015 shared task on shallow discourse parsing," in *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, (Beijing, China), pp. 1–16, Association for Computational Linguistics, July 2015.

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

[48] N. Xue, H. T. Ng, S. Pradhan, A. Rutherford, B. Webber, C. Wang, and H. Wang, "CoNLL 2016 shared task on multilingual shallow discourse parsing," in *Proceedings of the CoNLL-16 shared task*, (Berlin, Germany), pp. 1–19, Association for Computational Linguistics, Aug. 2016.

[49] A. Zeldes, D. Das, E. G. Maziero, J. Antonio, and M. Iruskieta, "The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection," in *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, (Minneapolis, MN), pp. 97–104, Association for Computational Linguistics, June 2019.

[50] S. Gopalan and S. Lalitha Devi, "BioDCA identifier: A system for automatic identification of discourse connective and arguments from biomedical text," in *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, (Osaka, Japan), pp. 89–98, The COLING 2016 Organizing Committee, Dec. 2016.

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

[52] D. Zeyrek and M. Kurfalı, "TDB 1.1: Extensions on Turkish discourse bank," in *Proceedings of the 11th Linguistic Annotation Workshop*, (Valencia, Spain), pp. 76–81, Association for Computational Linguistics, Apr. 2017.

[53] M. Ergin, *Türk Dil Bilgisi*. Istanbul: Boğaziçi Yayınları, 1975.

[54] J. Kornfilt, *Turkish*. Descriptive Grammars, London: Routledge, 1997.

[55] G. Lewis, *Turkish grammar*. Oxford University Press, 1967.

[56] U. Arıkan, O. Güngör, and S. Üsküdarli, "Detecting clitics related orthographic errors in Turkish," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, (Varna, Bulgaria), pp. 71–76, INCOMA Ltd., Sept. 2019.

[57] B. Aktaş, C. Bozsahin, and D. Zeyrek, "Discourse relation configurations in Turkish and an annotation environment," in *Proceedings of the Fourth Linguistic Annotation Workshop*, (Uppsala, Sweden), pp. 202–206, Association for Computational Linguistics, July 2010.

[58] N. Asher, *Reference to Abstract Objects in Discourse*. Dordrecht, Boston, and London: Kluwer, 1993.

[59] T. Givón, *Topic Continuity in Discourse: A quantitative cross-language study*. Typological Studies in Language, John Benjamins Publishing Company, 1983.

[60] J. Kornfilt, *Turkish*. Descriptive Grammars, London: Routledge, 1997.

[61] D. Dönük, "Additive enclitic suffix -da in turkish as a cohesive device," *Ankara Üniversitesi Dil Dergisi*, vol. 130, pp. 54–67, 2008.

[62] R. Schiering, "Morphologization in turkish: Implications for phonology in grammaticalization," in *Paper presented at 13th International Conference on Turkish Linguistics*, (Uppsala), 2006.

[63] B. Say, D. Zeyrek, K. Oflazer, and U. Özge, "Development of a corpus and a treebank for present-day written turkish," *Proceedings of the 11th International Conference of Turkish Linguistics (ICTL)*, pp. 183–192, 01 2002.

[64] T. Sezer and B. Sezer, "Ts corpus: Herkes İçin türkçe derlem," in *Proceedings 27th National Linguistics Conference*, pp. 217–225, 05 2013.

[65] T. Sezer, "Ts corpus project: An online turkish dictionary and ts diy corpus," *European Journal of Language and Literature*, vol. 3, pp. 18–24, 2017.

[66] H. Sak, T. Güngör, and M. Saraçlar, "Turkish language resources: Morphological parser, morphological disambiguator and web corpus," in *Advances in Natural Language Processing* (B. Nordström and A. Ranta, eds.), (Berlin, Heidelberg), pp. 417–427, Springer Berlin Heidelberg, 2008.

[67] Prasad, Rashmi and Webber, Bonnie and Lee, Alan and Joshi, Aravind, *Penn Discourse Treebank Version 3.0*. Abacus Data Network, 2019.

[68] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, p. 37, 1960.

[69] W. Spooren and L. Degand, "Coding coherence relations: Reliability and validity," *Corpus Linguistics and Linguistic Theory*, vol. 6, 10 2010.

[70] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[71] M. Straka, J. Hajič, and J. Straková, "UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, (Portorož, Slovenia), pp. 4290–4297, European Language Resources Association (ELRA), May 2016.

