PRONOMINAL ANAPHORA RESOLUTION IN TURKISH AND ENGLISH

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

BY

MELEK ERTAN

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

JANUARY 2023

Approval of the thesis:

PRONOMINAL ANAPHORA RESOLUTION IN TURKISH AND ENGLISH

submitted by **MELEK ERTAN** in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Science Department, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç
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. Ümit Deniz Turan
English Language Teaching, Anadolu University _____

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

Assoc. Prof. Dr. Barbaros Yet
Cognitive Science, METU _____

Date: 27.01.2023

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: Melek Ertan

Signature :

ABSTRACT

PRONOMINAL ANAPHORA RESOLUTION IN TURKISH AND ENGLISH

Ertan, Melek
M.S., Department of Cognitive Science
Supervisor: Prof. Dr. Deniz Zeyrek Bozşahin

January 2023, 64 pages

This research analyzes pronominal anaphora in a Turkish and English translated TED corpus, namely the TED-MDB (Zeyrek et al., 2020) and presents a heuristic-based resolution algorithm for resolving pronominal anaphora in these languages separately. The corpus has characteristics of spoken language and has 364 English sentences aligned with their Turkish counterparts. The research is divided into two stages. In the first stage, the data was annotated using a web-based annotation tool INcePTION (Klie et al., 2018). The second phase of the study involves a computational analysis, where the traditional knowledge poor algorithm by Mitkov (1998) was tested on the annotated corpus for Turkish and English separately. The results showed that pronominal anaphora can be detected in TED talks with an F1-score of 0.61 in English, and with 0.63 in their Turkish translations.

Keywords: anaphora resolution, computational model, pronominal anaphora, knowledge based model, Natural Language Processing

ÖZ

TÜRKÇE VE İNGİLİZCE'DE ADILSAL ÖN GÖNDERİM ÇÖZÜMLEMESİ

Ertan, Melek
Yüksek Lisans, Bilişsel Bilimler Bölümü
Tez Yöneticisi: Prof. Dr. Deniz Zeyrek Bozşahin

Ocak 2023 , 64 sayfa

Bu araştırma, adılsal öngönderimi analiz eder ve adılsal öngönderim için buluşsal tabanlı bir çözümleme algoritmasını Türkçe ve İngilizce'de çevirilmiş TED derlemi olarak bilinen TED MDB için (Zeyrek ve diğ., 2020) ayrı olarak sunar. Derlem, konuşma dili niteliğinde olup, Türkçe karşılıklarıyla hizalanmış 364 İngilizce cümle içermektedir. Araştırma iki aşamaya ayrılmıştır. İlk aşamada, veriler web tabanlı işaretleme aracı INcePTION (Klie ve diğ., 2018) kullanılarak işaretlendi. Çalışmanın ikinci aşaması adılsal öngönderim için kurala dayalı bir kompütasyonal analizdir. Mitkov (1998)'un geleneksel bilgi tabanlı algoritması TED derleminde İngilizce ve Türkçe için ayrı olarak test edildi. Sonuçlar, adılsal öngönderimin İngilizce'de 0.61 ve Türkçe çevirilerinde 0.63 F1 puanı ile TED konuşmalarında tespit edilebileceğini göstermiştir.

Anahtar Kelimeler: artgönderim çözümlemesi, kompütasyonal model, adılsal artgönderim, bilgi tabanlı model, Doğal Dil İşleme

To all women whose dreams were taken away...

ACKNOWLEDGMENTS

I would like to start by thanking my beloved instructors in Cognitive Science department for helping me improve myself in this field and teaching me at this graduate level.

I am grateful to my thesis advisor Prof. Dr. Deniz Zeyrek Bozşahin for her valuable contributions, leading me during this process, and her constant care and feedback.

I would like to also send my biggest respects and thanks to Prof. Dr. Ümit Deniz Turan and Assoc. Prof. Dr. Barbaros Yet for their valuable feedback and comments on my thesis.

Words cannot express my gratitude to Fırat Öter for his continuous support and assistance both academically and personally. I am also indebted to Umutcan Üstüntaş for his help during this long journey. This endeavour would not have been possible without their assistance.

I would like to also thank Prof. Dr. Bilal Kırkıcı and Prof. Dr. Martina Gracanin Yüksek whom I met during my bachelor's degree for helping me find my interest in linguistics field with their great knowledge as well as encouraging me to pursue a graduate level degree.

I would like to extend my sincere thanks to my beloved friends Mervenur Çetin, Canan Doğan, and Owaym Khan for their support and not letting me give up.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

NP	Noun Phrase
NLP	Natural Language Processing
S	Sentence
BT	Government Binding Theory
TED-MDB	Technology, Entertainment, Design Multilingual Discourse Bank
TED	Technology, Entertainment, Design
U	Utterance
U_n	Nth Utterance
$C_f(U_n)$	Forward Looking Center
$C_b(U_n)$	Backward Looking Center
C_p	Preferred Center
AR	Anaphora Resolution
BFP	Brennan, Friedman, Pollard
FDG	Functional Dependency Parser of English
FFNN	Feed Forward Neural Network
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
LSTM	Long-Short Term Memory
BERT	Bidirectional Encoder Representations
ID	Identity Document
PDTB	Penn Discourse Tree Bank
WIT3	Web Inventory of Transcribed and Translated Talks
EU	European Union
VP	Verb Phrase

PP	Prepositional Phrase
JSON	JavaScript Object Notation
POS	Part of Speech
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

CHAPTER 1

INTRODUCTION

Although language understanding and language generation are relatively effortless tasks for human beings (because language is largely an automatic process), these tasks are quite challenging for computers. Anaphora resolution is one of these tasks that seems substantially easy at first glance. However, when we look at the underlying linguistic properties, it is observed to be more intricate than our assumptions. This makes it difficult to build a complete, estimable, and creditable automatic resolution system. Thus, anaphora resolution becomes a problem that needs to be addressed as language use is taking up a huge proportion of today's technology (speech recognition, voice commands, human robot interaction, translation etc.).

1.1 The Goal and Scope of the Research

The broad aim of this research is an analysis of pronominal anaphora in two typologically different languages (Turkish and English), which are chosen for several reasons. First of all, two languages bear different linguistic properties in terms of the way they tackle pronouns and their resolution. For example, these languages are different in terms of the order of the constituent and head directionality (English is head initial and Turkish is head final). Furthermore, being an agglutinative and pro-drop language, Turkish makes use of suffixes to mark person grammatically and pronouns can be dropped. In English, the pronoun appears on the surface level of the sentence. As a result, these languages might show differences in terms of the way they deal with anaphora as well. With this in mind, we decided to analyze pronominal anaphora in both languages. The dataset we used were the aligned translations of TED talks in the TED-MDB Corpus (Ozer & Zeyrek, 2019) because we wanted to see how anaphora takes place in the same context for these two different languages.

English sentences and Turkish counterparts of the TED-MDB were annotated separately and manually according to a set of guidelines developed. The output of this manual annotation was used as the input for a computational model. This was a rule-based computational model based on Mitkov (1998)'s model created to extract the antecedents of pronouns in the texts.

1.2 Research Questions

This study tries to understand the following research questions:

- How does pronominal anaphora take place in English TED talks and their Turkish translations, i.e. in typologically different languages?
- Is it possible to create a Knowledge Poor algoritm which utilizes the same sets of heuristics for English and Turkish for pronominal anaphora resolution?

1.3 Contributions

This thesis tries to contribute to the field by attempting to create two different models using the same rules in different ways for pronominal anaphora resolution in English and Turkish. This is a step for creating a bilingual anaphora resolution for Turkish and English. Different from most of the anaphora resolution systems in the literature, it tries to resolve cataphoric relations as well.

1.4 Outline of the Thesis

This thesis consists of five chapters. Chapter 1 provides the goal and focus of the research and the outline of the thesis is presented.

In Chapter 2, the terminology related to the phenomenon of anaphora and the types of anaphora both in English and Turkish are presented respectively with their detailed explanations. Later on, the theories which are related to anaphora are reviewed.

Chapter 3 describes the methodology of the thesis. The annotation process, preprocessing of the data and the heuristics of the model are given in detail.

Chapter 4 reports the performance of the model by presenting the recall, precision and F1 score evaluation metrics for the Turkish and English pronominal anaphora resolution models.

Next, Chapter 5 discusses the performance of the models and provides the limitations of study.

Having explained the present study's major aspects, the next chapter reviews the background with a literature review.

CHAPTER 2

LITERATURE REVIEW

2.1 Discourse, Referentiality, and Anaphora Resolution

In this chapter, some of the basic concepts of discourse and referentiality are defined to be able to present the nature of the research. When we are dealing with anaphora resolution, we are also dealing with discourse, i.e., the unit of language above the sentence. Discourse is characterized by several important features, namely, coherence, cohesion and referentiality. Bublitz (2011) defines cohesion as reference to the relations that occur among the structural parts of units of language such as a word, phrase, clause, or sentence. Nonetheless, these intra-sentential relations are distinct because they are governed by phonological and grammatical norms. On the other hand, Coherence is a cognitive attribute that is reliant on the interpretation of the language user and is not a constant quality of conversation or text. Insufficient cohesive mechanisms may disrupt the hearer's or reader's comprehension of coherence. The sentences a and b of example (1) are not coherent because the reader cannot understand what the relationship between two utterances is. It is difficult to comprehend. Also, example (1) lacks cohesion because none of the linguistic cues in b connects to a. For example, it cannot be understood who the pronoun 'they' refers to. On the other hand, it is highly possible that if the sentences are uttered within a shared context, they might be inferred as coherent and cohesive.

(1) a. I will call you when I am home.
b. They did not win the race.

Reference is a significant concept that makes a text cohesive. The connection between the linguistic form of an entity in the real word and the entity itself is called reference. The real-world object is called the referent. There can be different modes of reference between the linguistic form and the real-world entity, as explained below.

• **Exophora** (outer reference) - is the type of reference that the referent of the linguistic form is not in the text, but it is out of it (Nemcık, 2006).

• **Endophora** (inner reference) - is the kind of reference where both the referent and the linguistic form can be found in the text's space.

As being a part of endophora, the word anaphora comes from Greek word that means carrying back (Mitkov, 2022). Similarly, Halliday and Hasan (1976) give the defi-

nition of anaphora as "a cohesion which points back to some previous item". The cohesion occurs between two parts called antecedent and an anaphor. The entity that refers to another item previously mentioned is called an anaphor, while the previously introduced item is the antecedent. The whole process of identifying the anaphor and connecting it to its antecedent is called anaphora resolution (Mitkov, 2014). When both anaphor and the antecedent refer to the same entity out of the text, they co-refer, and they are called coreferential. The connection between these two parts is called coreference. The instances of coreference between various types of statements in the text are called coreference chains. Consider the examples below:

(2) ***Umut*** was crying, but ***he*** stopped when he saw his mother.

In example (2), the pronoun 'he' is identified as an anaphor while the antecedent is 'Umut'. We can understand the one who cried and stopped are both the same person who is in the real world denoted with the linguistic form 'Umut'. The relationship between the anaphor and the antecedent can be coreferential as it is observed in these examples because their referent in the real world is the same. However, there are some cases where this relationship between the anaphor and the antecedent is not coreferential. See the following example:

(3) Alex has been looking at his brother's ***toy***. He wants ***one***.

In example (3), the indefinite pronoun anaphor 'one' is used instead of his brother's toy. However, the referent of the antecedent and the anaphor does not co-refer. The world knowledge tells us that the boy wants a toy of his own that is like his brother's. Even though anaphora seems like a subcategory of coreference, it may fail in cases such as the examples given above. The research field that is known as anaphora resolution is considered as a subfield of entity resolution and even though it has some common features with coreference, it differs in certain contexts. To be able to solve the underlying conditions of anaphora resolution, many different approaches have been taken by the researchers in the literature, such as discourse analysis, Natural Language Processing (NLP) methods exploiting traditional computational models, machine learning models, deep learning models and neural network models. Before we dive into the details of the computational background, an extensive definition of types of anaphora and relations should be made. Even though types of anaphora have been examined as well as the suitable metrics for its evaluation and preprocessing, there seems little consensus about various types of anaphora. They have been categorized in many ways such as the form of anaphor, the locations of the antecedent and anaphora, and many more. In this section, the definitions of anaphora types that are covered in the research will be provided with detailed examples from (Mitkov, 2014), (Sukthanker et al., 2020) and (Yıldırım, 2008).

Table 2.1: Different Classifications of Anaphora by Researchers

Sukthanker et al. (2020)	Mitkov (2014)	Yıldırım (2008)
A) Types of anaphora according to the type of anaphor		
1-Pronominal Anaphora -One anaphora - Indefinite pronominal - Definite pronominal - Adjectival pronominal 2- Demonstratives 3- Presuppositions 4- Discontunious Sets 5- Inferable and Bridging	1- Pronominal Anaphora - Personal pronoun -Reflexive pronoun - Demonstrative pronoun -Relative pronoun -Adverb anaphora 2-Lexical Noun phrases -Bridging anaphora 3- Noun Anaphora(one) 4-Verb Anaphora 5-Zero Anaphora - Zero pronominal -Zero noun -Zero verb -Zero verb phrase	1- Pronominal(adılsal) -Nominative(yalın) -Accusative(belirtme) -Dative(yönelme) - Dative (çıkma/ayrılma) -Genitive(ilgi/tamlayan) -Locative(bulunma) Reflexive Anaphora(dönüşlü) -Reciprocal 2- Lexical NP 3- Subordinate Verb Clause 4- Zero (boş anaphora)
	B) Direct vs. Indirect anaphora	
	C) Types of the location of anaphor -Intrasentential -Intersentential	
	D) Identity sense vs. Rerefence	

2.1.1 Types of Anaphora

2.1.1.1 Types of Anaphora according to the form of anaphor

1. Pronominal anaphora

This type of anaphora is the most common and studied form of anaphora. The pronominal anaphora occurs with pronouns of any kind. To put it more clearly, per-

sonal pronouns, possessive pronouns, reflexive pronouns, demonstrative pronouns, relative pronouns, zero pronouns, local and temporal pronouns and indefinite pronouns are in the scope of pronominal anaphora. Example (4) illustrates this:

(4) ***Ayşe and her sister*** love reading, but ***they*** do not like writing.

