

REFERENCE SELECTION IN TURKISH: A CORPUS-BASED APPROACH

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

BY

FARUK BÜYÜKTEKİN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COGNITIVE SCIENCES

JUNE 2025

REFERENCE SELECTION IN TURKISH: A CORPUS-BASED APPROACH

submitted by **Faruk Büyüktekin** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Cognitive Sciences Department, Middle East Technical University** by,





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: Faruk Büyüktekin

Signature :

ABSTRACT

REFERENCE SELECTION IN TURKISH: A CORPUS-BASED APPROACH

Büyüktekin, Faruk

Ph.D., Department of Cognitive Sciences

Supervisor: Assist. Prof. Dr. Umut Özge

June 2025, 127 pages

This thesis investigates reference selection in natural language, focusing on the mechanisms that shape the form of referring expressions. Drawing from both linguistic theory and data-driven computational methods, the study seeks to uncover how grammatical, discourse, and cognitive factors jointly influence referential form. As the target language, Turkish offers a testing ground to explore referential choices beyond the patterns observed in well-studied languages due to its typologically distinct characteristics specifically, its rich morphology, frequent use of null pronouns, and flexible word order. A central contribution of this work is the creation of a novel coreference corpus based on spontaneous, goal-directed dialog. Unlike prior research that has typically relied on semi-artificial or isolated sentences, or written texts, this study uses situated task-based interaction, capturing reference in real-time naturalistic speech. To facilitate this, a new annotation scheme was developed to represent the full range of referential forms, including full noun phrases, overt pronouns, and null pronouns, and their contextual and grammatical properties. The resulting corpus, which is the most comprehensive coreference corpus of Turkish dialogs to date, enables systematic and computationally viable analyses of referential phenomena. Building on this resource, the thesis conducts extensive statistical analyses and employs machine learning to evaluate the effects and interactions of multiple features on referential form. These include speaker role, turn-taking, grammatical role, competition, distance, topicality, and sentential position. Among the findings, competition and distance emerged as the most predictive features in model performance, while speaker role and turn-taking showed weaker but interpretable effects. Statistical tests confirmed that many of these factors significantly influence form choice, supporting and extending theoretical predictions of major theories and models related to referential form selection. By integrating corpus development, feature engineering, statistical modeling, and explanatory machine learning, this thesis offers a unified framework

for analyzing reference in Turkish. It not only contributes to theoretical accounts of referential choice in typologically diverse settings but also provides scalable tools for future research in natural language processing, cognitive modeling, and dialog systems.

Keywords: corpus, dialog, coreference, referring expression, referential form



ÖZ

TÜRKÇE'DE REFERANS SEÇİMİ: DERLEM TEMELLİ BİR YAKLAŞIM

Büyüktekin, Faruk

Doktora, Bilişsel Bilimler Bölümü

Tez Yöneticisi: Dr. Öğr. Üyesi. Umut Özge

Haziran 2025 , 127 sayfa

Bu tez, doğal dilde gönderim seçimini incelemekte ve bunu belirleyen mekanizma üzerine odaklanmaktadır. Dilbilim kuramları ve veri odaklı yöntemlerden yararlanarak, dilbilgisel, söylemsel ve bilişsel etmenlerin gönderim biçimini nasıl etkilediğini ortaya konmaya çalışmıştır. Zengin biçimbilgisi, örtük argüman kullanımı ve esnek sözdizimi gibi tipolojik olarak ayırt edici özellikleri ile hedef dil olan Türkçe, iyi çalışılmış dillere kıyasla gönderim tercihlerini araştırmak için zengin bir test ortamı sunmaktadır. Bu çalışmanın temel katkılarından biri, amaç odaklı diyaloglara dayanan özgün bir eşgönderim derlemi oluşturmaktır. Önceki araştırmaların çoğu yarı-yapay veya izole cümlelere ya da yazılı metinlere dayanırken, bu çalışma, gerçek zamanlı doğal konuşma verilerini esas alarak gönderimin söylem akışı içinde nasıl gerçekleştiğini incelemektedir. Bu doğrultuda, tam ad öbekleri, açık ve örtük adılar olmak üzere tüm gönderim biçimlerini ve bunların bağlamsal ve dilbilgisel özelliklerini içeren bir işaretleme şeması geliştirilmiştir. Bununla oluşturulan derlem, mevcut en kapsamlı Türkçe dialog eşgönderim veri setini teşkil ederek, gönderim olgusunun sistematik ve hesaplama elverişli bir şekilde analiz edilmesini mümkün kılmıştır. Derleme kapsamlı istatistiksel analizler ve makine öğrenmesi yöntemleri uygulanmış ve belirli özelliklerin gönderim biçimi üzerindeki etkileri ve etkileşimleri değerlendirilmiştir. İncelenen özellikler arasında konuşucu rolü, konuşma sırası, dilbilgisel rol, rekabet, mesafe, gönderim yoğunluğu ve sözdizimsel konum gibi ölçütler yer almaktadır. Bulgular arasında, söylem içi rekabet ve mesafe, model başarımını en çok etkileyen özellikler olarak öne çıkmıştır. Konuşmacı rolü ve sırasının etkisi ise daha zayıf olmakla birlikte yorumlanabilir düzeydedir. Yapılan istatistiksel testler, birçok etkenin gönderim biçimini anlamlı şekilde etkilediğini doğrulamış ve gönderim seçimi ile ilgili kuramsal yaklaşımların öngörülerini desteklemiş ve genişletmiştir. Bu tez, derlem geliştirme, özellik mühendisliği, istatistiksel modelleme ve

açıklanabilir makine öğrenmesini entegre ederek Türkçe’de gönderim analizi için bütüncül bir yaklaşım önermiştir. Çalışma tipolojik olarak farklı bir dil ile kuramsal referans seçimi çalışmalarına katkı yapmakla birlikte doğal dil işleme, bilişsel modelleme ve dialog sistemleri gibi alanlarda kullanabilecek araçlar sunmaktadır.

Anahtar Kelimeler: derlem, dialog, eşgönderim, gönderim ifadesi, gönderim ifadesi biçimi





To my family

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Asst. Prof. Dr. Umut Özge. His endless support, insightful guidance, and patient mentorship were truly instrumental in every stage of this research. His expertise, encouragement, and friendship not only shaped the course of this thesis but also made the entire journey more meaningful and rewarding. I could not have completed this work without him.

I extend my sincere appreciation to my defense committee members Assoc. Prof. Dr. Murat Perit Çakır, Prof. Dr. İsmail Sengör Altıngövde, Assoc. Prof. Dr. İsa Kerem Bayırlı, and Asst. Prof. Dr. Öykü Mançe Çalışır for their valuable time, rigorous feedback, and thoughtful suggestions that significantly strengthened this work. Their perspectives were very helpful in refining and developing this work.

I am particularly grateful to The Scientific and Technological Research Council of Türkiye (TÜBİTAK), for funding my research stay at the Collaborative Research Centre (CRC 1252) "Prominence in Language" , University of Cologne through the 2214-A International Research Fellowship Program. Additionally, I am also grateful to the Council of Higher Education (YÖK) for awarding me the 100/2000 PhD Scholarship in the field of Data Science, which provided support during my doctoral studies. Finally, I am thankful to the Middle East Technical University Scientific Research Projects Coordination for funding the annotation work necessary for my research under the grant number GAP-704- 2023-11066.

I would like to thank Prof. Dr. Klaus von Heusinger of House of Prominence at the University of Cologne and his research team for their generous hospitality and intellectual collaboration during my research stay. The discussions with the team and their work on Turkish linguistics fundamentally enriched both my research and perspective about reference. I am also thankful to Assoc. Prof. Dr. Cengiz Acartürk and Prof. Dr. Ruslan Mitkov for arranging my study stay at the University of Wolverhampton, where I gained invaluable methodological insights about natural language processing.

I would like to extend special thanks to Anıl Ögdül, Batuhan Karataş, and Derin Dinçer for their meticulous and dedicated work in the annotation process. Their attention to detail and consistency played a crucial role in ensuring the integrity and reliability of the dataset that underpins this research. I am deeply indebted to Korkut Koçak and Mehmet Ali Akyol. Their deep knowledge and expertise in statistics and data science guided me throughout the analysis process of the thesis. I am grateful to Alaz Aydın, who joined this reference journey as well and answered my questions whenever I had one. I am also thankful to Sibel Gülnar and Hakan Güler, and my friends at the Graduate School of Informatics, who helped me preserve and finish this work especially towards the end. Finally, I have been so lucky to have great people in my life. I am thankful to Evren, Coşkun, Beytullah, and Mehmet. Their friendship made this journey fun and bearable. And Korkut and Ferda... No words can fully

capture my gratitude to you.

This work belongs as much to you all as it does to me.



TABLE OF CONTENTS

ABSTRACT	iv
ÖZ	vi
ACKNOWLEDGMENTS	ix
TABLE OF CONTENTS	xi
LIST OF TABLES	xix
LIST OF FIGURES	xxi
LIST OF ABBREVIATIONS	xxiii
CHAPTERS	
1 INTRODUCTION	1
1.1 The Object of Study	2
1.2 Research Gaps and Motivation	2
1.3 Aims and Objectives	4
1.4 Significance of the Study	4
1.4.1 Empirical Contributions	4
1.4.2 Theoretical Contributions	5
1.4.3 Methodological Contributions	6
1.5 Research Questions	6
1.6 Thesis Structure	7

1.7	Conclusion	8
2	BACKGROUND	9
2.1	Information structure and givenness	9
2.2	Status-driven Models	10
2.2.1	Taxonomy of assumed familiarity	10
2.2.2	Continuity in discourse	12
2.2.3	Activation model	12
2.2.4	Accessibility Scale	13
2.2.5	Givenness Hierarchy	14
2.3	Coherence Models	16
2.3.1	Centering Theory	17
2.3.1.1	Centering theory in other languages	19
2.3.1.2	Corpus based evaluation of Centering Theory	19
2.3.2	The Bayesian model	20
2.4	Rational models	21
2.4.1	Rational Speech Act (RSA) Theory	22
2.5	Empirical Work	24
2.6	Coreference and Referential Selection	25
2.7	Coreference and anaphora	25
2.7.1	Basic terminology	26
2.7.2	Related work	27
3	MOTIVATION	31
3.1	Limitations of previous accounts of reference selection	31

3.1.1	Local Context in Reference Selection Research	32
3.1.2	Focusing on a Narrow Range of Referring Expressions	33
3.1.3	Controlled Experimental Settings in Reference Selection Research	34
3.1.4	Neglecting Pro-Drop and Flexible Word Order Languages	35
3.2	Situated Dialogues as Ecologically Valid Source of Data	36
3.2.1	Coreference as a Natural Domain for Reference Selection Research	37
3.2.2	Coreference Annotation Schemes Developed for Typologically Similar Languages	38
3.3	Limited Coreference Datasets and Computational Models for Most World Languages	41
3.4	Addressing the Limitations: Reference Selection in Turkish Dialogs	42
3.4.1	Developing a Novel Annotation Scheme and Constructing a Coreference Corpus	42
3.4.2	Feature Engineering	43
3.4.3	Systematic Analysis of Features	43
3.4.4	Advancing Reference Selection Research	44
	Summary of Contributions	45
4	CORPUS	47
4.1	Corpus creation	47
4.1.1	Genre selection	47
4.1.2	The source of our texts	48
4.1.2.1	Experimental setup	48
4.1.3	Text preparation	49

4.1.4	Tool choice	50
4.1.5	Training the annotators	50
4.2	Annotation scheme development	51
4.2.1	Scope	51
4.2.2	Markables	52
4.2.3	Non-markables	53
4.2.4	Various issues in coreference annotation	54
4.2.4.1	Embedded mentions	54
4.2.4.2	Mentions in compounds	54
4.2.4.3	Appositives	55
4.2.4.4	Genitive-possessive constructions	55
4.2.4.5	Grammatical coreference	55
4.2.4.6	Split anaphora	55
4.2.4.7	Null elements	56
4.2.5	Annotation procedure	56
4.3	Corpus statistics	56
4.3.1	Inter-Annotator Agreement	57
4.3.2	Descriptive statistics of the corpus	58
5	ANALYSES	61
5.1	Feature Engineering: Raw Data and Derived Features	61
5.1.1	Raw Data Extraction	62
5.1.2	Features	62
5.1.2.1	Linguistic Forms	63

5.1.2.2	Speaker and Dialog-Level Features	63
5.1.2.3	Turn-taking features	64
5.1.2.4	Grammatical Roles	65
5.1.2.5	Competition Metrics	66
5.1.2.6	Distance Metrics	68
5.1.2.7	Topicality Metrics	69
5.1.2.8	Word Order Features	69
5.2	Hypotheses	70
5.3	Dataset	72
5.4	The distribution of referring expressions according to the role of the discourse participants	73
5.4.1	Turn-Taking and Referential Form	75
5.5	Grammatical Role	77
5.5.1	Analysis of the Relationships Between Referential form and Grammatical Role	78
5.5.1.1	The Linguistic Form of the Antecedent	78
5.5.2	The Relationship Between <code>rform</code> and <code>rrole</code>	78
5.5.3	The Relationship Between <code>rform</code> and <code>aform</code>	79
5.6	Correlation Analysis of Continuous Features	80
5.6.1	General Patterns	81
5.6.1.1	Distance/Recency Measures	81
5.6.1.2	Word Order and Distance	81
5.7	Competition in Referring Expressions	82
5.7.1	Quantifying Competition	82

5.7.2	Distribution of Referring Expressions Across Competitor Men- tions	83
5.7.2.1	<code>rform mcomp0</code>	84
5.7.2.2	<code>rform mcomp2</code>	84
5.7.2.3	<code>rform mcomp3</code>	86
5.8	Distance/Recency	87
5.9	Sentential position	88
5.9.1	The word order of the antecedent and referential form	88
5.9.2	The word order of the referring expression and referential form	90
5.10	Discourse Density	91
5.10.1	Global Discourse Density (<code>mdensity_a</code>)	92
5.10.2	Local Discourse Density (<code>mdensity_m</code>)	93
5.11	Feature Selection and Preparation	94
5.11.1	Overview	94
5.11.2	Included Features and Motivations	95
	Linguistic Form: <code>aform</code>	95
	Grammatical Role: <code>arole</code> and <code>rrole</code>	95
	Turn-Taking: <code>tt_mention</code>	95
	Speaker Role: <code>speaker</code>	95
	Distance/Recency: <code>wdist</code>	95
	Competition Metrics: <code>mcomp3</code>	95
	Discourse Salience: <code>mdensity_a</code> , <code>mdensity_m</code>	96
	Word Order: <code>awo</code>	96

5.11.3	Excluded Features and Motivations	96
	Dialog Identifier: <code>dialog</code>	96
	Turn-Taking (Sentence Level): <code>tt_sentence</code>	96
	Entity Competition: <code>ecomp0–ecomp3</code>	96
	Additional Distance Metrics: <code>asent, rsent, sdist</code>	96
5.11.4	Data Preparation for XGBoost Modeling	96
5.11.5	XGBoost for Reference Prediction	97
5.11.6	Summary of Included Features	97
5.12	XGBoost Modeling and Results	97
5.12.1	Target Variable and Feature Selection	99
5.12.2	Modeling with XGBoost	100
5.12.3	Feature Importance	100
6	DISCUSSION AND CONCLUSION	103
6.1	Overview of Findings	103
6.1.1	Discourse Factors as Dominant Predictors	103
6.1.2	Grammatical Role and Word Order	104
6.1.3	Interactional Factors: Limited but Present Effects	105
6.1.4	Overt Pronouns as a Special Case	105
6.1.5	Theoretical Implications	105
6.2	Limitations and Future Directions	106
6.3	Conclusion	107
	REFERENCES	109

APPENDICES	117
A ANNOTATION SCHEME	119
A.1 Task	119
A.1.1 Roles of discourse participants:	119
A.1.2 Working space	119
A.1.2.1 Target shapes:	119
A.1.2.2 Shapes and their anchors	119
A.2 Markables	120
A.2.1 Span boundary	120
A.2.2 Linguistic form of candidate markables	121
A.3 Non-markables	121
A.4 Steps	122
A.5 Tags	122
A.6 How to annotate overt entities	122
A.6.1 Mentions in matrix clauses	122
A.6.2 Subordinate clauses	123
A.7 Null entities	124
A.7.1 Mentions in matrix clauses	124
A.7.2 Subordinate clauses	124
CURRICULUM VITAE	127

LIST OF TABLES

TABLES

Table 2.1	Summary of Issues in Status-Based Models	17
Table 3.1	Coverage of Typical Annotation Schemes(e.g., OntoNotes, AR- RAU) vs. Needs for Pro-Drop, Morphologically Rich, Free Word Order Languages	40
Table 4.1	Anchors for the entities	50
Table 4.2	Interannotator agreement scores	58
Table 4.3	Counts of the corpus	58
Table 5.1	Distribution of referring expression forms (rform) by turn-taking at sentence level (tt_sentence)	76
Table 5.2	Distribution of referring expression forms (rform) by turn-taking at mention level (tt_mention)	76
Table 5.3	Distribution of grammatical roles of referring expressions	77
Table 5.4	Distribution of grammatical roles across referential forms	78
Table 5.5	Cross-tabulation of Referring Expression Form (rform) and Gram- matical Role of the Referent (rrole)	79
Table 5.6	Cross-tabulation of Antecedent Form (aform) and Referring Ex- pression Form (rform)	79

Table 5.7 Descriptive Statistics of Mention Competition (<code>mcomp0</code>) by Referring Expression Form (<code>rform</code>)	84
Table 5.8 Descriptive Statistics of Antecedent-Side Mention Competition (<code>mcomp2</code>) by Referring Expression Form (<code>rform</code>)	85
Table 5.9 Descriptive statistics for antecedent word order (<code>awo</code>) by referring expression form (<code>rform</code>)	88
Table 5.10 Descriptive Statistics of Word Order <code>rwo</code> by Referring Expression Form <code>rform</code>	90
Table 5.11 Descriptive statistics of <i>mdensity_a</i> by referring expression form (<i>rform</i>)	92
Table 5.12 Descriptive statistics for local discourse density <code>mdensity_m</code> by referring expression form <code>rform</code>	94
Table 5.13 Included Features for XGBoost Model	98
Table 5.14 Confusion Matrix	100

LIST OF FIGURES

FIGURES

Figure 2.1	Taxonomy of the Assumed Familiarity (from Prince, 1981). . . .	10
Figure 2.2	Givon's continuity hierarchy of linguistic forms	13
Figure 2.3	Accessibility Scale	14
Figure 2.4	Givenness Hierarchy	15
Figure 2.5	Reference types (Halliday and Hasan, 1976, p.33)	27
Figure 4.1	Target Shapes	49
Figure 4.2	Annotation sample from the annotation tool	51
Figure 4.3	Grammatical Roles	59
Figure 4.4	Linguistic forms	60
Figure 4.5	Grammatical Roles	60
Figure 5.1	Speaker role	74
Figure 5.2	Correlation Analysis of Continuous Features	80
Figure 5.3	Distribution of referring expression forms (<code>rform</code>) across levels of local mention competition (<code>mcomp0</code>).	85
Figure 5.4	Distribution of referring expression forms (<code>rform</code>) across levels of antecedent mention competition (<code>mcomp2</code>).	86

Figure 5.5	Distribution of referring expression forms across different levels of sentence distance (<code>sdist</code>) between antecedent and referent.	87
Figure 5.6	Distribution of referring expression forms across antecedent po- sitions	89
Figure 5.7	Distribution of referring expression forms across referring ex- pression positions	91
Figure 5.8	Distribution of <code>mdensity_a</code> across referring expression forms (<code>rform</code>)	93
Figure A.1	Working space	120
Figure A.2	Target Shapes	120

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
RSA	Rational Speech Act Theory





CHAPTER 1

INTRODUCTION

Reference is fundamental to human communication. Through referring expressions, speakers can direct their interlocutors' attention to entities in the world or in discourse, establishing and maintaining coherence across utterances. Understanding the mechanisms that guide speakers' choices among referential forms, whether to use full noun phrases, pronouns, or null expressions, has long been a central concern in linguistic, cognitive, and natural language processing research (Ariel, 1990; Gundel et al., 1993; Prince, 1981).

Research over the past decades has demonstrated that referential choice is shaped by factors such as the grammatical role of the referent, its discourse salience and accessibility (Grosz et al., 1995), the recency of prior mentions, and the presence of competing referents in the local context. These insights have led to the development of influential models, including Accessibility Theory (Ariel, 1990), the Givenness Hierarchy (Gundel et al., 1993), and Bayesian models (Kehler & Rohde, 2013), and Rational Models (Degen, 2023), each proposing that speakers adjust referential form based on their assumptions about the addressee's mental representation of discourse entities.

The study of reference has long been a matter of investigation in the fields of philosophy of language, formal semantics, pragmatics, and cognitive science. Early theoretical approaches, such as Russell (1905)'s theory of definite descriptions and S. A. Kripke (1972)'s causal-historical account of proper names, focused primarily on logical form and truth-conditional interpretation. These frameworks sought to explain how linguistic expressions map to entities in the world, often relying on idealized, decontextualized examples. However, such approaches tend to abstract away from the dynamic, interactive nature of reference in actual discourse, where factors like speaker goals, common ground, and cognitive processing shape referential choice. In contrast to purely truth-conditional accounts, cognitive theories have emphasized the role of mental representation and discourse prominence in reference resolution. Models such as The Givenness Hierarchy (Gundel et al., 1993), Accessibility Theory (Ariel, 1990), and Centering Theory (Grosz et al., 1995) propose that speakers select referring expressions (e.g., full noun phrases, pronouns, or zero anaphora) based on the cognitive status of the referent in the hearer's mind. These theories posit scales of accessibility, where more reduced forms (e.g., pronouns) signal higher activation, while fuller forms (e.g., definite descriptions) indicate lower accessibility. Such accounts bridge linguistic form with cognitive processes, explaining patterns in naturally occurring discourse.

More recent advances in computational linguistics and probabilistic pragmatics have framed reference as a rational, goal-directed process. Rational Speech Act (RSA) models (Frank & Goodman, 2012) treat referential choice as a Bayesian decision problem, where speakers select forms that optimally balance informativeness and efficiency under uncertainty. These models capture not only individual production choices but also speaker-hearer coordination in interactive settings, predicting how conversational partners adapt to each others knowledge and expectations.

In this thesis, these diverse approaches jointly inform the selection and operationalization of features that influence referential form in situated dialog. By combining insights from formal theories with data-driven computational modeling, this work aims to provide a more comprehensive understanding of how reference operates in real-world communication.

1.1 The Object of Study

The unit of observation in this study is a referential act. A referential act is defined as a speaker’s attempt to refer to an entity in the discourse or the world. We call the entity that is referred to the **referent**, the expression used to refer to the referent the **mention** or **referring expression**, interchangeably. We further narrow down our focus to a special type of referential act where the reference is to an entity that is already mentioned at least once in the preceding discourse. We call this sub-type of referential act **coreference**. To illustrate over a minimal discourse:

(1) **Ali** eve geldi. Mutfakta **onu** bir sürpriz bekliyordu.

Here the expressions *Ali* and *onu* are both references to (or mentions of) the same entity, namely an individual named ‘Ali’. The relation, which we call **coreference** between these two mentions is our unit of observation. We call *onu* the referring expression and *Ali* the antecedent. In what follows when we want to speak about certain properties of these items, we will modify the terms by ‘r’(=reference) and ‘a’ (=antecedent). For instance, when we want to speak about the form of the antecedent we will use ‘aform’ and when we want to speak about the form of the reference we will use ‘rform’.

In the above example the reference *o* is a pronoun. Under other conditions the form of the reference could have been *adamı*, *Ali’yi*, or some other form. Our central question concerns the factors that influence the choice of the form of the reference in a coreference act. We call this problem **referential form selection** or **referential choice**.

1.2 Research Gaps and Motivation

Related to the problem of **referential form selection**, we have observed several gaps in the literature, which we aimed to address in this work. Despite substantial

advances, research on referential form choice has predominantly focused on *typologically similar languages*, especially English and other Indo-European languages. These languages typically exhibit *rigid word order*, *relatively limited morphology*, and *restricted use of null referential forms*, creating structural constraints that differ significantly from languages such as Turkish. Turkish, as a *morphologically rich, pro-drop language with flexible word order*, presents a markedly different context for studying reference production. The freedom to omit arguments, re-order constituents, and rely heavily on discourse-pragmatic cues makes Turkish an ideal case for investigating how linguistic, cognitive, and discourse factors interact to shape referential choices.

In addition to its typological distinctiveness, much prior research has relied on *controlled experimental designs* employing *isolated or semi-artificial sentences* to manipulate individual discourse variables. While such methods afford high internal validity, they often fail to capture the *dynamics of naturally occurring speech*. In contrast, task-based collaborative dialogs offer a rich environment where *reference production is situated within real-time, goal-oriented communication*. Such contexts introduce spontaneous disfluencies, planning pressures, and complex discourse structures that more closely approximate everyday language use. To better understand referential choice as it unfolds naturally, it is crucial to examine it in *situated interactional settings* rather than solely under experimental manipulation.

Another major limitation in the literature is the tendency to examine *individual factors in isolation* such as discourse salience, grammatical role, or competition rather than considering how *multiple factors jointly influence* the choice of referring expression. However, reference production is inherently *multi-factorial*, shaped by interacting discourse, syntactic, cognitive, and pragmatic pressures. Without modeling these interactions, explanations of referential behavior risk oversimplification. A comprehensive analysis must simultaneously account for these diverse influences to reveal how speakers integrate competing constraints when selecting referential forms.

Moreover, much of the existing work on discourse structure has limited its scope to *local discourse environments*, often considering only the immediate or immediately preceding sentence. While recency certainly facilitates referent accessibility, focusing exclusively on adjacent utterances ignores longer-range dependencies and cumulative discourse effects. A broader view that incorporates both *local and global discourse environments* is necessary to capture the true dynamics of salience and accessibility over extended dialogic interaction.

Finally, research often restricts attention to a *single referential form* (e.g., pronouns vs. noun phrases) or a *binary distinction* between reduced and full forms. Such simplifications overlook the rich array of options speakers have at their disposal. In pro-drop languages like Turkish, speakers not only choose between full noun phrases and pronouns, but also between *overt pronouns and phonologically null forms*. A full account of referential form selection must therefore consider the *entire range of available forms*, capturing the nuances across these forms.

In this study, we address these gaps systematically. Our approach integrates (i) a *typologically diverse language* (Turkish), (ii) *naturalistic, task-oriented dialog data*, (iii) *simultaneous analysis of multiple linguistic and discourse factors*, (iv) *local and*

global discourse context, and (v) *comprehensive inclusion of all referential forms*. Through the development of a novel annotated corpus and the application of both statistical and machine learning methods, we aim to offer a theoretically grounded, ecologically more valid, and empirically robust account of referential choice in dialogic interaction.

1.3 Aims and Objectives

Therefore the aim of this thesis is to provide a systematic and empirically grounded account of referential form selection in Turkish situated dialogs. It seeks to achieve this through the following objectives:

(2) *Objectives of the thesis:*

- a. To develop a detailed annotation scheme capturing full noun phrases, overt pronouns, and zero pronouns addressing the peculiarities of Turkish
- b. To construct a novel Turkish coreference corpus based on naturalistic, task-based dialogs.
- c. To derive a set of theoretically motivated features from the resulting corpus data, representing linguistic (e.g., grammatical roles), discourse (e.g., distance, competition, and topicality), and interactional (e.g., turn-taking) properties.
- d. To test hypotheses regarding how these factors interact to shape referential form, based on previous theories of accessibility, salience, and discourse management.
- e. To apply statistical and machine learning methods to model referential choice and determine the most predictive features.

1.4 Significance of the Study

In line with these aims, the present study offers contributions to reference research in several dimensions:

1.4.1 Empirical Contributions

- **Annotation Scheme Development:** As part of the corpus creation process, the study developed a novel and linguistically-informed annotation scheme tailored to the specific challenges of Turkish. This scheme systematically accounts for morphosyntactic properties, pro-drop characteristics, referential forms (full noun phrases, overt pronouns, and zero pronouns), grammatical roles, and discourse structure. By formalizing and applying this scheme, the research ensures consistent, replicable, and theory-neutral annotations, thus laying a methodological foundation for future studies on reference and coreference in Turkish

and other morphologically rich languages.

- **Corpus Development:** This study introduces first and the most comprehensive Turkish coreference corpus of situated dialogs, filling a gap in Turkish discourse and reference research. Unlike existing corpora that predominantly focus on written genres or limited dialogic samples, this corpus captures task-oriented spontaneous speech in collaborative tasks. The data reflect the natural production of referring expressions under problem solving conditions, where speakers coordinate actions and manage joint attention. Furthermore, the corpus includes detailed annotations at multiple levels—mention spans, coreference chains, referential forms, grammatical roles, and turn-taking cues—making it a rich resource for the study of coreference resolution, referential choice, and discourse processing in Turkish. The corpus is expected to serve as an essential empirical foundation for both theoretical investigations and the development of computational models of reference in Turkish and other morphologically rich languages.
- **Data-Driven Analysis:** The study advances the empirical investigation of referential choice by adopting a systematic and multi-level data-driven analysis framework. Building on the richly annotated corpus, it applies a series of advanced statistical methods, including descriptive and inferential statistics (e.g., chi-square tests, Kruskal-Wallis tests, post-hoc comparisons), correlation analyses, and machine learning models to examine the selection of referential forms across multiple dimensions. Specifically, the analysis investigates how grammatical roles, discourse-level features (such as recency, sentence and word distance, discourse density, competition among referents), word order positions, and interactional features (e.g., speaker role, turn-taking) shape the production of referring expressions. Importantly, the study does not limit itself to isolated factors but explores their interaction and combined effects, thus offering a holistic view of referential choice in naturalistic discourse. This integrative approach enables the detection of subtle patterns and dependencies that single-variable analyses might overlook, providing empirically grounded insights into how speakers manage salience, accessibility, and ambiguity during situated communication. Furthermore, by complementing traditional statistical approaches with machine learning techniques, the study enhances the robustness of its findings in discourse and reference research.

1.4.2 Theoretical Contributions

- **Integration of Factors:** The findings support multi-factorial accounts of referential choice, demonstrating that accessibility, competition, recency, and syntactic positioning jointly shape speakers' decisions.
- **Context Sensitivity:** The results provide evidence for context-sensitive referencing strategies in Turkish, refining our understanding of how competition and distance affect the use of reduced and explicit forms.

1.4.3 Methodological Contributions

- **Machine Learning Analysis:** The application of XGBoost to predict referential form highlights the complex cognitive phenomena, allowing for the discovery of feature interactions often inaccessible to traditional statistical methods.

Building on the frameworks outlined above, this thesis formulates a set of hypotheses concerning the factors influencing the choice of referential form. The hypotheses reflect assumptions grounded in these models, adapted to the structural characteristics of Turkish.

H1 (Speaker Role): Task-specific roles (e.g., operator vs. presenter) modulate referential strategies, as speakers with more discourse control may rely more on explicit referencing due to role requirements.

H2 (Turn-Taking): Referential form is influenced by whether the mention occurs within the same turn as its antecedent or across speaker boundaries. Cross-turn references are predicted to increase the likelihood of explicit forms, reflecting cognitive load and memory effects.

H3 (Grammatical Role): Referring expressions with subject antecedents will be more likely to take reduced forms (zero or overt pronouns) than those with object or oblique antecedents, consistent with theories that associate subjects with discourse salience.

H4 (Competition): In contexts where multiple discourse entities are salient (high competition), speakers will favor more explicit referring forms to reduce ambiguity.

H5 (Distance): As the distance (in words or sentences) from the antecedent increases, speakers will prefer more explicit forms (e.g., full noun phrases), aligning with models of accessibility decay.

H6 (Word Order): Referential expressions in syntactically prominent positions (e.g., sentence-initial subjects) are more likely to appear in reduced form, reflecting integration with the local syntactic structure.

These hypotheses are empirically evaluated through statistical and machine learning analyses of the dialog data to answer the research questions below.

1.5 Research Questions

This thesis seeks to advance our understanding of referential form selection in Turkish situated dialogs by systematically investigating how multiple linguistic, discourse, and interactional factors shape speakers' choices. Building upon theoretical models of accessibility and discourse processing, and utilizing both statistical and machine learning methods, the study addresses the following research questions:

- (3) a. How do grammatical roles at the antecedent and referent positions (a_{role} , r_{role}) influence the choice of referring expression form, and does the

influence differ when referring within versus across clauses and discourse segments?

- b. How does referential form selection interact with turn-taking dynamics (`tt_mention`, `tt_sentence`), particularly in terms of maintaining discourse continuity and re-establishing referents across speaker boundaries?
- c. To what extent does linear distance (`sdist`, `wdist`) between antecedent and referent affect referential form, and are there thresholds or gradient patterns in form choice across increasing distances?
- d. How do competition metrics at various discourse levels (`mcomp0`, `mcomp2`, `mcomp3`) impact the likelihood of using more explicit (full noun phrases) versus reduced forms (overt and zero pronouns)?
- e. What role does discourse density — both global (`mdensity_a`) and local (`mdensity_m`) — play in referential form selection, and how does it interact with recency and competition in shaping referential choices?
- f. How does sentence-level positioning of antecedents and referring expressions (`awo`, `rwo`) influence referential form, and to what extent do speakers prefer reduced forms in syntactically prominent positions?
- g. In what ways do task-specific speaker roles (presenter vs. operator; `speaker`) modulate referential choices, especially in terms of balancing explicitness and efficiency under asymmetrical information conditions?
- h. which factors emerge as the most powerful predictors predict referential form selection based on the aforementioned linguistic and discourse features, and ?