[72] M. Straka and J. Straková, "Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe," in *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, (Vancouver, Canada), pp. 88–99, Association for Computational Linguistics, Aug. 2017.

[73] M. Straka, "UDPipe 2.0 prototype at CoNLL 2018 UD shared task," in *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, (Brussels, Belgium), pp. 197–207, Association for Computational Linguistics, Oct. 2018.

[74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[75] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: a library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 08 2008.

[76] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[77] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifier," *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, vol. 5, 08 1996.

[78] M. Kurfalı, "Automatic sense prediction of implicit discourse relations in turkish," Master's thesis, Middle East Technical University, 2016.

[79] G. Valentini and F. Masulli, "Ensembles of learning machines," in *Neural Nets WIRN Vietri-2002, Series Lecture Notes in Computer Sciences*, vol. 2486, pp. 3–22, 05 2002.

[80] T. K. Ho, "Random decision forests," in *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, (M), pp. 278–282, IEEE Computer Society, 1995.

[81] A. Isaksson, M. Wallman, H. Goransson Kultima, and M. Gustafsson, "Cross-validation and bootstrapping are unreliable in small sample classification," *Pattern Recognition Letters*, vol. 29, pp. 1960–1965, 10 2008.

[82] D. Berrar, *Cross-Validation*, vol. 1, pp. 542–545. Elsevier, 01 2018.

# Appendix A

# CONFUSION MATRICES

Table 18: Logistic Regression Training I

| **True Value** \ **Predicted Value** | **DC** | **NDC** |
|---|---|---|
| DC | 63 | 15 |
| NDC | 14 | 60 |

Table 19: Logistic Regression Training II

| **True Value** \ **Predicted Value** | DC | NDC |
|---|---|---|
| DC | 62 | 17 |
| NDC | 10 | 63 |

Table 20: Logistic Regression Training III

| **True Value** \ **Predicted Value** | DC | NDC |
|---|---|---|
| DC | 57 | 21 |
| NDC | 13 | 61 |

Table 21: Support Vector Machine (SVM) Training I

| **True Value** \ **Predicted Value** | DC | NDC |
|---|---|---|
| DC | 58 | 20 |
| NDC | 10 | 64 |

Table 22: Support Vector Machine (SVM) Training II

| **True Value** \ **Predicted Value** | DC | NDC |
|---|---|---|
| DC | 58 | 21 |
| NDC | 5 | 68 |

Table 23: Support Vector Machine (SVM) Training III

| **True Value** \ **Predicted Value** | DC | NDC |
|---|---|---|
| DC | 53 | 25 |
| NDC | 6 | 68 |

Table 24: Random Forest Training I

| **True Value** \ **Predicted Value** | DC | NDC |
|---|---|---|
| DC | 65 | 13 |
| NDC | 13 | 61 |

Table 25: Random Forest Training II

| **True Value** \ **Predicted Value** | DC | NDC |
|---|---|---|
| DC | 59 | 20 |
| NDC | 9 | 64 |

Table 26: Random Forest Training III

| **True Value** \ **Predicted Value** | DC | NDC |
|---|---|---|
| DC | 57 | 21 |
| NDC | 8 | 66 |

# Appendix B

# CROSS-VALIDATION RESULTS FOR EVALUATION

Since cross validation does not yield reliable results in small data set, we have added two cases where it gives both higher results and lower results than standard measurement methods.

Table 27, 28 and 29 provide the results achieved with 5 fold cross validation that are lower than the presented results in Chapter 6.

Table 30, 31 and 32 provide the results achieved with 5 fold cross validation that are higher than the presented results in Chapter 6.

Table 27: Results of Logistic Regression with 5 fold Cross Validation-Lower Scores

| Folds / Confusion Matrices | Fold 1 | Fold2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|---|
| Accuracy | 0.80 | 0.74 | 0.73 | 0.90 | 0.73 |
| Precision | 0.86 | 0.80 | 0.73 | 0.86 | 0.85 |
| Recall | 0.76 | 0.70 | 0.73 | 0.92 | 0.66 |
| F1 | 0.81 | 0.75 | 0.73 | 0.89 | 0.75 |

Table 28: Results of Support Vector Machine (SVM) with 5 fold Cross Validation-Lower Scores

| Folds / Confusion Matrices | Fold 1 | Fold2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|---|
| Accuracy | 0.74 | 0.70 | 0.73 | 0.96 | 0.76 |
| Precision | 0.80 | 0.86 | 0.80 | 1.00 | 0.92 |
| Recall | 0.70 | 0.65 | 0.70 | 0.93 | 0.68 |
| F1 | 0.75 | 0.74 | 0.75 | 0.96 | 0.78 |