In example(4), the personal pronoun refers to the noun phrase 'Ayşe and her sister' and it is categorized as a pronominal anaphora. Sukthanker et al. (2020) gives different names for conjoined sets of noun phrases while Mitkov (2014) categorizes them as pronominal anaphora. As a result, this set of noun phrase 'Ayşe and her sister' is recognized as split anaphora (discontinuous sets) by Sukhtanker. However, sometimes these split sets can be observed in different parts of the sentences. Yet, they might be the antecedent of only one linguistic form together. The second type that is categorized differently by Mitkov is presuppositions. They are the indefinite pronouns used commonly as it can be seen in the example given below.

(5) ***Everyone*** has the right to achieve ***their*** dream.

In example (5), the pronoun refers to the indefinite pronoun 'everyone'. However, it is not clear who these people are.

2. Zero Anaphora

Zero anaphora is the type of anaphora that is invisible on the surface level of the sentence. The representation of anaphora is not overtly done, but it can be interpreted from other clues in the sentence.

Zero anaphora has some sub-categories such as zero pronominal anaphora, zero noun anaphora and zero verb phrase anaphora (ellipsis). It is usually shown with the sign ∅.

(6) ***I*** went to school and ∅ talked to my friends.

Zero noun anaphora occurs when the head of the noun phrase is dropped. Most of the time there is a modifier visible in the sentence.

(7) Mary got five ***books*** for herself, but Jane didn't get any ∅.

The last form of zero anaphora which is known as zero verb phrase anaphora is also called ellipsis. It arises when the verb in the sentence is omitted, and it refers to a verb or verb clause in the previous sentence.

(8) I have never ***seen penguins***, but my sister has ∅.

In example (8), the variable ∅ refers to the verb phrase 'seen penguins'.

3. Verb Anaphora

In the given example:

(9) They ***had a terrible time on vacation***, so ***did*** we.

the verb form did stands for 'had' in the sentence. This type of anaphora is called verb anaphora.

On the other hand, Yıldırım (2008) introduces a sub-category of verb anaphora, which is the subordinate verb clause anaphora. Even though this type is called 'subordinate verb clause', the equivalent in English is not just verb anaphora, but it also includes relative clauses. To provide a clearer example examine the Turkish sentence below.

(10) a. ***Öğretmenler*** ∅ okula erken gelmek istemiyor.
b. Songül ***tuttuğu evi*** çok beğendi.

In example (10-a), we can observe a zero pronoun that is the agent of the subordinate clause. The zero pronoun refers to 'öğretmenler'. Moreover, example (10-b) is categorized by Yıldırım (2008) as a subordinate clause anaphora as well. The suffix(-DIK) in thesubordinate clause agrees with the person it refers to, which is 'Songül'. [1]

4. Adverb Anaphora

Adverbs of time and place are also identified as different types of anaphora. They may be given in two forms: locative 'there' and temporal 'then'. Mitkov (2014) includes adverb anaphora under the category of pronominal anaphora.

(11) a. Is she going to ***the supermarket***? My friend will be ***there***, too.
b. ***During World War I***, a lot of people died. Nobody knew the exact number back ***then***.

As it is shown in example (11), the adverbs 'there' and 'then' refer to a place and the time period mentioned in the text. However, these two adverbs are frequently used deictically in spoken language. Therefore, not all forms of 'then' and 'there' can be marked as anaphora. The term deictic will be discussed in detail in the following parts.

[1] What Yıldırım (2008) defines here appears to be 'PRO' which is a null category. However, the term was not used in the original study.

2.1.1.2 Types of Anaphora according to the location of the anaphor

When the antecedent and the anaphor is used in the same sentence then this type is called intrasentential. Reflexives and pronouns are the main examples of this type of anaphora because they are used in the same sentence with their antecedents. On the other hand, if the antecedent and the anaphor are given in different sentences then they are called as intersentential anaphora. Most of the time the antecedents of intersentential anaphors are observed in the 2-3 preceding sentences. However, this span can be larger in spoken language or different types of texts (Hobbs, 1978).

2.1.1.3 Identity of sense Anaphora vs. Identity of Reference Anaphora

As stated in the previous parts of the chapter, not all varieties of anaphora are coreferential and this distinguishes anaphora resolution from coreference. When the antecedent and anaphor refer to the same entity in the real world, they are called coreferential. Coreferential anaphora denotes identity of reference anaphora because the discourse entity stands for the same item. On the other hand, it is possible that the relation between the anaphor and the antecedent might not stand for the same entity in the real world, but the sense of the real-world entity. What it means is that their form appearing in the sentence can be the same, but they do not refer to the same entity in real world. For instance,

(12) Merve had ***her nails*** done at the salon, Buket got ***them*** done in the same place, too.

At first glance, both the antecedent and theanaphor seems coreferential. However, what 'them' stands for in the second part of the sentence is not Merve's nails, but Buket's nails. Therefore, the anaphor 'them' and 'her nails' are not coreferential, but it is an example of identity of sense anaphora. This anaphora is commonly observed with 'one' anaphora.

2.1.1.4 Anaphora vs. Cataphora

Cataphora's classification as either a form of anaphora or a distinct type of entity resolution is still a matter of controversy. When the anaphor is used before the antecedent in the sentence, it is called cataphora. The relationship between the parts is defined as cataphoric. Simply, cataphora is the opposite of anaphora. In this research, instead of treating cataphora as a different entity resolution task, I will approach the cataphoric relation as a different type of anaphoric relation.

2.1.2 Types of Relations in Anaphora

The sorts of relations between the anaphor and the antecedent are similarly not a consensus-based area in anaphora resolution. While some regard them as different

types of anaphora or entity resolution tasks, some argue that they should be considered as types of relations between the anaphor and the antecedent. the two languages seem to handle the same type of anaphora in different ways. With this in mind, three types of relations between the antecedent and the anaphor given in the literature will be reviewed.

Anaphoric relations: This is the type of relation that occurs when the anaphor is used after the antecedent in the text.

Cataphoric relations: This is observed when the antecedent follows the anaphor in the text.

Ambiguous: This relationship occurs when the anaphor in a sentence has at least two or more plausible antecedents, and each of these antecedents produces a unique interpretation of the text.

2.1.3 Non-anaphoric uses of pronouns

Languages can be more complex than we have thought when we start analyzing them. Anaphora resolution as a part of language related task becomes more complex when we start observing non-anaphoric uses of pronouns. In other words, each pronoun we observe in the text may not be anaphoric and need an antecedent for the meaning to be conveyed. Therefore, I will cover the non-anaphoric uses of pronouns which are pleonastic 'it', deixis, and generic uses in this part.

Pleonastic: Third person pronoun 'it' can be non-anaphoric and it can be referred Lappin and Leass (1994). However, Celce-Murcia (1987) calls is 'prop it'. In these types of uses we see it used in the sentence mostly because of the syntactic necessities of English that require an overt subject in every sentence. Mitkov (2014) states that pleonastic 'it' appears in many different constructions such as structures with:

a. *adjectives like "it is enough. . . , it is significant. . . , it is clear. . . etc."*
b. *cognitive verbs like "it is thought..., it is considered. . . , it seems that. . . . etc."*
c. *weather related vocabulary items like "it is windy, it is rainy, it is snowy. . . etc."*
d. *time expressions like "it is 12 o'clock, it is about time, it is summer, . . . etc."*
e. *distance related expressions like "how close is it to Ankara, it is a long way to. . . etc."*
f. *idiomatic uses like "call it even, it is over. . . etc." . cleft constructions like "it was me that. . . .,it is Mrs. White who. . . etc."*

As it can be seen above, these pronouns used in these expression does not refer to anything in the real world or have an antecedent that they can get their meaning from. Therefore, they are non-anaphoric uses of it.

Generic uses: Similarly, some other indefinite pronouns or personal pronouns can be used non-anaphorically. Most of the generic uses are found in proverbs and sayings such as:

(13) ***He*** who dares wins.

In example (13) the pronoun do not refer to a person in the text anaphorically or cataphorically.

Deixis: Deictic uses are more common in spoken forms of language and the anaphor which can mostly be a personal pronoun, demonstrative, or temporal adverb which does not relate to anything previously addressed in the text, but rather to a particular moment, person, or location inside the discourse (Mitkov, 2014).

2.2 Different Approaches and Theories

2.2.1 Government Binding Theory

Noam Chomsky as an influential linguist proposed Binding Theory as a part of Principles and Parameters theory (Chomsky et al., 1982). Some syntactic limits on the coreference of noun phrases are introduced by Binding Theory. Later, Minimalism Theory brought these two together and provided a broader explanation. The part that is relevant to anaphora resolution lies in the Binding Theory (BT). This theory especially deals with the anaphors and what antecedents can or cannot take as a referent. The term of c-command should be well established before the introduction of the principles given in Binding Theory. One commonly assumed version of command is the tree-configurational relation of c-command (Reinhart, 1983).

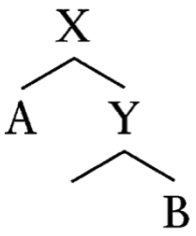

Figure 2.1: C-command tree configuration

Node A c-commands node B if and only if

a. A does not dominate B and B does not dominate A, and
b. the first branching node dominating A also dominates B.(Spencer et al., 1991)

The notations show the hierarchical relations in a syntactic tree. As it can be seen, the branching node X is in the highest hierarchy, and it dominates all the other nodes. The node that dominates A (which is X) also dominates B. Also, A does not dominate B. In this case B is c-commanded by A. However, if we look at the first branching note dominating B, it is given as Y. Since A is not dominated by Y, A is not in the c-command domain of B. This commanding relation is important when we are an-

alyzing principles of binding in the theory. The following is how BT differentiates between the three primary categories of NPs:

a. reflexives: , herself, kendine... [Principle A]
b. pronouns: we, they, siz... [Principle B]
c. full noun phrases: Ayşe, Mark... [Principle C] (Kurt, 2021)

The three principles of Binding Theory are also known as Principle A, B, and C (based on Chomsky et al., 1982):

a. An anaphor (reflexive or reciprocal) must be bound in its local domain.
b. A pronominal (non-reflexive pronoun) must not be bound in its local domain.
c. A non-pronoun (R-expression) must not be bound (Asudeh & Dalrymple, 2006).

Principle A: A reflexive pronoun in a sentence requires to have a close antecedent.

(14) $Mary_i$ is talking to $herself_i$.

In the example (14) given above, the reflexive pronoun can only refer to Mary and it cannot refer to anything else because a reflexive should be in the c-command domain of its antecedent and close to it. This also the same for reciprocal pronouns.

Principle B: All other non-reflexive pronominal pronouns cannot have an antecedent in their local domain in the sentence.

(15) $Ahmet_i$ hates him_j.

Example (15) has to follow condition B to be grammatical. In condition A, the reflexive required to be in the c-command domain of its antecedent. However, in the example above, the antecedent cannot be in the c-command domain of the pronoun. If it does, it becomes ungrammatical. The pronoun 'him' should refer to another person not Ahmet.

Principle C: A non-pronoun should be free in its local domain, and it cannot refer to a pronoun antecedent that c-commands itself.

(16) He_i believes $Mike_j$ will come back.

The example (16) above would be grammatical if and only if the pronoun refers to someone else, not Mike. If the pronoun refers to Mike, the non-pronoun is bound in its domain, and it violates Condition C. These principles in BT are considered in many models for anaphora resolution because they can eliminate some of the candidates violating the principles.

2.2.2 Centering Theory

The concepts of the "center" and "centering" were first introduced to specify an almost single mathematical approach to discourse interpretation in the work that Aravind Joshi and Steve Kuhn did in 1979 (A. K. Joshi & Kuhn, 1979). This was the beginning of what would later to be become known as Centering Theory. Centering Theory has grown out of computational linguistics and tries to explain how coherence takes place in the discourse as well as how interpretation occurs. It is the case that sometimes even if the sentences have the same propositional composition, the way they affect the coherence of the discourse might change tremendously. In a study, Sidner (1979) used 3 different Centering structures called discourse focus, actor focus and potential foci. The discourse focus stands for the topic about which the speaker seeks to make statements, while actor focus is the entity of discourse that is postulated as the agent of the occurrence in the utterance. On the other hand, potential foci or focus is a set of the substitute candidates for these two main foci. The focuses are notified with symbols and the descriptions of the symbols are given below.

a. U refers to the utterance and U_n is the n^{th} utterance in the discourse whereas U_{n+1} means the subsequent utterance.
b. $C_f(U_n)$ denotes the forward-looking centers in the utterance.
c. $C_b(U_n)$ marks the back-looking center of the utterance U_{n-1}. All sentenced except for the initial sentence has it.
d. C_f is ordered in terms of the syntactical role they get in the utterance.
e. If $C_f(U_n)$ has a higher rank in the utterance, it is highly possible to become the $C_b(U_{n+1})$. . The highest rank baring element of $C_f(U_n)$ is called the preferred center $C_p(U_n)$.
f. Four types of transition that ensues between each pair of utterances U_n and U_{n+1}.