1.6 Thesis Structure

The thesis is organized as follows:

- **Chapter 2 (p. 9)** reviews theoretical frameworks relevant to reference production and interpretation, including models of accessibility, salience.
- **Chapter 3 (p. 31)** discusses the main limitations of previous accounts of reference selection and presents how the current thesis addresses them to advance reference research.
- **Chapter 4 (p. 47)** presents the methodology, covering corpus construction, annotation scheme development and initial analysis of the resulting corpus.
- **Chapter 5 (p. 61)** reports the feature extraction process and the resulting features and the results of descriptive and inferential statistical analyses, as well as machine learning modeling.
- **Chapter 6 (p. 103)** discusses the implications of the findings for models of reference production and discourse processing and offers concluding remarks, highlighting the study's contributions and outlining directions for future research.

1.7 Conclusion

This chapter has outlined the theoretical, empirical, and methodological motivations behind the study of referential form selection in Turkish situated dialogs. The thesis addresses significant gaps in the literature by moving beyond English-centric models, single-factor analyses, and well-planned and experimental settings. It does so through the construction of a richly annotated Turkish coreference corpus, the integration of multiple discourse and grammatical features, and the use of both traditional statistical and modern machine learning approaches. By capturing naturalistic language production in collaborative dialog, this study contributes a more ecologically valid and typologically diverse account of reference. Its findings can inform not only linguistic theory but also applied domains such as discourse-aware NLP systems, cognitive modeling of language use, and cross-linguistic discourse analysis. The next chapter introduces the theoretical foundations in greater depth, focusing on existing models of reference, their predictions, and their applicability to the Turkish data.



CHAPTER 2

BACKGROUND

This chapter reviews the theoretical and computational works related to reference and referring expressions. Referential selection is a central concern in linguistics and cognitive science, focusing on the factors influencing how speakers choose particular referring expressions in discourse. Accordingly, this chapter offers an overview of key theories and models, beginning with status-driven accounts and proceeding through empirical work, Centering Theory, and Bayesian and Rational Speech Act (RSA) models. This describes how perspectives on reference selection have evolved from descriptive to formal and probabilistic accounts. To lay the groundwork, the chapter first explores traditional status-based models, which emphasize givenness and accessibility as central concepts in determining referring expression choice.

2.1 Information structure and givenness

The origins of research on referential form choice lie in theories of information structure, where givenness plays a pivotal role. The term "information structure" was first coined by Halliday, following the work of the Prague school. In his work Halliday and Hasan (1976), he introduced the notion of *givenness*, which is based on the idea of recoverability of an entity from the preceding discourse. If the information is *new*, the hearer naturally cannot recover it from the previous discourse segment presented by the speaker. However, if the hearer can recover it *anaphorically* or *situationally*, the information is considered *given* or *old*.

Similarly, Haviland and Clark (1977) suggested that certain processes in comprehension are a consequence of an implicit agreement about conversations as part of Grice's *cooperative principle* and *maxims* (Grice, 1975). People are cooperative in their conversations, and in particular they adhere to *the given-new contract*. By this agreement, speakers attempt to judge what their listeners do and do not know, and they construct their sentences accordingly.

Upon these foundational ideas, subsequent research deepened the analysis of givenness and cognitive statuses. Notably, Chafe (1976) offered a framework for understanding how cognitive processes shape linguistic choices in discourse. He considered linguistic forms as *linguistic packaging*, where the forms do not have anything to do with the content of the utterance. Rather, they are *cognitive statuses* the referent has based on the assessment of the speaker about the listener's mind. He examined how speakers manage information flow by considering the listener's knowledge and

attention and distinguished between "given" and "new" information, noting that given information, assumed to be known by the listener, is often conveyed using pronouns or reduced forms, while new information requires more explicit expressions. Chafe also explored how contrastiveness and definiteness influence the selection of referring expressions, and he investigated the roles of subjects and topics in structuring discourse. Additionally, he introduced the concept of *point of view* or *empathy*, regarding how speakers may align their narrative perspective with particular participants to guide the listener's interpretation.

2.2 Status-driven Models

Building directly on the notion of givenness, status-driven models formalized how assumed familiarity and discourse prominence shape referential choice.

2.2.1 Taxonomy of assumed familiarity

One of the earliest systematic status-driven models was offered by Prince (1981). In her work, she firstly elaborated on different conceptualizations of givenness: Predictability, salience, and shared knowledge. She showed how they were problematic in analyzing discourse referents. To eliminate these issues, she proposed "Assumed Familiarity" which she claimed to lack unclear connotations other conceptions suffer from. Rather than a binary distinction or continuum, she designed the model to be a taxonomy. She considered the *text* a set of instructions given by the speaker to the hearer to build a *discourse model*. The model consists of discourse entities, attributes and links between entities. An entity is a discourse object in the sense of Karttunen's (1976) discourse referent.

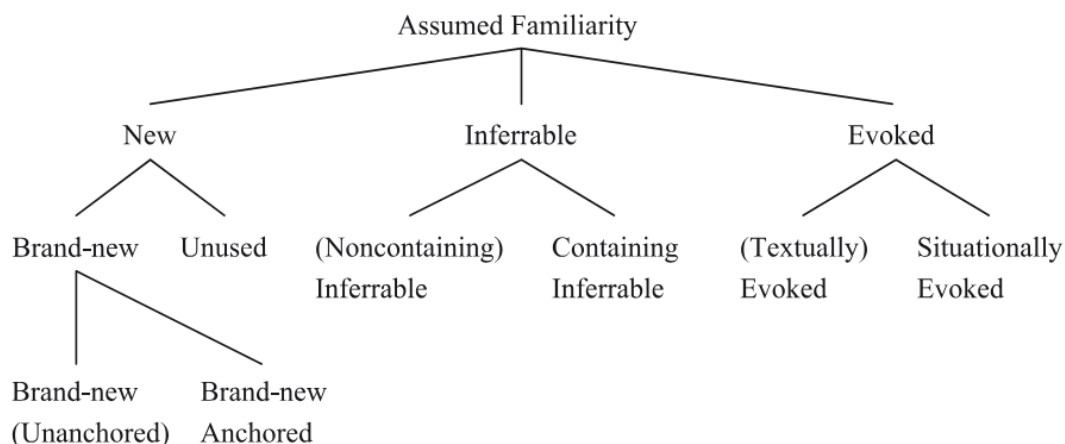


Figure 2.1: Taxonomy of the Assumed Familiarity (from Prince, 1981).

According to the taxonomy, when an entity introduced into dicourse, its considered NEW. The NEW entities are classified into two subcategories. Brand-new items must

be created anew by the hearer in the model, but unused items previously exist but are not activated. Evoked items are either anaphorically introduced items such as the pronoun *he* referring to a previously mentioned entity or situationally evoked such as the pronoun *you*. Inferrable entities can be inferred from the discourse because they have relations like part-of or subset.

In a later study, she applied the distinctions offered in the discourse model to investigate how subjects differ from non subjects in the text. She investigated the differences between subjects and nonsubjects with respect to one formal phenomenon, definiteness, and one discourse phenomenon, the information status of the entities that the subjects and nonsubjects represent. She found out that “NPs representing discourse entities” or “referring expressions” do not appear in their syntactic positions randomly. Subjects tend to be definite and old information whereas objects appear to be indefinite and new information. Li (1976) argued that the tendency of subjects to be old and definite is a categorical requirement for topics.

The particular claims she considered about the relationship between definiteness and information status on the one hand and subjecthood on the other. The canonical subjects tend to be definite to represent old information. The first question is which definition of old is related to subjecthood. The second question is whether it is possible to tease apart old information from definiteness. She stated that one can make a difference between formal definiteness vs informational definiteness for cross-linguistic and historical purposes. However, she took definiteness to be a formal category and other terms informational and tested the claim that subjects are more likely to be formally definite than nonsubjects in the ZPG letter.

Prince argued that old given and new information came to mean different things. She differentiated between three notions of ‘old new’ information. The first sense is focus-presupposition constructions (Chomsky et al., 1971), which was not found in the ZPG letter. The second sense was related to information which suggests entities/referents may be old/new with respect to the hearer’s beliefs. Hearer old entities are typically definite and hearer new entities are indefinite. While this distinction is typical, it is not perfect. Both definite and indefinite NPs could be inferable and some definites could behave like indefinites. Also, some indefinites represent hearer old entities. Therefore, instead of focusing on hearer’s beliefs, Prince suggested considering the entities from the perspective of the discourse model, where they can be discourse old or discourse new.

In addition to these binary distinctions, she added another status. Inferables are entities which are evoked when the speaker assumes that the hearer can infer their existence. She discussed the question whether it is possible to collapse inferables into one of the binary distinctions. However, she came to the conclusion that the linguistic form of the are not distinctive enough and continuum analysis is troubling on cognitive grounds. So she left inferables as a third category. She distinguished a fifth category of status called containing inferables. They are similar to inferables in that they require inferencing on the basis of certain background knowledge. The difference is that the entity which triggers is not in the prior discourse but within the NP itself.

2.2.2 Continuity in discourse

While Prince's taxonomy emphasized static familiarity categories, Givón (1983) approached referential choice as a dynamic process shaped by discourse continuity. His approach is based on the idea accessibility of topic noun phrases by focusing on the most relevant segment in a discourse, the thematic paragraph to discuss discourse continuity. There are three major aspects of continuity: Thematic continuity, Action continuity, and Topic/participants continuity. All three together play a role in bridging the gap between macro and micro organizational levels of language. However, Givon argued that there could be an implicational hierarchy between them:

(1)

Theme > Action > Topics/Participants

According to Givon, thematic continuity is the most macro-oriented category out of the three. Therefore, it is the most weakly coded in the grammar. Action continuity is primarily linked to both temporal sequentiality and temporal adjacency. Topic/participants continuity, which is encoded in the grammar in the sense of Chafe's information packaging, is the main concern of his work. Givon classified major topic functions within the thematic paragraphs. Chain initial topics are newly introduced referents. While chain medial topics are the most persistent entites, chain final topics do not last long. He also highlighted the factors which might affect the topic availability to speakers and hearers, namely length and absence from the register (discourse file), potential interference from other topics, availability of semantic information and thematic information.

Givon's approach to discourse continuity is motivated by certain observations of gestalt psychology. The general assumption he made is that what continues is more predictable and what is predictable is easier to process. And the opposite also holds. What is discontinuous is less predictable and therefore what is less predictable is difficult to process. He suggested performing the relevant measurements on texts rather than speakers or hearers because they reveal two types of information to correlate with accessibility, linguistic expressions which code topics/participants in discourse and their position. Although the syntactic coding of discourse function is not perfect, he argued it is shaped by efficiency. He offered a relative ranking of grammatical encoding with respect to accessibility:

Givon stated that the cross-linguistic analysis confirms this scale in general, but it is still too language-specific. To develop a more inclusive scale, other ways of syntactic coding such as word-order, morphology, intonation or phonological size should be considered as well.

2.2.3 Activation model

Further extending the cognitive perspective, Chafe (1987) proposed an activation-based model that captures fine-grained distinctions in accessibility. He introduced a model categorizing information into three states: *active*, *semi-active*, and *inac-*

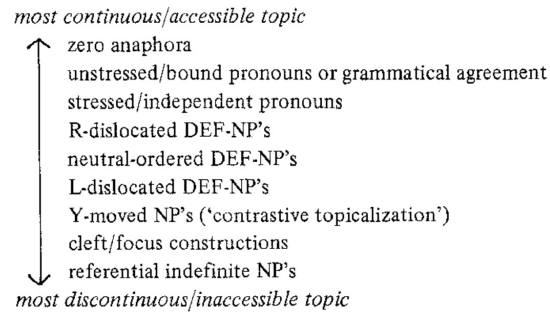


Figure 2.2: Givon's continuity hierarchy of linguistic forms

tive. *Active information* is currently in the speaker's focus of consciousness and is typically referenced using pronouns. *Semi-active information* resides in peripheral awareness and often requires full noun phrases for reference. *Inactive information* is not presently in consciousness and necessitates more elaborate expressions to introduce it into the discourse. He argued that these cognitive states influence syntactic and lexical choices, as speakers aim to present information in a manner that aligns with the listener's mental state, thereby facilitating effective communication.

2.2.4 Accessibility Scale

Moving from taxonomic classifications to graded scales, Ariel (1988) introduced the Accessibility Scale to capture the nuanced continuum of referent accessibility. For this, she first considered the work by Clark and Marshall (1981), which defined three context types to account for the use of referring expressions. They rely on mutual knowledge between a speaker and addressee. The mutual knowledge is established by *community membership*, *physical co-presence*, and *linguistic co-presence*. However, she concluded that their goal to correlate linguistic markers with the context type is not sustainable.

Then, she focused on Prince (1981)'s proposal, which is a more sophisticated approach to context type. The model assumes a grading of referring expressions as pointing to varying degrees of givenness or familiarity. However, she argued that it fails to explicitly specify categories of linguistic forms. They are just properties of discourse entities. In other words, the relationship between linguistic forms and discourse conditions is not clearly specified.

In the end, Ariel suggested that natural language provides speakers with means to code accessibility of the referent to the addressee. The three context types are ordered as to degree of accessibility to the addressee. In the unmarked case, general knowledge is not automatically accessible. Information processing from the physical surroundings is mentally represented with a higher degree of accessibility. Recent linguistic material is the most accessible source of information. Although the analogy to long term, short term and active memory systems seem obvious, Ariel claimed that current models are not rich enough to reflect the variety she argued for; therefore she suggested that context type theories of definite descriptions and psychological

accounts of anaphoric expressions should be integrated.

To provide evidence for her claim, she tested accessibility focusing on *distance*, the factor easiest to quantify. She suggested a rough three-way distinction between *high*, *mid* and *low accessibility* markers, but *accessibility* comes in a much richer variety of shade. *Low accessibility* markers are commonly uttered on the basis of general knowledge in unmarked initial retrievals (e.g. proper names and definite descriptions). *Mid accessibility* markers connect discourse to given entities from the physical surroundings. However, since most references in actual discourse are made to textual units, *mid accessibility* markers often also function anaphorically. *High accessibility* markers usually require a linguistic antecedent (e.g. pronouns, gaps).

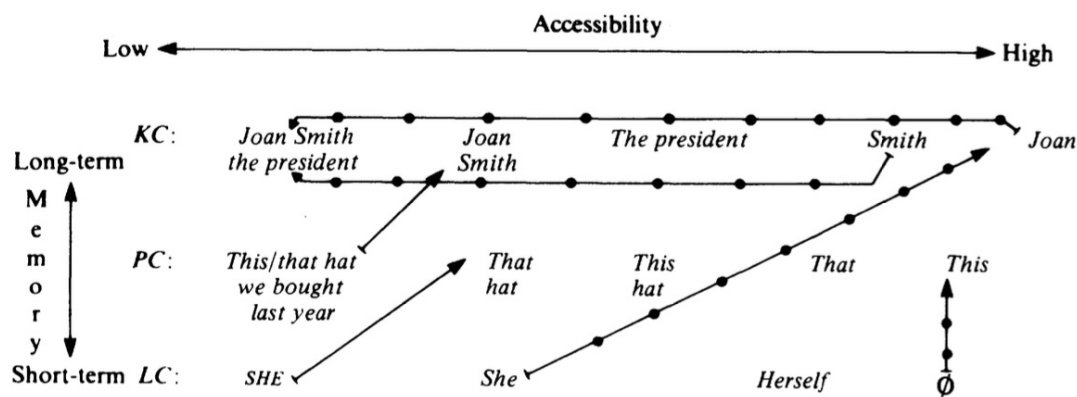


Figure 2.3: Accessibility Scale

2.2.5 Givenness Hierarchy

Complementing Ariel's work, Gundel et al. (1993) developed the Givenness Hierarchy, which links referential form to cognitive status through implicational relations. The hierarchy's main premise is that different determiners and pronominal forms conventionally signal different cognitive statuses (information about location in memory and attention state) thereby enabling the addressee to restrict the set of possible referents. Their main assumption is that the form of referring expressions like other aspects of language (word order and sentence intonation) depends on the assumed cognitive status of the referent on assumptions that a cooperative speaker can reasonably make regarding the addressee's knowledge and attention state in the particular context in which the expression is used. There are six cognitive statuses relevant to the form of referring expressions in natural language discourse:

In using a particular form, a speaker signals that the associated cognitive status is met. They thus agreed with Garrod and Sanford (1988) and Ariel (1988) that the different forms serve as processing signals to the addressee. However, while these authors viewed the statuses signaled by different forms as mutually exclusive, the statuses are implicationally related in the givenness hierarchy. Since each status entails all lower statuses, all lower statuses (statuses to the right) have been met. The statuses are thus

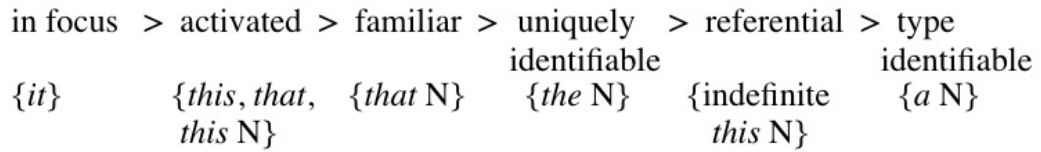


Figure 2.4: Givenness Hierarchy

ordered from the most restrictive (in focus) to the least restrictive (type identifiable) with respect to the set of possible referents they include.

In conclusion the general assumption of status based models is that there is a direct relationship between the form of a referring expression and the accessibility (or givenness or topicality) of the referent in the addressee's discourse model (e.g. (Ariel, 1990, 2001; Givón, 1983; Gundel, Hedberg, & Zacharski, 1993, 2012)). When the referent is highly accessible, the use of a short anaphoric form with limited semantic information suffices, but for inaccessible referents, a form with more semantic content is needed.

Despite their significant contributions to reference research, several conceptual and empirical points need to be considered, highlighting the complexity of modeling reference in discourse. One fundamental issue concerns how statuses are organized within models. Some proposals, such as Prince's Taxonomy of Assumed Familiarity, adopt a classification approach in which referential statuses are treated as discrete categories. Others, such as Ariel's Accessibility Scale and Gundel et al.'s Givenness Hierarchy, suggest a hierarchical or scalar organization, implying gradience or partial ordering among statuses. Whether referential statuses are best conceptualized as categorical, hierarchical, or as points along a continuum remains an open theoretical question. This distinction has important implications for how models capture the flexibility and nuance of natural discourse.

Another issue concerns the static nature of many status-based models. These frameworks typically define referential statuses at a single point in time, often overlooking the dynamic processes through which referents acquire, maintain, or lose status in discourse. For instance, a referent initially introduced as new may gradually become more accessible or familiar as discourse unfolds. Capturing these shifts is essential for developing more comprehensive models of reference. Relatedly, status-based models often differ in how they conceptualize incrementality. Some models appear to assume that referential statuses are updated on a per-utterance basis, consistent with turn-by-turn or sentence-by-sentence processing. Others suggest that referent accessibility may shift more fluidly and continuously across discourse. Clarifying whether referential status is best understood as discretely updated or as subject to continuous modulation remains an important challenge.

A central premise of status-based models is that referential status predicts the form of referring expressions. However, as noted earlier, the empirical relationship between status and form is often probabilistic rather than deterministic. Pronouns, definite de-

scriptions, and full noun phrases do not map neatly onto specific statuses, and speakers may choose forms strategically for reasons that transcend status. Thus, refining how models link cognitive or discourse status to linguistic realization is an ongoing concern.

Finally, the relationship between referential status and information structure (IS) requires careful attention. While statuses such as "given" or "new" may overlap with IS concepts like "Theme" and "Rheme," they are not synonymous. Not all given entities are thematically prominent, and not all themes are referentially given. This orthogonality complicates attempts to integrate status-based models with broader accounts of discourse structure and raises questions about how best to model the interplay between referential and informational status. These considerations point to areas for the refinement of status-based approaches. Addressing these issues is critical for developing models that not only account for referential form choice but also capture the dynamic and interactive nature of discourse in context

Despite their valuable contributions, status-driven models are not without limitations. First, various taxonomies of information status have been proposed, often differing in scope and definitions. This lack of consensus complicates comparisons across studies and raises questions about theoretical uniformity. Second, although status-based models have been used to explain the distribution of linguistic forms—such as pronominalization, nominal phrase selection, marked syntactic constructions, and prosody—these explanations are not always predictive. The relationship between form and information status is often probabilistic rather than deterministic, and speakers may deviate from expected patterns for various communicative reasons. Third, an important conceptual challenge concerns the relationship between information status and information structure. While there is a general tendency for entities referred to as Themes (i.e., discourse topics) to be given, familiar, or contextually bound, information status alone does not fully determine thematic assignment. In fact, particular information statuses are neither necessary nor sufficient conditions for Theme selection. These challenges suggest that status-based models, though insightful, are limited in their ability to account for the complex and dynamic nature of referential choice. Consequently, alternative approaches that emphasize discourse coherence and inferential processes have emerged. These coherence-based models are introduced in the following section.

2.3 Coherence Models

While status-driven models focus primarily on cognitive accessibility and assumed familiarity, coherence models emphasize the role of discourse organization and attentional dynamics in referential choices. These models propose that referring expressions are not only shaped by static cognitive statuses but also by the evolving relationships between utterances and the coherence relations that structure discourse. This section explores coherence-based approaches, beginning with Centering Theory and followed by probabilistic models that account for discourse expectations and inference.

Table 2.1: Summary of Issues in Status-Based Models

Model	Main Issues
Taxonomy of Assumed Familiarity	Lacks predictive power; does not clearly specify the relationship between linguistic forms and discourse categories; ambiguity in classifying inferables.
Continuity Model	Primarily descriptive; assumes but does not formalize cognitive mechanisms; hierarchy of continuity not universally applicable across languages.
Activation Model	Static categories (active, semi-active, inactive) may not fully capture fluidity in discourse; no account for strategic referential choices.
Accessibility Scale	Scalar approach is insightful but does not provide a deterministic link between accessibility level and referential form; variability remains unaccounted for.
Givenness Hierarchy	Assumes implicational relations but does not fully explain exceptions; overlap between statuses and forms in actual discourse use.

2.3.1 Centering Theory

A shift from static status-based models to dynamic models of discourse structure emerged with Grosz et al. (1994)'s Centering Theory. Centering theory was developed to account for attention, referring expression selection, and coherence of utterances in a discourse segment. It was part of a larger project which aimed to develop a general theory of discourse structure and meaning. It fits within the theory of discourse structure developed by Grosz and Sidner (1986). In the theory, they classified three elements of discourse structure: linguistic structure, intentional structure, and attentional state. Here, attentional state models the focus of attention of the participants. According to them, each discourse segment has local and global coherence. Local coherence is between adjacent utterances in a segment while global coherence deals with other segments in the discourse.

Centering theory focuses on local coherence and its relationship to attentional state. Grosz et al. (1994) conceptualized main center terms. Centers are entities which combines one utterance to another in a discourse segment. Therefore, they are discourse constructs and semantic objects. They distinguished between two types of centers: Forward-looking and Backward-looking centers. Each utterance in a discourse segment is assigned a set of forward-looking centers and each utterance succeeding the first utterance is assigned a single backward-looking center, $\mathbf{Cb}(\mathbf{U})$. The forward-looking centers in an utterance $\mathbf{U}$ are ranked in such a way to reflect the relative salience of the entities. The more salient an entity is in the ranking, the more likely it is to be the backward looking center in the following utterance. The backward-looking center connects with one of the forward-looking centers of the previous utterance $\mathbf{U}$. This connection can be of different types.

- $\mathbf{Cb}(\mathbf{U}_i)$: Backward-looking center of utterance $\mathbf{U}_i$; links the current utterance to the previous discourse.
- $\mathbf{Cf}(\mathbf{U}_i)$: Ordered list of forward-looking centers in $\mathbf{U}_i$; potential topics for subsequent discourse.
- $\mathbf{Cp}(\mathbf{U}_i)$: Preferred center of $\mathbf{U}_i$; highest-ranked element of $\mathbf{Cf}(\mathbf{U}_i)$.

Constraints

1. **Uniqueness:** Each utterance $\mathbf{U}_i$ has at most one $\mathbf{Cb}$.
2. **Realization:** Every element of $\mathbf{Cf}(\mathbf{U}_i)$ must be realized (i.e., mentioned) in $\mathbf{U}_i$.
3. **Continuation:** If $\mathbf{Cb}(\mathbf{U}_i)$ is defined, it must be the highest-ranked element of $\mathbf{Cf}(\mathbf{U}_{i-1})$ that is realized in $\mathbf{U}_i$.

Rules

Rule 1 (Pronoun Rule): If any element of $\mathbf{Cf}(\mathbf{U}_{i-1})$ is realized as a pronoun in $\mathbf{U}_i$, then so is $\mathbf{Cb}(\mathbf{U}_i)$.

Rule 2 (Transition Preference): Transition types are ordered by coherence preference:

Continue > Retain > Smooth Shift > Rough Shift

Transition Types

$\mathbf{Cb}(\mathbf{U}_i) = \mathbf{Cb}(\mathbf{U}_{i-1})$	$\mathbf{Cb}(\mathbf{U}_i) = \mathbf{Cp}(\mathbf{U}_i)$	Transition Type
Yes	Yes	Continue
Yes	No	Retain
No	Yes	Smooth Shift
No	No	Rough Shift

Centering theory makes some additional claims regarding referring expression selection and the nature of coherence in a discourse segment. First of all, each Un has only one backward-looking center. Secondly, forward-looking centers are partially ranked. However, the factors responsible for the order of these elements need to be determined. Thirdly, centering imposes some constraints on how entities are realized throughout the discourse segment (See Rule 1). Also, it makes certain preferences among different types of transitions (See Rule 2). The interpretation of an utterance might require the production of the following utterances. Therefore, incremental comprehension relies on partial information. Another claim is related to the locality backward-looking centers. They can only connect with forward-looking centers of the immediately preceding utterances. The final claim is that centers are determined as a result of a combination of syntactic semantic and pragmatic factors.

In a later review, M. A. Walker et al. (1998) gave an overview of these rules and the constraints and discussed the open issues in centering. They classified them into five groups. They analyzed the utterance level issues and the discussed if centering is a crosslinguistic phenomenon, looked at its relationship to information structure. While Centering Theory was initially developed for English, its influence extended across languages. Researchers subsequently examined its applicability and constraints in typologically diverse languages such as Japanese, Italian, and Turkish.

2.3.1.1 Centering theory in other languages

After its introduction, centering has been applied to several languages. One of the early attempts was in Japanese. M. Walker et al. (1996) investigated the role of centering in Japanese discourse processing to see if it can be used in machine translation systems and evaluate the syntactic factors in discourse processing. Another study was in Italian. Di Eugenio (1996) explored the correlation between centering and referring expression use in Italian, which is a pro-drop language. She showed that alternations between expressions are in line with centering transitions. In a similar vein, Turan (1995) applied centering to Turkish, another pro-drop language. The study focused on what kind of referring expressions can participate in a list of forward looking centers of an utterance and what factors are influential in determining the most salient one among them. Turan also tried to understand the functions of null and overt pronouns in discourse and proposed two additional rules, center promotion rule and discourse point of view. Strube and Hahn (1999) also dealt with the factors which determine the most salient entity and their analysis led them to propose functional information structure distinctions (i.e. hearer-old and hearer-new) should replace grammatical roles as ranking criteria.

2.3.1.2 Corpus based evaluation of Centering Theory

Theoretical proposals have been increasingly tested against empirical data. Corpus-based studies have refined Centering Theory's rules and tested its predictions about coherence and referring expression use. In an early study, Poesio et al. (2000) clarified the basic notions, rules, and constraints which were not specified enough in the theory. These include notions like previous utterance realization, and ranking. Their

aim was to specify them so that they can be reliably used during annotation and evaluation of a corpus of centering. After specifying them, Poesio et al. (2002) tested the claims of the theory in a corpus annotated for centering in English. They found out that Rule 1 are verified most of the time but Rule 2 are sensitive to how it is operationalized. Tetreault (2001) compared pronoun resolution algorithms and introduced a centering algorithm to test it if centering claims can improve pronoun resolution accuracy. Friedrich and Palmer (2014) conducted an extensive corpus analysis to examine to what extent centering theory can model discourse coherence. Although the results demonstrated a strong support for centering metrics, they showed that genre differences are influential in coherence and centering theory is unable to outperform a simple baseline.

2.3.2 The Bayesian model

While Centering Theory focuses on attentional states, Bayesian models introduced a probabilistic perspective that integrates discourse coherence with expectations about referential interpretation. Instead of reference research focusing on specific features of antecedents, Kehler et al. (2008) took a coherence-driven analysis into account. Their approach is based on Hobbs (1978), who claimed that pronoun interpretation by hearers occurs as a result of world knowledge and inference of hearers while establishing coherence relations. They explored the idea in terms of four biases, namely grammatical parallelism, thematic roles, implicit causality and subjecthood. The results provided them with evidence for the explanatory role of coherence. They found out that probabilistic expectations of hearers about coherence relations and next mentions are influential in pronoun interpretation.

In a following study, Kehler and Rohde (2013) set out to combine Centering Theory with Hobbs' (1978) theory of coherence and coreference. As mentioned earlier, Centering Theory aims to explain local coherence with attentional states and transitions between them. However, Hobbs does not include an independent mechanism for pronoun use as Centering Theory does. Pronoun interpretation occurs while establishing coherence relations. The factors influential in Centering Theory such as information structure and grammatical roles do not affect the process of pronoun use. To this end, they conducted a number of psycholinguistic experiments to explore the claims of both theories. The results revealed that pronoun use is affected by probabilistic expectations that listeners have about incoming coherence relationships and entities, in line with Hobbsian ideas. The results also indicated that topichood status of referents is influential in pronoun interpretation. Overall, the experiment suggested that pronoun use occurs as a result of the interplay between top-down and bottom up cognitive processes. Kehler and Rohde (2013) reconciled the findings with a probabilistic model which is shaped by predictions of both theories.

Rohde and Kehler (2014) strengthened this probabilistic stance of pronoun interpretation with further evidence coming from experimental work on semantic biases and topichood and evaluated the Bayesian model they offered previously (Kehler & Rohde, 2013). The model is able to predict how a hearer comprehends a pronoun with respect to next mention and production biases. The hearer basically reverse-engineers the intended referent in terms of following Bayesian principles:

$$P(\text{referent} \mid \text{pronoun}) = \frac{P(\text{pronoun} \mid \text{referent})P(\text{referent})}{\sum_{\text{referent} \in \text{referents}} P(\text{pronoun} \mid \text{referent})P(\text{referent})} \quad (2.1)$$

The equation represents the probabilistic model of pronoun interpretation, where the posterior term $P(\text{referent} \mid \text{pronoun})$ reflects the **pronoun interpretation bias**—the probability that the addressee will resolve the pronoun to a particular referent. The likelihood term $P(\text{pronoun} \mid \text{referent})$ denotes the **pronoun production bias**, indicating the likelihood that the speaker will use a pronoun to refer to an intended referent. The prior $P(\text{referent})$ represents the **next-mention bias**, the probability of a specific referent being mentioned next by the speaker, irrespective of the form of referring expression used. The model suggests that interpretation and production are not mirror processes, nor are the biases simply additive. Instead, pronoun interpretation biases arise from an integration of the listener’s *top-down predictions* (about who is likely to be mentioned next) and *bottom-up linguistic evidence* (the speaker’s choice to use a pronoun). This integration forms a more complex relationship between the production and interpretation models, highlighting the interaction between speaker intentions and listener expectations.

Recent advances in probabilistic modeling have further expanded this approach. Rational models such as the Rational Speech Act (RSA) framework build upon Bayesian principles to unify production and interpretation processes.

2.4 Rational models

Rational models formalize reference production and understanding by treating speakers and listeners as cooperative agents engaged in probabilistic inference. Within this framework, speakers select utterances that optimize communicative goals by balancing informativeness, brevity, and relevance. Listeners, on the other hand, infer intended referents by reasoning about the speaker’s likely communicative intent. The main assumption behind this is that speakers need to be both informative and efficient while uttering a certain referring expression (Arnold & Zerkle, 2019).

These constraints can be traced back to some initial works. Grice’s Cooperative Principle has been a foundational principle in pragmatics since its introduction. It posits that effective communication relies on interlocutors adhering to shared norms for successful communication. Grice formulated this principle as:

Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged (Grice, 1975).

He then operationalized this principle through four maxims: *Quantity*, *Quality*, *Relation*, and *Manner*. *The Maxim of Quantity*, in particular, requires speakers to provide information that is as informative as necessary, but not more so. This balance ensures

that contributions are neither lacking nor redundant, aligning with the cooperative nature of discourse. In line with Gricean pragmatics, rational models assume speakers as rational agents who consider the listener’s perspective, choosing utterances that optimize communicative efficiency by balancing informativeness and brevity (Degen et al., 2020).

However, Grice also highlighted that these maxims are not rigid rules but are often broken to deliver implicatures. For example, a speaker might intentionally provide less information, encouraging the listener to infer additional meaning based on context and common ground. This interplay between obeying and violating the maxims shows the nature of communication. The Cooperative Principle and its associated maxims provide a framework for understanding how speakers and listeners achieve the complexities of conversation.

Another core idea behind rational models has been Zipf’s principle of least effort (1935), which posits that language has evolved to minimize the cognitive and verbal effort required for communication. Cancho and Solé (2003) formalized this idea within an optimization framework, demonstrating that Zipf’s law emerges as a result of trade-offs between speaker and listener efforts. Pimentel et al. (2021) extended this line of work by refining communicative cost functions, showing that lexical choices reflect optimization strategies that balance predictability and complexity.