Table 29: Results for Random Forest with 5 fold Cross Validation-Lower Scores

| Folds / Confusion Matrices | Fold 1 | Fold2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|---|
| Accuracy | 0.74 | 0.74 | 0.73 | 0.83 | 0.80 |
| Precision | 0.80 | 0.80 | 0.73 | 0.86 | 0.85 |
| Recall | 0.71 | 0.66 | 0.71 | 0.92 | 0.72 |
| F1 | 0.75 | 0.77 | 0.68 | 0.89 | 0.77 |

Table 30: Results for Logistic Regression with 5 fold Cross Validation-Higher Scores

| Folds / Confusion Matrices | Fold 1 | Fold2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|---|
| Accuracy | 0.83 | 0.80 | 0.86 | 0.76 | 0.86 |
| Precision | 0.93 | 0.88 | 0.81 | 0.93 | 1.00 |
| Recall | 0.78 | 0.78 | 0.92 | 0.71 | 0.80 |
| F1 | 0.85 | 0.83 | 0.86 | 0.81 | 0.88 |

Table 31: Results for Support Vector Machine (SVM) with 5 fold Cross Validation-Higher Scores

| Folds / Confusion Matrices | Fold 1 | Fold2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|---|
| Accuracy | 0.87 | 0.80 | 0.86 | 0.76 | 0.86 |
| Precision | 1.00 | 0.88 | 0.93 | 0.93 | 1.00 |
| Recall | 0.80 | 0.78 | 0.83 | 0.71 | 0.80 |
| F1 | 0.88 | 0.83 | 0.88 | 0.81 | 0.88 |

Table 32: Results for Random Forest with 5 fold Cross Validation-Higher Scores

| Folds / Confusion Matrices | Fold 1 | Fold2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|---|
| Accuracy | 0.87 | 0.77 | 0.90 | 0.76 | 0.86 |
| Precision | 0.93 | 0.88 | 0.87 | 0.93 | 0.93 |
| Recall | 0.80 | 0.78 | 0.93 | 0.71 | 0.75 |
| F1 | 0.88 | 0.83 | 0.86 | 0.81 | 0.83 |

# Appendix C

# SAMPLE ANNOTATION AND PARSING MAPPING

Annotated data includes, ANT, dA, POST, TAG, SENSE, TDB-ANN, Level-1, Level-2, Level-3, RAW, RICH and CONTEXT as keywords.