Two of transitions in Centering Theory are continuation and retain. The other two are types of shifting which are smooth and rough shift (A. Joshi et al., 2005). There are also two rules that are given about the pronoun usage preference. The first rule that centers around the pronouns is significant for anaphora resolution. When one component of $Cf(Un)$ is utilized as a pronoun it is also used as a pronoun in $C_f(U_{n+1})$. The second rule is that the transitions preferred are continue, retain, smooth shift and rough shift respectively. The rank of the $C_f(U_n)$ for English are given as Subject > Direct Object > Indirect Object > Other subcategorized elements > Adjuncts by (Xiao, 2021). As we can see from the ranking, the grammatical structure of the sentence and roles decide the rank of $C_f(U_n)$. When it comes to Turkish, Turan (1998) gives a more detailed order, which is Empathy > Subject > Indirect Object > Direct Object > Others > Quantified Indefinite Subjects >Arbitrary Plural Null Pronominals.

To analyze the definitions and the transitions examine tables 2.3 and 2.4.

When both samples are analyzed, we can see that sample A is more coherent than sample B for many reasons. First of all, the center of the discourse is kept for all the utterances. If we look at U_1 and U_2, we can see that the higher ranked C_f is chosen for C_p and it is also held as a C_p in U_2. This transition is an example of continue. Since

Table 2.2: Transitions in Centering Theory adapted from A. Joshi et al., 2005

Transitions in centering	$C_b(U_{n+1})$ = $C_b(U_n)$ **or** $C_b(U_n)$ = **undefined**	$C_b(U_n + 1)$ =! $C_b(U_n)$
$C_b(U_{n+1})$ = $C_p(U_{n+1})$	Continue	Smooth shift
$C_b(U_{n+1})$ =! $C_p(U_{n+1})$	Retention	Rough shift

Table 2.3: *Centering Theory Sample A*

Sample A
U_1: Daniel works at a tech company. $Cf(U_1)$: {Daniel, tech company}, $C_p(U_1)$: {Daniel} and $C_b(U_1)$: {undefined}
U_2: He has completed more than 100 projects there. $C_f(U_2)$: {Daniel, projects, tech company} $C_p(U_2)$: {Daniel} and $C_b(U_2)$: {Daniel}
U_3: He likes his job because he loves coding. $C_f(U_3)$: {Daniel, job, coding} $C_p(U_3)$: {Daniel} and $C_b(U_3)$: {Daniel}

Table 2.4: *Centering Theory Sample B*

Sample B
U_5: Daniel works at a tech company. $C_f(U_5)$: {Daniel, tech company}, $C_p(U_5)$: {Daniel} and $C_b(U_5)$: {undefined}
U_6: The company finished many projects that he did. $C_f(U_6)$: {Tech company, projects, Daniel} $C_p(U_6)$: {tech company} and $C_b(U_6)$: {Daniel}
U_7 : He likes his job because he loves coding. $C_f(U_7)$: {Daniel, job, coding} $C_p(U_7)$: {tech company} and $C_b(U_7)$: {Daniel}

the U_1 has its C_b as undefined and it is different than U_2 a smooth shift is observed. On the other hand, the transition from U_5 to U_6 is different.This is because the C_p of U_6 is not the highest-ranking center in U_5. The center of utterance shifts from 'Daniel' to 'tech company'. Since both the C_b of the utterances are different and they have different C_p's, we observe a rough shift between U_5 and U_6.

2.2.3 Constraints for Anaphora Resolution

Humans use their innate knowledge of language to make the connection between the antecedent and the anaphor. However, when the computational models come into play, different questions and problems arise. A computer does not possess world knowledge, or it cannot grasp the semantic meaning of the words. Therefore, to be able to create a comprehensive and practical computational model, it is necessary to understand the underlying rules how anaphora resolution is carried out in the natural language. We need to determine a set of linguistic constraints (or hand-crafted rules) that might be useful for the identification and the determination of the antecedent among some possible candidates.

1. Gender Agreement

Any kind of anaphor and antecedent should agree on gender (masculine, feminine, neutral etc.). This is an important constraint in terms of ruling out many other noun phrases that has a different gender than the anaphor. English utilizes gender while Turkish does not. This means that, while this constraint can help the system to find the suitable antecedent of an anaphor in English, it may perform poorly in Turkish. As a result, only gender agreement will not be enough to identify the antecedent and some other constraints are needed.

2. Number Agreement

The antecedent and the anaphor should agree on the number as well. A singular antecedent can be referred by a singular pronoun or a plural noun can be referred with a plural pronoun. But metonymically used nouns or noun phrases may be problematic with number agreement. (See the example below)

(17) ***Facebook*** started to create a new platform for ***their*** new project.

According to number agreement, it would not be possible to identify the actual antecedent because it violates the number constraint. What is meant by 'Facebook' is 'the team in Facebook' and if we had a model that uses number agreement only, it would fail. This reveals that syntactic rules alone will not be enough.

3. Personal pronoun agreement

Based on this restriction, it seems that a personal pronoun and the antecedent to which it refers cannot exist together in a simple sentence [2](Kucuk & Yondem, 2007). See example (18):

(18) ***Mehmet*** saw ***him***.

Personal pronoun 'him' in example (18) cannot refer to 'Mehmet'. The pronoun must be free in its local domain according to Binding Theory. However, when this personal

[2] The term simple sentence is not defined in detail in the research. However, it seems like the term is used to refer to the resolution of anaphors in their local domain.

pronoun is reflexive, the pronoun has to be bound in its local domain. It cannot refer to another person. These are related to the principles of BT and mentioned above.

4. Grammatical Role

This syntactic constraint suggests that the subject of the sentence is given a higher priority than the noun in the object position. This can be given as a preference rather than a compulsory feature to be met. Under this category, we can also mention parallelism. It is more likely that an anaphor in the object position refers to an antecedent in the object position, whereas an anaphor in the subject position ismost likely to stand for an antecedent in the subject position.

5. Selectional Preferences

Selectional preferences of some words might demand the semantic information to be exploited, such as the animacy of the agent. Some verbs are required to occur with animate subjects. Therefore, animacy can be crucial during the identification of the antecedent.

6. Recency

The NP that is closest to the anaphor is given more salience. Proximity can be important when it comes to deciding among two possible candidates for resolution. However, this is not a compulsory condition. If two of the candidates happen to fulfill much more important constraints such as gender and number, the one that is adjacent to the anaphor is preferred.

7. Discourse Knowledge

Although there are many clues that help us to be able to eliminate the incompatible candidates during the identification of the antecedent such as semantic, syntactic and morphological clues, sometimes the knowledge of discourse is necessary for decision. The focus of the discourse lasts for a few sentences before it shifts to a different topic. Therefore, this can help us to identify the most suitable antecedent for an anaphor.

8. Repeated Mention

The NP that has been the focus of the discourse in the previous part of the text is given more salience. When a constituent is repeated throughout the context and there is another possible antecedent, the more mentioned candidate is preferred.

9. Syntactic Constraints

The process of anaphora resolution is inseparable from syntactic information in the text. C-command is a significant indicator of anaphora-antecedent matching. The anaphor should be in the c-command domain of the antecedent in normal cases. However, the cataphoric relation between the antecedent and the anaphor can be ruled out with this constraint.

10. World knowledge

World-knowledge or common sense is a big challenge for anaphora resolution systems. Since the computers are not capable of acquiring world knowledge like human

beings, it requires a lot of effort. It is quite possible that none of the constraints that are given above can narrow the possible antecedent candidates into one without world knowledge. These cases require attention and may affect the accuracy of the systems.

2.3 Computational Models

Anaphora resolution as an NLP task has attracted many scholars and it has been widely studied over the years. People had many different approaches on how to solve the task effectively and many of them were able to generate precise results. Some made use of lexical cues to handle the task while some of them preferred discourse based or syntactic based algorithms. In this section, diverse and influential approaches to the anaphora and pronoun resolution will be given in three main categories which are traditional approaches, machine learning and deep learning models.

2.3.1 Traditional Approaches

2.3.1.1 Hobbs Naïve Algorithm (Hobbs, 1978)

Hobbs' Naïve Algorithm was one of the most well-known traditional approaches in anaphora resolution (AR). This algorithm exploits a rule based, left to right breadth-search algorithm. Also, it utilizes syntactic parse trees to look for an antecedent for the pronoun. The algorithm traverses the tree and starts with the NP node of the pronoun. Then it moves up to the first S(sentence) node and looks for an NP node from left to right. If there is no NP node in this search, the nodes are pruned. In other words, the search for those nodes is frozen if they are not NP nodes. If there is an existing NP node on the search space, the algorithm searches if the node matches in terms of gender and number. The first node that matches the constraints and the selectional features are matched with the pronoun. The idea here indeed is related to the linguistic background of the phenomenon. When the search tree goes up into the S node and searches for a possible antecedent, it is looking for a node that c-commands the pronoun. Also, some constraints related to the principles of Binding Theory are included in this algorithm. Later on, Hobbs combined his naïve algorithm with a semantic approach where he created calculus axioms to represent the semantic information in the sentences and applies intersentence relation operation, predicate interpretation and bidirectional search that uses the naïve algorithm.

Even if the approach is one of the earliest algorithms to tackle pronoun resolution, Hobbs reports that the overall accuracy of the algorithm was 88.3% and when it is used with the selectional constraints, it rises up to 91.7%. However, he also states that half of the time there was only one suitable antecedent in the data points. Therefore, for 132 of more complex samples, the algorithm along with the selectional contstraints solves the 81.8% of the cases, which proves that the approach is quite successful. This algorithm has been used by many other researchers in Turkish, too. Tüfekçi and Kiliçaslan (2005) reformulate the steps in the algorithm and adds some other constraints for Turkish pronoun resolution. Later on, they compare the results of Hobbs' naïve algorithm with Mitkov's knowledge poor algorithm. They test the al-

gorithm on two different types of corpora and the results are in favor of the knowledge poor algorithm by 8.82% average for both corpora (Tüfekçi & Kiliçaslan, 2005).

2.3.1.2 Baldwin's COGNIAC

This algorithm is considered as one of the knowledge based algorithms and it is also a rule-based algorithm. The model requires sentence detection, part-of-speech tagger, noun phrase recognition, semantic information of the tokens such as gender, number and partial parse trees (Baldwin, 1997). When the algorithm decides that there is not a possible antecedent for the pronoun it does not match it. Since the algorithm does not possess extensive world knowledge, the ambiguity arises more compared to human beings. Like Hobbs' algorithm, COGNIAC also makes use of Binding Theory principles to look for antecedents in sentences that contain reflexive pronouns. They train the model on narrative text. The term 'possible antecedent' is used for the entities that follow the rules such as gender and number agreement. The algorithm rules are given below.

1- Unique in discourse: If there is one possible antecedent in the search scope, then pick it as the antecedent.

2- Reflexives: If there is a reflexive anaphor observed in the sentence, choose the closest possible antecedent as the antecedent.

3- Unique in prior and current: If first rules apply to the prior sentence and the read-in portion of the current sentence, pick the same antecedent.

4- Possessive pronoun: If the anaphor is a possessive pronoun, and the pronoun is observed in the previous sentence and unique, then pick it as the antecedent.

5- Unique current sentence: If there is one candidate available for the anaphor in the current sentence then choose it as the antecedent.

6- Unique subject/pronoun: The subject of the sentence has one possible antecedent in the previous sentence and the anaphor is in the subject position of the current sentence, choose the same antecedent.

How COGNIAC resolves anaphora can be given as follows. First of all, the pronouns are identified from left-to right. Then, the rules given above are applied in the order that they are given. When there is an antecedent that is identified, the next rules are not applied and necessary antecedent matching is done in the document. If the rule cannot identify an antecedent, the next rule is implemented until the end. If there is no match at the end of the last rule, the pronoun is left unresolved. The model reaches a significant precision result with 92% and 64% recall score in 200 pronouns resolved. This gives the idea that the model has been quite successful.

2.3.1.3 Lappin and Leass (1994)

The model proposed by the researchers themselves is called Resolution of Anaphora Procedure. It uses the syntactic parser of McCord's Slot Grammar and a simple dynamic model of attentional state (Lappin & Leass, 1994). The possible candidates in the previous part of the sentences are measured in their salience and given weights. At the end of the rule implementation, the one that is the most salient is chosen as the antecedent. Apart from being able to pair the anaphor and antecedent, the system can also identify pleonastic uses of 'it' and when it is identified, it stops and does not look for an antecedent. The candidates are eliminated based on the factors defined in the algorithms and they are eliminated by these rules. 7 factors are chosen and can be given as follows.

1. Recency: the candidates in the recent and most proximate position are preferred over the candidates that are far away.

2. Subjectivity: The grammatical role subject is preferred and given more salience. The order of salience in terms of the grammatical role is similar to the order given in Centering Theory.