Shannon’s Information theory introduced fundamental measures such as entropy, mutual information, and conditional entropy to quantify uncertainty and information transmission (Shannon, 1948). The theory has been instrumental in understanding human communication as an optimization process, where speakers encode messages efficiently while mitigating potential noise. Piantadosi et al. (2011) demonstrated that human lexicons optimize communicative efficiency by structuring word lengths according to their expected information content, rather than simple frequency. Similarly, R. P. Levy and Jaeger (2007) showed that speakers strategically adjust syntactic constructions to maintain uniform information density, supporting the hypothesis that linguistic structure is shaped by communicative pressures. The Rational Speech Act (RSA) framework, a pioneering rational model of pragmatics, is well-grounded in Bayesian cognitive models and pragmatic theory.

2.4.1 Rational Speech Act (RSA) Theory

The Rational Speech Act (RSA) framework, introduced by Frank and Goodman (2012, 2014), formalizes language use as a probabilistic reasoning process, in which speakers select utterances to maximize communicative utility while accounting for listener expectations. The model is grounded in Bayesian inference and recursive reasoning: speakers model the beliefs of listeners and adjust their utterances accordingly to optimize informativeness while minimizing cognitive effort.

The RSA framework has been applied to a wide range of linguistic phenomena, including scalar implicature (Frank & Goodman, 2012), gradable adjective interpretation (Qing & Franke, 2015), and overinformativeness in referential expressions (Degen et al., 2020). By relaxing traditional Boolean semantics in favor of probabilistic representations, RSA accounts for variability in referential choices, such as

why speakers tend to overmodify referents in complex visual scenes (Degen et al., 2020) or why full noun phrases are used more frequently when event progressions are unpredictable (Stegemann-Philipps et al., 2021).

Orita et al. (2015) extended RSA by incorporating dynamic discourse information to model referential choices in interactive contexts. Their results demonstrated that speakers integrate prior discourse probabilities when selecting referential forms, highlighting the importance of context sensitivity in reference production. Similarly, R. Levy (2018) proposed an iterative RSA model to explain predictability-sensitive optional function word omission, demonstrating that rational principles predict observed distributional patterns in corpus data.

In referential communication, RSA provides a unifying framework for understanding why speakers choose particular referring expressions over others. The model predicts that speakers select referring expressions that maximize informativeness given the listener’s knowledge state, while also considering production costs. Degen et al. (2020) showed that RSA successfully predicts variation in referential form selection, capturing tendencies such as:

1. The increased use of full noun phrases in surprising discourse contexts.
2. The preference for overmodification when referring to objects with atypical features.
3. The balancing of economy and clarity in choosing between pronouns and noun phrases.

Further, computational implementations of RSA have been used to model real-time language comprehension and production. G. Chen et al. (2023) applied an information-theoretic RSA model to spatial deixis, demonstrating that linguistic systems evolve to optimize the trade-off between informativity and complexity. Zaslavsky et al. (2018) extended RSA to semantic domains beyond reference, showing that cross-linguistic variation in lexical structures can be explained by efficiency-driven optimization processes.

The Rational Speech Act framework provides a theoretical model for explaining referential choice in human communication. By integrating probabilistic inference, information theory, and cognitive constraints, RSA offers a comprehensive account of how speakers select referring expressions to optimize communicative efficiency. Its application to referential form selection not only aligns with empirical data but also provides a computationally precise approach to modeling language use. As research progresses, further refinements of RSA, particularly in interactive and multimodal communication settings, are likely to yield deeper insights into the fundamental principles governing referential choice in natural language.

Similar to status based models, rational models, particularly the Rational Speech Act (RSA) framework, have advanced our understanding of referring expression generation by modeling communication as recursive social reasoning. Despite its successes, RSA and related rational models face some challenges when applied to real-world referential choice, which often involves suboptimal and heuristic processes. Rational

models often struggle with scalability when applied to complex, real-world scenarios. The recursive reasoning inherent in RSA can become computationally intensive, making it challenging to apply these models to large datasets or intricate visual scenes. For instance, Le et al. (2022) note that while RSA performs well on simple datasets, its application to large-scale problems remains largely unexplored. Moreover, Rational models assume that speakers and listeners share common prior beliefs and knowledge about the context. This assumption may not hold in real-world communication, where interlocutors often have differing backgrounds and perspectives. Such discrepancies can lead to misinterpretations that such models can not predict accurately. Critics argue that RSA models may overestimate the rationality of human communicators. In practice, speakers often produce underinformative or ambiguous expressions, deviating from the optimal behavior predicted by RSA. Monroe and Potts (2015) highlight that while RSA captures certain pragmatic behaviors, it may not fully account for the heuristic and sometimes suboptimal strategies employed by humans. While rational models consider context to some extent, they may not fully capture the dynamic and multifaceted nature of real-world contexts. Factors such as discourse history, speaker intentions, and environmental cues can significantly influence referring expression generation, and incorporating these elements into RSA remains a challenge. In summary, while rational models like RSA offer valuable insights into the mechanisms of referring expression generation, their limitations in scalability, assumptions about shared knowledge, overestimation of rationality, and contextual adaptability highlight areas for further research and model refinement.

2.5 Empirical Work

Theories and models of referential selection have provided a systematic framework for linking referring expressions to cognitive statuses, shaping both theoretical and computational approaches to reference resolution. In line with this theoretical development, the early research investigating the factors behind referring expression use focused on syntactic structure. They examined the effect grammatical role in the form of the referring expression. The role of subject position has become the main consideration of many studies.

Initial empirical work focused on syntactic factors, such as grammatical role. Later studies broadened the focus to include semantic and discourse-level influences, such as recency and thematic continuity. Following the work dealing with surface syntactic roles, later work looked at the effect of semantic relations such as verb classes and thematic roles.

As more corpora have become available, researchers have examined the effect of recency on the form and use of referring expressions. The distance has been calculated in terms of the number of words or sentences between an anaphor and its antecedent. Empirical research has validated many theoretical claims about referential choice, particularly the influence of syntactic roles, thematic continuity, and recency. However, several limitations persist, which we will address in the following chapter in detail.

2.6 Coreference and Referential Selection

To operationalize and empirically investigate referential choices in discourse, coreference annotation offers a crucial methodological tool. The study of reference in natural language fundamentally concerns how speakers and writers manage the introduction, maintenance, and reactivation of discourse entities. A central aspect of this process is the phenomenon of **referential selection**, which addresses how language users choose among alternative forms such as pronouns, demonstratives, definite descriptions, proper names, and null elements when referring to entities that are either already introduced or about to be introduced into discourse. This choice is shaped by a complex interplay of grammatical, cognitive, and discourse-pragmatic factors, including referent accessibility, salience, competition, and discourse structure.

At the heart of this dynamic lies **coreference**, which can be understood as the surface linguistic realization of reference selection in discourse. Coreference occurs when multiple referring expressions point to the same discourse entity, forming what are known as *coreference chains*. These chains provide a systematic and observable record of how referential forms change over time and across discourse contexts. Crucially, each link in a coreference chain embodies a *referential selection decision*, reflecting how the speaker assesses the current discourse conditions to select a form that balances efficiency, clarity, and coherence.

Studying coreference thus offers a natural and empirically tractable ground into the analysis of referential selection. Coreference data not only makes referential choices visible and measurable but also enables researchers to investigate patterns of form selection in relation to antecedent properties, discourse structure, and listener knowledge. For example, shifts from full noun phrases to pronouns typically indicate increasing discourse accessibility or topicality, while regressions to fuller forms may signal referent reactivation or the resolution of ambiguity.

Moreover, coreference is inherently sensitive to **discourse dynamics**. As interlocutors manage focus, shift topics, and negotiate common ground, their referential choices adapt accordingly. This makes coreference chains particularly valuable for studying referential selection in naturalistic and situated contexts, where referential decisions are influenced by factors beyond sentence-level syntax or semantics.

In sum, coreference offers both a *theoretical* and *methodological pathway* for advancing our understanding of referential selection. By analyzing how speakers and writers select referential forms as they build and maintain coreference chains in discourse, we can gain deeper insights into the cognitive, grammatical, and interactional principles that shape reference in language use. For this reason, coreference occupies a central position in the present study, serving as the empirical foundation for investigating reference selection in Turkish dialogues.

2.7 Coreference and anaphora

Coreference annotation and corpus research have attracted significant attention among Natural language processing (NLP) researchers, cognitive scientists, and linguists, as

understanding referring expressions and the relations between them is fundamental to natural language understanding. Numerous NLP tasks, including information retrieval, question answering, and summarization, require coreference resolution for effective performance. This need has resulted in an increase in the number of corpora annotated for coreference relations in recent decades, particularly with the success of data-driven techniques, especially for widely-studied languages like English (Uryupina et al., 2020; Weischedel et al., 2013; Zeldes, 2017) and German (Bourgonje & Stede, 2020; Lapshinova-Koltunski & Ferreira, 2022).

However, the majority of languages still remain low-resourced in this respect. Turkish, a member of the Turkic language family, is among these low-resourced languages, facing a scarcity of coreference-annotated datasets (Çöltekin et al., 2023). The available annotation schemes, predominantly designed for major languages, fall short when applied to morphologically rich and pro-drop languages like Turkish. Such languages exhibit complex inflectional morphemes and allow reduced or null forms when the referents are pragmatically inferable or morphologically cued by agreement.

In this connection, adapting existing annotation schemes to Turkish poses numerous challenges and it is particularly challenging to offer a universal scheme for all languages when the complexity of the anaphoric phenomena is taken into consideration as stated by Poesio (2004). For instance, the treatment of morphological information, such as suffixes that carry referential information, is often overlooked. Similarly, the handling of phonologically null elements, which are pervasive in Turkish, is not sufficiently addressed. This inadequacy can lead to a loss of critical information necessary for accurate coreference resolution. As a result, there is a need for developing a specialized annotation scheme that can accommodate the unique features of Turkish and similar languages, ensuring more robust and reliable coreference annotation. A deeper understanding of referential phenomena requires distinguishing between related concepts such as reference, anaphora, and coreference. The following subsections outline this conceptual landscape and review how annotation schemes have evolved to capture these distinctions.

2.7.1 Basic terminology

Coreference can be better understood within the larger picture of cohesion and concepts related to it. Cohesion itself is based on the idea that the spoken or written communication is usually a united whole rather than unrelated utterances or sentences. For cohesion to occur, the interpretation of some linguistic element in the discourse sometimes depends on previously mentioned items in the text (Halliday & Hasan, 1976). A closely related notion to cohesion is reference. It is the relationship between a linguistic expression and an entity in the world. There are two main types of reference. Exophoric reference refers to an entity which is outside the text. On the contrary, endophoric reference refers to another expression in the preceding discourse segment. Endophora is further divided into two types. Anaphora can be described as an item which relates back to a previous item in some way. The element which is referring back is called anaphor and the previously mentioned entity which then anaphor refers to or is related to is its antecedent. The process of linking the anaphor with its antecedent is called anaphora resolution. Cataphora, on the other

hand, points to an item in the following discourse segment.

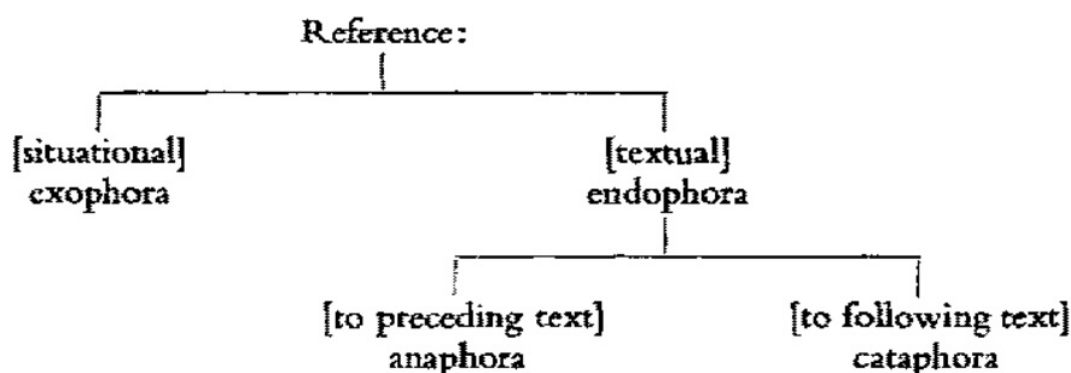


Figure 2.5: Reference types (Halliday and Hasan, 1976, p.33)

There are a variety of anaphora which are observed in written or oral language based on the form of the anaphor (Mitkov, 2014). Lexical noun phrase anaphora could appear as proper names and definite descriptions. Pronominal anaphora is one of the most studied and therefore understood type of anaphora in the literature. Anaphors in this type can be in the form of personal pronouns, possessive pronouns, reflexive pronouns, and demonstrative pronouns. Another type of anaphora is zero anaphora. It is considered to be one of the most challenging types of anaphora to resolve since they are not physically realized at the surface level. Although they are invisible, they do not damage the cohesion of the discourse but strengthen it. They are decoded by the reader or hearer without any loss during the comprehension of the discourse. If the anaphor and antecedent refer to the same entity, they are thought to be coreferential. This relation is also called identity anaphora, as in 2.

- (2) A man came. **He** brought a book.

An anaphor can be preceded by a number of expressions referring to the same entity and therefore they are said to form a coreference chain. Such theoretical work on reference and anaphora has become the foundation of the guidelines which have been prepared to create coreference corpora.

2.7.2 Related work

Efforts to annotate coreference have led to the development of various schemes tailored to specific languages and tasks. Here, we summarize the major initiatives, highlighting their guiding principles and limitations. The earliest attempts to develop annotation schemes for coreference annotation could be traced back to the Message Understanding Conference (MUC) information extraction tasks (Hirschman et al., 1997). The task was created to group all the mentions of an entity together and the scheme specified the basic task criteria, the markables to be annotated and the relations to be established in English. The task evolved with Automatic Content Ex-

traction (ACE) Program (Doddington et al., 2004) enriching the coverage with entity, relation, and event annotation in English, Chinese and Arabic. The MATE/GNOME proposals (Poesio, 2004) were geared towards being more linguistically oriented than previous schemes, making a discourse model assumption. It also included bridging anaphora in addition to identity relations.

The PoCoS-Potsdam Coreference Scheme (Krasavina & Chiarcos, 2007) claimed to adopt language independent principles during markable annotation. The scheme applied to German, English and Russian. The OntoNotes guidelines (Weischedel et al., 2011) includes several layers, one of which is coreference layer. It aimed to include all coreferential relations and specifically focuses on how to handle identity relations and appositives. Like the ACE scheme, it was applied to English, Arabic and Chinese. The later schemes have become more comprehensive, including different kinds of anaphora in addition to coreference and more fine grained subcategories like ARRAU (Uryupina et al., 2020).

However, some guidelines took a more psychological approach and considered coreference as part of information structure annotation. Nissim et al. (2004) developed a scheme to annotate coreference and information status relations in English dialogs. Götze et al. (2007) prepared guidelines for information status, topic, and focus annotation. They aimed for language independence, theory neutrality, reliable marking, and framed coreference under information status annotation in terms of givenness. The RefLex scheme (Riester & Baumann, 2017) was developed for referential and lexical analysis of spoken and written text. Coreferentiality has been at the referential level along with bridging relations in the scheme.

The development of annotation schemes have paved the way for the construction of many corpora in different languages. The initial products were naturally produced in the languages of the schemes mentioned above. One of the well-known and largest coreference corpora is the OntoNotes project (Weischedel et al., 2013). It consists of various genres such as news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows. It was annotated for syntax and predicate argument structure and word sense and coreference in English, Chinese, and Arabic. ARRAU (Uryupina et al., 2020) is another multi-genre corpus which contains around 350K tokens. Unlike many corpora, it accepts nonreferential NPs and singletons as markable. It was annotated for different kinds of anaphoric relations including coreference, bridging anaphora and discourse deixis.

Similar to OntoNotes, AnCora (Taulé et al., 2008) is also a multilingual corpus. It consists of 500k tokens of newspaper texts in Spanish and Catalan. The texts were annotated for morphological information, syntactic phrases, grammatical functions relations. ParCorFull (Lapshinova-Koltunski & Ferreira, 2022) is a parallel corpus of English and German with a total of 160K tokens. It was only annotated for coreference relations. The growing interest and need in coreference datasets triggered corpus development in other languages such as Czech (Nedoluzhko et al., 2016), Hungarian (Vincze et al., 2018), Polish (Ogrodniczuk et al., 2016), and Dutch (Hendrickx et al., 2008).

Some researchers have explored referring expressions as a means to enhance the naturalness and efficiency of interactions. As part of these efforts, several corpora focused

on referring expressions used in situated dialogs have been developed. For example, Di Eugenio et al. (2000) compiled the COCONUT corpus, which features referring expressions produced during a 2D interior design task. Similarly, the QUAKE corpus (Byron & Fosler-Lussier, 2006) and the SCARE corpus (Stoia et al., 2008) were created from expressions used in collaborative 3D tasks. In the same vein, Spanger et al. (2009) developed a collaborative problem-solving task to construct the REX-J corpus, which contains Japanese referring expressions (Spanger et al., 2012).

However, to the best of our knowledge, the only coreference corpus developed in Turkish was Marmara Turkish Coreference Corpus (Schüller et al., 2017). It is an annotation layer on top of the METU-Sabancı Treebank, which consists of 33 documents from various genres with 53925 tokens in total. The scheme prepared for corpus includes noun phrases, pronouns, and nominalized adjectives as markables, but it does not consider the role of morphological information and null elements in Turkish. The gold data obtained from several annotators resulted in 5170 mentions and 944 coreference chains. Arslan and Eryiğit (2023) reannotated the corpus to handle the dropped pronouns with the data representation scheme they proposed. However, their scheme only deals with how to represent third person singular agreement makers and possessive pronouns for dropped pronouns.

Due to this limited availability of Turkish coreference data, the computational work on coreference in Turkish is also rather limited and mostly have exploited rule-based and classical machine learning methods. Yıldırım et al. (2004) developed a rule-based system for anaphora resolution in Turkish. Their model depends on the theoretical framework of the Centering Theory. In a later study, Tüfekçi and Kılıçaslan (2007) presented a computational model for resolving pronominal anaphora. It is based on Hobbs' naïve algorithm (Hobbs, 1978), which traverses a parse tree to find the antecedent of a pronominal anaphora. The first learning-based approach to anaphora resolution is limited to pronoun resolution (Yıldırım & Kılıçaslan, 2006). They trained a decision tree on a corpus of popular child stories. Pamay and Eryiğit (2018) proposed the first coreference resolution system, which uses support vector machines with a mention-pair model. There are recent attempts to use deep learning methods for Turkish coreference resolution. Demir (2023) presented the first neural coreference resolution system and Arslan et al. (2023) introduced a neural multilingual coreference resolution model which makes use of morphological information. However, they remain limited due to data sparsity.

This chapter provides a comprehensive overview of theoretical and computational models that seek to explain how speakers select referring expressions in discourse. Beginning with status-driven models, the discussion describes the development of foundational frameworks grounded in concepts such as givenness, accessibility, and assumed familiarity. Models such as Prince's Taxonomy, Givón's Continuity model, Chafe's Activation model, Ariel's Accessibility Scale, and Gundel et al.'s Givenness Hierarchy have all sought to capture how cognitive and discourse factors influence the accessibility of referents and shape the forms speakers choose to refer to them. While these models offer valuable descriptive insights and have established important links between referential form and discourse status, they also face notable limitations. Among these are the challenges of capturing referential dynamics, accounting for probabilistic variation, and integrating information structure and reference more systematically. In response to these limitations, coherence-based models have shifted

the focus from static representations of status to dynamic processes governing discourse reference management. Centering Theory introduced key notions such as local coherence, attentional state, and transition types to model how discourse entities are maintained and shifted across utterances. Building on this, probabilistic models, including Bayesian approaches and Rational Speech Act (RSA) frameworks, have further emphasized the role of inference, expectations, and communicative rationality in referential choice. These models have demonstrated that speakers and listeners engage in complex reasoning about discourse coherence, next-mention biases, and informativeness when producing and interpreting referring expressions. Finally, this chapter highlights the critical role of empirical research in validating and refining theoretical claims. Corpus-based studies have not only tested the predictions of earlier models but also underscored the importance of factors such as grammatical role, recency, discourse structure, and cognitive accessibility in shaping referential practices. In this regard, coreference annotation has emerged as a particularly valuable tool, providing systematic and observable data for analyzing referential choices in natural language. Overall, while the models and approaches reviewed in this chapter differ in focus and methodology, they converge on the recognition that referential selection is a multi-factorial phenomenon shaped by cognitive, discourse, and communicative pressures. Together, they offer a rich foundation for establishing and advancing reference research, while also pointing to the need for integrative analysis of the features that can account for the complex and context-sensitive nature of referential choice in situated discourse. The subsequent chapters build on this foundation to address these challenges through the development of a novel annotation scheme, the construction of a comprehensive coreference corpus, feature engineering, statistical and machine learning analyses.

CHAPTER 3

MOTIVATION

This chapter presents the motivation of the present study, focusing on critical gaps and limitations in previous research on reference selection. Although significant theoretical and empirical advances have been made in understanding how speakers choose referential forms, much of this work has been shaped by a restricted set of linguistic contexts, expression types, and methodologies. In particular, prior studies have predominantly examined local context in languages with rigid word order and obligatory overt arguments, often in highly controlled experimental settings. Such approaches, while valuable, offer only a partial view of the complex and dynamic processes involved in referential choice in natural discourse. This chapter first outlines the limitations associated with these earlier accounts, addressing issues such as the neglect of discourse-level analysis, the narrow focus on certain referring expressions, and the underrepresentation of typologically diverse languages. It then introduces the rationale for the current research, which aims to overcome these challenges by investigating reference selection in Turkish situated dialogues through a comprehensive empirical approach. By doing so, this study seeks to contribute to a more ecologically valid, crosslinguistically generalizable, and theoretically sound understanding of how speakers select referring expressions in real-world communicative contexts.

3.1 Limitations of previous accounts of reference selection

Reference selection, the process by which speakers choose among competing forms to refer to discourse entities, has long been a topic of linguistic, cognitive, and psycholinguistic inquiry. Despite substantial progress in research on reference selection, existing theoretical models and empirical studies remain subject to important limitations that constrain the generalizability and explanatory power of their findings. Much of this work has been shaped by data and theoretical assumptions rooted in a narrow set of linguistic and discourse contexts. Specifically, research has often been restricted to *local context*, which focuses on sentence-internal or immediate context relations at the expense of discourse-level phenomena. In addition, studies have tended to focus disproportionately on *a limited range of referring expressions*, primarily third person pronouns, definite and indefinite descriptions, demonstratives, and proper names. Finally, a pervasive bias toward languages such as English and German has led to the relative neglect of pro-drop, morphologically rich, and flexible word order languages, whose referential systems pose serious challenges for theory-building.

3.1.1 Local Context in Reference Selection Research

A major limitation of existing work on referring expressions lies in its predominant focus on intra-sentential or immediate context analysis. While the investigation of referential choices and resolutions within the boundaries of individual sentences or preceding sentences offer methodological advantages, including the ability to control variables like sentence structure, syntax, and immediate linguistic context, they also introduce significant constraints that limit the generalizability and ecological validity of findings.

By concentrating on single sentences or local context, much of this research neglects the fact that natural language is inherently *discourse-based* sometimes far exceeding the limits of immediate or neighboring sentences. Referential expressions routinely function across sentence boundaries, connecting ideas and entities over extended stretches of text or dialogue. Limiting analysis to local contexts prevents from fully capturing how reference operates in larger discourse structures, where coherence and cohesion are maintained through complex and often ambiguous cross-sentential relationships.

Moreover, local context approaches tend to overlook how speakers and listeners manage and interpret *long-distance dependencies*. In spontaneous and natural discourse, pronouns, elliptical constructions, and deictic expressions frequently refer back to antecedents introduced in earlier sentences or even earlier parts of a conversation or text. These forms rely heavily on shared knowledge, discourse context, and pragmatic inferencing, dimensions that are not adequately addressed when analysis is restricted to the sentence level.

Beyond linguistic considerations, this limitation has implications for cognitive models of language processing. Understanding referential expressions requires integrating information across discourse units, a process that reflects how humans comprehend language in real time. Models that focus narrowly on intra-sentential reference risk misrepresenting the nature of language processing by ignoring how individuals track and retrieve referents across sentences to maintain discourse continuity.

This shortcoming is also evident in computational applications. Natural language processing (NLP) systems trained or evaluated primarily on such data may struggle with tasks requiring discourse-level understanding, such as document summarization, narrative comprehension, or dialogue modeling. Reference chains, sequences of referring expressions that point to the same discourse entity across multiple sentences, are critical for constructing coherent representations of text or speech. Intra-sentential or local analyses often fail to account for how such chains are formed, maintained, and interpreted.

Furthermore, local context research tends to underplay the role of pragmatic and thematic factors, such as speaker intentions, presuppositions, and topic management, which are inherently discourse-level phenomena. These aspects often guide referential choices and resolutions beyond what is immediately available within sentence boundaries.

While local context driven studies have yielded important insights, particularly con-

cerning syntactic and semantic constraints on reference, they provide only a partial view of how referring expressions function in natural language use. To develop more comprehensive models of reference selection, future research must systematically incorporate inter-sentential and discourse-level analyses. Only by doing so can we achieve a fuller understanding of how referential expressions contribute to discourse coherence and facilitate efficient communication. This narrow analytical scope is compounded by another issue: the tendency to focus on only a limited range of referring expressions.

3.1.2 Focusing on a Narrow Range of Referring Expressions

A further limitation in research on reference selection is the tendency to focus on a narrow subset of referring expressions, often treating them in isolation rather than as part of a dynamic system of options available to speakers. While this approach has generated valuable insights, particularly regarding pronoun resolution, it constrains our understanding of how different referential forms interact and compete in natural language use. A comprehensive model of reference selection must account for the full spectrum of referential expressions and the principles governing their use.

Among all forms, third person singular pronouns (e.g., *he*, *she*, *it*) have been the most intensively studied, particularly in psycholinguistic and computational work. Their high frequency, combined with their inherent ambiguity — lacking descriptive content and relying heavily on discourse context — make them ideal candidates for exploring context-sensitive reference resolution. In theoretical linguistics, they play a central role in syntactic models such as Binding Theory (Chomsky, 1981), which seeks to explain how pronouns find their antecedents within specific syntactic domains. However, this focus comes with limitations. Concentrating heavily on pronouns may obscure the broader cognitive and discourse-pragmatic factors involved in referential choice. Unlike pronouns, other referring expressions encode additional information (e.g., spatial cues, specificity, uniqueness) and serve different communicative functions. Thus, insights derived exclusively from pronoun studies may not generalize to the full range of referring expressions.

Demonstratives (e.g., *this*, *that*) constitute another important yet relatively under-explored class of referring expressions. Theoretical accounts such as Accessibility Theory (Ariel, 1990) highlight their sensitivity to the cognitive status of referents, suggesting that demonstratives are used for entities that are less accessible than those referred to with pronouns but more accessible than those requiring full descriptions. Despite their crucial role in deictic reference, discourse organization, and signaling speaker perspective, demonstratives have received limited attention compared to pronouns. This gap restricts our ability to develop models that reflect the range of strategies speakers use to manage discourse and signal referent accessibility.

Definite descriptions (e.g., *the man*) and indefinite descriptions (e.g., *a woman*) play vital roles in introducing and maintaining referents. These forms differ from pronouns and demonstratives in that they carry rich semantic information about referent identity, uniqueness, and presuppositions. For instance, definites presuppose familiarity or uniqueness, whereas indefinites introduce new discourse entities. Although these dis-

tinctions are well established in formal semantics and pragmatics (e.g., Heim, 1982; Kamp and Reyle, 1993), research on reference selection has often marginalized them in favor of simpler, more easily controllable pronoun-based studies. As a result, models tend to under-represent the ways in which referents are introduced, reintroduced, and maintained using descriptive forms.

Proper names (e.g., *John*, *Paris*) represent yet another dimension often sidelined in reference selection research. While names lack the descriptive content of definite descriptions, they carry unique identification and are often treated as directly referential in philosophical semantics (e.g., S. Kripke, 1980). In discourse, however, names interact with other referential forms, contributing to topicality, emphasis, and narrative structure. Proper names have received limited empirical attention in reference selection studies, despite their relevance for modeling referential competition and understanding the full repertoire of speakers' choices in discourse.

Focusing narrowly on individual referring expressions risks fragmenting our understanding of reference and obscuring the interactive and context-dependent nature of referential choice. In actual language use, speakers continuously weigh options among pronouns, demonstratives, descriptions, and names based on factors such as accessibility, salience, discourse function, and audience design (Arnold, 2010; Gundel et al., 1993). Only by systematically studying these forms together can we develop models that capture how referential forms compete and cooperate to achieve communicative goals.

3.1.3 Controlled Experimental Settings in Reference Selection Research

A significant limitation in much of the research on reference selection stems from the reliance on highly controlled, laboratory-based experiments, particularly within the field of experimental pragmatics. While such methodologies offer the advantage of isolating variables and ensuring replicability, they often lack ecological validity; that is, they may fail to capture the complexities of language use as it naturally occurs in everyday interaction.

In controlled environments, tasks are typically simplified, contexts are highly structured, and participants may engage in artificial communicative scenarios. This can lead to language production and comprehension processes that differ substantially from those employed in spontaneous, naturalistic discourse. For instance, speaker-hearer dynamics, which involve subtle coordination, negotiation of meaning, and context-sensitive reference resolution, are often flattened or absent in laboratory tasks. As a result, findings derived from these experiments may not generalize well to real-world communicative situations.

Moreover, natural interaction frequently involves multimodal cues, overlapping speech, informal constructions, and situationally grounded references, ungrammatical and incomplete utterances, elements that are typically absent in controlled settings. This discrepancy poses a challenge for developing comprehensive models of reference selection that aim to account for the full range of phenomena observed in natural language use. Without incorporating data from more ecologically valid contexts — such as dialogues, narratives, or naturally occurring conversations — our understanding of

how referring expressions are produced, interpreted, and negotiated remains incomplete. Critically, the neglect of naturalistic data is further amplified by a bias towards studying only certain languages.

3.1.4 Neglecting Pro-Drop and Flexible Word Order Languages

A further critical limitation in the literature on reference selection arises from the linguistic bias present in much of the foundational and empirical research. To date, most models and experimental studies have been developed using languages such as English and German, which share typological characteristics including relatively rigid word order, limited inflectional morphology, and the obligatory expression of subjects. While these languages offer certain methodological conveniences—such as clearer syntactic boundaries and fewer reduced forms—their properties do not generalize easily to typologically distinct languages.

This poses a particular problem when these models are applied to pro-drop languages with rich morphology and flexible word order, such as Turkish, Japanese, or Italian. In these languages, referential choices are shaped by additional and often more subtle mechanisms that are under-represented or entirely absent in the languages on which most models are based.

For example, Turkish exhibits extensive use of null pronouns (zero anaphora), especially in subject position, where referents can be omitted when they are pragmatically inferable or morphologically marked on the verb through subject-verb agreement. Such patterns challenge models based on languages where overt pronouns are obligatory and therefore treated as primary carriers of referential information. Moreover, Turkish allows considerable variation in word order, governed not by syntactic necessity but by discourse-pragmatic factors such as topicality, focus, and information structure. Existing models, which are primarily designed around subject-verb-object (SVO) languages with relatively fixed syntax, struggle to account for how referents are tracked and retrieved when their surface positions vary dynamically.

In addition, the rich inflectional morphology characteristic of Turkish encodes referential cues through case marking, possessive suffixes, and agreement features. These morphological signals can significantly reduce the need for overt referring expressions and can influence referential choice in ways not adequately captured by models derived from morphologically impoverished languages.

The reliance on models based on English and German, therefore, results in important theoretical and empirical gaps. These frameworks often fail to account for how referential forms are selected, interpreted, and resolved in languages where discourse-pragmatic factors and morphological marking play a more central role. Without incorporating data from pro-drop, morphologically rich, and free word order languages, reference selection models remain limited in cross-linguistic validity and explanatory power.

These limitations highlight the need for more inclusive and ecologically valid approaches to the study of reference. Addressing them is essential for developing models that can adequately capture the diversity and complexity of referential behavior

across different languages and discourse contexts.

In sum, research on reference selection has been shaped by theoretical and empirical choices that, while methodologically convenient, have also imposed important limitations on the scope of current models. The predominant focus on intra-sentential and local context analysis has sidelined critical discourse-level processes, while the narrow concentration on certain referential forms has overlooked the rich variability of referring expressions used in natural language. Furthermore, the emphasis on languages with rigid syntax and obligatory argument expression (e.g. subject, object, or adjunct) has left unresolved the challenges posed by pro-drop, morphologically rich, and free word order languages, where reference is often managed through subtle and interactionally grounded means.

Future work must aim to overcome these limitations by embracing more discourse-sensitive, typologically informed, and form-diverse approaches to reference selection. Expanding empirical coverage across different languages and communicative settings will not only improve the descriptive adequacy of reference models, but also enhance their cognitive and computational relevance. Only through such integrative efforts can the field advance toward a comprehensive and cross-linguistically robust account of how speakers choose referring expressions in real-world communication.

In designing our research on reference selection, we made two key decisions which laid the foundation of our study: using situated dialogues as the primary source of empirical data and focusing on coreference as the phenomenon of interest. These choices were motivated both by theoretical and methodological considerations, namely ecological validity and linguistic diversity, regarding the nature of reference selection.

3.2 Situated Dialogues as Ecologically Valid Source of Data

Reference selection is inherently a situated and dynamic phenomenon, shaped not only by the structure of language but also by the communicative intentions and shared knowledge of interlocutors. While much previous research has analyzed reference in monologic or decontextualized text types, there are compelling theoretical and empirical reasons to argue that situated dialogues provide the most ecologically valid and informative context for studying how speakers select referring expressions.