**ANT:** contains words occur before *dA*
**dA:** contains dA itself
**POST** contains words occur after *dA*
**TAG:** contains DC or NDC tags, accordingly to discourse usage annotation of *dA*
**SENSE** when the 'TAG' is 'DC', contains discourse relation conveyed by *dA*
**TDB-ANN:** contains the annotation of the sample if it were done by previous studies in TDB 1.1.
**Level-1** when 'TAG' is 'DC', contains PDTB Level-1 senses regarding the function of *dA*.
**Level-2:** when 'TAG' is 'DC', contains PDTB Level-2 senses regarding the function of *dA*.
**Level-3:** when 'TAG' is 'DC' and the relation has third level sense in PDTB sense hierarchy, contains PDTB Level-3 senses regarding the function of *dA*.
**RAW:** contains raw sample.
**RICH:** contains sample processed in UDPipe.
**CONTEXT:** contains positions as keys; lemmas and POS tags as values.

```
{'ANT': 'Bir ayağı fincanın dışında sanki ! İşte öbür ayağını', 'dA': 'da', 'POST': 'kenardan dışarıya attı !', 'T
AG': 'DC', 'SENSE': 'ADDITIVE', 'TDB-ANN': 'ann : EntRel | Arg1( 6314 . .6346 ) | Arg2( 6348 . .6391 )', 'Level-1'
: 'EXPANSION', 'Level-2': 'LEVEL-OF-DETAIL', 'Level-3': 'ARG2-AS-DETAIL', 'raw': 'Bir ayağı fincanın dışında sanki
! İşte öbür ayağını da kenardan dışarıya attı !', 'rich': [{'Id': '1', 'Word_Form': 'Bir', 'Lemma': 'bir', 'UPosTa
g': 'NUM', 'XPosTag': 'ANum', 'Feats': 'NumType=Card', 'Head': '2', 'DepRel': 'det', 'Deps': '_', 'Misc': '_'}, {'
Id': '2', 'Word_Form': 'ayağı', 'Lemma': 'ayak', 'UPosTag': 'NOUN', 'XPosTag': 'Noun', 'Feats': 'Case=Acc|Number=S
ing|Person=3', 'Head': '5', 'DepRel': 'obj', 'Deps': '_', 'Misc': '_'}, {'Id': '3', 'Word_Form': 'fincanın', 'Lemm
a': 'fincan', 'UPosTag': 'NOUN', 'XPosTag': 'Noun', 'Feats': 'Case=Gen|Number=Sing|Person=3', 'Head': '4', 'DepRel
': 'nmod:poss', 'Deps': '_', 'Misc': '_'}, {'Id': '4', 'Word_Form': 'dışında', 'Lemma': 'dış', 'UPosTag': 'ADJ', '
XPosTag': 'NAdj', 'Feats': 'Case=Loc|Number=Sing|Number[psor]=Sing|Person=3|Person[psor]=3', 'Head': '5', 'DepRel'
: 'amod', 'Deps': '_', 'Misc': '_'}, {'Id': '5', 'Word_Form': 'sanki!', 'Lemma': 'sanki!', 'UPosTag': 'ADV', 'XPos
Tag': 'Adverb', 'Feats': '_', 'Head': '11', 'DepRel': 'advmod', 'Deps': '_', 'Misc': '_'}, {'Id': '6', 'Word_Form'
: 'İşte', 'Lemma': 'işte', 'UPosTag': 'ADV', 'XPosTag': 'Adverb', 'Feats': '_', 'Head': '11', 'DepRel': 'advmod',
'Deps': '_', 'Misc': '_'}, {'Id': '7', 'Word_Form': 'öbür', 'Lemma': 'öbür', 'UPosTag': 'ADJ', 'XPosTag': 'Adj', '
Feats': '_', 'Head': '8', 'DepRel': 'amod', 'Deps': '_', 'Misc': '_'}, {'Id': '8', 'Word_Form': 'ayağını', 'Lemma'
: 'ayak', 'UPosTag': 'NOUN', 'XPosTag': 'Noun', 'Feats': 'Case=Acc|Number=Sing|Number[psor]=Sing|Person=3|Person[p
sor]=3', 'Head': '11', 'DepRel': 'obj', 'Deps': '_', 'Misc': '_'}, {'Id': '9', 'Word_Form': 'da', 'Lemma': 'da', '
UPosTag': 'CCONJ', 'XPosTag': 'Conj', 'Feats': '_', 'Head': '8', 'DepRel': 'advmod:emph', 'Deps': '_', 'Misc': '_'
}, {'Id': '10', 'Word_Form': 'kenardan', 'Lemma': 'kenar', 'UPosTag': 'ADJ', 'XPosTag': 'NAdj', 'Feats': 'Case=Abl
|Number=Sing|Person=3', 'Head': '11', 'DepRel': 'amod', 'Deps': '_', 'Misc': '_'}, {'Id': '11', 'Word_Form': 'dışa
rıya', 'Lemma': 'dışarı', 'UPosTag': 'NOUN', 'XPosTag': 'Noun', 'Feats': 'Case=Dat|Number=Sing|Person=3', 'Head':
'0', 'DepRel': 'root', 'Deps': '_', 'Misc': '_'}, {'Id': '12', 'Word_Form': 'attı!', 'Lemma': 'at', 'UPosTag': 'VE
RB', 'XPosTag': 'Verb', 'Feats': 'Aspect=Perf|Mood=Ind|Number=Plur|Person=1|Polarity=Pos|Tense=Past', 'Head': '11'
, 'DepRel': 'compound', 'Deps': '_', 'Misc': '_'}], 'context': {-8: ['bir', 'NUM'], -7: ['ayak', 'NOUN'], -6: ['fi
ncan', 'NOUN'], -5: ['dış', 'ADJ'], -4: ['sanki!', 'ADV'], -3: ['işte', 'ADV'], -2: ['öbür', 'ADJ'], -1: ['ayak',
'NOUN'], 0: ['da', 'CCONJ'], 1: ['kenar', 'ADJ'], 2: ['dışarı', 'NOUN'], 3: ['at', 'VERB']}}
```