3. Existential Emphasis: Existential constructs provide this priority to nominal predicates rather than others.

4. Accusative emphasis: Direct objects are preferred over others because they are the complements of the verbs.

5. Indirect vs. Oblique complement Emphasis: Indirect and oblique objects are given more salience after direct objects.

6. Head noun emphasis: The NPs that are not included in any of the other NPs are given preference.

7. Non-adverbial emphasis: The NPs that are not a part of the adverbial prepositional phrases are preferred.

As it can be understood from above, the grammatical role of the candidate affects its salience and gets a different weight that increases the chance of being selected.

This algorithm was tested with a corpus of computer manuals and 360 randomly selected pronouns were resolved. The algorithm performed with 86% accuracy on average. The accuracy of the algorithm was 89% for intrasentential pronouns and 72% for intersentential pronouns. When it is compared to Hobbs Naïve algorithm it performed 4% better on average. Yet, Naïve algorithm was more successful in solving intersentential pronouns.

2.3.1.4 BFP Algorithm

The name BFP stands for the names of the researchers who developed the algorithm. Brennan et al. (1987) created this algorithm by exploiting the rules in Centering Theory. As it was mentioned before, Centering Theory describes 4 types of transitions

between the utterances and the centers of these utterances. The researchers used these transitions and extended them to make them more specific. The shifting transitions were revisited, and they examined the shift in more detail and later on these shifts were named as rough and smooth shifts. When the backward-looking center (C_b) of the current utterance is not the same as the C_b of the previous utterance, and when the backward-looking center of the current utterance is the same as the preferred center of the current utterance, it is called a smooth shift. On the other hand, when the backward-looking center in the current sentence and the backward-looking center in the previous sentence are different and the preferred center of the current utterance is different from the backward-looking center of the current utterance, this is called a rough shift. This algorithm uses the ranking of the transitions in the Centering Theory, which is continue > retain > smooth shift > rough shift. The framework of the BFP algorithm can be given as follows.

- Compute the possible backward looking and forward-looking center combinations for the sentences.
- Utilize the rules and constraints given in the Centering Theory.
- Rate them by using the transition preference order. The aim of the algorithm was to provide conceptual clarity rather than efficiency. The plan was to add more constraints and preferences to easily extend the algorithm for more complex discourse structures.

2.3.1.5 Robust Knowledge Poor Algorithm

Mitkov's robust knowledge poor algorithm (Mitkov, 1998) was a very influential model and it has been also implemented in many different languages afterwards. They used computer manuals as the dataset for the resolution task. The idea of the model was that the input text was preprocessed by a part-of-speech tagger. When the algorithm started, it first looked for the pronouns. The next step was to identify the noun phrases in the current sentence. The extraction of NPs were carried out by grammatical rules. Only base NPs were identified. Complex or embedded NPs were not extracted.[3] Later, the search space was defined as the preceding 2 sentences before the current sentence where the anaphor is found. Different search scopes length as preceding 2,3 and 4 sentences were considered in different versions of the algorithm. The noun phrases were extracted as the possible antecedent candidates. The first constraint to eliminate the candidates was the gender and number agreement. The so-called antecedent indicators were applied to determine the most suitable antecedent (see below for details of the indicators). After all the antecedent indicators assigned a score to the NPs according to the definitions of the rules, the highest scoring candidate was determined as the antecedent of the anaphor. The algorithm does not resolve cataphoric relations. Similarly, Lappin Leass does not solve cataphoric relations, either. Knowledge poor algorithm eliminates the pleonastic uses of 'it'. The fully automatic anaphora resolution algorithm is called MARS. MARS uses Fuctional Dependency Parser of English known as FDG parser that can provide dependency relations for the words as well as providing lemmas and syntactic roles of the words. MARS also adds

[3] The grammatical rules for NP extraction are not explained in detail.

some more antecedent indicators to the original approach for better resolution results. In the original approach, Mitkov defines some antecedent indicators for the resolution process. After the elimination of the gender and number agreement, all possible antecedents are given a score between +2, +1, 0, and -1. The antecedent indicators can both decrease or increase the score of the NP. The indicators that assign a negative score has the impeding capacity, while the indicator that can increase the score has the boosting capacity. The antecedent indicators are:

1. First noun phrases: the very first NP in the sentence is given the score of +1. This gives the subject of the sentence more salience.

2. Indicating verbs: The NPs that are preceding a set of predefined verbs [4] are given the score of +1. Mitkov reports that NPs following these verbs are more salient according to empirical evidence.

3. Lexical iteration: The NPs which are observed twice, or more are given +2 while the NP that is observed once is given +1 in the paragraph that the pronoun appears. The mention of the NPs did not have to be in the same form. The NPs that had the same head are counted such as 'a bottle', 'the bottle', 'toner bottle'.

4. Section heading preference: If the NPs are also seen in the name of the section of the computer manual, they are given +1 score.

5. Collocation match: If the NP has the same collocation pattern with the pronoun, it is awarded with +2. Example (19) shows a collocation match.

(19) Press ***the key*** down and turn the volume up...Press ***it*** again.

6. Immediate reference: The NPs that are observed with a construction of and, or, after, until etc. are given a score of +2. The NP 'the printer' is given immediate reference score in example (20).

(20) To print the paper, you can stand ***the printer*** up or lay ***it*** flat.

7. Sequential instructions: A score of +2 is given to NPs in the NP1 position of constructions: *'To V1 NP1, V2 NP2. (Sentence). To V3 it, V4 NP4'* where the noun phrase NP1 is the possible antecedent of the anaphor 'it' and given the score of +2. In example (21), the first NP 'the video recorder' is given the score of +2.

(21) To turn on ***the video recorder***, press the red button. To programme ***it***, press the 'Programme key.

8. Term preference: The NPs that represents one of the terms in the genre of the text are given +1 score.

9. Indefiniteness: as being one of the impeding indicators, the indefinite NPs are

[4] Verb set = discuss, present, illustrate, identify, summarise, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore, assess, analyse, synthesise, study, survey, deal, cover

given -1 score.

10. Prepositional noun phrases: The NPs that are observed in the prepositional phrases are given a score of -1. Prepositional phrase in example (22) should be scored -1.

(22) Insert the casette ***into the VCR*** making sure it is suitable for the length of recording.

11. Referential distance: The distance of the NP to the pronoun can increase its chance to be the antecedent or not. Therefore, the NP that appears in the previous part of the same sentence with the pronoun is given the score of +2, the NPs in the previous sentence are given +1, The NPs that are before the previous sentence are given 0 and if the NPs are more distant, they are given -1.

The English model reached a success rate of 89.7% which is quite impressive. Later on, the model was implemented on different languages such as Polish and Arabic and they also generated similar and even better success rates, which proves the reliability of the model. Interestingly, the model was also implemented with bilingual corpora and translation of texts in French and English. The results were slightly poorer than only one language resolving models which achieved a success rate of 76.52%. The knowledge poor approach has also been applied in Turkish by Kucuk and Yondem (2007) and the results were promising.

2.3.1.6 Machine & Deep Learning Approaches

Mention Pair Models

Back in the 1980s, rule-based approaches and heuristics for the anaphora resolution taskswere popular and until now several of them have been discussed. When there was an increase in the corpus-based approaches for coreference resolution, the machine learning approaches also gained interest. One of the early and influential models was the mention-pair model. The first researchers who proposed it were Aone and William (1995) and later on many others contributed to the model. The mention-pair model tackles the coreference resolution task as a classification task and it is a supervised learning model. That means the coreferential chains between the NPs are given with their values and the model learns from these instances to be able to classify new instances. The model is trained to classify if two given NPs are coreferential or not. In the model, the NPs are represented with feature vectors including syntactic, morphological, semantic, and lexical information. The model has two main steps. Firstly, the NPs are classified as being coreferential or not. Later, the chains are created based on the positively classified pairs. The learning algorithms used were mostly decision-trees and later different learning models were used such as memory-based, support vector machines, maximum entropy learners, and Bayesian model.

Entity Mention Model

Mention-pair models would determine if one NP was coreferential with an antecedent

or not. It did not compare it with the other available antecedents. To overcome some of these disadvantages, the entity mention model was proposed. The idea was that the previous information about coreference was important for the decisions to be made in the upcoming parts of the text. The model tries to deal with the 'expressiveness'. To achieve this, the training instances are changed from NPs and the positive or negative coreference pairs to a pair of NP, cluster and a label showing if the designation of the cluster and the NP is positive or negative. The entity mention model was used on a dataset which was trained with a mention-pair model previously and they used decision tree classifiers and inductive logic programming. Even if the results of inductive logic programming showed significant increase, the overall the model could not perform as good as the mention-pair models because it was very difficult to represent clusters as features.

Mention Ranking Model

Another disadvantage of mention pair model was that it was using binary classifiers and the results for coreference would be either 'yes' or 'no'. The rule-based traditional approaches would generate some possible candidates as the antecedent and exploited the constraints or preferences to choose the best candidate with the highest possibility. With this in mind, Yang et al. (2008) created an order of importance for the constraints and then continued until the result converged into the best antecedent. The model was efficient and produced significant results.

Cluster Ranking Model

Even though mention ranking models performed well, the models were still not making use of the previous information for the resolution, and this is how cluster ranking models were introduced to solve the problem. These models try to bring the best of two worlds and use both mention ranking models and cluster ranking models. This model is similar to Lappin Leass' pronoun resolver. The model trained by Rahman and Ng (2011) with 39 features used a Support Vector Machine classifier. They compared the results of the cluster ranking model with mention-pair, entity mention, and mention ranking models and they achieved a better success rate except for one dataset.

Deep Learning Models

Most of these models that handle coreference or anaphora resolution need handcrafted features to be defined and this is rather time consuming. Also, the semantic dependency and the information about the context was not totally reachable in the previous models. This is where the deep learning models come into play. Deep learning models allow the words to be represented with vectors together with the contextual information. Many different types of deep learning models have been used in the literature for different languages. Some of these models are Feed Forward Neural Network (FFNN), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) models and Transformers. These models seem very promising for the fully automatized anaphora resolution systems. However, these types of deep learning models require a huge amount of data to be trained. Yet, some studies used the BERT model, which is a kind of transformer model, in languages such as Japanese and Korean for anaphora resolution. In a study that tries to resolve zero pronouns in Korean, researchers compared the BERT model with other machine learning models and the BERT model performed better (Kim et al., 2021).

In another study which uses English and Japanese translations as the data, the BERT model performs better when Japanese text is fed with English translations of the text (Umakoshi et al., 2021).

In summary, in this chapter, anaphora resolution as a task was described in detail. Later, the terms and concepts related to anaphora resolution were defined. Next, linguistic theories such as Binding Theory and Centering Theory which are closely related to anaphora resolution were described. Lastly, the computational models for anaphora resolution that were influential in the literature were reviewed.

In the next chapter (Chapter 3), the methodology of the study will be presented.

CHAPTER 3

METHODOLOGY

In this chapter, the methodology of the study will be provided. The description of the dataset, the process of annotation, the annotation tool, the annotation manual, the procedure of annotation, the reliability measurement, and the computational model with the preprocessing and the features used in our model will be explained in detail.

3.1 Outline of the Methodology

The process of this study is two-fold. Firstly, the data has been annotated by the author of this thesis (referred to as the annotator)who is a native Turkish speaker and proficient in English. Besides, she is trained in linguistics and discourse studies. The annotation procedure was carried out by using an annotation manual developed in the course of this thesis. Annotations were created by a freely-available annotation tool. Later, the annotated data was exported from the tool to create the dataset for the computational studies, which is the second phase of the study. This study aims to analyze how pronominal anaphora takes place in English sentences aligned with their Turkish counterparts and present a model that is able to resolve pronominal anaphora in English and Turkish separately for anaphoric relations and cataphoric relations for indefinite pronouns and relative pronouns. Annotations of zero pronouns were excluded and will be revisited in further research. By taking Mitkov's knowledge poor algorithm into consideration and the features described in Kucuk and Yondem (2007), a rule based computational model was designed to automatize the pronominal anaphora resolution and the predictions of the model were compared with the annotation result.

3.2 TED MDB Corpus

TED talks are independently organized events in many countries with the motto of "ideas worth spreading". The speakers are invited to deliver a speech on many different topics. Since it has been gathering a lot of attention from the public, these talks were translated into many different languages by volunteers for free. What makes TED Talks as a source of multilingual data is that these translations are checked and controlled by language coordinators from TED before they are published (Zeroual & Lakhouaja, 2020). There are many studies which use TED talks to create a multilingual corpus such as Cettolo (2016), Cattoni et al. (2021) and Kunchukuttan et al. (2017). the TED Multilingual Discourse Bank (Zeyrek et al., 2020) was born thanks

to the efforts of many researchers. The TED-MDB Corpus is a collection of the transcriptions of 6 TED talks in 6 different European languages which are English, German, Russian, European Portuguese, Polish and Turkish which is not a European language. It follows the PDTB (Penn Discourse Tree Bank) approach in Zeyrek et al. (2018). The transcriptions are acquired from the WIT3 corpus (Cettolo et al., 2012). Table 3.1 below shows the ID numbers of the talks together with the author and the title of the talk.

Table 3.1: *TED talks annotated in TED-MDB*

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

In this research, the data from the TED-MDB (Turkish and English languages) was taken with sentence alignments. The reason of choice for the languages was proficiency of the researcher in both languages in addition to the fact that these two languages are quite different from each other in terms of the sentence structures. The translations will allow us to capture how pronominal anaphora takes place in the same context for different languages. The total number of English sentences aligned with their Turkish counterparts is 364. The sentences with their alignments were taken from Ozer and Zeyrek (2019). Since the researcher is interested in the pronominal anaphora in the texts, the omitted sections of language were added for all the sentences with the variable ∅ that represents all types of zero anaphora. The pronominal anaphora and all types of pronouns that are markable were annotated in all the texts in both languages. The sentences in each document were given an ID before they were annotated. The IDs consist of the number of the document, the language and sentence number. The documents were stored as txt documents that is readable by the tool. Sample sentences with their IDs is shown below.