First and foremost, reference selection is fundamentally interactional. In dialogue, referring expressions are not chosen in isolation but in response to interlocutors' assumed knowledge, current attention, and prior discourse moves. Speakers constantly monitor and adapt to their addressees' informational states and expectations, adjusting referential forms accordingly (Clark & Brennan, 1991). This means that the choice between full noun phrases, pronouns, demonstratives, or ellipsis is shaped by an ongoing process of audience design and collaborative grounding, which only dialogues naturally instantiate.

Secondly, situated dialogues are uniquely characterized by immediacy and shared context. Unlike narratives or written expository texts, dialogues often unfold in contexts where referents are mutually perceptible, jointly attended, or physically co-present. This creates a rich environment for the use of reduced referring expressions,

such as pronouns, demonstratives, and zero anaphora. The reliance on shared situational context allows interlocutors to omit or minimize referential expressions when referents are pragmatically inferable (Enfield, 2003). Thus, studying dialogues captures referential practices that are otherwise difficult to observe in decontextualized text genres.

In addition, dialogues exhibit a high degree of incrementality and discourse dynamism. Turn-taking, repair, clarification requests, and topic shifts introduce complex referential challenges that compel speakers to manage referential coherence locally and adaptively. Referring expressions in dialogue are sensitive to recency, competition among potential referents, and the salience shifts that arise as topics emerge and recede in conversational flow (Grosz & Sidner, 1986). This incremental and emergent structure makes dialogues an ideal site for examining how referential form choices track discourse accessibility and topicality in real time.

Furthermore, from a cognitive and psycholinguistic perspective, situated dialogue offers a more naturalistic approach to the processing of reference. In interactive contexts, listeners rely heavily on discourse cues, prosody, gaze, and co-speech gestures to resolve referring expressions (Goodwin, 1981). These multimodal and context-sensitive mechanisms, crucial for reference resolution, are poorly captured in written or monologic texts. Studying reference in situated dialogues therefore enables researchers to better align models of referential choice with how language is processed and understood in everyday communication.

Finally, situated dialogues have crucial relevance for computational modeling and NLP applications. Contemporary dialogue systems, conversational agents, and human-robot interaction frameworks require models that can handle the complexities of real-time reference in interaction. Training and evaluating such models demands data that reflects the true nature of reference selection as it occurs in dialogue. Without studying situated interactions, computational approaches risk remaining overly simplistic or inapplicable to real-world communicative scenarios.

In summary, situated dialogues are not only a natural environment for referential practices but also a necessary one for building empirically grounded, cognitively plausible, and computationally robust models of reference selection. By capturing the interplay of discourse dynamics, shared context, and interlocutor coordination, situated dialogues offer unparalleled insight into how speakers choose referring expressions in communicative action.

3.2.1 Coreference as a Natural Domain for Reference Selection Research

The study of reference selection seeks to explain how speakers choose among alternative forms—such as pronouns, demonstratives, definite descriptions, and proper names—when referring to discourse entities. While this decision is influenced by a range of factors including salience, grammatical role, and discourse structure, it is within the phenomenon of coreference that these choices are most systematically observable. Consequently, coreference provides a crucial empirical and theoretical domain for investigating the mechanisms of reference selection.

Coreference phenomena offer a direct window into how speakers maintain and reintroduce referents across discourse. Each time a referent recurs, speakers must select the most appropriate form, balancing between reduced forms (e.g., pronouns or null forms) when referents are highly accessible, and fuller expressions (e.g., definite descriptions or proper names) when referents are less accessible, ambiguous, or reintroduced after gaps (Ariel, 1990; Gundel et al., 1993). This selection process reflects fundamental principles of discourse organization and accessibility, where referential form choice encodes cognitive assumptions about the hearer’s mental representation of discourse entities.

In particular, coreference chains—sequences of expressions referring to the same discourse entity—provide natural units for analyzing patterns of referential form choice over time. Shifts in referential forms within these chains reveal how speakers modulate expression based on factors such as topicality, salience, and the need to disambiguate. For example, Centering Theory (Grosz et al., 1995) suggests that reduced forms such as pronouns are preferred when referents maintain a central discourse status, whereas more explicit forms are employed when discourse focus shifts or referent reactivation is necessary.

From a cognitive perspective, coreference is tightly linked to memory and processing mechanisms involved in language comprehension. Listeners must resolve referring expressions incrementally and integrate them into mental discourse representations. Psycholinguistic research has shown that reduced forms like pronouns are processed more efficiently when referents are highly accessible, whereas fuller forms impose greater processing demands (Arnold, 2010). Thus, analyzing coreference allows researchers to explore how referential form selection aligns with cognitive constraints and processing dynamics.

Moreover, from a computational standpoint, coreference resolution is one of the most challenging tasks in natural language processing (NLP). Developing systems capable of accurately resolving coreference requires understanding when speakers use reduced versus explicit forms, and what discourse or contextual cues guide these choices (Poesio et al., 2018). As such, insights from reference selection research directly inform computational models, while annotated coreference data provides a critical resource for training and evaluating algorithms tasked with referential interpretation.

In sum, coreference sits at the intersection of linguistic form, discourse coherence, cognitive processing, and computational modeling. It is a phenomenon where reference selection decisions are repeatedly and systematically enacted, offering a unique perspective on the principles that govern natural language use. For these reasons, coreference must be regarded as central to any comprehensive and cross-disciplinary account of reference selection.

3.2.2 Coreference Annotation Schemes Developed for Typologically Similar Languages

An important limitation in current research infrastructure for reference and coreference resolution lies in the design of coreference annotation schemes, many of which

have been developed with typologically similar European languages such as English, German, and Dutch. These languages share core characteristics, including relatively rigid word order, limited inflectional morphology, and obligatory expression of subjects. As a result, the annotation frameworks built for these languages often reflect assumptions and categories suited to their particular referential systems, while failing to adequately accommodate the properties of pro-drop, morphologically rich, and free word order languages.

In languages like Turkish, Japanese, Italian, and Spanish, referential expressions are shaped by interaction among morphological, syntactic, and pragmatic factors. For instance, the phenomenon of null anaphora (or zero pronouns) is prevalent in these languages, particularly in subject positions where referents can be omitted when they are inferable from the discourse context or recoverable via verbal agreement morphology. However, many existing annotation schemes designed for languages where overt subject expression is obligatory tend to lack explicit guidelines or categories for marking null references. This often results in either inconsistent annotation practices or the exclusion of such forms altogether, leading to an incomplete representation of referential relations.

Additionally, in morphologically rich languages, information about case, number, person, and possessiveness is frequently encoded on nominal expressions and verbs through inflectional morphemes. These morphological cues can play a crucial role in reference resolution, sometimes reducing the need for overt referring expressions. However, annotation schemes oriented towards morphologically impoverished languages often give little attention to how these markers contribute to reference tracking, thus overlooking essential features of referential interpretation in these languages.

The issue is further complicated in languages with free or flexible word order, such as Turkish and Russian, where constituent positioning is influenced more by discourse-pragmatic factors (e.g., topicality, focus) than by rigid syntactic rules. In these cases, referential salience and prominence may be signaled through position shifts rather than through explicit lexical or pronominal forms. Annotation schemes designed with fixed word order languages in mind are often ill-equipped to capture these discourse-level cues, potentially leading to inadequate or misleading representations of referential structure.

These design biases limit not only the descriptive adequacy of annotation efforts, but also the performance of computational models trained on such resources. Models developed using datasets annotated according to schemes optimized for European languages may generalize poorly when applied to typologically diverse languages. As a result, the field risks perpetuating a narrow and language-specific view of reference and coreference phenomena, undermining the goal of building universal or cross-linguistically robust models of discourse processing.

While OntoNotes introduced some coverage of dropped subjects (e.g., in Chinese), the treatment remains shallow and biased towards languages with obligatory subjects and clearer surface realization of referents. ARRAU offers more detailed annotation of reference phenomena, including bridging and discourse-old/new status, but is still largely designed with English discourse properties in mind and does not fully address morphologically-driven reference or free word order effects. As this tables illus-

Feature / Phenomenon	Typical Annotation Schemes	Requirements
Null Anaphora / Zero Pronouns	Limited or inconsistent treatment; often excluded or only partially annotated (mainly subjects).	Systematic annotation of null references in all syntactic positions, including subjects, objects, and adjuncts.
Morphological Cues (person, number, case agreement)	Morphological features generally not integrated into coreference annotations.	Explicit annotation of morphological agreement and case marking as referential signals.
Word Order Variability and Discourse Functions	Fixed word order assumptions; discourse-prominent positions often not annotated.	Capture discourse functions and pragmatic roles associated with word order variations (e.g., topicality, focus).
Definite/Indefinite Noun Phrases and Referentiality	Definite NPs included; indefinite and generic references often excluded.	Richer categorization of referential status, including indefinite and generic expressions.
Anaphoric vs. Non-Anaphoric Expressions	Focus on anaphoric relations; non-anaphoric salient mentions often ignored.	Inclusion of discourse entities such as cataphora, bridging, and inferable referents.
Ellipsis and Discourse-Omitted Elements	Often ignored or inconsistently annotated.	Systematic treatment of elliptical constructions and omitted arguments recoverable from context.
Proper Names, Pronouns, and Definite Descriptions	Detailed annotation and analysis.	Continued coverage integrated with reduced/null forms and morphology-sensitive resolution mechanisms.

Table 3.1: Coverage of Typical Annotation Schemes(e.g., OntoNotes, ARRAU) vs. Needs for Pro-Drop, Morphologically Rich, Free Word Order Languages

trates, existing coreference annotation schemes provide coverage for languages with rigid syntax and explicit referential expressions. However, they fall short in capturing the full range of referential strategies employed in pro-drop, morphologically rich, and free word order languages. For truly cross-linguistic and typologically sensitive models of reference, annotation schemes must be adapted to systematically account for null anaphora, morphological cues, discourse-driven word order variation, and implicit referentiality.

To address these shortcomings, future annotation frameworks must be expanded and adapted to account for features such as null referents, morphological reference cues, and discourse-based word order variation. Doing so is essential to achieving more inclusive and theoretically grounded models of reference that reflect the full range of strategies available across the world’s languages.

3.3 Limited Coreference Datasets and Computational Models for Most World Languages

One of the most pressing challenges in advancing the study of reference and coreference in typologically diverse languages is the limited availability of annotated datasets. To date, the majority of coreference corpora have been developed for languages with relatively rigid word order, limited inflectional morphology, and obligatory subject expression, such as English, German, and Dutch. In contrast, languages that are pro-drop, morphologically rich, and characterized by flexible word order, such as Turkish, Japanese, Italian, and Korean, remain significantly underrepresented in available resources.

The absence of large-scale, systematically annotated corpora for these languages presents a major limitation for both empirical research and computational modeling. Coreference resolution systems rely heavily on annotated data to learn patterns of referential expression and resolution. In the absence of such data, it becomes exceedingly difficult to train, evaluate, or even design models that can handle the unique properties of these languages. Existing models, largely developed and optimized for English and similar languages, are often ill-suited for transfer or direct application to typologically distinct languages without substantial adaptation.

This gap in resources has several important consequences. First, it restricts our empirical understanding of how reference operates in languages where null anaphora is common, where referential information is often encoded morphologically rather than lexically, and where word order variation plays a crucial role in signaling discourse functions like topicality and focus. Without annotated datasets that capture these phenomena, theoretical models risk remaining speculative or biased towards features observable in better-resourced languages.

Second, the lack of datasets severely limits the development of computational models that can accurately process and interpret reference in naturalistic settings. State-of-the-art coreference resolution systems, including neural and transformer-based approaches, require substantial amounts of annotated data to achieve acceptable levels of performance. In the case of pro-drop and morphologically complex languages,

model architectures may also need to incorporate language-specific features, such as morphological parsers or discourse-level features. However, without benchmark datasets, progress in designing and validating such models remains stalled.

Third, this scarcity of resources impedes cross-linguistic research and typologically informed theories of reference. The current bias towards European languages not only limits the ecological validity of computational models but also hinders our ability to test hypotheses about the universality or language-specificity of referential strategies. A more inclusive set of annotated corpora would facilitate comparative studies, enable the training of multilingual and typologically-aware models, and contribute to the development of genuinely cross-linguistic theories of reference.

Addressing this limitation requires dedicated efforts to construct high-quality coreference datasets for pro-drop, morphologically rich, and free word order languages. Such initiatives must pay special attention to phenomena often overlooked in existing schemes, including zero anaphora, elliptical constructions, discourse-driven word order variation, and morphological cues for reference resolution. Without these resources, the field will continue to rely on models and theories that inadequately reflect the diversity and complexity of reference across the world’s languages.

3.4 Addressing the Limitations: Reference Selection in Turkish Dialogs

To overcome the major limitations identified in prior research on reference selection, this thesis undertook a systematic and multi-layered approach grounded in the investigation of Turkish, a pro-drop, morphologically rich, and free word order language. By focusing on this typologically distinct linguistic environment, our research directly addresses critical gaps related to the restriction to intra-sentential or local context analysis, the focus on a limited number of referring expression forms, and the general neglect of languages with complex discourse and morphological patterns.

3.4.1 Developing a Novel Annotation Scheme and Constructing a Coreference Corpus

The first step in our research was the development of a novel coreference annotation scheme, tailored specifically to capture the distinctive referential phenomena found in Turkish dialogues. Existing annotation frameworks, typically designed for languages such as English and German, proved insufficient for addressing key issues such as:

- Zero anaphora (null references), which are frequent in Turkish due to its pro-drop nature and are often recoverable via verbal morphology or discourse context.
- Morphological reference cues, including subject-verb agreement, case marking, and possessive suffixes, which play a crucial role in signaling referents.
- Word order variability, which reflects discourse-pragmatic functions such as topicalization and focus, and influences referent accessibility and salience.

Our annotation scheme systematically incorporated these elements, providing guidelines for annotators to identify and annotate all referring expressions, overt and covert, and to link them appropriately to their antecedents across sentence and turn boundaries.

Using this scheme, we constructed a coreference corpus of situated dialogues, which forms the empirical basis of the study. The dialogs were collected from a naturalistic experimental setting where interlocutors jointly solved tasks, ensuring that referential choices were shaped by interactional demands and shared situational context. This corpus includes:

- A wide range of referential forms, including full noun phrases, overt pronouns, demonstratives, null forms, and proper names.
- Dialogic structure, including turn-taking, topic shifts, and grounding activities.
- Rich morphosyntactic and discourse-pragmatic annotation layers.

The annotation scheme development and corpus construction procedures are described in detail in the Corpus chapter.

3.4.2 Feature Engineering

Building on the annotated corpus, we then derived a set of theoretically and empirically motivated features to analyze the factors influencing reference selection. These features were informed by prior research in discourse coherence, information structure, accessibility theories, and coreference resolution. They cover a wide range of linguistic and discourse dimensions:

- Discourse and topicality features, such as the distance from the antecedent, mention density, and the recency of referents.
- Syntactic and grammatical features, such as the grammatical role of the antecedent (subject, object) and word order position.
- Competition features, capturing the number of competing entities in the discourse at the point of reference.

Feature engineering was carried out carefully to ensure compatibility with both statistical analysis and machine learning models, converting categorical variables as necessary and normalizing continuous measures.

3.4.3 Systematic Analysis of Features

To investigate the individual and combined effects of these features on referential form choice, we conducted a multi-step quantitative analysis. First, we applied descriptive statistics to examine the overall distribution of referential forms and feature

values. This initial step revealed important patterns in the frequency and contextual use of different referring expressions. Next, we employed correlation analyses to identify relationships between features and to assess multicollinearity, which informed subsequent modeling decisions. We used Chi-square tests to investigate associations between categorical features and referential form choice, and Kruskal-Wallis tests to evaluate differences in distributions of continuous features across referential forms, especially when normality assumptions were not met. These tests allowed us to establish which features showed statistically significant variation by referential form.

To complement and extend the statistical analysis, we adopted a data-driven machine learning approach, using XGBoost, a powerful and interpretable tree-based model, to predict referential form based on the engineered features. The XGBoost model allowed us to capture non-linear interactions between features, evaluate the predictive power of the feature set, assess feature importance, identifying which variables contributed most strongly to model predictions. Feature importance analysis revealed which discourse, syntactic, and morphological factors played primary roles in referential form selection. This analysis was crucial in validating theoretical claims and uncovering new insights into the interplay of linguistic and discourse features in situated language use with spontaneous naturalistic speech.

3.4.4 Advancing Reference Selection Research

Through the methodological innovations and empirical analyses presented in this study, we offer several key contributions that advance the field of reference selection research, particularly for typologically underrepresented languages such as Turkish.

First and foremost, we proposed a **novel annotation scheme for coreference in Turkish**, explicitly designed to capture the language’s distinctive referential characteristics. Unlike existing schemes, our framework systematically addresses challenges posed by *zero anaphora*, *morphological reference markers*, and *discourse-driven word order variation*. This enables a more accurate and theoretically grounded representation of referential relations in Turkish.

Building on this scheme, we have created **the first and most comprehensive coreference corpus of situated dialog** currently available for Turkish. The corpus, annotated according to our scheme, provides a rich empirical resource not only for analyzing referential practices but also for future work in natural language processing, discourse analysis, and cognitive modeling.

A further innovation of this study is the focus on **situated dialogues** as the primary source of data. By analyzing naturally occurring interactions rather than monologic or written texts (e.g. news, stories, papers), our research captures reference selection as it occurs in *spontaneous speech production*, shaped by real-time discourse dynamics, speaker-hearer coordination, and shared situational context.

Moreover, our study provides **comprehensive empirical coverage of referring expressions**. Rather than limiting our analysis to a narrow set of forms, we systematically include the full range of referring expressions, such as *full noun phrases*,

overt pronouns, demonstratives, null forms, and proper names. This broader scope enables us to investigate referential choice as a dynamic process sensitive to multiple interacting factors.

In addition, we derived a set of **core features informed by theoretical and empirical research** on reference selection. These features reflect dimensions such as discourse accessibility, morphological marking, grammatical role, and competition, providing a multi-layered perspective on referential choice.

Finally, we advance the field methodologically by **systematically testing the effects and interactions of these features**. Employing a combination of descriptive statistics, correlation analysis, Chi-square and Kruskal-Wallis tests, and machine learning models (XGBoost), we provide an integrated analysis that reveals not only which factors shape referential choice, but also how they interact in situated discourse.

Summary of Contributions

- Proposing a novel annotation scheme for coreference in Turkish, tailored to handle zero anaphora, morphological cues, and word order variation.
- Creating the most comprehensive coreference corpus for Turkish, based on naturally occurring situated dialogues.
- Using situated dialog data to ensure natural and spontaneous speech production is reflected in the analysis.
- Including the full range of referring expressions (full NPs, proper names, pronouns, demonstratives, null forms) for comprehensive analysis.
- Deriving theoretically and empirically motivated features that capture the complexity of reference selection.
- Testing the effect and interaction of these features systematically through statistical analyses and machine learning models (XGBoost).

Through these contributions, we aim to provide a typologically sensitive, methodologically rigorous, and ecologically valid account of reference selection. Our work thus advances the empirical foundations and theoretical understanding of reference selection, while offering a framework applicable to other typologically diverse and under-represented languages.



CHAPTER 4

CORPUS

This chapter describes the development of the coreference corpus that forms the empirical basis of this study. The chapter begins by presenting the rationale behind genre selection, followed by a detailed explanation of the experimental setup and the characteristics of the collected data. It then describes the annotation scheme developed to capture coreference phenomena in Turkish situated dialogs, including both overt and null referring expressions. Finally, the chapter reports corpus statistics and inter-annotator agreement, demonstrating the reliability and scope of the corpus. Overall the chapter illustrates the contribution of the corpus to the study of reference in typologically diverse and ecologically valid contexts.

4.1 Corpus creation

The creation of a reliably annotated coreference corpus constitutes a critical foundation for the empirical study of referential phenomena in discourse. In this thesis, the goal was to build a corpus specifically designed to capture how Turkish speakers select among alternative referring expressions in situated, task-oriented dialog. While much of the existing research on coreference has relied on monologic and often written text genres (e.g., news, narrative, and articles), dialog presents distinct opportunities and challenges. It introduces features such as turn-taking, spontaneous speech phenomena (e.g., disfluencies, repairs, ungrammaticalities), and shared task contexts that shape referential choices in ways not typically observed in written language. Moreover, Turkish, as a morphologically rich, pro-drop language with flexible word order, presents unique patterns of referential omission and realization that necessitate data-driven investigation. In this section, we describe the methodology used to create the corpus, including the selection of dialog data, text preparation, annotation procedures, and the steps taken to ensure the accuracy and consistency of the annotations. This corpus forms the empirical basis for all subsequent analyses reported in this thesis.

4.1.1 Genre selection

We selected situated dialogs as the genre for our corpus. Most existing coreference corpora started with texts like news texts, and continued with articles, and stories (Uryupina et al., 2020). However, in this study, we chose to annotate situated dialogs

involving spontaneous speech. The language in this genre exhibit certain features. The utterances/sentences are relatively short compared to the genres like news and articles and therefore syntactically less complicated. Speakers might often produce ungrammatical forms and disfluencies and add fillers, which is associated with cognitive load and planning.

Our decision to focus on situated dialogs as our texts was motivated by several reasons. First, situated dialogs provide rich contextual information, including the nature of the task, the setting, discourse participants, entities in the physical context, and shared knowledge among participants. Additionally, situated dialogs possess the spontaneity and complexity of natural language interaction absent in experimental stimuli or controlled experiments. This makes them valuable for testing cognitive and linguistic theories and hypotheses. Moreover, they offer diverse linguistic structures since they are produced in a situated context. Analyzing them can help investigate how these forms evolve throughout the text.

4.1.2 The source of our texts

Our dialogs were taken from an experimental study where pairs were expected to solve tangram puzzles (Mançe-Çalışır, 2018). The study was designed to investigate the social cognitive skills of individuals diagnosed with Autism Spectrum Disorder. The task setting included a task where participants with specific roles were asked to solve a tangram puzzle where they were to construct a target shape using geometric shapes called tans. The pairs were exposed to the problem through a tangram simulator similar to the one used by Spanger et al. (2009) and Tokunaga et al. (2010). The simulator displays two distinct areas: a part for the target shape and another part for the working space, as in Figure 3.2. The working space enables the operator to move, flip, and rotate the pieces shown on the computer screen.

4.1.2.1 Experimental setup

Apparatus:

- Eyetrice Eye Trackers: (2 pieces 60 Hz)
- Chin rest
- 17 inch monitor (2 pcs)
- Laptop (4 pcs)
- Router
- Mouse (2 pieces)
- Keyboard (2 pieces)

Participants: Forty-seven participants were recruited for the experiment. While 23 of them have been diagnosed with ASD diagnosed according to DSM-V criteria, 24

of the participants were adults typically developed non-clinical controls (NCC). The participants coming in pairs perform the first four tangrams with their relatives or friends and the last four tangrams with the experimenter. The participants without partners performs only the last 4 tangrams with the experimenter. Eye movements and speech of each participant and their partners during the task were recorded together with an eye movement tracking device.

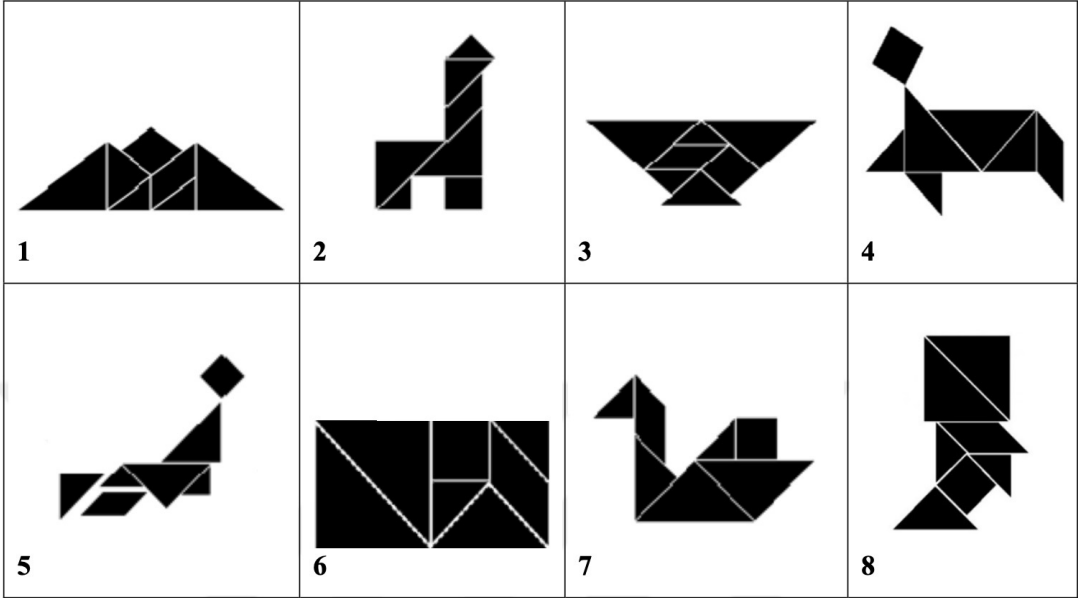


Figure 4.1: Target Shapes

Task: The participants were asked to solve tangram puzzles together. The task requires them to build a target shape by manipulating seven geometric shapes through a computer simulator. They are seated face to face and perform the tasks through shared screens. The separator between the tables prevents them from seeing each other. They are assigned specific roles, which aims to promote real-life language production. The presenter has access to the target shape and is expected to give instructions to the other participant about how to build the shape and the operator cannot see the shape but has control over the mouse to manipulate the geometric shapes to achieve the goal. Before the experiment, the participants went through a training session where they asked to achieve a target shape. Once they understood their roles, the task, and how to manipulate the shapes with the mouse, the experiment started. Each pair solved eight target shapes with similar difficulty levels. The geometric shapes are given in distinct colors so that they can use colors to refer to these objects.

4.1.3 Text preparation

For corpus data, we included only the sessions conducted with adults typically developed non-clinical controls. This amounts to 60 tangram solving tasks. We firstly transcribed the speech between the participants in the form of dialogs, indicating the roles of the pairs (ie. presenter and operator). Then, we manually split them into

Table 4.1: Anchors for the entities

Entities	Anchors
Presenter	presenter
Operator	operator
Pink triangle	pinktr
Green triangle	greentr
Yellow triangle	yellowtr
Red triangle	redtr
Blue triangle	bluetr
Black square	blacksq
Grey parallelogram	greyp

sentences and added punctuation where necessary. We added anchors for the entities available in the physical context at the beginning of each dialog. These are discourse participants and seven geometric including shapes two small triangles, one middle triangle, and two big triangles, small square, small parallelogram (see Table 4.1). We then encoded the dialogs in JSON format.

4.1.4 Tool choice

As a result of our evaluation of various annotation tools, we decided to use Labelbox (2024). It offers inherent templates for conversational texts, relatively easy annotation, and most importantly allows morpheme and character selection to capture morphological and null elements. It was also the most suitable tool to work with dyadic dialog data.

4.1.5 Training the annotators

We hired three graduate students, who are native speakers of Turkish with the necessary linguistic background for annotation. We conducted training sessions with materials that were not included in the corpus to familiarize them with the coreference task and test our annotation scheme. We detected the challenging issues and specified how to handle them.

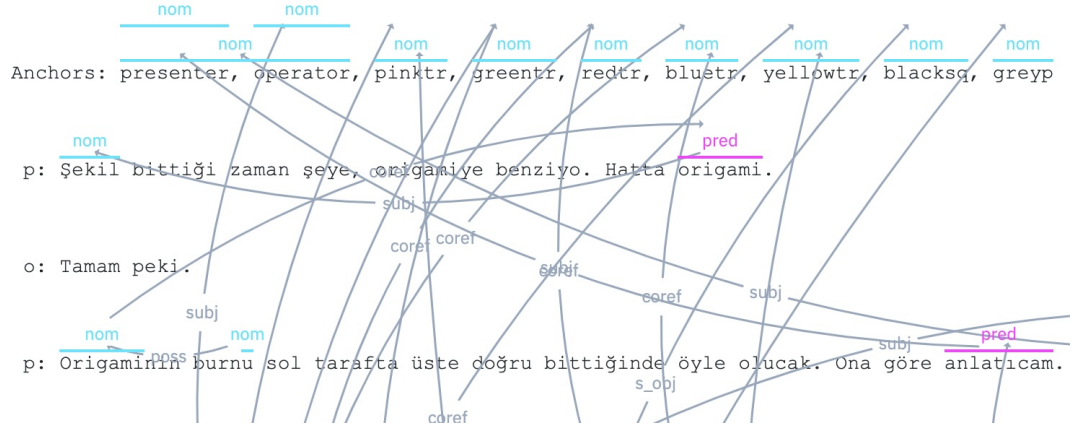


Figure 4.2: Annotation sample from the annotation tool

4.2 Annotation scheme development

We evaluated available schemes reviewed in the second chapter and found out that they were lacking the devices that are required for the annotation of sub-word morphological units and null elements. Therefore, we developed a scheme which is comprehensive enough to handle all realizations of mentions at different levels of Turkish structure. In this way, the scheme could be a model for morphologically rich low-resourced languages which frequently utilize null elements, especially Turkic languages.

4.2.1 Scope

Our annotation scheme mainly focuses on how to annotate co-referring expressions. Our definition of coreference follows van Deemter and Kibble (2000): “ α_1 and α_2 corefer if and only if $\text{Referent}(\alpha_1) = \text{Referent}(\alpha_2)$ ”

However, certain anaphoric relations fall beyond the scope of this work. Discourse deixis (Webber, 1988) can be associated with coreference but discourse deictic expressions may refer to preceding or succeeding discourse segments, such as clauses or sentences. Given that the antecedents in these cases are non-nominal expressions (Çallı, 2012), we do not include them in our dataset. Additionally, there exists another type of relationship between an anaphor and its antecedent beyond identity relations. Bridging anaphora (Clark, 1975), also known as associative anaphora (Hawkins, 2015), requires the hearer or reader to establish an indirect connection between the anaphor and its antecedent, drawing on their world knowledge. In sentence 1, the cover functions as the anaphor and *a book* as its antecedent. The reader infers the relationship because it is commonly understood that covers are parts of books.

- (1) The man brought a book. **The cover** has a nice illustration.

This brings us to the central aspects of the present study:

- referentiality
- strict coreference.

The present work is limited to the annotation of referential noun phrases. The operational test we employ for referentiality is case-marking. In this regard, we ignore all the nominal expressions that come in non-case-marked positions (see below for examples and exceptions).

We limited ourselves to strict coreference between referents, leaving out looser linking relations like discourse deixis and bridging anaphora.

Our scheme also involves annotating the grammatical roles of the mentions with the embedding level (matrix or subordinate) of their occurrences.

4.2.2 Markables

Our scheme restricts the class of mentions which are to be annotated as referential noun phrases and their manifestations as agreement markers on predicates, possessive suffixes, and null elements. We correlate referentiality of a referring expression with case-marking (Öztürk, 2004). Although it is problematic at times especially at the conceptual level, it provides a strong basis for decision-making during annotation. The other condition is that the noun phrase should refer to another expression with an identity relation either anaphorically or cataphorically. Therefore, a mention qualifies as a markable only if it is case-marked and part of a coreferential chain. We annotate the full span of the overt entities due to maximal projection principle, which is established in most schemes. Here are the types of Turkish markables included in the present work:

Overt nominals:

- (2) **Bir kitap** okuyorum.
a book read.1sg
'I am reading **a book**.' (Indefinite NP)
- (3) **Kitap** okuyorum.
book read.1sg
'I am reading.' (Bare noun)
- (4) **Kitabı** okuyorum.
book.Acc read.1sg
'I am reading **the book**.' (Definite NP)
- (5) Adamın okuduğu kitap
man.Gen read.Rel.Agr book
'The book that the man read' (modified NPs)

- (6) Proper names, pronouns and demonstratives and demonstrative NPs.

Null nominals:

- (7) Kitap geldi. $\emptyset$ Eskiymi.
book.Nom came **it** old.Past.3sg
'The book came. **It** was old.' (Subject drop)
- (8) Elma vardı. Ali $\emptyset$ yedi.
apple.Nom existed Ali **it** eat.Past.3sg
'There was an apple. Ali ate **it**.' (Object drop)
- (9) $\emptyset$ ev-i güzel.
his/her house-Poss beautiful.Cop
'**His/her** house is beautiful.' (Possessor drop)
- (10) Ben $\emptyset$ **okurken** uyudum. (Converbs)
I I/she read.Conv slept
'I fell asleep, while I/she was reading.'

4.2.3 Non-markables

We did not include and therefore annotate the following categories:

Singletons: We left out mentions that occur only once throughout the text and therefore do not take part in a coreferential chain. Zhu et al. (2023) showed that incorporating singleton information along with entity type and information status could help coreference models generalize better. However, since our main focus was on the choice of specific referring expression form, including such mentions could noise in our data, let alone

Predicatives: Predicatives are usually complements of a linking verb or a copula and state a property of the subject. Some schemes accept them as markables such as the Gum corpus (Zeldes, 2017), but we do not annotate them because they are not discourse entities themselves but properties so they cannot pass our referentiality criterion.

- (11) Ali **öğretmen** oldu.
Ali **teacher** become.Past.
'Ali became **teacher**.'

Abstract Entities: We left out reference to abstract objects like propositions, state-of-affairs, and other sort of such entities discussed by Asher, 1993, as their inclusion immensely complicates the annotation task when handled along with conceptually simpler type of referents we aimed to capture in the present study. (See Zeyrek et al. (2010) for abstract object annotation in Turkish Discourse Bank (Zeyrek et al., 2013)).