Figure 4: A Sample Annotation and Parsing Mapping in Hierarchical CSV Format (TDB 1.1, file 00002113)


```
{'ANT': 'Öğütme taşları çok büyük  ama öğütme taşlarının bir kısmı hem öğütme işleminde  hem', 'dA': 'de', 'POST':
'belki taş işçiliğinde kullanılıyor .', 'TAG': 'NDC', 'SENSE': '-', 'Level-1': '-', 'Level-2': '-', 'Level-3': '-'
, 'raw': 'Öğütme taşları çok büyük  ama öğütme taşlarının bir kısmı hem öğütme işleminde  hem de belki taş işçiliğ
inde kullanılıyor .', 'rich': [{'Id': '1', 'Word_Form': 'Öğütme', 'Lemma': 'öğüt', 'UPosTag': 'VERB', 'XPosTag': '
Verb', 'Feats': 'Aspect=Perf|Case=Nom|Mood=Ind|Polarity=Pos|Tense=Pres|VerbForm=Vnoun', 'Head': '2', 'DepRel': 'nm
od:poss', 'Deps': '_', 'Misc': '_'}, {'Id': '2', 'Word_Form': 'taşları', 'Lemma': 'taş', 'UPosTag': 'NOUN', 'XPosT
ag': 'Noun', 'Feats': 'Case=Nom|Number=Plur|Number[psor]=Sing|Person=3|Person[psor]=3', 'Head': '4', 'DepRel': 'ns
ubj', 'Deps': '_', 'Misc': '_'}, {'Id': '3', 'Word_Form': 'çok', 'Lemma': 'çok', 'UPosTag': 'ADV', 'XPosTag': 'Adv
erb', 'Feats': '_', 'Head': '4', 'DepRel': 'advmod', 'Deps': '_', 'Misc': '_'}, {'Id': '4', 'Word_Form': 'büyük',
'Lemma': 'büyük', 'UPosTag': 'ADJ', 'XPosTag': 'Adj', 'Feats': '_', 'Head': '9', 'DepRel': 'amod', 'Deps': '_', 'M
isc': '_'}, {'Id': '5', 'Word_Form': 'ama', 'Lemma': 'ama', 'UPosTag': 'CCONJ', 'XPosTag': 'Conj', 'Feats': '_', '
Head': '4', 'DepRel': 'cc', 'Deps': '_', 'Misc': '_'}, {'Id': '6', 'Word_Form': 'öğütme', 'Lemma': 'öğüt', 'UPosTa
g': 'VERB', 'XPosTag': 'Verb', 'Feats': 'Aspect=Perf|Case=Nom|Mood=Ind|Polarity=Pos|Tense=Pres|VerbForm=Vnoun', 'H
ead': '7', 'DepRel': 'nmod:poss', 'Deps': '_', 'Misc': '_'}, {'Id': '7', 'Word_Form': 'taşlarının', 'Lemma': 'taş'
, 'UPosTag': 'NOUN', 'XPosTag': 'Noun', 'Feats': 'Case=Gen|Number=Plur|Number[psor]=Sing|Person=3|Person[psor]=3',
'Head': '9', 'DepRel': 'nmod:poss', 'Deps': '_', 'Misc': '_'}, {'Id': '8', 'Word_Form': 'bir', 'Lemma': 'bir', 'UP
osTag': 'NUM', 'XPosTag': 'ANum', 'Feats': 'NumType=Card', 'Head': '9', 'DepRel': 'det', 'Deps': '_', 'Misc': '_'}
, {'Id': '9', 'Word_Form': 'kısmı', 'Lemma': 'kısım', 'UPosTag': 'NOUN', 'XPosTag': 'Noun', 'Feats': 'Case=Nom|Num
ber=Sing|Number[psor]=Sing|Person=3|Person[psor]=3', 'Head': '18', 'DepRel': 'nsubj', 'Deps': '_', 'Misc': '_'}, {
'Id': '10', 'Word_Form': 'hem', 'Lemma': 'hem', 'UPosTag': 'CCONJ', 'XPosTag': 'Conj', 'Feats': '_', 'Head': '12',