1. 1927_EN_1 The world is changing in some really profound ways, and I worry that investors aren't paying enough attention to some of the biggest drivers of change, especially when it comes to sustainability.

2. 1927_TR_1 Dünya gerçekten birçok yönden değişiyor; ∅ endişem o ki yatırımcılar değişimin en büyük faktörlerinden bazılarına yeterince dikkat etmiyorlar, özellikle de iş sürdürülebilirliğe gelince.

3.3 Annotation Process

All sub-types of pronominal anaphora were annotated while all other types of anaphora were excluded from the annotation process, and they were not tagged. The pronouns that were annotated are personal pronouns, reflexive pronouns, reciprocal pronouns, demonstratives, zero pronouns, indefinite pronouns, and relative pronouns. Our annotation manual will be provided in detail in section 3.3.1.

3.3.1 Annotation Manual

The manual for the annotation has been adapted from Lapshinova-Koltunski and Hardmeier (2018) and Guillou et al. (2014) version that were used for coreference annotation on TED talks and EU texts. However, since we were interested in pronominal anaphora, some changes were made by taking the definitions of pronominal anaphora and non-anaphoric uses of pronouns into account according to their detailed explanations in Mitkov (2014). The manual will be given in four parts as markables, unmarkables, the relations, and antecedents.

Markables

In this part, the definition of the segments of language that should be included and tagged in the annotation are defined.

•**Indefinite pronouns:** Indefinite pronouns such as anybody, anyone, nothing, nowhere etc. have been marked if they refer to a part in the text. Examine example (1) which is a markable indefinite pronoun.

(1) Now a fair question might be, what if ***all this sustainability risk stuff*** is exaggerated, overstated, it's not urgent, ***something*** for virtuous consumers or lifestyle choice?

•**Personal pronouns:** If the pronoun refers to another word in the text, it should be marked. All other forms of referring personal pronouns such as reflexives, reciprocal etc. should be included. Zero pronouns are also marked if they are to be matched with an antecedent. To rule out the zero noun and verb phrases, if the dropped word in the sentence can be replaced with both a pronoun and a noun, it is to be annotated. Example (2) below is an example of markable personal pronoun.

(2) I remember asking her what she thought of ***those early works***.If you didn't know ***they*** were hers, you might not have been able to guess.

•**Demonstratives:** Demonstrative pronouns such as 'this', 'that', 'these' and 'those' are marked if they refer to an existing part in the text. Example (3) illustrates a markable demonstrative pronoun.

(3) The word 'hobo' conjures up ***an old black and white image of a weathered***

old man covered in coal, legs dangling out of a boxcar, but ***these*** photographs are in color, and they portray a community swirling across the country, fiercely alive and creatively free, seeing sides of America that no one else gets to see.

•Temporal vs locative adverbs: Temporal and locative adverbs 'here', and 'then' are annotated under the category of demonstratives if they have an antecedent. No example of markable temporal and locative adverb was observed in our dataset.

•Relative clauses: Relative pronouns and reduced relative pronouns are to be marked. Example (4) shows a markable relative pronoun sample from our dataset.

(4) It didn't sound like a complaint, exactly, but just a way to let me know, a kind of tender admission, to remind me that he knew he was giving himself over to a voracious, ***unfinished path that*** always required more.

Unmarkables

In this part, the segments of language that should be excluded from the annotation are defined.

•Other anaphora types: All other types of anaphora such as lexical anaphora, noun anaphora and verb anaphora should be excluded because they are not in the scope of our research. Example (5) below is a zero noun anaphora and it is not markable as pronominal anaphora.

(5) One ∅ held a half-eaten ice cream cone in one hand and arrows in the left with yellow fletching.

•Non-anaphoric uses of pronouns: Pleonastic uses of 'it', deictic uses of personal pronouns, generic uses, and noun clauses are excluded because they are not considered as types of anaphora. If the dropped pronoun can only be replaced with a noun, it should not be annotated because it is also out of our scope. Since TED talks are spoken language texts and they include extensive use of deictic 'I', 'you', 'we', 'my' etc., they are considered as deictic, and they were not annotated as the approach taken by Lapshinova-Koltunski and Hardmeier (2018). Temporal and locative adverbs that are considered to be deictic are not annotated. Example (6) indicates deictic uses of pronouns.

(6) ***I*** mean, let ***me*** clarify something right ***here***.

•Substitution & ellipsis: Substitution occurs when a previously mentioned part of a sentence is replaced with a different word such as 'do'. On the other hand, ellipsis arises when a previously mentioned part of the sentence is omitted. Both substitutions and ellipsis should be excluded from the annotation process. Example (7) includes substitution and ellipsis and it is considered unmarkable.

(7) It's gotten smaller, it's got less detail,∅ less resolve.

The antecedent

The part of language that the anaphor refers to should be marked with its part of speech and connected with a coreference chain that shows the relation between the antecedent and the anaphor. For the antecedent the grammatical category of the word was used. That is, noun phrases (tagged as 'NP'), prepositional phrases ('PP'), verb phrases ('VP'), and sentences (tagged as 'sentence') were used as the tags of annotated text.

Relations: The relations that are annotated can be categorized into 3 groups.

Anaphoric: If the antecedent is observed before the use of the pronoun, the relation between the anaphor and the antecedent is annotated as anaphoric.

Cataphoric: If the anaphor appears before the antecedent the relationship should be annotated as cataphoric.

Ambiguous: If the anaphor has more than one possible antecedent and this leads to a different semantic interpretation of the anaphor, the relationship between the pronoun and the possible antecedents are marked as ambiguous.

3.3.2 Annotation Tool

There are various different open-source annotation tools used for coreference and anaphora and two of the most commonly used tools are MMAX2 and WebAnno. The developers of WebAnno created a new web-based annotation tool INcePTION (Klie et al., 2018). The easy use of the interface and the customization property were the reasons that this tool was chosen as the annotation tool. The data was imported into the tool in txt format, and it was exported in JSON format to be preprocessed. Figure 3.1 is a sample from the tool interface.

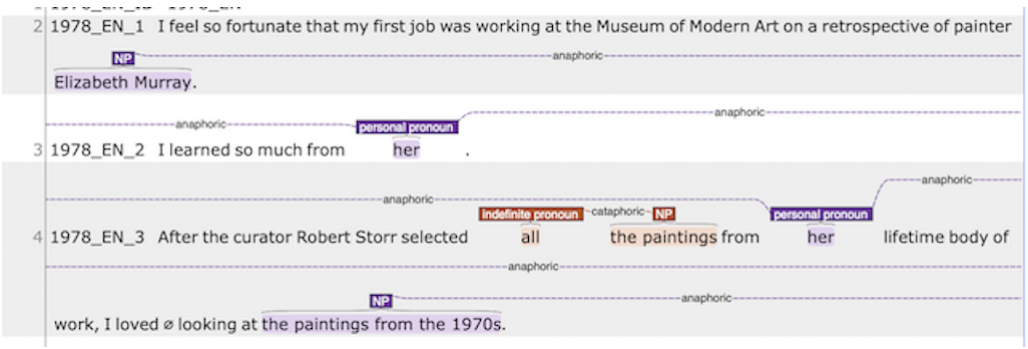

Figure 3.1: INCEptiON user interface

3.3.2.1 The Annotation Procedure

The whole annotation process was carried out by one annotator. The data consisted of 6 documents for English and 6 documents for Turkish. The annotator annotated each document for each language separately for pronominal anaphora according to the annotation manual mentioned in section 3.3.1. After finishing the annotation of all the documents, the annotator took a break of 1 month. Next, two of the English documents and their Turkish counterparts were chosen for reliability measurement. The annotator re-annotated these two documents in both languages. 185 sentences for English and their aligned counterparts in Turkish which were also 185 sentences were annotated again. This was equal to slightly more than the 50% of the total number of sentences. After finishing the annotation for reliability measurements, the annotator exported the final version of the annotation from the tool. To be able to provide agreement table for reliability measurements, all the annotations in the documents were aligned with their anaphor, antecedent and relation information and categorized as 'annotated'. All the excluded pronouns and deictics were calculated and categorized as 'excluded'. In the end, a two class table for agreement was created. Table 3.2 shows the representative values for agreement table.

Table 3.2: Representative agreement table between annotators

Annotator 1 / Annotator 2	Annotated	Excluded
Annotated	a	b
Excluded	c	d

3.3.3 Reliability Measurement

Reliability is the measurement which defines the consistency of the rater on a given task. Since one rater who is referred as annotator in our study completed the whole annotation process, we used an intra-rater reliability measurement. To calculate the intra-rater reliability, we decided to use a widely accepted reliability measurement in the literature which is Cohen's Kappa coefficient (Cohen, 1960). Cohen's Kappa is a statistical evaluation method that is used for categorical or qualitative items for inter-rater or intra-rater reliability. The Cohen's Kappa equation is given below, where P_o represents the observed agreement and P_e denotes probability of chance agreement. The Cohen's Kappa coefficient formula is as follows :

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

P_o is calculated as:

$$\frac{a + d}{a + b + c + d}$$

P_e is calculated as:

$$(\frac{a+b}{a+b+c+d} \times \frac{a+c}{a+b+c+d}) + (\frac{b+d}{a+b+c+d} \times \frac{c+d}{a+b+c+d})$$

Table 3.3 shows the agreement values for the annotation.

Table 3.3: Agreement and Disagreement Table between Annotators

Annotator 1 / Annotator 2	Annotated	Excluded
Annotated	639	30
Excluded	47	992

Based on the formula given above P_o is calculated as:

$$\frac{1631}{1708} = 0.95$$

P_e is calculated as:

$$(\frac{669}{1708} \times \frac{686}{1708}) + (\frac{1022}{1708} \times \frac{1039}{1708}) = 0.51$$

$$\kappa = \frac{0,95 - 0,51}{1 - 0.51} = 0.89$$

Table 3.4 shows the interpretation of the Kappa coefficient.

Table 3.4: Cohen's Kappa Interpretation Table

Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

For discourse studies the reliability standart is given by Spooren and Degand (2010) as 0.70 which is interpreted as substantial agreeement. The reliability measurements for our study was calculated above this standart and according to the table 3.4, it is interpreted as near perfect agreeement.

3.4 Computational model

Previously in Chapter 2, the linguistic background of the study and the computational approaches to anaphora resolution were provided. Based on these proposed approaches in the literature, we decided to test the traditional knowledge based approach which is knowledge-poor algorithm by Mitkov (1998). We used the translated and aligned English-Turkish corpus to see how it performs in the same domain (TED talks) in the original and translated texts with a rule based model. Similar to the original approach, the raw data was preprocessed with a tool for grammatical analysis which is UDPipe. Later on, the NPs were extracted and the rules applied one by one for the pronouns to match them with the highest ranking antecedent NP as a referee. The details of the features and the part of speech tagger will be given in detail in the following sections.

3.4.1 Data Preparation

All the process for the computational model were implemented in Python programming language (Van Rossum & Drake, 2009) by using Jupyter notebook. The output of the annotation tool was exported in JSON format and it was used as the input for the model. However, the data had to be prepared for the model to apply the rules. The dictionary format that consists of sub-dictionaries with keys and values was used to be able to capture the hierarchical structure of the data. Figure 3.2 shows the flowchart of how the new data structure for modelling is created.

First of all, we parsed the JSON data into two parts as the annotated relations and raw sentences. The raw sentences were stored with their previously given ID and the sentences themselves. Figure 3.3 is a sample of the data structure for the sentences.

The relations consisted of the antecedent, POS tag of the antecedent, form of the anaphor, the POS tag of the anaphor and the relation that holds between them with their index number. Figure 3.4 is an example of relation data point.

3.4.1.1 Grammatical Analysis

The next step was the linguistic analysis of the raw sentences to extract some properties so that we could create the rules for our model. We used UdPipe 2.0 as the parser and lemmatizer (Straka, 2018). UDPipe is a multilingual parser, lemmatizer, dependency parser and tagger which provides different parser models that were previously trained with different treebanks. We used the ‘english-partut-ud-2.10-220711’ model as the English parser and the ‘turkish-kenet-ud-2.10-220711’ as the Turkish parser.

Figure 3.2: Data Preparation

```
'1927_EN_29': 'So the views of CEOs are clear.',
'1927_EN_30': "There's tremendous opportunity in sustainability.",
```

Figure 3.3: sample sentence from the data structure

```
'3': {'type': 'anaphoric',
 'to': {'form': 'the emissions', 'pos': 'NP', 'sentence_id': '1927_EN_7'},
 'from': {'form': 'that',
  'pos': 'relative pronoun',
  'sentence_id': '1927_EN_7'}},
```

Figure 3.4: Sample of annotated relation

These two models were chosen because their POS tagging were more accurate along with rich feature information for the tokens. The UDPipe models for Turkish and English were imported from the server and the sentences (of our corpus) were parsed one by one. With the aim of providing the parser with less noise, the colon and the semicolon which are recognized as sentence splitters by UDPipe were replaced with comma values. Also, the variable that was used to represent zero anaphora was eliminated from the raw texts before it was processed by the parser because UDPipe recognized it as a punctuation mark and it caused noise.