Local adverbial and verbal demonstratives: We annotated only the nominal type out of the three major types of demonstratives. There are three types of demonstratives (Dixon, 2003), leaving local adverbial and verbal demonstratives out, because they constitute either a reference to an abstract entity or are not referential.

(12) **Buraya** gel.
here come
 'Come **here**.'

(13) **Böyle** yap.
like this do
 'Do **like this**.'

4.2.4 Various issues in coreference annotation

The annotation process revealed a number of issues and challenges that, in our opinion, might be of help for researchers planning to build similar corpora for languages like Turkish.

4.2.4.1 Embedded mentions

One issue that complicated the annotation process was the annotation of embedded mentions. As a principle stated above, we annotate the whole noun phrase but sometimes the phrase can consist of other mentions. For instance in a form like in 14, the markable *the book* is embedded in *the man who brought the book*. We annotated the embedded mention along with the larger one, in cases where there is a reference back to the embedded markable in the text.

(14) [[Kitabı]_{M2} getiren adam]_{M1} gitti.
 book-Acc bring.Rel man.Nom left
 'The man who brought the book left.'

A similar issue arose with coordinated noun phrases where there are separate references to both the entire NP and individually to its components.

4.2.4.2 Mentions in compounds

Turkish noun-noun compounds do not involve case marking on the dependent noun. Although this Such phrases pose a problem to our referentiality criterion. They do not have a case marking but still could be an entity in the real world. We apply the criterion here we use for the embedded mentions and annotate them if there is a reference to that item.

- (15) Mevlana türbesi
Rumi tomb
'The tomb of Rumi.'

4.2.4.3 Appositives

We include the appositives like *Istanbul, Turkey's most crowded city* in the markable of the nominal expression they attach to. Our rationale for doing this is the possible significance of this modification type for modelling efforts of coreference phenomena which might be conducted on the corpus in the future (See also Weischedel et al. (2011) and Hirschman et al. (1997) for discussion of appositives).

4.2.4.4 Genitive-possessive constructions

Turkish makes extensive use of genitive-possessive agreement both on a type and a token basis. There are 3 major constructions that depend on the agreement of a genitive marked noun phrase and a possessive marked head: Genitive-possessive NPs, object relative clauses, and subordinate clauses. The genitive marked possessor can be dropped in all these constructions. Therefore, it is imperative for a coreference corpus to systematically handle these constructions. In this regard, we annotated all the possessive suffixes as markables and linked them to their null and overt possessors.

4.2.4.5 Grammatical coreference

We left out all coreference relations that are governed by syntax rather than discourse, such as control structures, Turkish versions of *want*-type constructions, reflexive binding, and so on. Our aim here was to simplify the annotation process, as the mentioned dependencies can be automatically discovered by accurate syntactic parsing in the future.

4.2.4.6 Split anaphora

In split anaphora, which is a rather rare case of anaphora slightly more complex than standard anaphora (Yu et al., 2021), the antecedent of the anaphor can be the addition of previous discourse entities, which is also called aggregation. These cases are included in our dataset.

- (16) Ali Ayşe'yi bekliyor. **Onlar** birlikte gelecek.
'Ali is waiting for Ayşe. **They** will come together.'

4.2.4.7 Null elements

Turkish is a pro drop language, where zero pronouns are abundant in both spoken and written text, and get involved in coreference chains. In cases where a null markable has an overt morphosyntactic agreement like a verbal inflection or a possessive suffix, we annotated the corresponding suffix in lieu of the markable itself. However, when it comes to null objects there is no overt agreement correlate. It is still a matter of discussion how to treat such cases in annotation. For instance, Pradhan et al. (2012) inserted a small *pro* into the place the null element is omitted, but the detection of the correct place is also problematic on its own. We employed a convention of marking the space character just before the head predicate to represent a dropped object. The information concerning both types of null anaphora is recovered during post-processing, abstracting away from the conventions we employed in the annotation process.

4.2.5 Annotation procedure

Our scheme basically requires detecting a mention, assigning a grammatical role to it, and establishing a link with its antecedent. Although it might look complicated, we clearly defined the steps which our annotators need to follow.

1. Identify the markable.
2. Check whether it is a referential phrase or not. Case marking is an important indicator here.
3. If referential, continue.
4. Check whether it is a singleton or not.
5. If not singleton, continue.
6. Check whether it is realized at the subordinate or matrix clause level.
7. Assign its grammatical role and syntactic level accordingly.
8. Connect the markable with its closest antecedent.

4.3 Corpus statistics

Each text in the corpus has been independently annotated by two annotators. They identified the mentions in the texts and established the identity relations between them. This provided us with the unique entities and their realizations in the texts, in other words, their mentions. They labeled these mentions with the grammatical information with the categories subject, object, and other.

We built a custom tool in Python that (i) exports the annotated texts from our annotation tool, LabelBox, (ii) compared the annotations and calculated inter-coder agreement, (iii) extracted a graph representation of the coreference patterns of the dialogs, (iv) performed basic statistics, and (v) encoded the corpus in JSON format.

4.3.1 Inter-Annotator Agreement

Coreference annotation has been traditionally associated with two subtasks (reference?). *Mention annotation* involves detecting the mentions and their boundaries and *relation annotation* requires creating a link between an anaphor and its antecedent. Our annotation workflow also involves detecting mentions and establishing relations. Cohen’s κ (Cohen, 1960) and Krippendorff’s α (Krippendorff, 1970) are two widely used coefficients to measure inter-annotator agreement reliability in NLP annotation tasks (Artstein & Poesio, 2008). Cohen’s κ has been developed to measure inter-annotator reliability between two annotators for nominal data taking chance factor into account. Fleiss κ (Fleiss, 1971) is an extension which can measure the agreement between two or more coders. Similarly, Krippendorff’s α can measure the agreement between two or more coders, but can be applied to different metrics (eg. nominal, ordinal, interval, and etc.).

However, these coefficients are not the best candidates for coreference annotation because mentions and relations are not fixed and the negative cases are unknown (Deleger et al., 2012). Under such circumstances, it has been shown that the agreement between annotators can be measured with standard measures like precision, recall, and F-score (Brants, 2000; Hripcsak & Rothschild, 2005). We took one of the annotations to be predictions and the other one to be our gold standard to calculate F1 score to measure the agreement between our annotators for each text using the formulae below. We adopted the basic metrics introduced in Sang and De Meulder (2003) and implemented a strict evaluation based on the exact matches between both mentions and relations.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (4.1)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (4.2)$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.3)$$

Our annotators achieved high precision, recall and F1 scores (0.96) for mentions and (0.90) for relations on average, which is quite satisfactory for coreference annotation task (See Table 4.2 for interannotator agreement scores).

Table 4.2: Interannotator agreement scores

	Precision	Recall	F1
Mentions	0.96	0.96	0.96
Relations	0.90	0.90	0.90

4.3.2 Descriptive statistics of the corpus

We annotated 60 situated dialogues of participants solving a puzzle. Our dialogues have an average of 306 tokens. The dialogue with minimum number of words has 127 words and the one with maximum number of words has 961 (See Table 4.3 for average number of words in our dialogues).

Table 4.3: Counts of the corpus

	tokens	mentions	entities
mean	306	102	12
min	127	43	7
max	961	280	19
std	167	48	2.5
total	18360	6120	720

Our analysis indicated that there is an average of 12.3 entities and 102.4 mentions per dialogue, which means that each entity is mentioned approximately 8.2 times on average throughout a dialog.

We also looked at the grammatical functions of the mentions. We found out that 50.5% of the mentions occupy a subject position in a sentence. 49.5% of the mentions occupy an object position or part of a genitive possessive construction. (See Figure 4.3 for the percentage of grammatical roles of mentions)

We also analyzed the form of our referring expressions. We observed a relatively close distribution of null and overt form in our mentions. The percentage of mentions which have overt forms is 57.3% while the percentage of null forms is 42.7% (See Figure 4.4 for the percentage of forms of the mentions).

We aligned the grammatical function of the referring expressions along with their forms to see if the grammatical function has a relation with the form. When we looked at the mentions which occupy the subject position, we observed that 61.91% of the expressions have null forms. However, when we looked at the non-subject positions including objects and all other positions, our analysis showed that only 23.34% of referring expressions have null forms (See Figure 4.5 for null forms in subject and non-subject positions). Consequently, our data indicated that there can be a strong

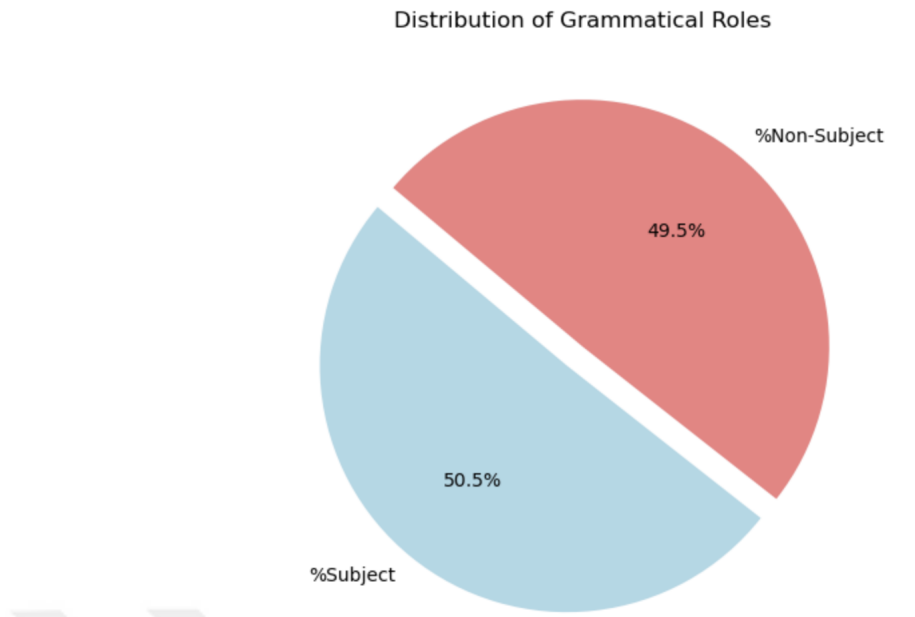


Figure 4.3: Grammatical Roles

relationship between subjecthood and linguistic form of the mentions.

In conclusion, we proposed a novel comprehensive annotation scheme about how to annotate coreferential phenomena including both overt and null mentions in a morphologically rich and pro drop Turkic language. The high inter-annotator agreement showed that our scheme can be reliably applied to languages similar to Turkish in the relevant respects. We created a new corpus of situated dialogs manually annotated for mentions and coreference relations. Our work has made novel contributions in a number of ways. Our dataset comprises 60 conversational texts. To our knowledge, it has been the first dialog corpus, which has been annotated for mentions and coreference relations in Turkish. Another significant contribution is that it includes null elements, agreement markers, and possessive suffixes as realizations of entities in text in addition to overt noun phrases and pronouns.

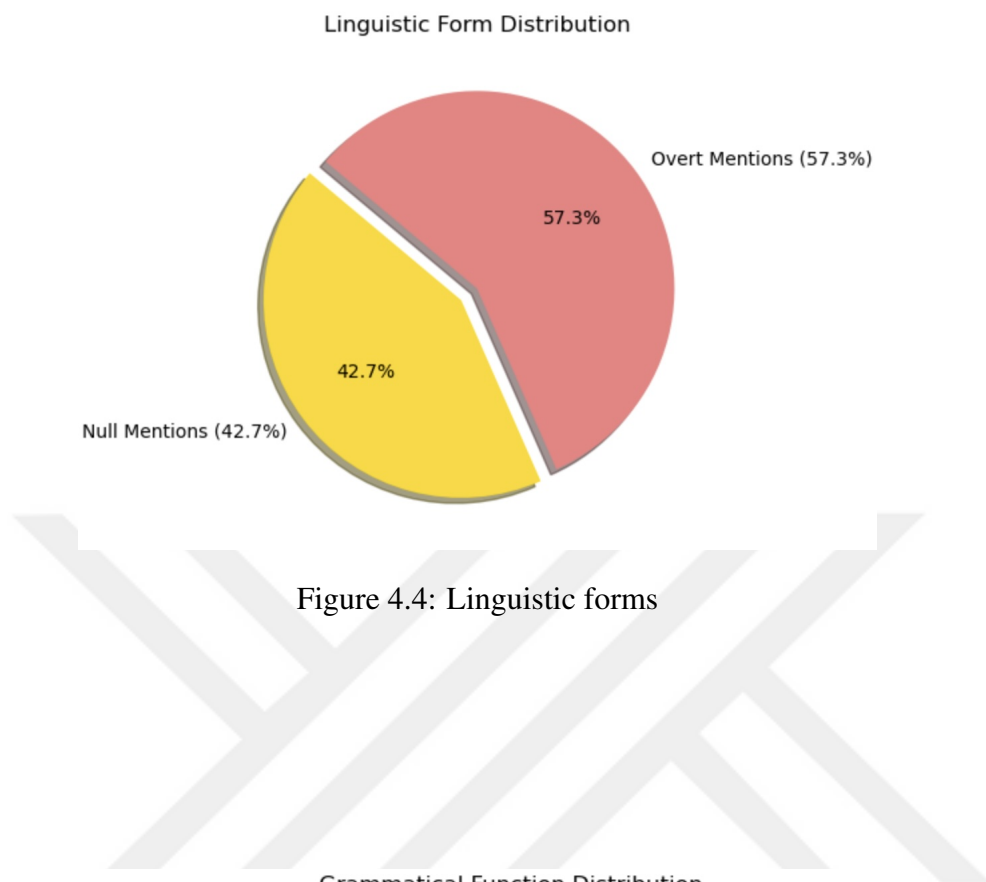


Figure 4.4: Linguistic forms

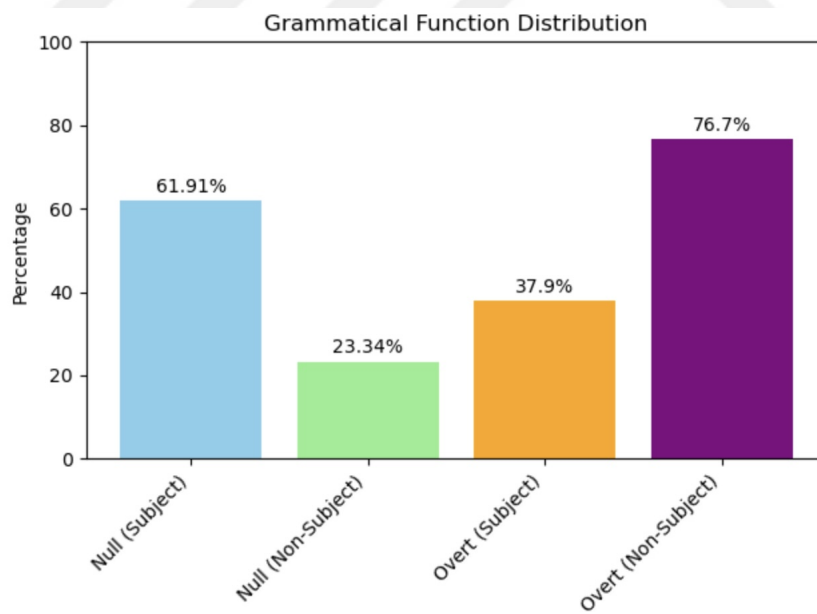


Figure 4.5: Grammatical Roles

CHAPTER 5

ANALYSES

This chapter analyzes the Turkish coreference corpus developed based on the scheme outlined in the second chapter. Current research on reference often analyzes mostly individual factors that affect the selection of referring expressions. Although these studies have highlighted the influence of discourse features on referential choices, a more thorough understanding requires examining the complex interaction between the nature of the task, the task role of discourse participants, grammatical roles, recency/distance, word order while speakers simultaneously making a reference choice. To advance this understanding, the present chapter systematically investigates these interrelated factors using both statistical and machine learning approaches. By integrating analyses of speaker roles, turn-taking, grammatical functions, competition, recency, discourse density, and syntactic positioning, this chapter aims to reveal how each dimension — individually and in combination — shapes referential form choice in situated dialogs. The following sections present the results of these analyses, beginning with the preparation of the dataset and the extraction of relevant linguistic and discourse features for quantitative modeling.

5.1 Feature Engineering: Raw Data and Derived Features

This section outlines the process of feature engineering, which involved extracting raw data from the annotated Turkish coreference corpus and deriving variables to investigate the factors influencing reference selection. Both the raw data and the engineered features represent linguistic, cognitive, and contextual properties crucial to the analysis.

Following the creation of our corpus, which involved the meticulous annotation of mentions and coreference relations, we conducted an extensive information extraction process to transform the annotated data into a structured format suitable for our analysis. This step was crucial for performing deeper investigations about the linguistic and cognitive phenomena underlying coreference in Turkish dialogs. Specifically, we extracted a range of information from the annotated dialogs, encompassing both raw data and derived features.

5.1.1 Raw Data Extraction

The raw data include main information about expressions such as ids of mentions, their spans/boundaries, grammatical role tags, and coreference relations. Building upon this foundational layer, we identified coreference chains, which represent the connections between referring expressions as constructed by participants. Additionally, we cataloged the entities referred to by these expressions, providing a comprehensive map of the referential landscape within the dialogs. For all these tasks we used Python scripts utilizing `networkx` library.

The annotated corpus, therefore, provided foundational data containing detailed information about mentions, including their IDs, spans, referents, tags, relations, and other coreference-specific annotations. From this raw data, various layers of information were extracted to form the basis for computational and statistical analyses. Key components extracted from the corpus include:

- **Raw Data:** Information about mentions, including linguistic forms, tags, spans, and relationships.
- **Chains:** Coreference chains in dialogs, indicating the referring expressions refer to the same entity.
- **Entities:** The entities/referents that co-referring expressions point to.
- **Features:** Variables derived from raw data, reflecting linguistic, cognitive, and pragmatic factors influencing reference selection.

5.1.2 Features

To facilitate detailed analysis, we derived a diverse set of features describing the linguistic and contextual properties of mentions and their antecedents. These features capture both the surface-level and structural aspects of coreference, such as the linguistic forms of mentions, their grammatical roles, their syntactic positions within the sentence, and the competitive context surrounding them. Moreover, contextual measures, such as the distance between mentions and antecedents, were incorporated to quantify their relative placement in the dialog. These extracted features provide a robust framework for exploring referential form selection and modeling linguistic phenomena in Turkish, a morphologically rich, pro-drop language.

The features derived from the raw data can be grouped into major categories, each representing critical aspects of reference selection: speaker/dialog-level variables, grammatical features, proximity metrics, competition factors, topicality measures, and sentential position. These features were meticulously operationalized to encapsulate the multifaceted nature of reference selection in natural discourse:

5.1.2.1 Linguistic Forms

These features capture the type of linguistic expression used for the mention and its antecedent. These features are central to our analysis, providing insight into how speakers choose referring forms based on contextual and grammatical factors.

- **aform**: The linguistic form of the antecedent, which includes:
 - Full noun phrases **fullnp**: Complete noun phrases that include determiners, modifiers, and heads.
 - * e.g. **kırmızı üçgen** “the red triangle”
 - * e.g. **soldaki küçük üçgen** “the small triangle on the left”
These phrases might include modifiers (e.g., *kırmızı* , *küçük*), determiners or locative specifiers (e.g., *soldaki*), and head nouns (e.g., *üçgen*).
 - Overt Pronouns **overt**: Overt pronouns referring to entities.
 - * e.g. **o** “he/she/it”
 - * e.g. **bu** “this”
These are phonologically realized pronouns that are used when disambiguation or emphasis is required.
 - Null pronouns **zero**: Implicit pronouns and agreement markers commonly used in Turkish due to its pro-drop nature.
 - * e.g. **(Ø) Başladım.** “(I) have begun.”
 - * e.g. **(Ø) Çevirdi(Ø).** “(He/She) flipped (it).”
In these examples, the subjects and the object are omitted and only recoverable through verbal morphology and preceding context .
 - **rform**: The linguistic form of the referent include the same categories under **aform**, but occur at the point of referential act rather than antecedence. For instance:
 - **fullnp**: *küçük üçgen* “the small triangle”
 - **overt**: *o* “he/she/it”
 - **zero**: *(Ø) Yerleştirdi(Ø).* “(He/She) placed (it)”
- These are analyzed in terms of their morphosyntactic form and their correspondence with antecedents.

5.1.2.2 Speaker and Dialog-Level Features

These variables account for contextual and pragmatic influences on reference selection. The speaker’s role and the specific dialog may shape the linguistic form of referring expressions due to task-specific and situational constraints.

- **speaker**: The task role of the speaker who uttered the referent (e.g., presenter, operator).
 - The **presenter** is expected to use more explicit linguistic expressions to minimize ambiguity while giving instructions to the other participant:

“*Kırmızı üçgeni al.*”
“Take the red triangle.”
 - e.g. The **operator** is supposed to produce reduced forms more while realizing the instructions given by the **presenter**:

“(Ø) yerleştirdim.”
“(I) placed (it).”
- **dialog**: The dialog in which the referent was introduced. **Dialogs** produced by different speakers might include use of certain linguistic forms.

5.1.2.3 Turn-taking features

Turn-taking features specify if the mention or sentence was produced in a new turn.

- **tt_mention**: Indicates whether the referring expression was produced in a new speaker turn. (Binary: 1 = change of turn at the moment of mention, 0 = within the same turn.)
 - **Example (tt_mention = 1):**

Presenter: Şimdi kırmızı üçgeni al.
“Presenter: Now take the red triangle.”
Operator: (Ø) Aldım.
“Operator: (I) took (it).”
Explanation: The referring expression is a zero pronoun (Ø), produced in a new turn. Since the referent (*kırmızı üçgen*) was introduced in the previous speaker’s turn, **tt_mention** = 1.
 - **Example (tt_mention = 0):**

PRese: Kırmızı üçgeni al ve onu sağ tarafa yerleştir.
“Take the red triangle and place it to the right side.”
Explanation: Both the antecedent and referring expression (*onu*) occur within the same speaker turn. Thus, **tt_mention** = 0.
- **tt_sentence**: Indicates whether the sentence containing the referring expression was initiated in a new speaker turn. (Binary: 1 = sentence starts a new turn, 0 = sentence occurs within the same turn.)
 - **Example (tt_sentence = 1):**

Presenter: Bir kare var mı?
“Presenter: Is there a square?”
Operator: Evet, (Ø) görüyorum.

“Operator: Yes, (I) see (it).”

Explanation: The sentence containing the referring expression ($\emptyset$) starts a new turn. Therefore, `tt_sentence = 1`.

– **Example (`tt_sentence = 0`):**

Presenter: Sağda bir üçgen var, onu seç ve sağ köşeye koy.

“There’s a triangle on the right, select it and place it in the corner.”

Explanation: The referring expression *onu* appears in the second clause of a multi-sentence turn. The sentence does not start a new turn. Hence, `tt_sentence = 0`.

5.1.2.4 Grammatical Roles

The following features are related to the grammatical roles/functions of antecedents and referents in discourse. They provide insight into how grammatical accessibility influences the choice of referring expressions, with subject positions generally favoring reduced forms like pronouns.

- `arole`: The grammatical role of the antecedent (i.e., subject, object, other).
 - **Subject antecedent (subj):**
O kareyi aldı. ($\emptyset$) Sonra üçgeni çevirdi.
"He took the square. Then (he) flipped the triangle."
Here, the antecedent *O* appears in the subject position.
 - **Object antecedent (obj):**
Presenter: Kırmızı üçgeni karenin üstüne yerleştir.
"Presenter: Place (the red triangle) on top of the square."
Operator: Tamam, üçgeni koydum.
"Operator: Okay, I placed (the triangle)."
The antecedent *Kırmızı üçgeni* appears as the object of the sentence.
 - **Other antecedent (e.g. possessive):**
Bu, Adamın kafası.
This is *The man's head*.
The referent is part of a possessive noun phrase.
- `rrole`: The grammatical role of the referent (i.e., subject, object, other).
 - **Subject mention (subj):**
($\emptyset$) *Onu* yerleştiriyorum şimdi.
“(I) am placing it now.”
The referent occurs in subject position and is realized as an overt pronoun.
 - **Object mention (obj):**
Onu sağ tarafa taşı.
“Move it to the right side.”
The overt pronoun *onu* marks the referent in object position.
 - **Other mention (e.g., possessive):**
O rengine karar veremedi.

“He/She couldn’t decide on (its) color.”

The referent is embedded in a possessive structure (*rengine*) and is marked with genitive suffixation *i*.

5.1.2.5 Competition Metrics

Competition metrics quantify the degree of competition a referent faces from other mentions and entities in the discourse. They reflect the likelihood of ambiguity in reference interpretation.

Mention-based competition:

- **mcomp0**: The number of competitor mentions in the sentence where the referent is introduced.

- **Example (mcomp0 = 2):**

Üçgeni aldım ve onu kareye yapıştırdım.

“I took the triangle and placed it next to the square.”

Explanation: The sentence with the referent *onu* contains two other competing mentions: *üçgeni* and *kareye*.

- **mcomp1**: The number of competitor mentions in the sentence where the referent appears and the preceding sentence.

- **Example (mcomp1 = 3):**

Presenter: Kırmızı üçgeni taşı ve kareye yaklaştır.

“Presenter: Move the red triangle and put it next to the square.”

Operator: Onu kareye göre hizaladım.

“Operator: I aligned it in line with the square.”

Explanation: The current and previous sentences contain four mentions (*Kırmızı üçgeni*, *kareye*, *Onu*, *kareye*). The referring expression *onu* occurs in a high-competition environment, but reduced form is used due to strong focus continuity.

- **mcomp2**: The number of competitor mentions in the sentence where the antecedent is introduced.

- **Example (mcomp2 = 2):**

Kareyi mavi üçgen ile öbür üçgene yaklaştır.

“Move the square between the blue triangle and the other triangle.”

Explanation: Two mentions (*mavi üçgen*, *öbür üçgene*) compete with the antecedent (*Kareyi*), which may reduce its accessibility in the discourse model.

- **mcomp3**: The number of competitor mentions across the sentences spanning the antecedent and referent.

- **Example (mcomp3 = 3):**

Kareyi al.

“Grab the square.”

Sonra üçgeni al. Sağ köşeye (Ø) koy.

“Then take the triangle. Put it in the corner.”

Kareyi biraz daha sola alabilir misin?

“Can you move the square a bit more to the left?”

Explanation: Four mentions appear in the span including the antecedent (*Kareyi*) and mention(*Kareyi*). In total, three other mentions (*Kareyi*, *üçgeni*, *(Ø)*) compete with the mention.

Entity-based competition:

- **ecomp0:** The number of unique competitor entities in the sentence where the referent is introduced.

- **Example (ecomp0 = 2):**

Onu üçgene yaklaştır ve kareyi al.

“Move it next to the triangle and take the square.”

Explanation: Two unique entities—*üçgene* and *kare*—compete with the mention *onu*, raising ambiguity and requiring context-dependent resolution.

- **ecomp1:** The number of unique competitor entities in the sentence where the and the preceding sentence appear.

- **Example (ecomp1 = 1):**

Presenter: Kareyi üçgene yaklaştır.

“Presenter: Move the square next to the triangle.”

Operator: Onu oraya hizaladım.

“Operator: I aligned it there.”

Explanation: Unique entities mentioned include only *üçgene*, which create only one referential alternative for *onu*.

- **ecomp2:** The number of unique competitor entities in the sentence where the antecedent is introduced.

- **Example (ecomp2 = 1):**

Üçgeni çevirdim, kareyi aldım ve (Ø) yerleştirdim .

“I rotated the triangle, took the square, and placed it.”

Explanation: One unique entity (*Üçgeni*) competes in the antecedent sentence, potentially creating referential ambiguity.

- **ecomp3:** The number of unique competitor entities across the sentences spanning the antecedent and referent.

- **Example (ecomp3 = 5):**

Kareyi al.

“Grab the square.”

Sonra üçgeni al. Sağ köşeye (Ø) koy.

“Then take the triangle. Put it in the corner.”

Kareyi biraz daha sola alabilir misin?

“Can you move the square a bit more to the left?”

Explanation: One unique entity (*üçgeni*) appear in the discourse span between the antecedent and mention (*Kareyi, Kareyi*), prompting reuse of the full noun phrase to help resolve ambiguity.

5.1.2.6 Distance Metrics

Distance metrics quantify the distance between antecedents and mentions in discourse. These variables are vital for examining how linguistic forms are affected by distance/recency, with shorter distances often favoring reduced forms and longer distances favoring explicit forms like full noun phrases.

- **sdist:** The number of sentences between the antecedent and the referent.

- **Example (sdist = 1 — short distance):**

Üçgeni seçtim.

“I selected the triangle.”

Onu sağ tarafa yerleştiriyorum.

“I am placing it on the right.”

Explanation: The referent and antecedent are in adjacent sentences (**sdist** = 1), and an overt pronoun (*onu*) is used.

- **Example (sdist = 3 — longer distance):**

Bir üçgen aldım.

"Take a triangle."

Rengi fark etmez.

"The color doesn't matter."

Karenin yanına koy.

"Placed it next to the square."

Üçgeni tekrar alabilir misin?

"Can I take the triangle again?"

Explanation: The antecedent *Bir üçgen* is introduced first and referred to again later with the mention *Üçgeni*, requiring a more explicit reference(*Üçgeni*).

- **wdist:** The number of words between the antecedent and the referent.

- **Example (wdist = 2 — short distance):**

Üçgeni al ve onu yerleştir.

“Take the triangle and placed it.”

Explanation: The word distance between *Üçgeni* and *onu* is small (**wdist** = 2), and a reduced form is used.

- **Example (wdist = 20 — long distance):**

Sağ köşedeki kareyi aldım. Sonra büyük üçgeni aldım, onun şekline göre sonra diğer küçük üçgeni birkaç kez daha denedim. En son, tekrar yine kareyi yerleştirdim.

“I took the square in the right corner. Then I took the large triangle; based

on its shape, I tried the other small triangle a few more times. Finally, I placed the square again.”

Explanation: The antecedent *Sağ köşedeki kareyi* is mentioned earlier and not referenced again for 20 words until the mention *kareyi*. The speaker reuses the full noun phrase to ensure clarity due to the intervening number of words.

5.1.2.7 Topicality Metrics

Topicality metrics quantify how densely a referent is used in the discourse.

- **mrnk:** The number of times the referent has been mentioned so far in the discourse, reflecting its persistence throughout the text.
- **mdensity_a:** The global mention density calculated as the total number of mentions of the entity up to the current mention divided by the number of sentences since the first mention. This provides a global measure of referential salience.
- **mdensity_as:** A variant of **mdensity_a** where, instead of counting all mentions, only the number of sentences in which the entity has been mentioned is considered. This value is divided by the number of sentences since the first mention, offering a sentence-level perspective on cumulative salience.
- **mdensity_m:** The local mention density computed as the number of mentions of the entity in the most recent n sentences, divided by n . This reflects the local recency and salience of the entity in the immediate discourse context.
- **mdensity_s:** The local sentence density calculated as the number of sentences in the most recent n sentences in which the entity has been mentioned, divided by n . This focuses on the distribution of mentions across sentences in the local discourse window.

5.1.2.8 Word Order Features

To quantify the linear position of referring expressions and their antecedents within sentences, two normalized word order measures were derived. Both features aim to capture gradient word order effects that are suggested to influence the form of referring expressions. Earlier positions often correlate with topic-like or syntactically prominent roles (e.g., subjects), while later positions may be more associated with focus or less accessible referents (e.g., objects and adjuncts). By normalizing to sentence length, these measures abstract away from sentence-specific variation and allow for cross-sentence comparison of positional patterns.

- **awo:** This variable captures the relative position of the antecedent within its sentence. It is computed by dividing the index of the first character of the

antecedent span by the total number of characters in the sentence. The resulting value ranges from 0 (indicating sentence-initial position) to 1 (indicating sentence-final position). This measure reflects how early or late the antecedent appears in the sentence, with lower values indicating prominence due to early positioning.

- r_{wo} : This variable encodes the relative position of the referring expression within its sentence. It is calculated in the same way as a_{wo} , by dividing the character index of the first token of the referring expression by the total character length of the sentence. This normalized index allows comparison across sentences of different lengths and captures the syntactic and informational structure of referring expression production.

The derived features provide a comprehensive set of the linguistic, cognitive, and contextual factors influencing reference selection. By including variables that span speaker role, turn-taking, grammatical role, competition, distance, topicality, and sentential position, the allows for exploration of the mechanisms underlying referential choice. These features were subsequently used in both statistical tests and machine learning analysis to uncover patterns and test hypotheses about reference selection. The data-driven approach, combined with cognitive and linguistic insights, ensures the robustness and validity of the analysis. This transition from dataset creation to structured data extraction bridges the gap between raw annotations and systematic analyses, laying the groundwork for theoretical and computational investigation.

5.2 Hypotheses

This section presents the hypotheses formulated to guide the statistical and machine learning analyses in the thesis. Each hypothesis reflects predicted relationships between the form of the referring expression and a range of discourse, grammatical, and cognitive variables. These expectations are grounded in previous theoretical frameworks, models, and prior empirical findings, as well as patterns observed in the current dataset. The hypotheses are organized according to key categories of factors found to influence referential choice, including speaker role, turn-taking, grammatical role, competition, distance, and sentential position.

Speaker role:

1. **H:** Given the task-specific roles of the participants (presenter and operator), The presenter, who is responsible for task planning and providing instructions, is likely to produce more explicit forms (e.g., full noun phrases) to minimize ambiguity.
2. **H:** The operator, who relies on shared context and responds to the instructions, tends to use implicit forms (e.g., zero pronouns) more frequently.