'DepRel': 'cc', 'Deps': '_', 'Misc': '_'}, {'Id': '11', 'Word_Form': 'öğütme', 'Lemma': 'öğüt', 'UPosTag': 'VERB',
'XPosTag': 'Verb', 'Feats': 'Aspect=Perf|Case=Nom|Mood=Ind|Polarity=Pos|Tense=Pres|VerbForm=Vnoun', 'Head': '12',
'DepRel': 'nmod:poss', 'Deps': '_', 'Misc': '_'}, {'Id': '12', 'Word_Form': 'işleminde', 'Lemma': 'işlem', 'UPosTa
g': 'NOUN', 'XPosTag': 'Noun', 'Feats': 'Case=Loc|Number=Sing|Number[psor]=Sing|Person=3|Person[psor]=3', 'Head':
'18', 'DepRel': 'obl', 'Deps': '_', 'Misc': '_'}, {'Id': '13', 'Word_Form': 'hem', 'Lemma': 'hem', 'UPosTag': 'CCO
NJ', 'XPosTag': 'Conj', 'Feats': '_', 'Head': '17', 'DepRel': 'cc', 'Deps': '_', 'Misc': '_'}, {'Id': '14', 'Word_
Form': 'de', 'Lemma': 'de', 'UPosTag': 'CCONJ', 'XPosTag': 'Conj', 'Feats': '_', 'Head': '13', 'DepRel': 'advmod:e
mph', 'Deps': '_', 'Misc': '_'}, {'Id': '15', 'Word_Form': 'belki', 'Lemma': 'belki', 'UPosTag': 'ADV', 'XPosTag':
'Adverb', 'Feats': '_', 'Head': '18', 'DepRel': 'advmod', 'Deps': '_', 'Misc': '_'}, {'Id': '16', 'Word_Form': 'ta
ş', 'Lemma': 'taş', 'UPosTag': 'NOUN', 'XPosTag': 'Noun', 'Feats': 'Case=Nom|Number=Sing|Person=3', 'Head': '17',
'DepRel': 'nmod:poss', 'Deps': '_', 'Misc': '_'}, {'Id': '17', 'Word_Form': 'işçiliğinde', 'Lemma': 'işçilik', 'UP
osTag': 'NOUN', 'XPosTag': 'Noun', 'Feats': 'Case=Loc|Number=Sing|Number[psor]=Sing|Person=3|Person[psor]=3', 'Hea
d': '18', 'DepRel': 'obl', 'Deps': '_', 'Misc': '_'}, {'Id': '18', 'Word_Form': 'kullanılıyor', 'Lemma': 'kullan',
'UPosTag': 'VERB', 'XPosTag': 'Verb', 'Feats': 'Aspect=Prog|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Polite=Infm
|Tense=Pres|Voice=Pass', 'Head': '0', 'DepRel': 'root', 'Deps': '_', 'Misc': '_'}, {'Id': '19', 'Word_Form': '.',
'Lemma': '.', 'UPosTag': 'PUNCT', 'XPosTag': 'Punc', 'Feats': '_', 'Head': '18', 'DepRel': 'punct', 'Deps': '_', '
Misc': '_'}], 'context': {-13: ['öğüt', 'VERB'], -12: ['taş', 'NOUN'], -11: ['çok', 'ADV'], -10: ['büyük', 'ADJ'],
-9: ['ama', 'CCONJ'], -8: ['öğüt', 'VERB'], -7: ['taş', 'NOUN'], -6: ['bir', 'NUM'], -5: ['kısım', 'NOUN'], -4: ['
hem', 'CCONJ'], -3: ['öğüt', 'VERB'], -2: ['işlem', 'NOUN'], -1: ['hem', 'CCONJ'], 0: ['de', 'CCONJ'], 1: ['belki'
, 'ADV'], 2: ['taş', 'NOUN'], 3: ['işçilik', 'NOUN'], 4: ['kullan', 'VERB'], 5: ['.', 'PUNCT']}}
```