UDPipe gives us forms, the universal part of speech tags, lemmas, features, depen-

dency relations, dependency head and the range of the tokens. Figure 3.5 is an example of a data point after parsing through UDPipe.

```
(data["sentences"]["1927_EN_111"]["ga"][0])
✓ 0.4s
{'position': 0,
 'form': 'Thank',
 'pos': 'VERB',
 'lemma': 'thank',
 'dep_head': -1,
 'dep': 'root',
 'features': {'Mood': 'Ind',
  'Number': 'Sing',
  'Person': '1',
  'Tense': 'Pres',
  'VerbForm': 'Fin'},
 'score': 0,
 'antecedent': None}
```

Figure 3.5: UDPipe parser output sample

UDPipe provides the parse trees for the sentences. The parse trees show the dependency relations between the tokens (words) and their heads in the sentences. Figure 3.6 is a sample of a parse tree from UDPipe.

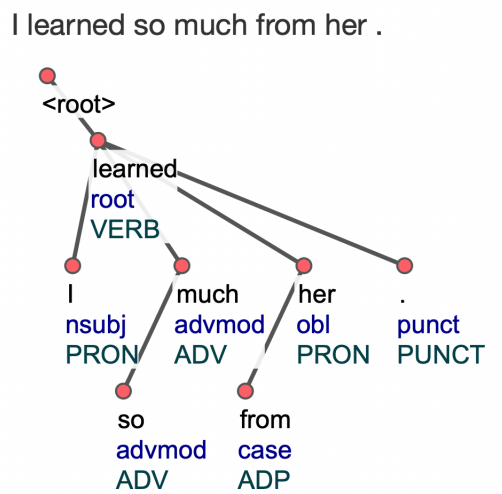

Figure 3.6: Parse tree output from UDPipe

After the implementation of the parser, we kept the results of the analysis as a sub-dictionary in the sentences as their ‘grammatical analysis’.

3.4.2 Filtering of Noise

After parsing the data there were two more steps before the implementation of the heuristics. The first task was the elimination of the non-markable pronouns, which are the deictics and the pleonastic uses of 'it'. In the process of annotation, we observed that the uses of first person singular pronouns, first person plural pronouns, second person singular pronouns and second person plural pronouns were used deictically to a great extent because these pronouns refer to the speaker and the adressee in the context of TED talks. When these pronouns were used anaphorically, they were commonly typed with quotation marks. Similarly, Kucuk and Yondem (2007), used quotation marks for the preference of antecedents in Turkish pronouns. We decided to use this heuristic for the elimination of the deictic uses of personal pronouns. This heuristic is used for both languages.

RULE 1: Deictic Pronouns

The rule searches for a token that has the POS tag of the 'pronoun' and checks the feature output of the parser to identify the type of the pronoun. If the algorithm encounters a demonstrative pronoun and it is the first or the second person singular or plural and it is used with a quotation mark, the pronoun is marked as 'eligible' for search.

The examples (8) and (9) are samples from the dataset for the use of quotations.

(8) Now, ***I*** do speak to ***a lot of investors*** as part of ***my*** job, and not all of ***them*** see it this way.

(9) Often ***I*** hear, "***We*** are required to maximize returns, so ***we*** don't do that here," or, "***We*** don't want to use the portfolio to make policy statements."

In order to identify the unmarkable pleonastic 'it' by our algorithm, we defined 3 different rules for the filtering of the pleonastic uses of 'it' by taking three of the definitions from Mitkov (2014). The uses of 'it' with adjectives, in passive constructions and cleft sentences explained in 1, 2, and 6 were taken into consideration and they were eliminated from the search space. The parser keeps the values of auxiliaries, the types of pronouns and passive constructions of the tokens with the key of 'feature' and the part of speech tags of nouns, adjectives and pronouns with the key of 'POS' in a dictionary format.

RULE 2: Pleonastic 'it'

The rule searches for the token in the form of 'it' and when it is found, it checks:

a. If the pronoun is followed with a token with the feature of 'auxiliary' preceding a token with the POS tag of 'adjective', mark the pronoun as 'not eligible'.
b. If the pronoun is followed with a token with the feature of 'passive', mark the pronoun as 'not eligible'.
c. If the pronoun is followed with a token with the feature of 'auxiliary' preceding a token with the POS tag of 'noun', mark the pronoun as 'not eligible'.

See the following example from the dataset which is marked as 'not eligible'.

(10) I think ***it***'s reckless to ignore these things, because doing so can jeopardize future long-term returns.

In order to capture the deictic uses of demonstratives, we took some sample sentences from the data and searched if there were any patterns of uses as deictic. It was observed that when demonstratives were followed with a verb and they were in the subject position, they were more likely to be deictic.

RULE 3: Deictic uses of demonstratives

The demonstrative pronoun 'this' in example (11) is used deictically. When the demonstrative pronoun is followed with a verb and it is in the subject position, it is marked as 'not eligible'. The type of the pronoun and the dependency relations were detected by UDPipe parser. Algorithm was defined as follows:

When a token with the POS tag of 'pronoun' is identified follow these steps:

- Check the feature of the token.
- If the value is 'demonstrative', check the 'dependency relation' of the token.
- If the value is 'subject', check the following token.
- If the following token has the POS tag of 'verb', mark the pronoun as 'not eligible'.

See the example (11) from the dataset which is marked as 'not eligible'.

(11) ***This*** is why we are concerned with ESG.

3.4.3 Extracting NPs

The extraction of the NPs were necessary because we wanted to search for only the antecedent NPs. The parser provided us with the part of speech tags of the tokens. Therefore, an NP extractor algorithm was necessary for our model. The extraction of the noun phrases was the last step before the antecedent-anaphor search. We used the dependency relations from the UDPipe output to extract the NPs with an iterative search algorithm that traverses the parse tree and finds all the child nodes of a noun and creates a string that consists of all the child nodes of the noun. NP extractor algorithm was designed by the researcher.

RULE 4: NP Extractor

For a 'POS' tag of 'noun' or 'pronoun';

- Search the child nodes from the dependency tree by checking their 'dependency head'

- Add each child node by using their index numbers to create a string if their parent is in the list and the child is not added into the string.

- If there is a preposition with the dependency relation as 'adp' in the beginning, delete it.

Figure 3.7, shows an example of an NP from the dataset. The algorithm encounters the POS tag of 'noun' for the token 'ways' and starts to search for its child nodes. The algorithm adds the tokens 'in', 'some', 'really', 'profound', and 'ways' into a string to create the NP. Since there is a preposition in the beginning with the 'dependency relation' of 'adp', it is removed from the string to reach the NP.

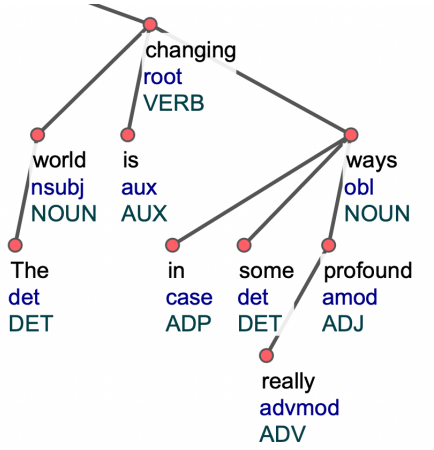

Figure 3.7: NP example

3.4.4 Heuristics for Search: Filtering and Ranking Candidates

Feature engineering, which is often known as feature extraction, refers to a collection of techniques for executing desired tasks for computational models. The purpose of feature extraction is to turn textual characteristics into values that the algorithm can interpret (Vajjala et al., 2020). By taking the proposed constraints and preferences in Kucuk and Yondem (2007) and linguistic theories which were also mentioned in Chapter 2, we designed two different sets of features. The first type of feature used was the eliminating sets of constraints mentioned in 3.4.4.1 which had to be satisfied. If the antecedent candidates could not satisfy these constraints, they were removed from the candidate list. The second set of features were defined as preferences described in 3.4.4.2 that did not eliminate the candidates, but they assigned a score based on the properties of the tokens. The search space for the antecedent search was defined as the sentence where the pronouns were found, and the prior two sentences. This search space was mentioned by Hobbs (1978) and implemented also in Mitkov (1998) and Kucuk and Yondem (2007).

3.4.4.1 Constraints for Candidate Filtering

Constraints are the eliminating features used in our model. These features had to be satisfied for the candidates, otherwise they were removed from the antecedent candidate list. The antecedent candidate list consisted of the extracted NPs (by the NP extractor) and has the form as strings. They were kept in a dictionary list.

Gender and Number

For a pronoun that was eligible for search after filtering, the gender and number features of the pronoun and the antecedent candidates heads were compared. Collective nouns defined in the collective noun list were excluded from number agreement[1]. If the antecedent candidate NPs and the pronoun had different values in terms of their number and gender, they were removed from the antecedent candidate list.

RULE 5: Gender and Number Agreement

When the the algorithm encounters a token with the POS tag of 'pronoun', it checks if it is marked as eligible. If it is marked as 'eligible', the search for its antecedent candidate starts and the steps below are followed:

- Check if the NP string head is in the list of collective nouns.
- If the head of the NP is in the collective noun list, stop the search.
- If it is not in the collective noun list, compare the 'gender' and 'number' values for the 'feature' key of the token and the pronoun.
- Remove all the NPs that do not match in terms of 'gender' and 'number' from the list of the candidates.

See example (12) where the pronoun and the candidate matches in terms of gender and number.

(12) Elizabeth Murray surprised me with ***her*** admission about ***her*** earlier paintings.

Personal Pronoun

According to principle B of Binding Theory, personal pronouns should be free in their local domain. Therefore, if there was a pronoun in the object position of the sentence, the subject of the sentence was removed from the candidate list.

RULE 6: Personal Pronoun Constraint

- For a token that has the POS tag of 'pronoun' marked as 'eligible' and has the feature of 'personal pronoun', check the dependency relation of the pronoun.

[1] See Appendix A for the list of these collective nouns

- If dependency relation is ‘object’, then remove the token which has the dependency relation ‘subject’ from the candidate list.

For the antecedent search for the pronoun ‘her’ which is a personal pronoun and in the object position, pronoun ‘I’ in the subject position is removed from the candidate list in the example (13).

(13) It’s what I have to imagine Elizabeth Murray was thinking when ***I*** saw ***her*** smiling at those early paintings one day in the galleries.

Reflexive Pronoun

Principle A of Binding Theory suggests that the reflexive pronouns should be bound in their domain. Thus, the reflexive pronoun that is eligible for search should refer to the closest gender and number matching NP string on its left.

RULE 7: Reflexive Pronoun Constraint

For a token that has the POS tag of ‘pronoun’ and marked as ‘eligible’;

- Check the feature of the token and if it is ‘reflexive’, find the closest NP string on its left after applying gender and number constraint.
- Assign the NP string as the antecedent of the pronoun.

For an antecedent search for the reflexive pronoun ‘himself’ in the example (14), the NP string ‘he’ is assigned as its antecedent candidate.

(14) ... and ***he himself*** was that Adam with his finger outstretched and not quite touching that God’s hand.

Syntactic Constraints

The antecedent identification of relative pronouns and indefinite pronouns were implemented under the syntactic constraints category. For English, the antecedent was chosen as the closest NP on the left of the relative pronoun while for Turkish anaphors, the antecedent was chosen as the rightmost NP string. The antecedent of the indefinite pronouns were chosen as the rightmost NP string for both languages.

RULE 8: Indefinite and Relative Pronoun Constraint

For a token that has the POS tag of ‘pronoun’ and marked as ‘eligible’;

- Check if the feature of the token and if it is ‘indefinite’ or ‘relative’ then search for the closest NP string on its left for English.
- Assign the NP string as its antecedent for the pronoun.
- Check if the feature of the token and if it is ‘indefinite’ or the dependency relation is ‘adjective clause’ then search for the closest NP on its right for Turkish.

- Assign the NP string as the antecedent for the pronoun.

The antecedent of relative pronoun 'which' in example (15) is chosen as 'a multi-material prosthetic socket' which is the closest NP on its left.

(15) We use a 3D printer to create ***a multi-material prosthetic socket which*** relieves pressure where needed on the anatomy of the patient.

3.4.4.2 Scoring for Ranking Candidates

After the elimination of the antecedent candidates according to the constraints, all the antecedent candidates which were left in the list were given scores based on these preferences. The features and the scores for the base recency score, the subject position score, and the repetition score were adapted from Kucuk and Yondem (2007), as explained below.

Recency Score

If the antecedent candidate in the search space was closer to the anaphor, it was given a higher score. The base score for the recency was given as +2.15. The further the position the NP string had, the less the score was given. This gave the closer NP strings more salience.

RULE 9: Recency Preference

For a token that has the POS tag of 'pronoun' and marked as 'eligible';

- Check the index number of each antecedent candidate.
- Calculate the recency score which is between +2.15 and +1.90 for each NP string on the left.
- Assign the calculated recency score to the NP string.

The NP string 'these photographs' in (16) is given the recency score of +2.15 because it is the closest NP string on the left of the pronoun where the token 'color' is eliminated with the number constraint.

(16) ... ***these photographs*** are in color, and ***they*** portray a community swirling across the country...

Subject Position Score

In Centering Theory, it is described that the subjects are more salient for pronoun reference. Therefore, the NP antecedent candidates that were in the subject of position in the search space were given a score of +1.85. This scoring was named as 'the first NPs in the sentence' in the original approach of Kucuk and Yondem (2007).