Turn-Taking:

3. **H:** Referring expressions produced at the beginning of new speaker turns (i.e., when a sentence follows a turn transition) are more likely to be realized as full noun phrases compared to those produced within the same turn.
4. **H:** Referring expressions produced immediately after a turn change (i.e., when the mention itself is aligned with a turn transition) are more likely to take the form of full noun phrases, whereas those produced within the same speaker turn are more likely to be realized as reduced forms (overt and zero pronouns).

Grammatical role:

5. **H:** Subject positions favor reduced referring expressions.
6. **H:** Non-subject positions increase the likelihood of full noun phrases.

Syntactic position:

7. **H:** Referring expressions occurring in earlier positions within the clause are more likely to take the reduce form of noun phrases.
8. **H:** Referring expressions occurring in later positions within the clause are more likely to be realized using full noun phrases forms.
9. **H:** The relationship between word order and referring expression form is expected to be gradient, such that the likelihood of full noun phrases increases from early to late positions in the sentence.

Competition:

10. **H:** Higher competition increases the likelihood of full noun phrases
11. **H:** Lower competition favors reduced forms
12. **H:** Competition increases the use of overt pronouns over null pronouns
13. **H:** Full noun phrases are preferred for re-introducing referents in competitive contexts
14. **H:** The use of null pronouns decreases as competition increases
15. **H:** Pronouns are more likely in low-competition but high salience contexts

Recency:

16. **H:** Increased distance leads to more explicit referring expressions
17. **H:** Decreased distance favors reduced referring forms
18. **H:** Pronouns are preferred for intermediate distances
19. **H:** Null pronouns are likely when mentions are consecutive
20. **H:** Full noun phrases are used to reintroduce distant referents

Density:

21. **H:** Higher discourse density (both global and local) will increase the likelihood of reduced referring expressions, such as overt and zero pronouns, due to enhanced referent accessibility.
22. **H:** Lower discourse density will be associated with a greater use of full noun phrases, reflecting the need for more explicit and informative referential expressions when referent accessibility is lower.
23. **H:** Local density measures (e.g., `mdensity_m`, `mdensity_s`) will have a stronger influence on referring expression form than global density measures (e.g., `mdensity_a`, `mdensity_as`), as local salience is expected to play a more immediate role in referential choice.

These hypotheses reflect a multi-level perspective on referential selection, integrating discourse-level, grammatical, and interactional factors. By testing these predictions using both traditional statistical analyses and machine learning models, the thesis seeks to reveal systematic patterns in referential form selection. The following section presents the resulting dataset and results of these analyses, examining to what extent each factor and their interactions shape the realization of referring expressions in situated dialogs.

5.3 Dataset

The initial dataset comprised a total of 5732 observations. In this study, each *observation* corresponds to a single instance of referential expression production within the dialog corpus. Specifically, an observation is defined as a collection of:

- the referring expression (e.g., full noun phrase, overt pronoun, or zero pronoun),
- its associated linguistic, syntactic, and discourse-level features (e.g., grammatical role, distance from antecedent, word order position),
- and its antecedent and referent properties, including discourse context and speaker information.

Therefore, each observation captures a complete referential act within a dialog, reflecting a speaker’s choice of referential form under specific cognitive, grammatical, and interactional conditions. As such, the unit of analysis is not the referent or entity, but the referential form decision made at a particular point in the dialog and the collection. This framing allows for the investigation of the factors influencing the speaker’s selection of referring expression form across varying discourse environments. However, mentions referring to discourse participants (experiment participants) were excluded from the analysis. This decision was made because discourse participants play an asymmetrical role in dialog creation, primarily contributing to discourse planning and management, rather than being part of the core referential structure under our investigation. Moreover, a substantial proportion of these mentions appeared in null

form, which further limited their utility for analyzing explicit coreference patterns. By excluding these mentions, we ensured that the dataset focused on referential expressions central to the experimental design.

The dataset for analysis was refined to include only mentions referring to target shapes, geometric shapes, and their parts, which constitute the core elements of the referential task. The dataset initially consisted of 2796 observations and 27 features.

In the dataset, we analyzed the overall distribution of referring expressions (REs) across three linguistic forms: full noun phrases (fullnp), overt pronouns (overt), and zero pronouns (zero). The results are as follows:

- Full noun phrases (fullnp): 1375 occurrences
- Overt pronouns (overt): 443 occurrences
- Zero pronouns (zero): 978 occurrences]

The distribution of referring expression forms in our dataset was not uniform. Full noun phrases ($n = 1375$, 49%), zero pronouns ($n = 978$, 35%), and overt pronouns ($n = 443$, 16%) showed substantial differences in frequency.

These results indicate that full noun phrases constitute the largest proportion of referring expressions, accounting for the most explicit form of reference. Zero pronouns are also frequently used, reflecting the reliance on implicit referential strategies in contexts where referents are pragmatically inferable or clearly established in the discourse. Overt pronouns appear less frequently, suggesting a more limited need for explicitly stated pronouns in the given task.

This overall distribution highlights the varying degrees of referential explicitness in the dialogs and suggests that task-specific demands and shared context play a significant role in determining the choice of referring expression form. This suggests that speakers preferred explicit referential forms unless discourse salience conditions favored omission or pronominalization. In our dialogs, speakers prefer using full NPs almost half the time when referring to entities. This suggests that either salience is not high enough or ambiguity needs to be resolved explicitly quite often. Zero pronouns are still quite frequent, showing that null anaphora are highly natural for Turkish as a pro-drop language. Overt pronouns are relatively less frequent, which is also expected because in Turkish, overt pronouns are often reserved for contrastive focus, disambiguation, or emphasis rather than as a default strategy unlike English, which uses overt pronouns a lot more often.

5.4 The distribution of referring expressions according to the role of the discourse participants

The dialogs in our dataset originate from an experimental setting where participants collaborate to solve tangram puzzles (Mançe-Çalışır, 2018). In this task, pairs of participants work together to build a target shape by manipulating seven geometric

shapes using a computer simulator. The participants are seated face-to-face, but a physical separator prevents them from seeing each other. This setup requires them to rely solely on verbal communication. The participants are assigned specific roles to simulate natural language production in a goal-oriented context: the presenter, who has access to the target shape, provides instructions to the operator, who controls the mouse to manipulate the shapes but cannot see the target shape. Given these role differences, we examined the distribution of referring expressions (REs) produced by each participant type, categorized into three linguistic forms:

- Full noun phrases (fullnp): Explicit, descriptive noun phrases.
- Overt pronouns (overt): Pronouns explicitly stated in the dialog.
- Zero pronouns (zero): Implicit pronouns that rely on contextual inference.

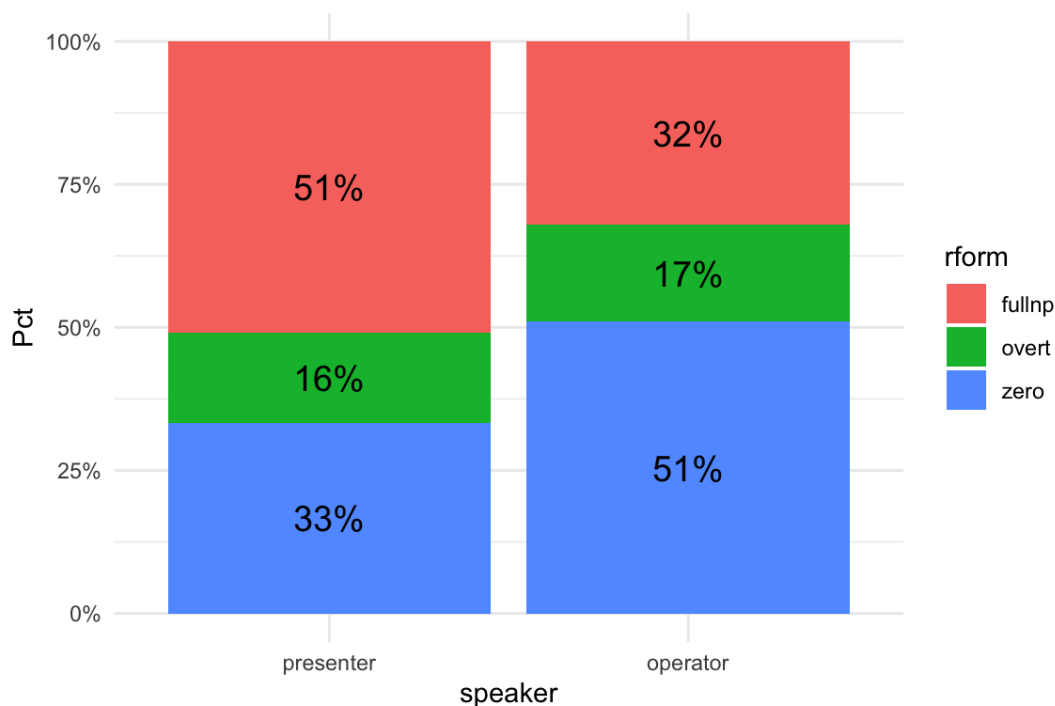


Figure 5.1: Speaker role

The results show distinct patterns based on participant roles:

- **Presenter:** The presenter, whose task involves detailed instructions, produced the majority of full noun phrases (51%). These explicit expressions likely help ensure clarity in describing the target shape and its components. The presenter also used zero pronouns in 33% of their REs, likely relying on shared context established earlier in the dialog. Overt pronouns accounted for 16% of their REs, likely for referencing previously mentioned entities.
- **Operator:** The operator, whose role is to follow the presenter's instructions and execute tasks, relied heavily on zero pronouns, which made up 51% of

their REs. This reflects the operator's dependence on the shared understanding established by the presenter's instructions. Full noun phrases accounted for 32% of their REs, likely used when seeking clarification or confirming specific components of the task. Overt pronouns were used in 17% of the operator's REs.

This asymmetry highlights the influence of role-specific tasks on the linguistic forms chosen in communication. The presenter's focus on planning and managing the task necessitates explicit references, while the operator, responding within an established context, relies more on implicit forms like zero pronouns. This difference underscores the interplay between task demands and language production in dialog-based interaction.

5.4.1 Turn-Taking and Referential Form

Turn-taking is a fundamental characteristic of conversational interaction and has been shown to shape linguistic choices in real time. As speakers alternate turns, they must manage referential continuity to ensure successful communication. Prior research suggests that turn boundaries introduce discourse discontinuities which may affect the accessibility of referents and, consequently, the form of referring expressions (). Specifically, speakers tend to use reduced forms such as overt and zero pronouns when maintaining reference within the same turn, where discourse coherence and topicality are preserved. In contrast, referential expressions produced across turn boundaries, particularly when the interlocutor intervenes or when discourse continuity is disrupted, are more likely to take the form of full noun phrases to re-establish referent identity for the listener. The current analysis aims to investigate how referring expression forms (full noun phrases, overt pronouns, and zero pronouns) are distributed across different turn-taking contexts, with particular attention to the distinction between turns based on sentential and mention boundaries. We defined two features to capture turn-taking patterns in relation to referring expressions, namely sentence and mention level features.

Relationship Between Referring Expression Form and Turn-Taking at Sentence Level

The distribution of referring expression forms varied systematically as a function of turn-taking at the sentence level (Table 5.1). Full noun phrases were more likely to be produced at the beginning of new speaker turns compared to overt and zero pronouns, whereas reduced forms were generally favored in same-turn contexts.

A chi-square test of independence confirmed that this association was statistically significant, $\chi^2(2) = 26.06$, $p < .05$. Although the effect size, as indicated by Cramer's V ($V = 0.097$), was small, the pattern suggests that turn transitions increase the likelihood of using more explicit referring expressions. These findings are consistent with discourse continuity accounts, which propose that speakers tend to use full noun phrases at turn boundaries to re-establish referent identity, while reduced forms such

Table 5.1: Distribution of referring expression forms (rform) by turn-taking at sentence level (tt_sentence)

tt_sentence	Full NP	Overt Pronoun	Zero Pronoun
0 (Same Turn)	1074	319	674
1 (New Turn)	301	124	304

as overt and zero pronouns are preferred in contexts where referential continuity is maintained within the same turn.

Relationship Between Referring Expression Form and Turn-Taking at Mention Level

The distribution of referring expression forms showed a strong association with turn-taking at the mention level (Table 5.2). Full noun phrases were markedly more frequent when the referential mention occurred at a new speaker turn (589 out of 1375 full NPs), whereas overt and zero pronouns were strongly associated with same-turn contexts. Specifically, overt pronouns exhibited a strong preference for same-turn production (369 out of 443 occurrences), while zero pronouns also occurred predominantly within the same turn (768 out of 978 occurrences).

Table 5.2: Distribution of referring expression forms (rform) by turn-taking at mention level (tt_mention)

tt_mention	Full NP	Overt Pronoun	Zero Pronoun
0 (Same Turn)	786	369	768
1 (New Turn)	589	74	210

A chi-square test of independence confirmed that this relationship was statistically significant, $\chi^2(2) = 173.14$, $p < .05$. The effect size, as indicated by Cramer's V ($V = 0.249$), was small-to-moderate but notably larger than that observed at the sentence level. This suggests that turn transitions at the point of mention has a stronger influence on referential form choice. Full noun phrases are more likely to be used when a new turn is initiated, presumably to facilitate referent identification across speaker boundaries. By contrast, overt and zero pronouns are favored when discourse continuity is maintained within the same speaker turn.

The analyses of turn-taking at both the sentence and mention levels revealed systematic patterns in referring expression form selection. At the sentence level, full noun phrases were somewhat more likely to be used when sentences began with a new speaker turn, whereas overt and zero pronouns tended to be produced within the same turn. Although this relationship was statistically significant, the effect size was

relatively small, suggesting a modest influence of sentence-level turn transitions. In contrast, turn-taking at the mention level showed a much stronger association with referential form. Full noun phrases were substantially more frequent at new turn beginnings, while overt and zero pronouns were overwhelmingly produced within the same turn. This pattern, supported by a larger effect size, indicates that referential form choice is particularly sensitive to speaker change at the point of mention. Together, these findings suggest that turn transitions, especially those aligned with the production of referring expressions themselves, promote the use of more explicit forms to enhance referential clarity across speaker boundaries, whereas reduced forms are preferred in contexts where discourse continuity and referent accessibility are maintained.

5.5 Grammatical Role

The grammatical roles of mentions were analyzed to identify their distribution across syntactic positions. The roles include objects (`obj`), subjects (`subj`), and subordinate subjects (`s_subj`) and objects(`o_subj`), possessives (`poss`). The results are presented in Table 5.3:

Table 5.3: Distribution of grammatical roles of referring expressions

Grammatical Role	Frequency (n)
Object (<code>obj</code>)	1467
Possessive (<code>poss</code>)	707
Secondary Object (<code>s_obj</code>)	66
Secondary Subject (<code>s_subj</code>)	108
Subject (<code>subj</code>)	448

The distribution of grammatical roles across referring expressions reveals important patterns regarding syntactic functions in discourse. Object positions (`obj`) were the most frequent, accounting for 1467 instances, followed by possessive constructions (`poss`) with 707 occurrences. Subject positions (`subj`) were comparatively less frequent (448), while subordinate objects (`s_obj`) and subjects (`s_subj`) were relatively rare, with 66 and 108 instances respectively.

This pattern reflects the task-oriented nature of the dialogs, where actions upon objects and references to possessed items were particularly salient. The relatively lower frequency of subjects suggests that referential expressions often targeted entities being manipulated or described rather than agents performing actions. Moreover, the limited occurrence of secondary grammatical roles highlights the tendency for discourse to favor core syntactic positions, such as objects and possessors, in referential choice.

5.5.1 Analysis of the Relationships Between Referential form and Grammatical Role

To understand the relationship between the referential form of grammatical role of referring expressions, and we conducted a Pearson’s Chi-squared test along with additional visualizations. Below are the results and interpretations.

5.5.1.1 The Linguistic Form of the Antecedent

We explored the relationship between `arole` (`obj`: object, `s_subj`: subordinate subject, `subj`: subject) and the `rform` (full noun phrases, overt pronouns, and zero pronouns).

Table 5.4: Distribution of grammatical roles across referential forms

	<code>obj</code>	<code>poss</code>	<code>s_obj</code>	<code>s_subj</code>	<code>subj</code>
fullnp	847	250	16	42	220
overt	249	99	5	9	81
zero	555	235	14	26	148

A Pearson’s Chi-squared test revealed a statistically significant association between referring expression form (`rform`) and the grammatical role of the antecedent (`arole`), $\chi^2(8) = 16.50$, $p = .036$. However, the strength of this association was very weak (Cramér’s $V = 0.054$; Contingency Coefficient = 0.077), suggesting that while grammatical role exerts some influence on the choice of referring form, it is likely to be one of several factors shaping referential decisions.

5.5.2 The Relationship Between `rform` and `rrole`

We also investigated the relationship between `rrole` (`subj`: subject) `obj`: object, `s_subj`: subordinate subject, `subj`: subject) and the `rform` (full noun phrases, overt pronouns, and zero pronouns).

Table 5.5 presents the distribution of referring expression forms across different grammatical roles of the referent. Full noun phrases were most frequently used when the referent served as an object (745 instances) or a possessor (344 instances), while zero pronouns were common for both objects (456 instances) and possessors (308 instances). Overt pronouns appeared less frequently across all roles but showed a relatively higher presence for objects (266 instances) and subjects (87 instances). Overall, the grammatical role of the referent showed some influence on the choice of referring expression form, with objects and possessors being the most prominent roles for all forms.

Table 5.5: Cross-tabulation of Referring Expression Form (`rform`) and Grammatical Role of the Referent (`rrole`)

Referring Form	obj	poss	s_obj	s_subj	subj
Full NP	745	344	40	60	186
Overt Pronoun	266	55	15	20	87
Zero Pronoun	456	308	11	28	175

A Pearson’s Chi-squared test revealed a statistically significant association between referring expression form (`rform`) and the grammatical role of the referent (`rrole`), $\chi^2(8) = 81.00$, $p < .05$. Although the association was statistically robust, the effect size was weak (Cramér’s $V = 0.12$; Contingency Coefficient = 0.168), indicating that while grammatical role influenced referring form choice, its contribution was modest relative to other discourse factors.

The results indicate that the grammatical role of the referent (`rrole`) shows a stronger association with referring expression form than the grammatical role of the antecedent (`arole`). While both associations were statistically significant, the effect size for `rrole` (Cramér’s $V = 0.12$) was notably larger than that for `arole` (Cramér’s $V = 0.054$), suggesting that the grammatical role at the point of reference — when the speaker produces the referring expression — exerts greater influence on form choice than the grammatical role of the previously mentioned antecedent.

5.5.3 The Relationship Between `rform` and `aform`

Then, we analyzed the relationship between `aform` (full noun phrases, overt pronouns, and zero pronouns) and the `rform` (full noun phrases, overt pronouns, and zero pronouns).

Table 5.6: Cross-tabulation of Antecedent Form (`aform`) and Referring Expression Form (`rform`)

Antecedent Form	Full NP	Overt Pronoun	Zero Pronoun
Full NP	1033	167	322
Overt Pronoun	109	114	172
Zero Pronoun	233	162	484

Table 5.6 shows the distribution of referring expression forms as a function of antecedent form. When the antecedent was a full noun phrase, the referent was also

most likely to be realized as a full noun phrase (1033 instances), although a notable number of cases shifted to zero pronouns (322 instances). When the antecedent was an overt pronoun, referents were more evenly distributed across full noun phrases (109 instances), overt pronouns (114 instances), and zero pronouns (172 instances). Similarly, antecedents that were zero pronouns often led to full noun phrase referents (233 instances) or remained as zero pronouns (484 instances). Overall, while there is a tendency for form continuity between antecedents and referring expressions, shifts — especially toward zero pronouns — were frequent, suggesting flexible referential strategies in discourse.

A Pearson’s Chi-squared test revealed a highly significant association between antecedent form (`aform`) and referring expression form (`rform`), $\chi^2(4) = 498.96$, $p < .05$. The strength of the association was close to moderate (Cramér’s $V = 0.299$; Contingency Coefficient = 0.389), indicating that the form of the antecedent plays a meaningful role in shaping the form of subsequent referring expressions, with both form continuity and form shifts being characteristic features of discourse reference patterns.

5.6 Correlation Analysis of Continuous Features

This section presents a detailed correlation analysis among the continuous features using Spearman’s rank correlation coefficients. Spearman’s correlation was chosen given the ordinal nature and potential non-normality of discourse variables. The resulting correlation matrix (Figure 5.2) reveals systematic relationships among the variables, which are discussed in detail below.

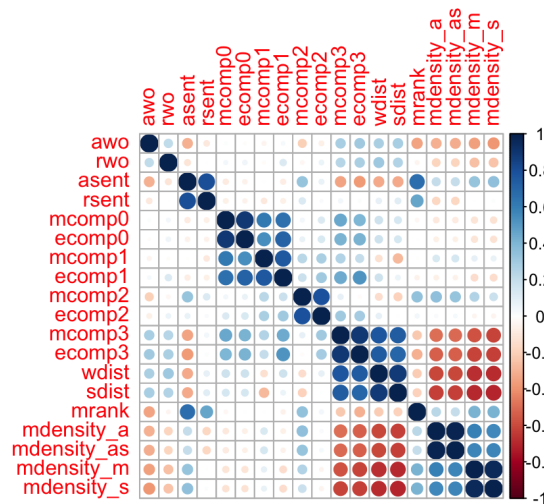


Figure 5.2: Correlation Analysis of Continuous Features

5.6.1 General Patterns

The correlations ranged from weak ($r < .3$) to very strong ($r > .8$), revealing several tightly connected clusters. These patterns support a theoretically motivated grouping of the features into cognitive-contextual domains: *competition*, *distance/recency*, *salience*, and *word order*.

5.6.1.1 Distance/Recency Measures

- **Sentence distance (sdist)** and **word distance (wdist)** showed a very strong positive correlation ($r \approx 0.80$ – 0.90), as expected in natural discourse.
- **Antecedent sentence (asent)** and **referent sentence (rsent)** indexes were also highly correlated ($r \approx 0.80$ – 0.90), reflecting global discourse progression.

Competition Metrics

- Mention-based competition metrics (`mcomp0`–`mcomp3`) and entity-based competition metrics (`ecomp0`–`ecomp3`) were strongly correlated at each level ($r \approx 0.85$ – 0.95), reflecting that more mentions generally introduce more discourse entities.
- Correlations between different levels of competition were slightly weaker, suggesting a distinction between local and global referential competition, suggesting a Centering theoretic analysis.

Density Metrics (Density and Mention Rank)

- Density metrics (`mdensity_a`, `mdensity_as`, `mdensity_m`, and `mdensity_s`) were strongly positively correlated ($r \approx 0.70$ – 0.95), reflecting shared discourse salience mechanisms.
- Mention rank (`mrank`) showed moderate correlations with density metrics ($r \approx 0.30$ – 0.50), as it reflects cumulative mention frequency rather than recency per se.

5.6.1.2 Word Order and Distance

- Antecedent word order (`awo`) and referent word order (`rwo`) were strongly positively correlated ($r \approx 0.80$), indicating positional continuity across referential links.
- Both `awo` and `rwo` were moderately negatively correlated with density measures ($r \approx -0.30$ to -0.50), supporting the view that salient referents tend to appear earlier in the clause.

Cross-Domain Relationships

- Competition metrics and distance measures exhibited very weak correlations ($r \approx 0.00$ – 0.20), indicating that local ambiguity and discourse distance function largely independently.
- Density metrics (density, mention rank) were weakly negatively correlated with competition metrics ($r \approx -0.20$ to -0.30), suggesting that higher competition reduces referential salience.

Overall, the correlation structure supports the theoretical distinction between competition, distance, salience, and word order pressures. The high within-domain correlations and low cross-domain correlations justify the inclusion of all feature groups in the subsequent modeling phase, also highlighting features with high colinearity. Importantly, these results confirm that referential form selection is influenced by multiple complementary mechanisms operating simultaneously in discourse.

5.7 Competition in Referring Expressions

The concept of competition in referring expressions refers to the presence of other mentions (competitors) within the discourse context that may create ambiguity or require more explicit reference to clarify the intended referent. In our dataset, we analyzed the competition dynamics by examining the frequency and distribution of referring expressions (REs) under competitive conditions. These conditions are quantified based on the number of competitor mentions and competitor entities co-occurring in specific sentences or across dialog segments.

When examining competition, we observed that the linguistic form of REs (full noun phrases, overt pronouns, and zero pronouns) interacts with the level of competition in the discourse. For instance, full noun phrases are often used in high-competition contexts to disambiguate the target referent from other entities or mentions. Overt pronouns are more likely to appear when competition is moderate, as they provide enough clarity while still being concise. Zero pronouns are predominantly used in low-competition contexts, where the referent can be easily inferred due to the shared discourse context.

This distribution aligns with the general principle that the explicitness of a referring expression increases with competition to ensure clarity and avoid ambiguity.

5.7.1 Quantifying Competition

We used several measures to capture the competition levels in the discourse:

- `mcomp0`: The number of competitor mentions in the same sentence as the referent

- `mcomp1`: The number of competitor mentions in the sentence of the referent and the preceding one
- `mcomp2`: The number of competitor mentions in the sentence containing the antecedent
- `mcomp3`: The number of competitor mentions spanning the sentences between the antecedent and referent

We also measured unique entities using:

- `ecomp0`: The number of unique competing entities in the same sentence as the referent
- `ecomp1`: The number of unique competing entities in the sentence of the referent and the preceding one
- `ecomp2`: The number of unique competing entities in the sentence containing the antecedent
- `ecomp3`: The number of unique competing entities between the antecedent and referent

To assess the relationship between mention-level and entity-level competition, Spearman's rank correlations were computed between mention counts and unique entity counts across several discourse windows. Within the referent's sentence, mention competition (`mcomp0`) and entity competition (`ecomp0`) were extremely strongly correlated $\rho = 0.89$, $p < .05$, indicating that increased mention density was closely associated with an increase in the number of unique entities. A similarly strong correlation was observed between mention and entity competition across the discourse between the antecedent and referent (`mcomp3` and `ecomp3`, $\rho = 0.90$, $p < .05$), suggesting that intervening discourse typically introduced new entities rather than repeating existing ones. In the referent's sentence combined with the preceding sentence (`mcomp1` and `ecomp1`), the correlation was slightly weaker $\rho = 0.76$, $p < .05$, and a comparable pattern was found within the antecedent's sentence (`mcomp2` and `ecomp2`, $\rho = 0.79$, $p < .05$). Overall, these results indicate a strong coupling between mention frequency and entity diversity, although the relationship slightly weakens over larger discourse windows. Given the high degree of collinearity among these measures, only mention-based were used separately in subsequent models to avoid multicollinearity issues.

5.7.2 Distribution of Referring Expressions Across Competitor Mentions

To analyze the effect of competition on the choice of referring expressions (REs), we examined the distributions of linguistic forms—full noun phrases (`fullnp`), overt pronouns (`overt`), and zero pronouns (`zero`)—across competitor mentions. Key descriptive statistics for each form are presented below.

5.7.2.1 `rform mcomp0`

Table 5.7 presents descriptive statistics for mention competition (`mcomp0`) across different referring expression forms. Full noun phrases were associated with the highest average number of competitor mentions ($M = 1.73$, $SD = 1.22$), followed by overt pronouns ($M = 1.33$, $SD = 1.11$) and zero pronouns ($M = 1.26$, $SD = 1.03$). The distribution of mention competition values was positively skewed for all forms, with overt pronouns exhibiting the highest skewness (1.221) and kurtosis (2.418), suggesting that while most instances involved few competitors, there were occasional cases with substantially higher competition levels. Overall, referring expressions differed systematically in the degree of local competition they were associated with.

Table 5.7: Descriptive Statistics of Mention Competition (`mcomp0`) by Referring Expression Form (`rform`)

<code>rform</code>	<code>n</code>	Mean	SD	Med	Min	Max	Skew	Kurt
Full NP	1375	1.73	1.22	2	0	7	0.998	1.564
Overt Pronoun	443	1.33	1.11	1	0	6	1.221	2.418
Zero Pronoun	978	1.26	1.03	1	0	7	1.098	1.979

Figure 5.3 shows the distribution of referring expression forms as a function of local mention competition (`mcomp0`). When no competitors were present, zero pronouns were the most frequent form (43.82%), followed by full noun phrases (36.27%) and overt pronouns (19.92%). As the number of competing mentions increased, the use of full noun phrases rose sharply, reaching 74.29% when five competitors were present. Conversely, the use of zero pronouns and overt pronouns declined steadily as competition intensified. These results suggest that in contexts with higher referential competition, speakers favored more explicit referring forms, likely to maintain clarity and reduce ambiguity.

A Kruskal-Wallis test revealed a significant effect of referring expression form on local mention competition (`mcomp0`), $\chi^2(2) = 109.49$, $p < .05$, with a small-to-moderate effect size ($\varepsilon^2 = 0.039$). Post-hoc pairwise comparisons with Bonferroni correction indicated that full noun phrases were associated with significantly higher local competition compared to both overt pronouns ($Z = 6.58$, $p < .05$) and zero pronouns ($Z = 9.81$, $p < .05$). There was no significant difference between overt and zero pronouns ($Z = 0.89$, $p = 1.00$).

5.7.2.2 `rform mcomp2`

The analysis of `mcomp2`, the number of competitor mentions in the sentence containing the antecedent, reveals how competition influences the choice of referring expressions (full noun phrases, overt pronouns, and zero pronouns). Below are the key findings based on descriptive statistics and the accompanying bar plot. Descriptive

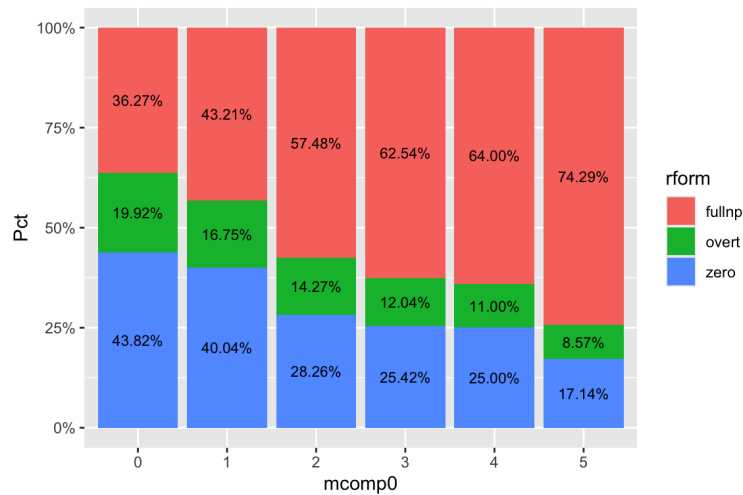


Figure 5.3: Distribution of referring expression forms (`rform`) across levels of local mention competition (`mcomp0`).

statistics for mention competition in the antecedent's sentence (`mcomp2`) by referring expression form are presented in Table 5.8. Zero pronouns were associated with the highest average antecedent-side competition ($M = 2.39$, $SD = 1.08$), followed by overt pronouns ($M = 2.16$, $SD = 1.03$) and full noun phrases ($M = 2.09$, $SD = 1.23$). All distributions were positively skewed, and zero pronouns showed slightly higher kurtosis, indicating more extreme competition values.

Table 5.8: Descriptive Statistics of Antecedent-Side Mention Competition (`mcomp2`) by Referring Expression Form (`rform`)

<code>rform</code>	<code>n</code>	Mean	SD	Med	Min	Max	Skew	Kurt	SE
Full NP	1375	2.09	1.23	2	1	8	1.223	1.506	0.03
Overt Pronoun	443	2.16	1.03	2	1	6	1.015	1.242	0.05
Zero Pronoun	978	2.39	1.08	2	1	8	1.291	2.739	0.03

Figure 5.4 shows the distribution of referring expression forms as a function of mention competition in the antecedent's sentence (`mcomp2`). When antecedent competition was low, full noun phrases were the most frequent form (67.28%), followed by overt pronouns (14.22%) and zero pronouns (18.50%). As competition increased, the proportion of full noun phrases decreased gradually to 51.33%, while the proportion of zero pronouns increased to 32.30%. Overt pronoun usage remained relatively stable across competition levels. These results suggest that antecedent-side competition exerted a modest influence on subsequent referential form choice, in contrast to the stronger effects observed for referent-side competition.

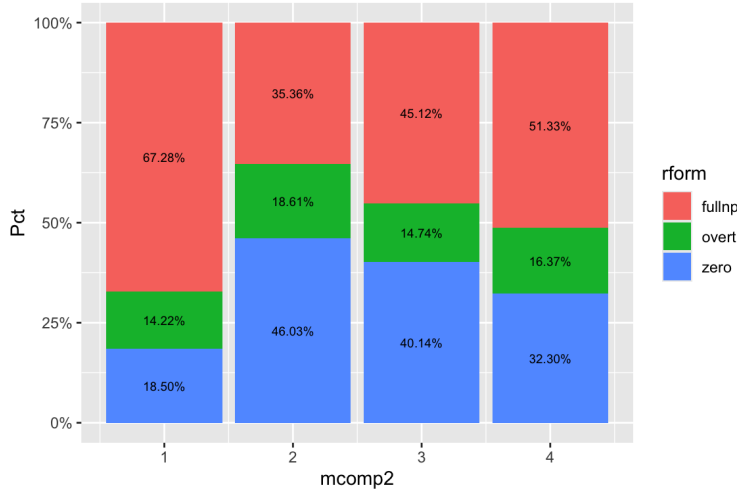


Figure 5.4: Distribution of referring expression forms (*rform*) across levels of antecedent mention competition (*mcomp2*).

A Kruskal-Wallis test revealed a significant effect of referring expression form on antecedent-side competition, $\chi^2(2) = 73.53$, $p < .05$, with a small effect size ($\epsilon^2 = 0.026$). Post-hoc pairwise comparisons with Bonferroni correction showed that zero pronouns were associated with significantly higher antecedent competition than both full noun phrases ($Z = -8.57$, $p < .05$) and overt pronouns ($Z = -3.82$, $p < .05$). Additionally, overt pronouns showed slightly higher antecedent competition compared to full noun phrases ($Z = -2.56$, $p = .031$).