Figure 5: A Sample Annotation and Parsing Mapping in Hierarchical CSV Format (TDB 1.1, file 00005121)

# Appendix D

# ML RESULTS FOR TRAINING CYCLES

Table 33: Evaluation Results for Training Cycle I

| **Training I** | | | |
|---|---|---|---|
| **Classifiers** / **Parameters** | **Logistic Regression** | **SVM** | **Random Forest** |
| Accuracy | 0.81 | 0.80 | **0.83** |
| Precision | 0.82 | **0.85** | 0.83 |
| Recall | 0.81 | 0.74 | **0.83** |
| F1 | 0.81 | 0.79 | **0.83** |

Table 34: Evaluation Results for Training Cycle II

| **Training II** | | | |
|---|---|---|---|
| **Classifiers** / **Parameters** | **Logistic Regression** | **SVM** | **Random Forest** |
| Accuracy | 0.82 | **0.83** | 0.81 |
| Precision | 0.86 | **0.92** | 0.87 |
| Recall | **0.78** | 0.73 | 0.75 |
| F1 | **0.82** | **0.82** | 0.80 |

Table 35: Evaluation Results for Training Cycle III

| **Training III** | | | |
|---|---|---|---|
| **Classifiers** / **Parameters** | **Logistic Regression** | **SVM** | **Random Forest** |
| Accuracy | 0.78 | 0.80 | **0.81** |
| Precision | 0.81 | **0.90** | 0.88 |
| Recall | **0.73** | 0.68 | **0.73** |
| F1 | 0.77 | 0.77 | **0.80** |

# Appendix E

# FEATURE WEIGHTS

Table 36: Weighted POS Tag Features for Logistic Regression

| POS tags \ Positions | -3 | -2 | -1 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|
| **ADP** | 0.408887007645108 | 0.5017936407524844 | 0.9239990442412912 | - | 0.45278479744209704 | 0.13449612556239465 |
| **ADV** | 0.2783210083345106 | 0.17640041607423557 | 1.2945546511169927 | 0.5420541003886239 | 0.12367037816686148 | 0.2275539840467775 |
| **AUX** | 0.6096128393937984 | 0.2787039378813854 | 0.19265553140640548 | 0.2489035484162773 | 0.018194706429185737 | 0.28883041592874803 |
| **CCONJ** | 0.1062409572509504 | 0.49294094926645987 | 1.11634970752153 | 0.0841034987804888 | 0.8920180453560627 | 0.12327931996309258 |
| **DET** | 0.16838273564850453 | 0.2657341257985865 | - | 0.5913623145093948 | 0.3378462453667147 | 0.2999065394725927 |
| **INTJ** | 0.09461319702841589 | - | 0.12377366140580126 | - | 0.0040831354016965075 | 0.10534105442002281 |
| **NOUN** | 0.35837654273673064 | 0.07954157425807772 | 0.17415071701678508 | 0.3200842747037121 | 0.05421848632327868 | 0.019656748470755395 |
| **NUM** | 0.44764853035744123 | 0.0795415742580772 | 0.1594800380719563 | 0.2573518208256289 | 0.04719960315758933 | 0.47945874778349795 |
| **PRON** | 0.08578236073196338 | 0.49862980409453017 | 0.41395570582017305 | 0.1778705509737208 | 0.2481948371154636 | 0.3416110739713394 |
| **PROPN** | 0.2325790970267306 | 0.3103081174922168 | 0.37286109705747733 | 0.42627065194231245 | 0.020202611350022544 | 0.1040384262950799 |
| **PUNCT** | 0.16326313477996487 | 0.07244403408333402 | - | 0.6021189553015142 | 0.46120084565071345 | 0.1771158149674601 |
| **VERB** | 0.22781558025933932 | 0.4601976982549534 | 2.32233407786395 | 0.34284026100160025 | 0.08233543363511808 | 0.03750592993990729 |

Table 37: Weighted POS Tag Features for SVM

| POS tags \ Positions | -3 | -2 | -1 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|
| **ADP** | 0.16770048183795416 | 0.2345710777985584 | 0.3951130349243309 | - | 0.1338257099532357 | 0.09702429539218627 |
| **ADV** | 0.11346470107673179 | 0.0014525334606652618 | 0.470981495970183 | 0.20365477497105727 | 0.07993712691735944 | 0.07997677609228421 |
| **AUX** | 0.3877901805448941 | 0.10957441055075588 | 0.20814493283793778 | 0.1596025560074375 | 0.0634264267119228 | 0.14824876561847283 |
| **CCONJ** | 0.07376334143383914 | 0.16569853301462328 | 0.29393695076372967 | 0.15816955092816093 | 0.5358931347159852 | 0.02044752157160616 |
| **DET** | 0.03301548644045399 | 0.18719340560821682 | - | 0.26171916532685646 | 0.16424372492722475 | 0.22353601544705504 |
| **INTJ** | 0.0659677024793734 2 | - | 0.08315128214954058 | - | 0.013037621785821601 | 0.10966271297733415 |
| **NOUN** | 0.06566043511576425 | 0.18698418027412328 | 0.03160250004629224 | 0.11678870049992729 | 0.060287233163114906 | 0.013985762906963878 |
| **NUM** | 0.26262730392125455 | 0.10365972173083574 | 0.11838644817789226 | 0.15022378047861298 | 0.03482527533418575 | 0.25697918079984533 |
| **PRON** | 0.02647159959125138 | 0.1980281336667442 | 0.18895971162172942 | 0.03027478888447699 | 0.13369196043673887 | 0.16420874914857292 |
| **PROPN** | 0.032951370625787596 | 0.13514067708815133 | 0.2053162753127135 | 0.19399641259289307 | 0.11236437967458507 | 0.0465236012299568 85 |
| **PUNCT** | 0.00038571647412061205 | 0.013004892290102349 | - | 0.5839730150450166 | 0.2325699107908476 | 0.07602810336319654 |
| **VERB** | 0.03710534024347306 | 0.21229375083700347 | 0.8223526111392216 | 0.17516958771315205 | 0.1023406749360154 | 0.027441629329130 37 |