RULE 10: Subject Position Preference

For a token that has the POS tag of 'pronoun' and marked as 'eligible';

- Check the 'dependency relation' of each antecedent candidate in the search space.
- If the 'dependency relation' is 'subject' then give +1.85 for each NP string head.

The NP string 'their prosthetic sockets' in example (17) gets a score of +1.85 because the dependency relation is 'subject'.

(17) The reason, I would come to find out, was that ***their prosthetic sockets*** were painful because ***they*** did not fit well.

Object Position Score

The object position score was not given in Kucuk and Yondem (2007). Therefore, we decided that a score of +1.50 that is less than the subject position score and more than the first NP score should be given for the candidates in the object position because they were described as less salient than the subjects in Centering Theory.

RULE 11: Object Position Preference

For a token that has the POS tag of 'pronoun' and marked as 'eligible';

- Check the 'dependency relation' of each antecedent candidate in the search space.
- If the 'dependency relation' is 'object' then give +1.50 for each NP string head.

Example (18) is an example of an NP string that gets a score of +1.50 for the antecedent search for the pronoun 'it'.

(18) So one day, when I met professor Hugh Herr about two and a half years ago, and he asked me if I knew how to solve ***this problem***, I said, " No, not yet, but I would love to figure ***it*** out."

Repetition Score

The NP heads that repeat more than once in the search space were given a score of +1.20. To include different forms of nouns, we used the lemmas for assigning a repetition score. Therefore, singular forms, plural forms and other types of words that has the same lemma are considered as repeating words.

RULE 12: Repetition Preference

For a token that has the POS tag of 'pronoun' and marked as 'eligible';

- Check the 'lemma' of each NP string head.
- If the 'lemma' repeats more than once give a score of +1.20 to each NP string head.

The algorithm checks the lemma of the NP string heads 'socket' and 'sockets' in example (19). Both of the NP string heads have the same lemma and they are given a score of +1.20 for each.

(19) The reason, I would come to find out, was that their ***prosthetic sockets*** were painful because they did not fit well. ***The prosthetic socket*** is the part in which the amputee inserts their residual limb, and which connects to the prosthetic ankle.

3.4.5 Preparations for Evaluation Metrics

We decided to implement classification metrics to calculate the accuracy of our model over English and Turkish texts of our corpus. The annotated relations were extracted from the files and kept separately with their anaphor and antecedent information. The antecedent-anaphor pairs which were manually annotated were compared with the anaphor-antecedent pairs matched by the algorithm. We used the 'top *N*' evaluation metric for our model. The first evaluation was calculated with 'top *N* = 1'. We decided to use this method because complex NPs can be very difficult to extract and different NPs possessing the same NP head can be accepted as the antecedents of the anaphors. The pseudo-code in 1 shows an illustration of the code for top *N*. These two evaluation metrics are reported separately in Chapter 4.

Algorithm 1 Compare *annotation result = model prediction*

Require: *anaphor annnotated = anaphor identified by model*
Ensure: *top N = 1*
  if $antecedent(ant) =$ *highest ranking candidate(Top 1)* **then**
    $True \leftarrow Prediction$
  else
    $False \leftarrow Prediction$
  end if
Ensure: *Top N = 3*
  if $antecedent(ant) \in$ *3 highest ranking candidates(Top 3)* **then**
    $True \leftarrow Prediction$
  else
    $False \leftarrow Prediction$
  end if

In this method, the antecedent-anaphor pairs from the annotation were compared with the highest ranking antecedent candidates in the candidate list in our algorithm. If the highest ranking candidate is the same with the antecedent which was manually annotated, it was considered as true. In the second approach we used the 'top *N* = 3' approach. In this method, if the antecedent is among the first three highest

scoring candidates in the antecedent candidates list, they were considered as correct predictions.

In Chapter 3, the research method was explained in detail. Procedure of annotation, annotation manual and tool, the reliability measurements were presented. Later, the process of computational modelling for pronominal anaphora with preprocessing, grammatical analysis and filtering of noise, extracting NPs and heuristics for search were defined thoroughly. The next chapter, reports the findings of the research and the performance of the pronominal anaphora resolution models.

CHAPTER 4

RESULTS

In this chapter, the evaluation metrics we used to evaluate our classification, namely, precision, recall and F1 score will be introduced.

Figure 4.1 below represents the classifications of predictions and truth values for a classification problem presented as a confusion matrix.

		Actual	
		Positive	Negative
Predicted	Positive	**True Positive**	**False Positive**
	Negative	**False Negative**	**True Negative**

Figure 4.1: A sample of confusion matrix

Precision

This metric is called precision and it is calculated by dividing the number of correct predictions (True Positives, or tp) with the sum of correct predictions and the number of incorrect positive predictions (False Positives, or fp) for a class. Formally it is given:

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

In our study, precision was calculated as follows:

$$\textbf{Precision} = \frac{tp\text{= } correctly\ resolved\ pronouns}{tp\text{= } correctly\ resolved\ pronouns + fp \text{ = } incorrectly\ resolved\ pronouns}$$

Recall

Recall is the division of the positive predicted instances to the sum of positive predictions and negative predictions that should have been classified as positive. It is defined as:

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

In our study recall was calculated as follows:

$$\textbf{Recall} = \frac{\textit{(tp = correctly resolved pronouns)}}{\textit{(tp = correctly resolved pronouns) + (fn = pronouns not predicted)}}$$

F1-score

F1-score, which is the harmonic mean of precision and recall, is a popular metric that combines precision and recall metrics. It is calculated as:

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

4.1 Performance of the Model

The heuristics mentioned in 3.4.4.1 were used for the pronominal anaphora resolution. Later, the antecedent-anaphor pairs matched by the algorithm were compared with the annotation results. The performance of the pronominal anaphora resolution models for English and Turkish were evaluated in two different ways. The first evaluation was carried out by comparing the actual antecedents with the antecedents chosen by the model where the antecedent chosen by the model is the highest scoring candidate in the candidate list (top N = 1) for each anaphor. Table 4.1 indicates the recall, precision and F1 score results for highest ranking antecedent candidates in the models.

Table 4.1: Top N = 1 results

Language	**Recall**	**Precision**	**F1 score**
Turkish	0.44	0.68	0.54
English	0.41	0.47	0.44

As it can be observed, the Turkish resolver performed slightly better than English reaching an F1 score of 0.54. The number of pronouns recognized and attempted for resolution for English and Turkish were 348 and 340 respectively. However, the number of pronouns which were not identified by the algorithm was higher in Turkish reaching 122 pronouns. This number was reported as 54 for English.

For the second evaluation, if the antecedent of the anaphor was among the top three candidates (top N = 3) in the antecedent candidates list, the prediction was considered as true. Table 4.2 shows the recall, precision and F1 scores for the models performance.

Table 4.2 suggests that the performance improved in the second evaluation metric. Similar to the first evaluation result, Turkish algorithm performed slightly better than the English one. Yet, it is important to note that the distibution of anaphoric and cataphoric relation resolved changed extensively between the languages. The anaphoric

Table 4.2: Top *N*= 3 Results

Language	**Recall**	**Precision**	**F1 score**
Turkish	0.52	0.81	0.63
English	0.57	0.66	0.61

relations were observed more frequently in English while cataphoric relations were more common in Turkish. Table 4.3 below shows the distribution of anaphoric and cataphoric relations from the annotation.

Table 4.3: Number of Relations in the Data

Language	**#Anaphoric Relations**	**# of Cataphoric Relations**
Turkish	111	229
English	348	79

Even if our model resolved cataphoric relations for indefinite pronouns and relative pronouns, it performed very well in Turkish by resolving 136 cataphoric relations out of 158 pronouns identified. For a detailed distribution of the resolved pronouns see Appendix C.

In general, the model performs better in 'top *N*= 3' evaluation method for both languages. When the performances of the languages were compared, the Turkish pronominal anaphora resolution model performs better than English pronominal anaphora resolver.

In this chapter, the results of the computational studies were given in two parts for Turkish and English pronominal anaphora resolution. For both languages, two different evaluation approaches were reported by their recall, precision and F1-score evaluation metrics. In the next chapter, the discussion of the results and error analysis will be presented.

CHAPTER 5

DISCUSSION

Chapter 5, provides a detailed discussion of the results and error analysis for the pronominal anaphora resolvers for Turkish and English. There are many studies that implement the Knowledge Poor algorithm (Mitkov, 1998) in many different languages and it was tested in many different languages. However, the datasets which were used were different from each other. Therefore, the results cannot be compared with each other. The rule-based approach presented by our study differs from its predecessor in the following ways:

- The data used in our research was taken from spoken language. Therefore, it includes extensive use of deictics.
- The algorithm includes the filtering of deictics, and pleonastic uses of 'it'.
- The model tries to resolve all types of pronominal anaphora for anaphoric relations and cataphoric relations for indefinite pronouns and relative pronouns.
- The algorithm tries to use similar constraints and preferences for both languages separately.

The features and scores for preferences used in our study were taken from a previous Knowledge Poor algorithm that resolves the third person singular pronouns in Turkish (Kim et al., 2021). The model performs better in Turkish version compared to English version.

Quote/unquote preference implemented in the original study used by Kucuk and Yondem (2007) was used for the elimination of the deictic uses of pronouns in our study. Two of the preferences in the original study that were not used in our study were the antecedent of zero pronoun and punctuation preferences. There were different reasons for this decision. The antecedents of zero anaphora preference was not used in our study because zero anaphora was excluded from our computational model and will be revisited in further studies. The punctuation preference for comma was not used in our study because the use of comma in our dataset has been inserted by TED talk transcribers and it does not follow the traditional rules. The first constraints used in our study were gender and number agreement for English and number agreement for Turkish because gender is not marked in Turkish language. Only number agreement is implemented in our model.

Another implemented constraint was syntactic constraints for the identification of relative pronoun and indefinite pronoun antecedents. These constraints were not given

in the original study of Kucuk and Yondem (2007) since it does not resolve cataphoric relations.

The other constraints used in our study were personal pronoun constraint and reflexive pronoun constraint. This constraints were based on the principles of Binding Theory and c-command domain explained in detail in Chapter 2.

As for scoring of the candidates, the first NP preference or subject preference was used in Mitkov (1998), Kucuk and Yondem (2007). This preference is also based on Centering Theory where the subjects are given more salience over the objects. To be able to address the preference of the objects over other NPs in the sentence, we added a preference that assigns a score to the objects of the sentences. The scores assigned to the objects in the sentence were less than the score assigned to the subjects. In this way, the subjects were more salient than the objects in the sentence. The recency and repetition preferences were implemented frequently by Lappin and Leass (1994), Mitkov (1998), Trouilleux (2002) and Kucuk and Yondem (2007). Kucuk and Yondem (2007) assigned the same recency score to the nouns that appeared in the same sentence. However, we decided to optimize the recency score so that the more distant the NP was, the less score is assigned to the NP. In this way, the closest NP to the pronoun was given more salience.

5.1 Discussion of the Performance Results

The reason of failure in English can be because of the number of the features implemented in our research. Kucuk and Yondem (2007) used these features in Turkish and they created promising results. When the same features were used, the English counterparts of the Turkish sentences performed slightly less accurately than the Turkish dataset. This shows that our dataset needs more heuristics for the resolution task to be more accurate such as semantic information and discourse knowledge. Another reason that the model might have failed can be because of the dataset that was used in our research. The dataset which was used in our study is based on spoken language. In written text the distance between the anaphor and the antecedent is shorter compared to spoken language according to Hobbs (1978). However, in our study, the search space was the sentence where the pronoun is found and the preceding two sentences. This might be a reason of poor performance in our dataset.

Another observation showed that, the datasets that were used in many other studies were samples from computer manuals, book sections or written forms of language and they do not include the deictic uses of the pronouns often. On the other hand, the deictic uses of pronouns were observed frequently in spoken language datasets. This caused a lot of noise in our data because they were not supposed to be included in the search space. Therefore, we tried to filter these deictic pronouns with different sets of heuristics. However, this filtering is very data specific and it is very likely to fail in different datasets. Another point is that long complex sentences might be fewer in datasets that were used in previous work, compared to long sentences in speeches. This creates a difficulty of accurately identified dependency relations between the constituents of the sentences such as subjects and objects. Complex sentences and sentences connected with commas created very big parse trees with less accurate

dependency relations.

Thirdly, the parser and lemmatizer used in our study was UDPipe. The output of this parser was not always accurate and the output provided for the tokens was missing valuable information for the implementation of constraints such as gender and number. For example, most of the possessive adjectives did not posses the number information for the pronoun. This made the gender and number constraints to become irrelevant for some cases. As a result, the elimination did not take place. Another drawback of the parser was that it did not provide morphological analysis of the words which is very important for Turkish because it is an agglutinative language.

5.2 Limitations of the Study

The features and the preferences used in our study were based on the linguistic theories and computational approaches in the literature. However, Kucuk and Yondem (2007) had a different scope of research. The scope of the original research was the resolution of third person singular pronouns. They tried to find the antecedents of these pronouns by providing information about the proper names. On the other hand, our model had a larger scope. We included all types of pronominal anaphora in our study. It is highly possible that the preferences and constraints used in our study were not enough to acquire promising results. Another limitation of our study was that we could not include zero anaphora into our research. To be able to include zero anaphora, a different set of data should be prepared and annotated once again. Therefore, we excluded zero anaphora from our computational model. Besides, the cataphoric relations were included in our research for the resolution of indefinite and relative pronouns with a syntactic constraint. The constraint performed well for identification of cataphoric relations. However, the inclusion of cataphora in an anaphora resolution systems with the same features is controversial.