5.7.2.3 *rform mcomp3*

Our descriptive analyses revealed that the number of competitor mentions between antecedent and referent sentences (*mcomp3*) was strongly skewed and kurtotic, with most referents encountering few intervening mentions but some outliers experiencing extreme competition (skew = 3.13, kurtosis = 11.50). The mean competitor mentions were higher for full noun phrases ($M = 18.70$, $SD = 21.66$) compared to overt pronouns ($M = 5.41$, $SD = 9.64$) and zero pronouns ($M = 4.37$, $SD = 4.82$).

A Kruskal-Wallis test indicated a highly significant effect of referring expression form on intervening competition, $\chi^2(2) = 817.34$, $p < .05$, with a moderate effect size ($\epsilon^2 = 0.088$). Post-hoc pairwise comparisons with Bonferroni correction showed that full noun phrases were associated with significantly more intervening competitor mentions than both overt pronouns ($Z = 19.49$, $p < .05$) and zero pronouns ($Z = 26.03$, $p < .05$), whereas overt and zero pronouns did not differ significantly from each other ($p = 1.00$). These findings suggest that full noun phrases tend to be used in contexts where referential continuity is more disrupted by intervening mentions.

5.8 Distance/Recency

To examine whether there are significant differences in the number of intervening sentences (`sdist`) based on the form of referring expressions (`rform`: full noun phrases, overt pronouns, and zero pronouns), we generated a stacked percentage bar plot (Figure 5.5).

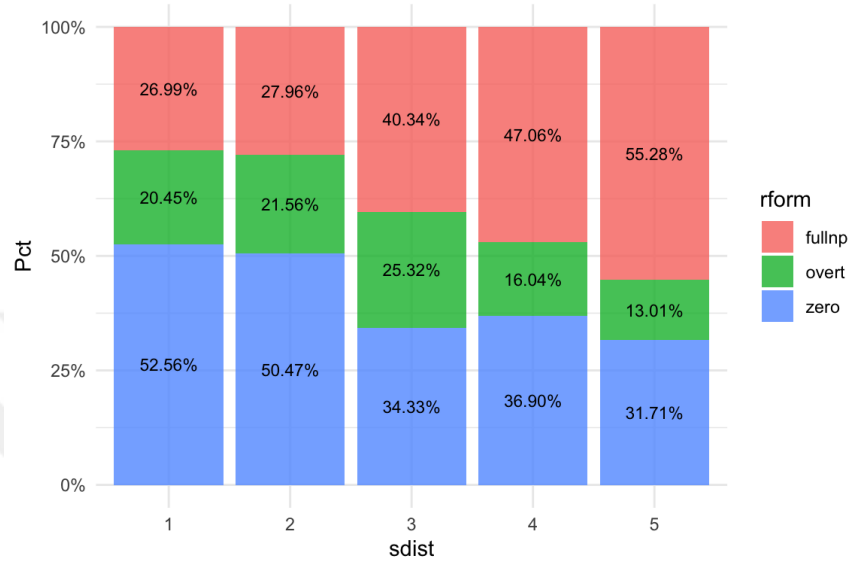


Figure 5.5: Distribution of referring expression forms across different levels of sentence distance (`sdist`) between antecedent and referent.

The plot shows how the distribution of referring forms changes as the sentence distance between antecedent and referent increases. When the referent follows closely after the antecedent (`sdist` = 1 and 2), zero pronouns dominate, comprising over half of all referring forms. However, as sentence distance increases (`sdist` = 3 and 5), the use of full noun phrases becomes more prevalent, while the proportion of zero pronouns decreases. This pattern suggests that greater discourse distance is associated with a shift toward more explicit referring expressions, likely reflecting the reduced accessibility of the antecedent across longer spans. To statistically assess whether the observed differences in sentence distance across referring forms are significant, a Kruskal-Wallis test was conducted comparing `sdist` values across the three referring expression types.

A Kruskal-Wallis test revealed a highly significant effect of referring expression form (`rform`) on the number of sentences between the antecedent and referent (`sdist`), $\chi^2(2) = 704.96$, $p < .05$. The effect size was large ($\epsilon^2 = 0.252$), suggesting that referential form is strongly influenced by the temporal distance from its antecedent. Post-hoc pairwise comparisons with Bonferroni correction showed that full noun phrases occurred at significantly greater sentence distances than both overt pronouns ($Z = 16.48$, $p < .05$) and zero pronouns ($Z = 24.98$, $p < .05$), and overt pronouns also occurred at slightly greater distances than zero pronouns ($Z = 2.52$, $p = .035$).

These findings indicate that as sentence distance increases, speakers increasingly opt for fuller, more explicit referring expressions, a pattern consistent with accessibility-based models of reference. Zero and overt pronouns, by contrast, tend to be used when the antecedent remains highly accessible within a close discourse span.

5.9 Sentential position

Another related factor influencing the referential form is the syntactic position of the referring expressions within the sentence. In languages with flexible word order, earlier sentence positions often correlate with greater syntactic prominence and discourse accessibility due to information structure, which may in turn influence how referents are later encoded. To investigate this, we examined the relationship between the word order of antecedent and anaphor (encoded as a normalized score ranging from 0 for sentence-initial to 1 for sentence-final positions) and the form of the referring expression (full noun phrases, overt pronouns, and zero pronouns). This analysis aims to find out whether referential form vary according to syntactic position in line with status-based theories of reference production.

5.9.1 The word order of the antecedent and referential form

To explore how word order of antecedents affects the choice of referential form by speakers, we first examined the distribution of antecedent position within the sentence (*awo*) across different referring expression types.

The table 5.9 summarizes how the antecedent’s position (scaled between 0 for sentence-initial and 1 for sentence-final) varies by referring expression form. Full noun phrases (*fullnp*) had the highest mean *awo* ($M = 0.39$), indicating they are more likely to refer back to later-occurring antecedents. Overt pronouns (*overt*) had a moderate mean ($M = 0.29$), while zero pronouns (*zero*) were associated with antecedents in the earliest positions ($M = 0.21$). The increasing skewness from full NPs (0.20) to zero pronouns (1.13) further reflects that zero pronouns overwhelmingly refer to antecedents in sentence-initial positions. These trends suggest that reduced forms prefer antecedents that appear earlier and are likely more prominent in the sentence structure.

Table 5.9: Descriptive statistics for antecedent word order (*awo*) by referring expression form (*rform*)

rform	n	Mean	SD	Med	Min	Max	Skew	Kurt	SE
fullnp	1375	0.39	0.29	0.38	0.00	0.96	0.20	−1.10	0.01
overt	443	0.29	0.28	0.23	0.00	0.97	0.72	−0.58	0.01
zero	978	0.21	0.27	0.08	0.00	0.98	1.13	0.12	0.01

Figure 5.6 provides a detailed visualization showing that referring expression forms varies systematically across antecedent word order. Full noun phrases `fullnp` tend to occur later in the sentence. The median value for full NPs is approximately 0.40, which suggests that they are frequently produced after the sentence-initial part. The relatively wide interquartile range (IQR) indicates that full NPs are quite flexible and can appear across various sentence positions, though they show a tendency towards later slots. Overt pronouns `overt` have a lower median value (about 0.23) and a narrower spread compared to full NPs. This suggests that overt pronouns are more likely to appear towards the earlier half of the sentence, though they are still distributed across various positions. Their range and spread are moderate. Zero pronouns `zero` are clearly the most skewed towards the beginning of the sentence. The median is very low (around 0.08), and the IQR and lower whisker indicate that the majority of zero pronouns appear very early, often at the start of the sentence. There are some outliers in higher sentence positions, but the general pattern indicates a preference for sentence-initial use.

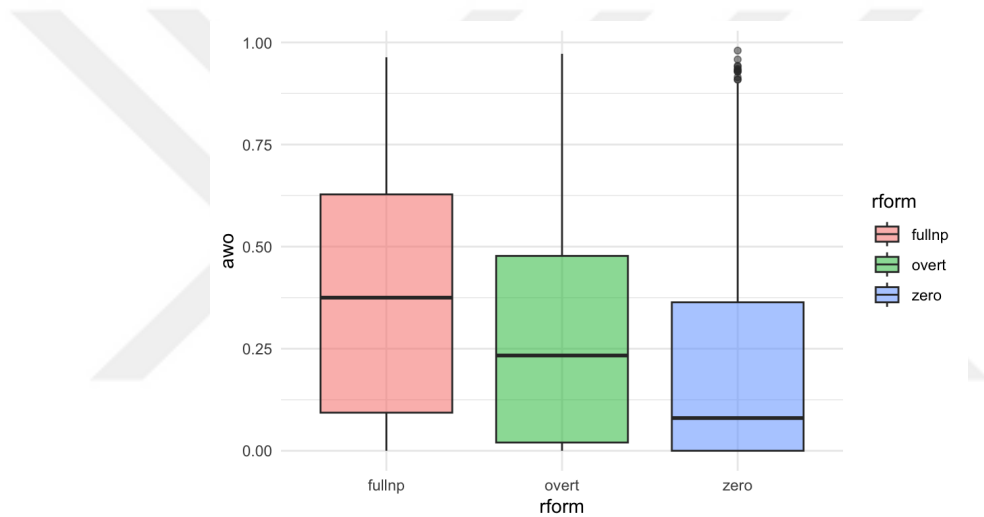


Figure 5.6: Distribution of referring expression forms across antecedent positions

A Kruskal-Wallis test confirmed that the differences among referring forms were statistically significant, $\chi^2(2) = 245.17$, $p < .05$, with an epsilon squared ($\epsilon^2 = 0.088$) indicating a moderate effect size. Post-hoc pairwise comparisons with Bonferroni correction revealed that full noun phrases were significantly later in the sentence than overt pronouns ($Z = 6.64$, $p < .05$) and zero pronouns ($Z = 15.56$, $p < .05$). Additionally, overt pronouns appeared later than zero pronouns ($Z = 5.03$, $p < .05$), though this difference was comparatively smaller.

Taken together, these findings indicate that word order systematically affects referential form choice, with fuller forms (i.e., full noun phrases) tending to occur in later positions, while more reduced forms (i.e., zero pronouns) are favored at sentence beginnings. This pattern aligns with the idea that reduced forms are more likely to be used when they can rely on sentence-initial salience or topic continuity.

5.9.2 The word order of the referring expression and referential form

To further explore how referring expression form `rform` interacts with word order, we analyzed the distribution of forms of referring expressions along the sentence structure using the word order of expression `rwo` itself. This measure captures the linear position of a mention within the sentence, enabling us to assess whether different forms (i.e. `fullnp`, `overt`, and `zero`) exhibit systematic tendencies towards earlier or later positions.

Table 5.10: Descriptive Statistics of Word Order `rwo` by Referring Expression Form `rform`

<code>rform</code>	<code>n</code>	Mean	SD	Med	Min	Max	Skew	Kurt
<code>fullnp</code>	1375	0.26	0.20	0.22	0.00	0.90	0.72	-0.24
<code>overt</code>	443	0.27	0.28	0.17	0.00	0.97	1.05	-0.10
<code>zero</code>	978	0.13	0.24	0.00	0.00	0.97	1.79	2.17

The descriptive statistics reveal distinct patterns across referring expression forms regarding their word order positions. Full noun phrases (`fullnp`) exhibit a moderate mean word order position ($M = 0.26$) and relatively low variability ($SD = 0.20$), suggesting they tend to occur towards earlier-to-mid positions in the sentence but still with some spread. The skewness (0.72) indicates a slight tendency towards earlier positions, but not extreme. Overt pronouns (`overt`) have a similar mean ($M = 0.27$) but higher variability ($SD = 0.28$) and higher skewness (1.05). This implies that while their average position is also early-to-mid, they are more often positioned early, and there is a stronger skew towards the sentence-initial part. Zero pronouns (`zero`) display the lowest mean ($M = 0.13$) and a very strong positive skew (1.79), alongside leptokurtic distribution (Kurtosis = 2.17). This indicates that zero pronouns overwhelmingly occur at the very beginning of the sentence (many zeros at 0), with few exceptions scattered towards later positions.

Figure 5.7 displays the distribution of referential word order positions (`rwo`) for different referring expression forms (`rform`), including full noun phrases (`fullnp`), overt pronouns (`overt`), and zero pronouns (`zero`). Overall, zero pronouns are strongly associated with early positions in the word order. This is indicated by their lower median and compact interquartile range near the bottom of the scale, despite having several outliers extending to higher positions. This suggests that zero pronouns mostly occur in initial or early sentence positions but occasionally also appear later. Overt pronouns, in contrast, exhibit a slightly higher median `rwo` and a wider spread than zero pronouns, reflecting their more flexible distribution across the sentence. However, they still lean towards earlier positions overall. Full noun phrases display the widest spread and a higher median compared to both zero and overt pronouns. This pattern indicates that full NPs are more likely to occur later in the word order, while also showing variability in their placement.

A Kruskal-Wallis test confirmed that the effect of `rform` on `rwo` was statistically sig-

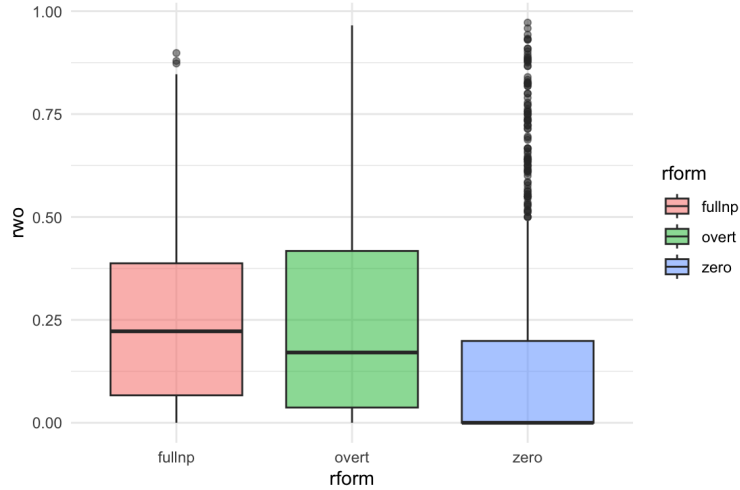


Figure 5.7: Distribution of referring expression forms across referring expression positions

nificant, $\chi^2(2) = 595.11$, $p < .05$. Pairwise comparisons indicated that zero pronouns differed significantly from both overt pronouns and full noun phrases. However, the difference between overt pronouns and full noun phrases did not reach statistical significance after correction. The effect size, as measured by Epsilon squared ($\epsilon^2 = 0.213$), suggests a moderate-to-large relationship between referring expression form and word order.

In summary, the relationship between referring expression form (`rform`) and relative word order position (`rwo`) reveals a systematic pattern that reflects the syntactic distribution of referential forms. As illustrated in the figure, *zero pronouns* predominantly occur at the beginning of the sentence (i.e., low `rwo` values), whereas **full noun phrases (fullNPs)** tend to appear in relatively later positions. *Overt pronouns* show an intermediate distribution, occupying positions between zeros and full noun phrases.

5.10 Discourse Density

In addition to linear positioning measures such as word order, discourse-level density metrics offer critical insights into the contextual conditions that shape referential choices. In particular, referential density reflects the extent to which an entity is mentioned or remains salient across discourse segments. Higher density typically indicates increased discourse prominence or competition, which may influence the form of referring expressions used by speakers.

In this thesis, four complementary density measures were calculated. The first two metrics, `mdensity_a` and `mdensity_as`, capture the global density of mentions and mention-bearing sentences, respectively, since the first mention of the entity. These measures reflect cumulative discourse salience and accessibility. In contrast, `mden-`

sity_m and *mdensity_s* provide local estimates of density based on the most recent discourse window (i.e., mentions or mention-bearing sentences within the last *n* sentences).

By examining the distribution of referring expression forms (*full noun phrases*, *overt pronouns*, and *zero pronouns*) across these density metrics, we aim to explore whether the choice of referring expression is sensitive to global and/or local discourse density. The following sections report the results of this analysis.

5.10.1 Global Discourse Density (*mdensity_a*)

This section examines how *mdensity_a* varies across different referring expression forms (*rform*), specifically full noun phrases (*fullnp*), overt pronouns (*overt*), and zero pronouns (*zero*). By investigating this relationship, we aim to better understand how the cumulative discourse prominence of a referent influences speakers' referential choices.

rform	n	Mean	SD	Med	Min	Skew	Kurt
fullnp	1375	0.27	0.34	0.145	0.00	1.701	3.139
overt	443	0.57	0.50	0.444	0.00	1.699	4.496
zero	978	0.59	0.49	0.500	0.02	1.622	4.156

Table 5.11: Descriptive statistics of *mdensity_a* by referring expression form (*rform*)

The descriptive statistics of *mdensity_a*, which reflects the global density of mentions for each referent since its first mention, revealed notable differences across referring expression forms. Zero pronouns and overt pronouns tend to occur in contexts characterized by higher mention density compared to full noun phrases. Specifically, the mean *mdensity_a* values for zero ($M = 0.59$) and overt pronouns ($M = 0.57$) were substantially higher than that of full noun phrases ($M = 0.27$). Similarly, median values followed this pattern, with zero pronouns having the highest median (0.50), overt pronouns in the middle (0.44), and full noun phrases the lowest (0.14). The relatively high skewness across all forms, particularly for full noun phrases and overt pronouns (both 1.7), suggests that *mdensity_a* distributions are right-skewed. This indicates that while most referring expressions occur at lower density levels, there is a non-negligible proportion produced under very high-density conditions.

The Kruskal-Wallis rank sum test indicated a statistically significant difference in antecedent mention density (*mdensity_a*) across referring expression forms, $\chi^2(2) = 490.21$, $p < .05$. The effect size, measured by Epsilon squared ($\epsilon^2 = 0.175$), suggests a moderate to large effect, indicating that the form of referring expression is meaningfully related to antecedent mention density. Post-hoc pairwise comparisons with adjusted *p*-values revealed that full noun phrases (full NPs) exhibited significantly lower *mdensity_a* values than both overt pronouns ($Z = -14.32$, $p < .05$) and zero pronouns ($Z = -20.57$, $p < .05$). In contrast, no significant difference was found between overt pronouns and zero pronouns ($Z = -1.37$, $p = .516$).

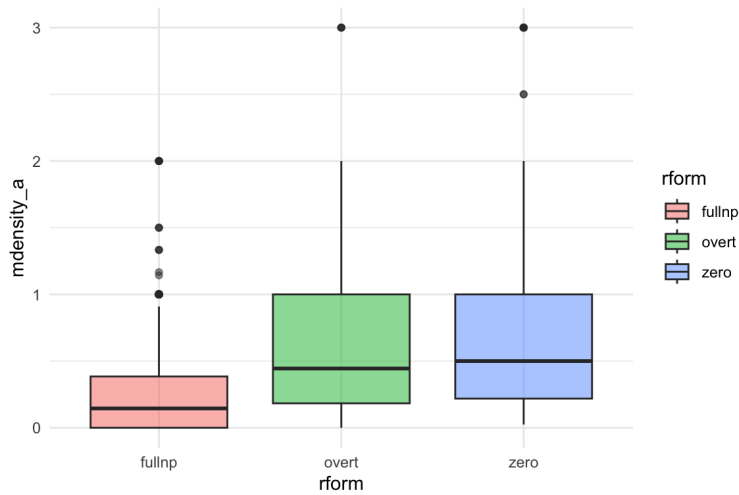


Figure 5.8: Distribution of `mdensity_a` across referring expression forms (`rform`)

Overall, these findings suggest that full noun phrases tend to be used in contexts with lower antecedent mention density, likely reflecting a need for higher informativeness and explicitness. In contrast, overt and zero pronouns are more frequent when antecedent mentions are more dense, indicating that pronouns are preferred when referents are already salient and easily retrievable from the discourse. Notably, overt and zero pronouns did not differ significantly from each other in terms of antecedent density, implying that other factors, such as syntactic or discourse constraints, may influence the choice between them.

5.10.2 Local Discourse Density (`mdensity_m`)

In addition to global measures of discourse density, the local density of mentions may play a critical role in shaping referential choice. To investigate this, we examined `mdensity_m`, which reflects the local discourse density calculated as the number of mentions within the last n sentences divided by n . This measure captures how frequently an entity has been referred to in the immediately preceding discourse and serves as an index of local salience. In this analysis, we explored the relationship between `mdensity_m` and the referring expression form (`rform`), comparing the distributions of this density measure across full noun phrases, overt pronouns, and zero pronouns. Descriptive statistics for each category are presented in Table 5.12.

Descriptive statistics for local discourse density `mdensity_m` across referring expression forms are presented in Table 5.12. Zero pronouns exhibited the highest local density values ($M = 0.44$, $SD = 0.26$), followed by overt pronouns ($M = 0.38$, $SD = 0.24$), while full noun phrases were associated with substantially lower local density ($M = 0.15$, $SD = 0.22$). The median values indicate that both overt and zero pronouns tend to occur when the local discourse contains relatively dense mention of the referent (Median = 0.40), whereas full noun phrases are more likely to occur in sparser local contexts (Median = 0.00). Moreover, the distributions show noticeable differences in skewness. Full noun phrases are positively skewed ($skew = 1.756$),

rform	n	Mean	SD	Med	Min	Max	Skew	Kurt
Full NP	1375	0.15	0.22	0.00	0.0	1.2	1.756	3.262
Overt Pronoun	443	0.38	0.24	0.40	0.0	1.2	0.547	0.122
Zero Pronoun	978	0.44	0.26	0.40	0.0	1.4	0.540	-0.060

Table 5.12: Descriptive statistics for local discourse density `mdensity_m` by referring expression form `rform`

reflecting frequent use in contexts with very low local mention density. In contrast, overt and zero pronouns display more symmetrical distributions ($skew \approx 0.5$), suggesting a more even spread across density values, though still favoring higher density contexts.

According to the Kruskal-Wallis test, the distributions differed significantly ($\chi^2 = 848.78$, $p < .05$, $\epsilon^2 = 0.304$), indicating a substantial effect size. Pairwise comparisons further revealed that both overt pronouns ($p < .05$) and zero pronouns ($p < .05$) were associated with significantly denser discourse environments compared to full noun phrases. Furthermore, zero pronouns also tended to occur in slightly denser contexts than overt pronouns, as evidenced by a significant albeit smaller difference ($p = .010$). These findings suggest that local discourse density is closely related to referential form selection: zero pronouns are preferentially used in highly dense discourse environments, followed by overt pronouns, while full noun phrases tend to be reserved for contexts with lower discourse density.

Overall, these findings are consistent with the expectation that pronouns are used preferentially when referents are more accessible and prominent in local discourse, whereas full noun phrases are favored in conditions of lower local mention density, where additional explicitness is beneficial for referent identification.

5.11 Feature Selection and Preparation

5.11.1 Overview

The process of selecting features for the XGBoost model was guided by theoretical, empirical, and statistical considerations. The ultimate goal was to identify a set of predictors that could best capture the cognitive and discourse factors underlying referential form selection, while avoiding redundancy, multicollinearity, and noise. The following criteria were applied systematically:

- **Theoretical relevance:** Features were retained if they aligned with models of reference production, such as accessibility, competition, and discourse salience.
- **Empirical findings:** Features showing statistically significant relationships with

referential form in exploratory analyses (e.g., chi-square, Kruskal-Wallis) were prioritized.

- **Statistical independence and efficiency:** Highly collinear variables and redundant measures were excluded to improve model interpretability and avoid issues in learning.

5.11.2 Included Features and Motivations

The final feature set for the XGBoost model comprised a diverse range of variables reflecting linguistic form, grammatical role, discourse structure, salience, and competition. The selection of these features was motivated by both statistical evidence and theoretical rationale, as outlined below.

Linguistic Form: `aform` Antecedent form was included due to its robust empirical association with referring expression choice (Cramér’s $V = 0.299$, $p < .05$). Theoretically, form continuity or shifts (e.g., NP $\rightarrow$ NP, NP $\rightarrow$ pronoun) play a crucial role in referential processing and production.

Grammatical Role: `arole` and `rrole` Although the observed association was moderate (Cramér’s $V = 0.12$), grammatical role is theoretically central to salience-based models. Subjecthood often enhances accessibility, favoring reduced forms.

Turn-Taking: `tt_mention` The turn-taking feature at the mention level was retained due to its significant and substantial association with referential form ($\chi^2(2) = 173.14$, $p < .05$, Cramér’s $V = 0.249$). Turn transitions influence the need for explicitness in reference.

Speaker Role: `speaker` Although the effect was smaller, speaker role reflects asymmetries in task-based language use. The presenter and operator roles introduce different discourse constraints that influence reference production.

Distance/Recency: `wdist` Word distance between antecedent and referent exhibited one of the strongest relationships with referring expression form ($\epsilon^2 = 0.252$). As discourse distance increases, speakers tend to choose more explicit referring forms.

Competition Metrics: `mcomp3` Competition metrics were carefully selected based on their theoretical relevance and empirical impact. `mcomp0` reflects local competition, `mcomp2` captures antecedent accessibility, and `mcomp3` quantifies intervening discourse complexity. Since `mcomp3` is correlated with other competition we included only it ($\epsilon^2 = 0.088$).

Discourse Salience: `mdensity_a`, `mdensity_m` Global (`mdensity_a`) and local (`mdensity_m`) density measures were retained based on their moderate to large effects on referential form ($\epsilon^2 = 0.175$ and 0.304 , respectively).

Word Order: `awo` Antecedent word order was included as it captures topicality and sentence-level salience effects. Empirical findings indicated moderate effects ($\epsilon^2 = 0.088$).

5.11.3 Excluded Features and Motivations

Several candidate features were excluded after careful evaluation to improve model parsimony and avoid redundancy.

Dialog Identifier: `dialog` This feature was excluded as it only indicated the conversation session. It does not have theoretical relevance to referential form selection and could introduce unwanted variance related to session-level idiosyncrasies.

Turn-Taking (Sentence Level): `tt_sentence` Although sentence-level turn-taking had a statistically significant relationship with referring form ($\chi^2(2) = 26.06$, $p < .05$), the effect size was small (Cramér's $V = 0.097$) and largely overlapped with the more powerful `tt_mention` feature.

Entity Competition: `ecomp0–ecomp3` Although entity competition measures were theoretically relevant, their extremely high correlations with mention-level competition measures (Spearman's $\rho > 0.75$) indicated redundancy. To avoid multicollinearity, only the strongest mention-based competition feature `mcomp3` was retained.

Additional Distance Metrics: `asent`, `rsent`, `sdist` These were excluded because they were strongly correlated with `wdist`, which was retained for its theoretical interpretability and strong empirical relationship with referring expression form.

5.11.4 Data Preparation for XGBoost Modeling

Following feature selection, data preparation steps were implemented to ensure compatibility with the XGBoost algorithm:

- **Categorical Variables:** `aform`, `arole`, `rrole`, `speaker`, and `tt_mention` were converted to factors and then one-hot encoded using the `model.matrix()` function.

- **Continuous Variables:** All numeric predictors were included directly without transformation.
- **Target Variable:** The dependent variable (`rform`) was encoded as an integer factor with three levels corresponding to full noun phrases (0), overt pronouns (1), and zero pronouns (2).
- **XGBoost Input Format:** The prepared predictor matrix and target vector were assembled into a `DMatrix` format for efficient modeling and memory management.

5.11.5 XGBoost for Reference Prediction

We selected XGBoost for the following reasons:

- **Handling Mixed Data Types:** XGBoost handles categorical (after encoding) and continuous features naturally and robustly.
- **Capturing Non-linear Relationships:** Reference selection is influenced by complex interactions among discourse, syntax, and salience; XGBoost is well-suited to learn such patterns.
- **Robustness to Class Imbalance:** The data has an imbalanced distribution across referential forms. XGBoost allows weighting and loss function tuning to handle this.
- **Interpretability via Feature Importance:** The algorithm provides interpretable feature importance measures, which can be aggregated to examine the relative influence of cognitive and linguistic predictors.

5.11.6 Summary of Included Features

To highlight the conceptual organization of the selected features, Table 5.13 presents the variables grouped by their theoretical motivations. This organization emphasizes the multidimensional nature of reference production, spanning cognitive, discourse, and interactional factors.

5.12 XGBoost Modeling and Results

Our dataset contains a diverse set of variables, including categorical features (e.g., `arole`, `rrole`), continuous features (e.g., `sdist`, `mrnk`), and competition metrics (`mcomp0`, `mcomp2`). We used XGBoost from the family of Gradient Boosting trees (T. Chen & Guestrin, 2016), which effectively handles such mixed data types and uncovers interactions between them. Reference selection involves non-linear relationships and complex interactions between linguistic, discourse, and contextual features. XGBoost

Table 5.13: Included Features for XGBoost Model

Theoretical Dimension	Di-	Feature	Empirical and Theoretical Motivation
Linguistic Form and Grammatical Role		aform	Antecedent form continuity or shift (Cramér’s $V = 0.299$)
		arole	Grammatical role salience (Cramér’s $V = 0.12$)
		rrole	Grammatical role salience (Cramér’s $V = 0.12$)
Turn-Taking and Role		tt_mention	New turn marking and referential explicitness (Cramér’s $V = 0.249$)
		speaker	Task-based role asymmetries (presenter vs operator)
Discourse	Dis-	wdist	Word distance between antecedent and referent ($\varepsilon^2 = 0.252$)
Competition and Discourse Crowding		mcomp3	Intervening discourse complexity ($\varepsilon^2 = 0.088$)
Discourse Density	Den-	mdensity_a	Global discourse density ($\varepsilon^2 = 0.175$)
		mdensity_m	Local discourse density ($\varepsilon^2 = 0.304$)
Word Order		awo	Antecedent sentence position ($\varepsilon^2 = 0.088$)

is very appropriate for modeling these relations through its boosted decision tree architecture, which builds successive trees to correct errors from previous ones. In our dataset, certain reference types (i.e., fullnps and zero pronouns) may occur more frequently than others. XGBoost includes mechanisms such as weighted loss functions to handle class imbalance, ensuring that less frequent classes are adequately represented in the model’s predictions. XGBoost provides detailed feature importance metrics, making it possible to identify the most influential variables in reference selection. These insights align with theoretical models of salience, topicality, and accessibility in linguistic research. XGBoost is optimized for both accuracy and computational efficiency.

5.12.1 Target Variable and Feature Selection

The goal of the present analysis was to model and predict the referential form selection in Turkish situated dialog. The dependent variable was the **Referential Form** (`rform`), consisting of three categories:

- **Full Noun Phrase (fullnp)**
- **Overt Pronoun (overt)**
- **Zero Pronoun (zero)**

The independent variables were selected based on theoretical motivations and empirical evidence. These variables represent discourse, cognitive, linguistic, and interactional dimensions:

- **aform**: Referential form of the previous mention (categorical)
- **arole**: Grammatical role of the antecedent (categorical)
- **rrole**: Grammatical role of the mention (categorical)
- **tt_mention**: Turn-taking status (within/across turns)
- **speaker**: Speaker identity (categorical)
- **wdist**: Word distance from the previous mention
- **mcomp3**: Competition metrics
- **mdensity_a**, **mdensity_m**: Density and persistence metrics
- **awo**: Sentential position of the antecedent

Categorical variables were converted to factors and then one-hot encoded for processing via XGBoost.

5.12.2 Modeling with XGBoost

The referential form prediction was conducted using an XGBoost classifier with the `multi:softprob` objective and `mlogloss` evaluation metric. The model was optimized using 5-fold cross-validation and early stopping. The final model was trained with the optimal number of rounds.

Overall Performance

- **Accuracy:** 70% (95% CI: 66.4% – 74.1%)
- **Kappa:** 0.50
- **P-value (Accuracy > No Information Rate):** < .05
- **McNemar's Test:** $p < .05$

Confusion Matrix

Table 5.14: Confusion Matrix

	FullNP	Overt	Zero
Predicted FullNP	219	23	29
Predicted Overt	6	16	8
Predicted Zero	44	55	158

Class-wise Statistics

- **FullNP:** Sensitivity = 0.81, Balanced Accuracy = 0.81
- **Zero Pronoun:** Sensitivity = 0.81, Balanced Accuracy = 0.76
- **Overt Pronoun:** Sensitivity = 0.17, Balanced Accuracy = 0.57

The model successfully captured the patterns for *full noun phrases* and *zero pronouns*, whereas *overt pronouns* remained more challenging to predict.

5.12.3 Feature Importance

Feature importance analysis was conducted using Gain scores from the trained model. Since one-hot encoding introduced multiple dummy variables, the importance scores were aggregated by parent features to improve interpretability.

The results indicated that discourse-related features were dominant:

- **Competition and Distance** emerged as the strongest predictors, supporting the view that referential choice is heavily influenced by discourse accessibility and the presence of competitors.
- **Linguistic factors** such as *word order* and *previous referential form* had moderate contributions to reference selection.
- **Interactional variables** such as *turn-taking* and *speaker identity* showed relatively low importance.

The XGBoost modeling results provide evidence that referential form selection in Turkish dialog is best predicted by discourse-level variables, particularly *competition* and *distance*. Linguistic factors play a secondary role, while interactional variables contribute minimally. These findings align with current theories that emphasize the role of discourse structure and accessibility in reference production.



CHAPTER 6

DISCUSSION AND CONCLUSION

This chapter summarizes the findings of the thesis and situates them within the broader theoretical and empirical landscape of reference selection research. The goal is to interpret the results, evaluate them in light of prior research, and discuss their implications for models of referential choice, particularly in morphologically rich, pro-drop languages such as Turkish. Furthermore, the chapter concludes with a critical reflection on the limitations of the study and avenues for future research.