Table 38: Weighted POS Tag Features for Random Forest

| POS tags \ Positions | -3 | -2 | -1 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|
| **ADP** | 0.0020433475109508303 | 0.0024015611206070274 | 0.008870471722575084 | - | 0.00276425107105718 | 0.0024328348639301163 |
| **ADV** | 0.003647732634261141 | 0.0017964225542523415 | 0.019575547018439193 | 0.002982947227259479 | 0.0012591461321925083 | 0.0024048514466901362 |
| **AUX** | 0.0013188866944429526 | 0.0007592109564414323 | 0.0003737439629329 1803 | 0.0004594838605719 0286 | 0.0004340827253770388 | 0.0004077421898218529 |
| **CCONJ** | 0.0022524733542080575 | 0.0027319214061196596 | 0.026502573477574894 | 0.0006039537234319824 | 0.002855707474939253 | 0.0022904052804000 05 |
| **DET** | 0.0013473875051191282 | 0.0022825584152057128 | - | 0.004398667946265015 | 0.0014814635137707696 | 0.0018493426151162703 |
| **INTJ** | 0.0000505778688492803 64 | - | 0.0002848781443548857 | - | 0.0003613573328436226 | 0.0001085535122870660 6 |
| **NOUN** | 0.0063948943575251 88 | 0.0059496534466031 2 | 0.013171123642333273 | 0.006378201792447827 | 0.004887231607519476 | 0.0049404460309878 98 |
| **NUM** | 0.0016630411726822 68 | 0.0022160294747426835 | 0.0006349380628495173 | 0.002017122861627 5984 | 0.0018753174587021052 | 0.0017295277435872961 |
| **PRON** | 0.0013304113765107 08 | 0.0036912006176355195 | 0.0035283002227058857 | 0.005213313444160169 | 0.0019919278607922115 | 0.0024083434808898777 |
| **PROPN** | 0.0023837573412952727 | 0.004577310332757777 | 0.0020354850885976297 | 0.000990692337918755 | 0.0010771316879823693 | 0.0018638155217245204 |
| **PUNCT** | 0.004555297698277329 | 0.0032773094558917 27 | - | 0.0018436523753695582 | 0.005252016645660191 | 0.0052099559010402 73 |
| **VERB** | 0.004333935367230767 | 0.0038538739453708 99 | 0.08077436878625686 | 0.005244191408517804 | 0.004257629973642881 | 0.0045710725827883 24 |

**TEZ İZİN FORMU /** THESIS PERMISSION FORM

**ENSTİTÜ /** INSTITUTE

**Fen Bilimleri Enstitüsü** / Graduate School of Natural and Applied Sciences ☐

**Sosyal Bilimler Enstitüsü** / Graduate School of Social Sciences ☐

**Uygulamalı Matematik Enstitüsü** / Graduate School of Applied Mathematics ☐

**Enformatik Enstitüsü** / Graduate School of Informatics ☐

**Deniz Bilimleri Enstitüsü** / Graduate School of Marine Sciences ☐

**YAZARIN /** AUTHOR

**Soyadı** / Surname : .....
**Adı** / Name : .....
**Bölümü** / Department : .....

**TEZİN ADI /** TITLE OF THE THESIS (**İngilizce** / English) : .....
.....
.....
.....
.....

**TEZİN TÜRÜ /** DEGREE**:** **Yüksek Lisans** / Master ☐ **Doktora** / PhD ☐

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ☐
2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two year. *** ☐
3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. ***** ☐

***** *Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir.*
*A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.*

**Yazarın imzası** / Signature ..... **Tarih** / Date .....