An additional limitation of the research is that our data size is very small. If we could expand our data size, it would be possible to create a more advanced anaphora resolution model using Transformers, Recurrent Neural Network(RNN) or Long-Short Term Memory(LSTM) approaches.

The last limitation of the study was the optimization of the scores assigned to the candidates by the preferences. The scores were implemented directly, but they were not optimized. This means the scores for the preferences that created the most efficient performance were not searched. The fine tuning of the preferences scores could be crucial for the performance of the model.

In summary, in Chapter 5, the general discussion of model performance was provided. Later on, the reasons of failure for the performance were explained. The next chapter, will conclude the thesis.

CHAPTER 6

CONCLUSION

In this chapter, a general overview of the thesis is presented and future work suggested.

This thesis consists of two stages. In the first stage of the study 6 documents containing 364 sentences in English with their Turkish counterparts were annotated according to a annotation manual. Later on, the intra-rater reliability of the annotation was calculated with the Cohen's Kappa coefficient. The second stage of the study included a computational model that tries to resolve pronominal anaphora in English and Turkish separately in the same domain. We implemented a promising approach used in the literature for many different languages including Turkish. The model tried to eliminate deictics and pleonastic uses of 'it' through linguistic rules. The models both for English and Turkish used the same features in different ways depending on the typological differences between the languages. The results of the computational model evaluation metrics showed that the features and preferences in our study performs better in the Turkish dataset in general compared to English with a better recall, precision and F1-scores.

The first contribution of the study is that it provided reliable annotation results for TED-MDB Corpus (Ozer & Zeyrek, 2019) in terms of pronominal anaphora for both Turkish and English. The results of the annotation can be used for cross-linguistic studies.

The second contribution of the study is that the research presented rule-based models for pronominal anaphora resolution in English and Turkish separately. This research is the first step of creating a bilingual anaphora resolution algorithm for these two languages. The performances can be improved with the use of more preferences and features with optimized scores.

In the future, the scope of the model can be extended and implemented on a different dataset to be able to generalize the performance of the model. Some other future studies can be given as:

- A model that includes zero anaphora can be designed. However, the data should be preprocessed with an overt representantion of these zero anaphors to be able to parse the data without causing noise.

- The output of the parser can be enriched with a more detailed analysis of tokens and their features such as gender and number to improve performance because

third person possessive adjectives were missing the gender and number information in UDPipe.

- Morphological analysis can be implemented for the words in Turkish since it provides valuable information. For example, the token 'bazıları' was used instead of 'bazı insanlar'. Morphological analysis could help us solve such cases.
- Semantic information can be included as a feature as will probably improve the performance of the system. In this way, some semantic features such as animacy can be included.
- Ultimately, the sentence alignments with their translated counterparts can be used to feed the model where one language lacks. For example, English is not a pro-drop language and overt pronoun uses in English can be used to feed the Turkish model to extract the features of the zero pronouns in Turkish.

REFERENCES

References

Aone, C., & William, S. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. *33rd Annual Meeting of the Association for Computational Linguistics*, 122–129.

Asudeh, A., & Dalrymple, M. (2006). Binding theory. *Encyclopedia of Language Linguistics*. https://doi.org/10.1016/B0-08-044854-2/01955-6

Baldwin, B. (1997). Cogniac: High precision coreference with limited knowledge and linguistic resources. *Operational factors in practical, robust anaphora resolution for unrestricted texts*.

Brennan, S. E., Friedman, M. W., & Pollard, C. (1987). A centering approach to pronouns. *25th Annual Meeting of the Association for Computational Linguistics*, 155–162.

Bublitz, W. (2011). Cohesion and coherence. *Discursive pragmatics*, *8*, 37–49.

Cattoni, R., Di Gangi, M. A., Bentivogli, L., Negri, M., & Turchi, M. (2021). Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, *66*, 101155.

Celce-Murcia, M. (1987). A comprehensive grammar of the english language. randolph quirk, sidney greenbaum, geoffrey leech, & jan svartvik. new york: Longman, 1985. pp. x+ 1, 779. *Studies in Second Language Acquisition*, *9*(1), 109–111.

Cettolo, M. (2016). An arabic-hebrew parallel corpus of ted talks. *arXiv preprint arXiv:1610.00572*.

Cettolo, M., Girardi, C., & Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. *Conference of european association for machine translation*, 261–268.

Chomsky, N., et al. (1982). *Some concepts and consequences of the theory of government and binding*. MIT press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., & Webber, B. (2014). Parcor 1.0: Pronoun coreference annotation guidelines. *Edinburgh, Uppsala*.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english*. Longman.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, *44*(4), 311–338.

Joshi, A., Prasad, R., & Miltsakaki, E. (2005). Anaphora resolution: A centering approach. *Encyclopedia of language and linguistics*.

Joshi, A. K., & Kuhn, S. (1979). Centered logic: The role of entity centered sentence representation in natural language inferencing. *Proceedings of the 6th international joint conference on Artificial intelligence-Volume 1*, 435–439.

Kim, Y., Ra, D., & Lim, S. (2021). Zero-anaphora resolution in korean based on deep language representation model: Bert. *ETRI Journal*, *43*(2), 299–312.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., & Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation [Event Title: The 27th International Conference on Computational Linguistics (COLING 2018)]. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. http://tubiblio.ulb.tu-darmstadt.de/106270/

Kucuk, D., & Yondem, M. T. (2007). Automatic identification of pronominal anaphora in turkish texts. *2007 22nd international symposium on computer and information sciences*, 1–6.

Kunchukuttan, A., Mehta, P., & Bhattacharyya, P. (2017). The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Kurt, B. G. (2021). Binding theory and a closer look at the anaphoric expression kendisi in turkish. *Dil ve Edebiyat Araştırmaları*, (23), 119–145.

Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational linguistics*, *20*(4), 535–561.

Lapshinova-Koltunski, E., & Hardmeier, C. (2018). Coreference corpus annotation guidelines.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Mitkov, R. (2014). *Anaphora resolution*. Routledge.

Mitkov, R. (2022). *The oxford handbook of computational linguistics*. Oxford University Press.

Nemcık, V. (2006). *Anaphora resolution* (Doctoral dissertation). Master's thesis, Faculty of Informatics, Masaryk University.

Ozer, S., & Zeyrek, D. (2019). An automatic discourse relation alignment experiment on TED-MDB. *Proceedings of the 2019 Workshop on Widening NLP*, 31–34. https://aclanthology.org/W19-3612

Rahman, A., & Ng, V. (2011). Coreference resolution with world knowledge. *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 814–824.

Reinhart, T. (1983). Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy*, 47–88.

Sidner, C. L. (1979). *Towards a computational theory of definite anaphora comprehension in english discourse.* (tech. rep.). Massachusetts Inst of Tech Cambridge Artificial Intelligence lab.

Spencer, A., Goodluck, H., Wardhaugh, R., Blakemore, D., Kenstowicz, M., Scbiffrin, D., Clark, J., Yallop, C., Tsujimura, N., & Borsley, R. D. (1991). Blackwell textbooks in linguistics.

Spooren, W., & Degand, L. (2010). Coding coherence relations: Reliability and validity.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 197–207. https://doi.org/10.18653/v1/K18-2020

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and coreference resolution: A review. *Information Fusion*, *59*, 139–162.

Trouilleux, F. (2002). A rule-based pronoun resolution system for french. *4th Discourse Anaphora and Anaphor Resolution Colloquium*, 1.

Tüfekçi, P., & Kiliçaslan, Y. (2005). A computational model for resolving pronominal anaphora in turkish using hobbs' naıve algorithm. *WEC (5)*, *2005*, 13–17.

Turan, U. (1998). Ranking forward-looking centers in turkish: Universal and language-specific properties. *Centering theory in discourse*, *8*, 139–160.

Umakoshi, M., Murawaki, Y., & Kurohashi, S. (2021). Japanese zero anaphora resolution can benefit from parallel texts through neural transfer learning. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1920–1934.

Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). *Practical natural language processing: A comprehensive guide to building real-world nlp systems*. O'Reilly Media.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Xiao, C. (2021). A literature review on centering theory. *Studies in Literature and Language*, *22*(3), 5–12.

Yang, X., Su, J., & Tan, C. L. (2008). A twin-candidate model for learning-based anaphora resolution. *Computational Linguistics*, *34*(3), 327–356.

Yıldırım, S. (2008). Türkçe derlemlerdeki artgönderimlerin tümdengelimli ve tümevarımlı yöntemlerle çözümlenmesi.

Zeroual, I., & Lakhouaja, A. (2020). Multed: A multilingual aligned and tagged parallel corpus. *Applied Computing and Informatics*.

Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., & Ogrodniczuk, M. (2020). Ted multilingual discourse bank (ted-mdb): A parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, *54*(2), 587–613.

Zeyrek, D., Mendes, A., & Kurfali, M. (2018). Multilingual extension of pdtb-style annotation: The case of ted multilingual discourse bank. *Proceedings of the 11th Language Resources and Evaluation Conference-LREC'2018*, 1913–1919.

Appendix A

COLLECTIVE NOUN LIST

A.1 Collective Noun list: English

"hesta","pentair","company","society", "family", "group", "crowd", "gang", "crew", "staff", "choir", "orchestra", "panel", "board", "stack", "series", "class", "jury", "audience"

A.2 Collective Noun List: Turkish

"şirket", "toplum", "aile", "grup", "kalabalık", "çete", "tayfa", "kadro", "koro", "orchestra", "panel", "kurul", "yığın", "seri", "sınıf", "jüri", "seyirci"

Appendix B

OTHER RESULTS

Document No	Recall	Precision	F1 score
1927	0.25	0.43	0.32
1971	0.39	0.64	0.48
1976	0.5	0.74	0.59
1978	0.5	0.69	0.58
2009	0.54	0.85	0.66
2150	0.44	0.75	0.55

Table B.1: Results for Top $N = 1$ in Turkish

Document No	Recall	Precision	F1 score
1927	0.40	0.67	0.50
1971	0.46	0.76	0.57
1976	0.60	0.69	0.72
1978	0.57	0.80	0.67
2009	0.54	0.85	0.66
2150	0.5	0.85	0.63

Table B.2: Results for Top $N = 3$ in Turkish

Document No	Recall	Precision	F1 score
1927	0.49	0.56	0.52
1971	0.36	0.45	0.36
1976	0.35	0.38	0.36
1978	0.39	0.43	0.41
2009	0.46	0.62	0.53
2150	0.41	0.53	0.46

Table B.3: Results for Top $N = 1$ in English

Document No	Recall	Precision	F1 score
1927	0.69	0.79	0.74
1971	0.43	0.54	0.48
1976	0.53	0.57	0.55
1978	0.60	0.66	0.63
2009	0.53	0.70	0.60
2150	0.5	0.65	0.56

Table B.4: Results for Top $N = 3$ in English

Appendix C

THE DISTRIBUTION OF ANNOTATED ANAPHORA-REFERENT PAIRS AND THEIR PREDICTION

C.1 English Distribution

Document No	Total pairs	Zero anaphora (excluded)	True match	False match	No prediction
1927	135	56	Top(1)= 39 Top(3)= 55	Top(1)= 30 Top(3) = 14	10
1971	43	13	Top(1)= 11 Top(3)=13	Top(1)= 13 Top(3)= 11	6
1976	94	17	Top(1)= 27 Top(3)= 41	Top(1)= 44 Top(3)= 30	6
1978	162	128	Top(1)= 51 Top(3)=78	Top(1)= 66 Top(3)=39	11
2009	56	24	Top(1)= 15 Top(3)=17	Top(1)= 9 Top(3)=7	8
2150	61	5	Top(1)=23 Top(3)=28	Top(1)= 20 Top(3)=15	13

Table C.1: Distribution of annotated anaphora-referent pairs and their predictions in English

C.2 Turkish Distribution

Document No	Total pairs	Zero anaphora (excluded)	True match	False match	No prediction
1927	177	115	Top(1)= 16 Top(3)= 25	Top(1)= 21 Top(3) = 12	25
1971	50	22	Top(1)= 11 Top(3)=13	Top(1)= 6 Top(3)=4	11
1976	104	46	Top(1)= 29 Top(3)= 35	Top(1)= 10 Top(3)=4	19
1978	184	106	Top(1)= 39 Top(3)=45	Top(1)=17 Top(3)=11	22
2009	91	47	Top(1)= 24 Top(3)=24	Top(1)= 4 Top(3)=4	16
2150	77	7	Top(1)=31 Top(3)=35	Top(1)= 10 Top(3)=6	29

Table C.2: Distribution of annotated anaphora-referent pairs and their predictions in Turkish

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 [X]

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences [ ]

YAZARIN / AUTHOR

Soyadı / Surname : ERTAN
Adı / Name : MELEK
Bölümü / Department : Enformatik Enstitüsü, Bilişsel Bilimler/ Informatics Institute, Cognitive Science

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English) : Pronominal Anaphora Resolution in Turkish and English/ Türkçe ve İngilizcede Öngönderim Çözümlemesi

TEZİN TÜRÜ / DEGREE**:** **Yüksek Lisans** / Master [X] **Doktora** / PhD [ ]

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. [X]
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 .....