6.1 Overview of Findings

The analyses presented in Chapter 5 collectively revealed that referential form selection in Turkish situated dialog is shaped by a complex interplay of linguistic, discourse, and interactional factors. Importantly, discourse-related variables, especially competition and recency measures, emerged as the most robust predictors of referential form. Global density and word order also influenced referential choices to a great extent. Grammatical roles and interactional factors such as speaker role and turn-taking exhibited relatively weak effects.

Across various analyses, the following key findings were observed:

- Full noun phrases were used predominantly in high-competition contexts, at longer discourse distances, and at turn boundaries, reflecting a need for explicitness and clarity.
- Zero pronouns were strongly associated with high discourse density, sentence-initial word order, and minimal competition, suggesting that reduced forms are favored when referents are salient and accessible.
- Overt pronouns were the least predictable form and occurred across intermediate values of competition, distance, and density, reflecting their pragmatic and context-sensitive usage.

6.1.1 Discourse Factors as Dominant Predictors

A central finding of the study was the dominance of discourse-level variables in predicting referential form. These include:

1. Competition measures, particularly total number of competitor mentions between the mention and antecedent (i.e. `mcomp3`)
2. Distance metrics (i.e. `wdist`)

They emerged as the most influential features in the XGBoost model. This result supports accessibility-based models of reference (Ariel, 1990; Gundel et al., 1993), which argue that referential choices are primarily determined by the cognitive status of discourse entities.

The complementary effects of competition and distance observed in this study provide further elaboration to these theories. While competition increased the likelihood of full noun phrases to resolve ambiguity, lower distance was associated with reduced forms, especially zero pronouns, suggesting that cumulative mention patterns enhance referent accessibility. This dual pattern requires models that distinguish between discourse accessibility and referential ambiguity as separate cognitive pressures on referential choice.

These results also align with Centering Theory (Grosz et al., 1995), where the attentional state of discourse is central. High mention density may signal continued focus on an entity (i.e., the center), licensing reduced forms such as zero pronouns. In contrast, high competition introduces referential ambiguity, which requires explicit marking, often through full noun phrases, reduce ambiguity which is in line with rational models (Frank & Goodman, 2012). This alignment with rational models underscores the adaptive nature of referential form selection, where speakers balance the need for efficiency with the demands of resolving ambiguity.

6.1.2 Grammatical Role and Word Order

Grammatical roles of mentions also influenced referential form, though with relatively weak effects. Subjects and subordinate subjects were more likely to be referred to using reduced forms, whereas objects and possessives tended to attract fuller forms. This finding aligns with traditional assumptions about the relationship between grammatical function and discourse prominence (Givón, 1983), yet the modest effect sizes indicate that influence of grammatical roles is not as strong as expected.

Similarly, word order effects revealed that antecedents occurring earlier in the sentence were more likely to be referred to by zero pronouns, while those occurring later were more often associated with full noun phrases. This pattern reflects the interaction between syntactic prominence and discourse accessibility, supporting proposals from status-based models (e.g., Centering Theory, Grosz et al., 1995) that sentence-initial positions are more accessible and favor reduced forms.

This pattern also aligns with the idea of syntactic prominence as a cue to topicality, as proposed in accessibility hierarchies (Ariel, 1990). In Turkish, the sentence-initial position—regardless of grammatical function—can mark topic-like status, especially in spoken discourse. This further explains the association between initial position and zero pronouns.

6.1.3 Interactional Factors: Limited but Present Effects

Speaker role and turn-taking variables exhibited relatively weak, yet statistically significant, effects on referential form choice. Full noun phrases were more frequent at turn boundaries, particularly when mentions occurred at the beginning of new turns. This pattern likely reflects the need to re-establish referential identity across speaker transitions, as predicted by discourse continuity models (Sacks et al., 1974).

Speaker role effects indicated that presenters produced more full noun phrases, whereas operators relied more on zero pronouns. This asymmetry reflects task-related pragmatic needs: presenters, responsible for instruction, prioritized explicitness, while operators, engaged in task execution, relied more on contextually inferred references. Nevertheless, the effect sizes were modest, suggesting that while interactional factors modulate referential choice, they do so in conjunction with discourse-level pressures.

These findings invite further integration with theories of grounding and common ground in dialog, where referential clarity is managed dynamically in relation to speaker roles and turn structure. Especially in goal-oriented interaction, speakers adapt their referring expressions not only to discourse salience but also to their current epistemic responsibilities.

6.1.4 Overt Pronouns as a Special Case

An important and recurrent pattern was the difficulty of predicting overt pronouns. Both statistical analyses and XGBoost modeling revealed that overt pronouns were less strongly associated with discourse and linguistic features compared to full noun phrases and zero pronouns. This suggests that overt pronouns in Turkish may fulfill specialized discourse functions, such as signaling contrast, emphasis, or disambiguation (Enç, 1986) that are not fully captured by general accessibility metrics. The relative unpredictability of overt pronouns points to the need for future research that explicitly models such pragmatic and interactional nuances.

These results challenge scalar accessibility models that assume a smooth mapping from salience to form. Instead, the behavior of overt pronouns supports a non-linear or categorical approach, where certain forms may be reserved for marked discourse purposes regardless of accessibility level. This lends support to hybrid or rational models (Frank & Goodman, 2012) that incorporate speaker goals and pragmatic inference.

6.1.5 Theoretical Implications

The findings of this study contribute to theoretical models of reference in several important ways:

- (1) a. First, the strong predictive power of discourse competition and distance measures supports models emphasizing referent accessibility and salience

- over purely syntactic or semantic accounts.
- b. Second, the partial but consistent effects of grammatical role and word order suggest that syntactic prominence interacts with, but does not override, discourse-level pressures.
- c. Third, the differentiated behavior of overt pronouns highlights the limitations of purely accessibility-based models and calls for the integration of pragmatic functions into theories of referential form selection.

These observations converge on a hybrid architecture of referential choice, where reference is jointly determined by salience, competition, grammatical structure, and pragmatic function. Models such as the Rational Speech Act framework provide a useful formalization of this view, integrating probabilistic reasoning about listener inferences with accessibility-based cues.

Taken together, these results support a hybrid view of referential choice in which discourse accessibility, referential competition, syntactic structure, and pragmatic functions jointly shape the selection of referring expressions.

6.2 Limitations and Future Directions

While the current study provides evidence on the determinants of referential form selection in Turkish situated dialog, several limitations require discussion.

- (2) a. The dataset, although rich and carefully annotated, is limited to task-oriented dialog and may not generalize to other genres, such as narratives or casual conversations.
- b. The feature set, while comprehensive, does not include measures of information structure (e.g., focus, contrast) or fine-grained pragmatic annotation, which may account for variation, especially in overt pronouns.
- c. While the task-oriented nature of the dialog provides ecological validity and controlled reference scenarios, it still differs from more open-ended or spontaneously evolving discourse. The artificiality of the computer-based setting may constrain certain interactive behaviors, particularly non-verbal cues such as gestures, which are known to play a role in reference production and resolution.
- d. The use of a computer interface excludes gestural and multimodal forms of reference, which may limit the applicability of the findings to real-world face-to-face interactions where deixis, pointing, and spatial referencing are available to the speaker.
- e. The dataset does not include proper names or indefinite noun phrases, and nearly all full noun phrases appear to be definite. This linguistic limitation constrains the generalizability of findings across referential types and prevents analysis of form choice in contexts where identifiability and definiteness vary.
- f. The category of null pronouns, while treated uniformly in the main analyses, likely comprises distinct phenomena, including subject pro-drop and object ellipsis. These should ideally be disaggregated to capture finer-

grained patterns in Turkish reference production, and their treatment as a single category may obscure important syntactic and discourse-level differences.

- g. Finally, the study does not model individual differences of speakers in referential behavior. Speakers may vary in their use of reduced vs. explicit forms due to cognitive, stylistic, or strategic preferences. While some variation may be absorbed by discourse features, capturing speaker-specific tendencies remains an open area for future research.
- h. The XGBoost model, while powerful, does not provide a generative account of reference production and is limited in its interpretability regarding underlying cognitive mechanisms.

In addition to expanding the dataset and annotations, future work could incorporate experimental methods (e.g., eye tracking or brain imaging) to better probe the cognitive mechanisms underlying referential choice. Computationally, hierarchical Bayesian models and RSA-based simulations can offer deeper insights into how speakers integrate multiple constraints in real-time reference planning.

Future work should address these limitations by incorporating additional pragmatic and information-structural variables, expanding the corpus to diverse discourse genres, and exploring more cognitively motivated modeling approaches, such as hierarchical Bayesian models or Rational Speech Act frameworks.

6.3 Conclusion

This thesis investigated the determinants of referential form selection in Turkish situated dialog using a combined approach of corpus annotation, statistical analysis, and machine learning. The results demonstrate that referential choice is primarily governed by discourse-level factors, particularly competition and salience, with syntactic and interactional variables playing secondary roles. While zero pronouns and full noun phrases conformed to accessibility-based predictions, overt pronouns emerged as a pragmatically flexible category requiring further investigation.

Overall, the findings contribute to an enriched understanding of referential choice in pro-drop languages and provide empirical support for hybrid models that integrate discourse, syntax, and pragmatics. By bringing together detailed corpus analysis with computational modeling, this study advances our knowledge of how speakers dynamically manage referential coherence in dialog and lays the groundwork for future research on reference in natural language use.



REFERENCES

- Ariel, M. (1988). Referring and accessibility. *Journal of linguistics*, 24(1), 65–87.
- Ariel, M. (1990). Accessing noun-phrase antecedents. *Linguistics*, 113–118.
- Ariel, M. (2001). Accessibility theory: An overview. In T. J. M. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 29–87). John Benjamins. <https://doi.org/10.1075/hcp.8.04ari>
- Arnold, J. E. (2010). How speakers refer: The role of accessibility. *Language and Linguistics Compass*, 4(4), 187–203.
- Arnold, J. E., & Zerkle, S. A. (2019). Why do people produce pronouns? pragmatic selection vs. rational models. *Language, Cognition and Neuroscience*, 34(9), 1152–1175.
- Arslan, T. P., Acar, K., & Eryiğit, G. (2023). Neural end-to-end coreference resolution using morphological information. *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, 34–40.
- Arslan, T. P., & Eryiğit, G. (2023). Incorporating dropped pronouns into coreference resolution: The case for Turkish. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 14–25.
- Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4), 555–596.
- Asher, N. (1993). *Reference to abstract objects in discourse*. Springer.
- Bourgonje, P., & Stede, M. (2020). The potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1061–1066.
- Brants, T. (2000). Inter-annotator agreement for a german newspaper corpus. *LREC*.
- Byron, D. K., & Fosler-Lussier, E. (2006). The osu quake 2004 corpus of two-party situated problem-solving dialogs. *LREC*, 395–400.
- Çallı, A. B. S. (2012). Demonstrative anaphora in Turkish: A corpus based analysis. *First workshop on language resources and technologies for turkic languages*, 33.

- Cancho, R. F. I., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3), 788–791.
- Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and points of view. In C. N. Li (Ed.), *Subject and topic* (pp. 25–55). Academic Press.
- Chafe, W. L. (1987). Cognitive constraints on information flow. In R. S. Tomlin (Ed.), *Coherence and grounding in discourse* (pp. 21–51, Vol. 11). John Benjamins.
- Chen, G., Same, F., & van Deemter, K. (2023). Neural referential form selection: Generalisability and interpretability. *Computer Speech & Language*, 79, 101466.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.
- Chomsky, N. (1981). *Lectures on government and binding*. Foris Publications.
- Chomsky, N., Steinberg, D., & Jakobovits, L. (1971). Deep structure, surface structure, and semantic interpretation. *1971*, 183–216.
- Clark, H. H. (1975). Bridging. *Theoretical issues in natural language processing*.
- Clark, H. H., & Brennan, S. E. (1991). Grounding in communication.
- Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge University Press.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Çöltekin, Ç., Doğruöz, A. S., & Çetinoğlu, Ö. (2023). Resources for Turkish natural language processing: A critical survey. *Language resources and evaluation*, 57(1), 449–488.
- Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, 9(1), 519–540.
- Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to “overinformative” referring expressions. *Psychological review*, 127(4), 591.
- Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., Stoutenborough, L., Kouril, M., Marsolo, K., Solti, I., et al. (2012). Building gold standard corpora for medical natural language processing tasks. *AMIA Annual Symposium Proceedings*, 2012, 144.

- Demir, Ş. (2023). Neural coreference resolution for turkish. *Journal of Intelligent Systems: Theory and Applications*, 6(1), 85–95.
- Di Eugenio, B. (1996). Centering in Italian. *arXiv preprint cmp-lg/9608007*.
- Di Eugenio, B., Jordan, P. W., Thomason, R. H., & D MOORE, J. (2000). The agreement process: An empirical investigation of human–human computer-mediated collaborative dialogs. *International Journal of Human-Computer Studies*, 53(6), 1017–1076.
- Dixon, R. M. (2003). Demonstratives: A cross-linguistic typology. *Studies in Language. International Journal sponsored by the Foundation “Foundations of Language”*, 27(1), 61–112.
- Doddington, G., Mitchell, A., Przybicki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004, May). The automatic content extraction (ACE) program – tasks, data, and evaluation. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of the fourth international conference on language resources and evaluation (LREC’04)* (pp. 837–840). European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf>
- Enç, M. (1986). Topic switching and pronominal subjects in Turkish. *Studies in Turkish linguistics*, 195, 209.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 378.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998. <https://doi.org/10.1126/science.1218633>
- Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, 75, 80–96. <https://doi.org/10.1016/j.cogpsych.2014.08.002>
- Friedrich, A., & Palmer, A. (2014). Centering theory in natural text: A large-scale corpus study.
- Garrod, S., & Sanford, T. (1988). Thematic subjecthood and cognitive constraints on discourse structure. *Journal of pragmatics*, 12(5-6), 519–534.
- Givón, T. (Ed.). (1983). *Topic continuity in discourse: A quantitative cross-language study* (Vol. 3). John Benjamins. <https://doi.org/10.1075/tsl.3>
- Götze, M., Weskott, T., Endriss, C., Fiedler, I., Hinterwimmer, S., Petrova, S., Schwarz, A., Skopeteas, S., & Stoel, R. (2007). Information structure. *Interdisciplinary studies on information structure: ISIS*, (7), 147–187.
- Grice, H. P. (1975). Logic and conversation. *Syntax and semantics*, 3, 43–58.

- Grosz, B. J., Joshi, A. K., & Weinstein, S. (1994). Centering: A framework for modelling the coherence of discourse.
- Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse (J. Hirschberg, Ed.). *Computational Linguistics*, 21(2), 203–225. <https://aclanthology.org/J95-2003/>
- Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3), 175–204.
- Gundel, J. K., Hedberg, N., & Zacharski, R. (2012). Underspecification of cognitive status in reference production: Some empirical predictions. *Topics in cognitive science*, 4(2), 249–268.
- Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69(2), 274–307. <https://doi.org/10.2307/416535>
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Routledge. <https://doi.org/10.4324/9781315836010>
- Haviland, S. E., & Clark, H. H. (1977). Comprehension and the given-new contract. In R. O. Freedle (Ed.), *Discourse production and comprehension* (pp. 1–40). Ablex.
- Hawkins, J. (2015). *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. Routledge.
- Heim, I. (1982). *The semantics of definite and indefinite noun phrases in English*. [Doctoral dissertation, University of Massachusetts, Amherst].
- Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.-M., Van Der Vloet, J., & Verschelde, J.-L. (2008). A coreference corpus and resolution system for dutch. *LREC*.
- Hirschman, L., Robinson, P., Burger, J., & Vilain, M. (1997). Automating coreference: The role of annotated training data. *Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, 118–121.
- Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, 44(4), 311–338.
- Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3), 296–298.
- Kamp, H., & Reyle, U. (1993). *From discourse to logic*. Kluwer.

- Karttunen, L. (1976). Discourse referents. In J. D. McCawley (Ed.), *Syntax and semantics 7: Notes from the linguistic underground* (pp. 363–385). Academic Press.
- Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, 25(1), 1–44. <https://doi.org/10.1093/jos/ffm018>
- Kehler, A., & Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1-2), 1–37.
- Krasavina, O., & Chiarcos, C. (2007). Pocos–potdam coreference scheme.
- Kripke, S. (1980). *Naming and necessity*. Blackwell.
- Kripke, S. A. (1972). Naming and necessity: Lectures given to the princeton university philosophy colloquium. In *Semantics of natural language* (pp. 253–355). Springer.
- Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30, 61–70.
- Labelbox. (2024). Labelbox [Online; accessed 2024].
- Lapshinova-Koltunski, E., & Ferreira, P. A. (2022). *Parcorfull2. 0: A parallel corpus annotated with full coreference*. Saarländische Universitäts-und Landesbibliothek.
- Le, H., Daryanto, T., Zhafransyah, F., Wijaya, D., Coppock, E., & Chin, S. (2022). Referring expressions with rational speech act framework: A probabilistic approach. *arXiv preprint arXiv:2205.07795*.
- Levy, R. (2018). Communicative efficiency, uniform information density, and the rational speech act theory.
- Levy, R. P., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, 849–856. <https://papers.nips.cc/paper/3129-speakers-optimize-information-density-through-syntactic-reduction>
- Li, C. (1976). Subject and topic: A new typology of language. *Subject and topic*.
- Mançe-Çalışır, Ö. (2018). *Geniş otizm fenotipi gösteren erişkinlerde sosyal biliş: Bir göz izleme çalışması [social cognition in adults with broad autism phenotype: An eye-tracking study]* [Doctoral dissertation, Institute of Health Sciences].
- Mitkov, R. (2014). *Anaphora resolution*. Routledge.

- Monroe, W., & Potts, C. (2015). Learning in the rational speech acts model. *arXiv preprint arXiv:1510.06807*.
- Nedoluzhko, A., Novák, M., Cinková, S., Mikulová, M., & Mírovský, J. (2016). Coreference in prague czech-english dependency treebank. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 169–176.
- Nissim, M., Dingare, S., Carletta, J., Steedman, M., et al. (2004). An annotation scheme for information status in dialogue. *LREC*.
- Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., & Zawisławska, M. (2016). Polish coreference corpus. *Human Language Technology. Challenges for Computer Science and Linguistics: 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7-9, 2013. Revised Selected Papers 6*, 215–226.
- Orita, N., Vornov, E., Feldman, N. H., & Daumé III, H. (2015). Why discourse affects speakers' choice of referring expressions. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, 1639–1649. <https://doi.org/10.3115/v1/P15-1158>
- Öztürk, B. (2004). *Case, referentiality and phrase structure*. Harvard University.
- Pamay, T., & Eryiğit, G. (2018). Turkish coreference resolution. *2018 Innovations in Intelligent Systems and Applications (INISTA)*, 1–7.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526–3529.
- Pimentel, T., Meister, C., Salesky, E., Teufel, S., Blasi, D., & Cotterell, R. (2021). A surprisal–duration trade-off across and within the world's languages. *arXiv preprint arXiv:2109.15000*.
- Poesio, M. (2004). The mate/gnome proposals for anaphoric annotation, revisited. *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, 154–162.
- Poesio, M., Cheng, H., Hitzeman, J., Stevenson, R., & Di Eugenio, B. (2002). A corpus-based evaluation of centering theory.
- Poesio, M., Cheng, H., Henschel, R., Hitzeman, J., Kibble, R., & Stevenson, R. (2000). Specifying the parameters of centering theory: A corpus-based evaluation using text from application-oriented domains. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 400–407.
- Poesio, M., Grishina, Y., Kolhatkar, V., Moosavi, N. S., Roesiger, I., Roussel, A., Simonjetz, F., Uma, A., Uryupina, O., Yu, J., et al. (2018). Anaphora resolution

with the arrau corpus. *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, 11–22.

- Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012, July). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In S. Pradhan, A. Moschitti, & N. Xue (Eds.), *Joint conference on EMNLP and CoNLL - shared task* (pp. 1–40). Association for Computational Linguistics. <https://aclanthology.org/W12-4501>
- Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics* (pp. 223–255). Academic Press.
- Qing, C., & Franke, M. (2015). Variations on a bayesian theme: Comparing bayesian models of referential reasoning. *Bayesian natural language semantics and pragmatics*, 201–220.
- Riester, A., & Baumann, S. (2017). The reflex scheme-annotation guidelines.
- Rohde, H., & Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, 29(8), 912–927. <https://doi.org/10.1080/01690965.2013.854918>
- Russell, B. (1905). On denoting. *Mind*, 14(56), 479–493.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4), 696–735.
- Sang, E. F., & De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.
- Schüller, P., Cingilli, K., Tunçer, F., Sürmeli, B. G., Pekel, A., Karatay, A. H., & Karakaş, H. E. (2017). Marmara Turkish coreference corpus and coreference resolution baseline. *arXiv preprint arXiv:1706.01863*.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), 379–423.
- Spanger, P., Yasuhara, M., Iida, R., & Tokunaga, T. (2009). A japanese corpus of referring expressions used in a situated collaboration task. *Proceedings of the 12th European workshop on natural language generation (ENLG 2009)*, 110–113.
- Spanger, P., Yasuhara, M., Iida, R., Tokunaga, T., Terai, A., & Kuriyama, N. (2012). Rex-j: Japanese referring expression corpus of situated dialogs. *Language Resources and Evaluation*, 46(3), 461–491.
- Stegemann-Philipps, C., Butz, M. V., Winkler, S., & Achimova, A. (2021). Speakers use more informative referring expressions to describe surprising events. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43).

- Stoia, L., Shockley, D. M., Byron, D. K., & Fosler-Lussier, E. (2008). Scare: A situated corpus with annotated referring expressions. *LREC*.
- Strube, M., & Hahn, U. (1999). Functional centering–grounding referential coherence on information structure. *Computational linguistics*, 25(3), 309–344.
- Taulé, M., Martí, M. A., & Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. *Lrec*, 2008, 96–101.
- Tetreault, J. R. (2001). A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4), 507–520.
- Tüfekçi, P., & Kılıçaslan, Y. (2007). A computational model for resolving pronominal anaphora in Turkish using hobbs-naive algorithm. *International Journal of Computer and Information Engineering*, 1(5), 1402–1405.
- Turan, Ü. D. (1995). *Null vs. overt subjects in Turkish discourse: A centering analysis*. University of Pennsylvania Pennsylvania.
- Uryupina, O., Artstein, R., Bristot, A., Cavicchio, F., Delogu, F., Rodriguez, K. J., & Poesio, M. (2020). Annotating a broad range of anaphoric phenomena, in a variety of genres: The arrau corpus. *Natural Language Engineering*, 26(1), 95–128.
- van Deemter, K., & Kibble, R. (2000). On coreferring: Coreference in muc and related annotation schemes. *Computational linguistics*, 26(4), 629–637.
- Vincze, V., Hegedűs, K., Sliz-Nagy, A., & Farkas, R. (2018). Szegedkoref: A hungarian coreference corpus. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Walker, M., Iida, M., & Cote, S. (1996). Japanese discourse and the process of centering. <https://arxiv.org/abs/cmp-lg/9609006>
- Walker, M. A., Joshi, A. K., & Prince, E. F. (1998). Centering in naturally occurring discourse: An overview. *Centering theory in discourse*, 128.
- Webber, B. (1988). Discourse deixis: Reference to discourse segments. *26th Annual Meeting of the Association for Computational Linguistics*, 113–122.
- Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., et al. (2013). Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23, 170.
- Weischedel, R., Pradhan, S., Ramshaw, L., Palmer, M., Xue, N., Marcus, M., Taylor, A., Greenberg, C., Hovy, E., Belvin, R., et al. (2011). Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, 17.

- Yıldırım, S., & Kılıçaslan, Y. (2006). A machine learning approach to personal pronoun resolution in Turkish. *Computational Linguistics*, 27(4), 521–544.
- Yıldırım, S., Kılıçaslan, Y., & Aykaç, R. E. (2004). A computational model for anaphora resolution in Turkish via centering theory: An initial approach. *International Conference on Computational Intelligence*, 124–128.
- Yu, J., Moosavi, N. S., Paun, S., & Poesio, M. (2021). Stay together: A system for single and split-antecedent anaphora resolution. *arXiv preprint arXiv:2104.05320*.
- Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31), 7937–7942.
- Zeldes, A. (2017). The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3), 581–612.
- Zeyrek, D., Demirsahin, I., Sevdik-Calli, A. B., Balaban, H. Ö., Yalçinkaya, İ., & Turan, U. D. (2010). The annotation scheme of the turkish discourse bank and an evaluation of inconsistent annotations. *Proceedings of the fourth linguistic annotation workshop*, 282–289.
- Zeyrek, D., Demirşahin, I., & Çallı, A. B. S. (2013). Turkish discourse bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue & Discourse*, 4(2), 174–184.
- Zhu, Y., Peng, S., Pradhan, S., & Zeldes, A. (2023). Incorporating singletons and mention-based features in coreference resolution via multi-task learning for better generalization. *arXiv preprint arXiv:2309.11582*.
- Zipf, G. K. (1935). *The psycho-biology of language*. Houghton Mifflin.



Appendix A

ANNOTATION SCHEME

The objective of this annotation scheme is to annotate referring expressions which co-refer to the same referent in Turkish.

A.1 Task

Our dialogs were taken from an experimental setting where pairs were expected to solve tangram puzzles (Mançe-Çalışır, 2018). The task requires them to build a target shape by manipulating seven geometric shapes through a computer simulator. They are seated face to face and perform the tasks through shared screens. The separator between the tables prevents them from seeing each other. They are assigned specific roles, which aims to promote real-life language production. The presenter has access to the target shape and is expected to give instructions to the other participant about how to build the shape and the operator cannot see the shape but has control over the mouse to manipulate the geometric shapes to achieve the goal.

A.1.1 Roles of discourse participants:

- **Presenter:** The participant who has access to the target shape and gives instructions to build the shape.
- **Operator:** The participant who has control over the mouse which manipulates the shapes in the working shapes to build the target shape.

A.1.2 Working space

A.1.2.1 Target shapes:

A.1.2.2 Shapes and their anchors

- Small triangles: pinktr, and greentr
- Middle triangle: yellowtr
- Big triangles: bluetr, and redtr

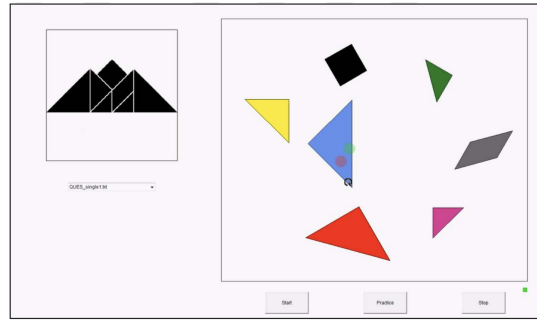


Figure A.1: Working space

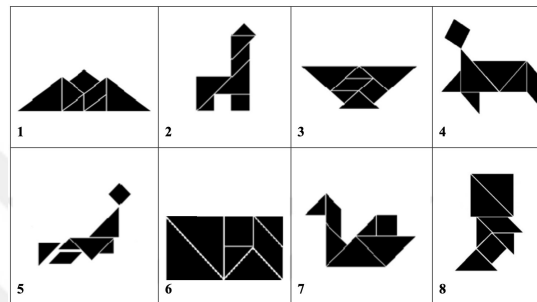


Figure A.2: Target Shapes

- Square: blacksq
- Paralelogram: greyp

A.2 Markables

A referring expression or mention qualifies as a *markable* only if it is:

- case-marked (i.e. referential)
- part of a coreferential chain.

A.2.1 Span boundary

Span types:

- **Overt mentions:** We annotate the full span of the overt entities excluding case marker if there is any attached to the expression.
- **Null mentions:**

Subjects: The predicate or the predicate to which the agreement marker is attached to.

Objects: The space before the word with the agreement marker of the predicate.

A.2.2 Linguistic form of candidate markables

- Indefinite descriptions:
Bir kitap okuyorum.
- Definite descriptions:
Kitabı okuyorum.
- Bare nouns:
Kitap okuyorum.
- Named entities:
For example: ODTÜ, Türkiye, Orhan Veli
- Pronouns:
O, Onlar, Siz ...
- Null Elements:
Adam eve geldi. [] Yorgundu.
- Suffixes:
Possessive suffixes: Adam-ın ev[i], Adam-ın okuduğ[u] kitap
- Agreement markers:
Okula gidiyor[lar].
- Converbs:
gelirken, gidince

A.3 Non-markables

We do not annotate the following categories:

- Singletons: Discourse referents which occur only once throughout the text and do not participate in a coreferential or bridging chain.
- Idiomatic expressions:
Damlaya damlaya göl olur.

- Predicatives
Ali [öğretmen] olmuş.
Buna [ev] denir.
- Abstract Entities
Herkes birlikte çalıştı. [Bu] bizi başarıya ulaştırdı.

A.4 Steps

- Check the noun phrases and agreement morphology: Referential or not?
- Check the entity: Singleton, Part of a coreferential chain
- Check the sentence: Matrix and/or Subordinate
- Check the role of the entity: Subject, Object, Gen-poss, etc.

A.5 Tags

- Span tags:
 - NOM:
 - PRED:
 - EXO:
- Link tags:
 - SUBJ:
 - OBJ:
 - SSUBJ:
 - SOBJ:
 - POSS:
 - COREF:

A.6 How to annotate overt entities

A.6.1 Mentions in matrix clauses

We annotate the mentions in the matrix clauses which corefer to a previous mention in the following ways:

If the mention is a **subject** of a matrix sentence, mark the whole NP as NOM and then link it with the nearest antecedent with the tag SUBJ.

[Ali] geldi. [O] burda.

nom ←subj nom

If the mention is a **object** of a matrix sentence, mark the whole NP as NOM (excluding case marker if any) and then link it with the nearest antecedent with the tag OBJ

[Ali] geldi. [O]nu gördüm.

nom ←obj nom

A.6.2 Subordinate clauses

We annotate the overt mentions in the subordinate clauses which corefer to a previous mention in the following ways:

If the mention is a subject of a syntactically uncontrolled subordinate clause, mark the whole NP as NOM excluding case marker if any and then link it with the nearest antecedent with the tag SSUBJ.

[Ali] gelmiş. [O] gidince çıkarız.

nom ←ssubj nom

If the mention is a subject of a subordination clause with genitive-possessive construction, mark the possessive suffix of the clause as PRED and the possessor as NOM and then link the suffix to the possessor with SSUBJ.

[Ali]'nin gelme[si]ni bekliyor.

nom ←ssubj pred

If the mention is a subject of a relative clause with genitive-possessive construction, mark the possessive suffix of the clause as NOM and the possessor as NOM and then link the suffix to the possessor with SSUBJ.

[Ali]'nin okuduğ[u] kitap eski.

nom ←ssubj nom

If the mention is an object of a subordinate clause, mark the whole NP as NOM (excluding case marker if any) and link it with the nearest antecedent and with the tag SOBJ.

[Kitap] güzel. Ali'nin [o]nu okuduğunu biliyorum.

nom ←sobj nom

If the mention is a possessed of a genitive-possessive phrase, mark the possessive suffix of the phrase as NOM and the possessor as NOM and then link the suffix to the possessor with the tag POSS.

[Ali]'nin kitab[ı]

nom $\leftarrow$ poss nom

A.7 Null entities

A.7.1 Mentions in matrix clauses

We annotate the null mentions in the matrix clauses which corefer to a previous mention in the following ways:

If the mention is a **subject** of a matrix sentence, mark the word with the agreement marker of the predicate as PRED and then link it with the nearest antecedent with the tag SUBJ.

[Ali] geldi. Kitabı [getirdi].

nom $\leftarrow$ subj pred

[Ali] geliyor. Kitabı [getiriyor].

nom $\leftarrow$ subj pred

[Ali] geldi. İyi [çocuk].

nom $\leftarrow$ subj pred

[Ali] geldi. İyi çocuk [değil].

nom $\leftarrow$ subj pred

[Ali] gelmedi. Burda [yok].

nom $\leftarrow$ subj pred

[Sen] gelirsin. Kitabı da getirir [misin]?

nom $\leftarrow$ subj pred

If the mention is an **object** of a matrix sentence, mark the space before the predicate as NOM and then link it with the nearest antecedent with the tag OBJ.

Ali [kitab]ı getirdi. Kütüphaneye[]bıraktı.

nom $\leftarrow$ obj nom

A.7.2 Subordinate clauses

We annotate the null mentions in the subordinate clauses which corefer to a previous mention in the following ways:

If the mention is a subject of a syntactically uncontrolled subordinate clause, mark the

predicate as PRED and then link it with the nearest antecedent with the tag SSUBJ.

[Ali] geldi. Kitabı [bırakınca] çıktık.

nom ←ssubj pred

If the mention is a subject of a subordination clause with genitive-possessive construction, mark the possessive suffix of the clause as PRED and link the suffix to the nearest possessor with th tag SSUBJ.

[Ali] geldi. Kitabı getirme[si]ni bekliyorduk. nom ←ssubj pred

If the mention is a subject of a relative clause genitive-possessive, mark the possessive suffix of the clause as NOM and then link the suffix to the possessor with the tag SSUBJ.

[Ali] geldi. Getirdiğ[i] kitabı bekliyorduk. nom ←ssubj nom

If the mention is an object object of a subordinate clause, mark the space before the predicate as NOM and then link it with the nearest antecedent and with the tag SOBJ.

Ali [kitab]ı getirdi. Kütüphaneye[]bırakmasını bekliyorduk. nom ←sobj nom

If possessed of a genitive-possessive phrase, mark the possessive suffix of the clause as NOM and then link the suffix to the nearest antecedent with POSS

[Ali] geldi. Kitab[ı] güzel. nom ←poss nom

If the relation between an anaphor and antecedent is none of the above, link the anaphor to the antecedent with the tag COREF.

[Ali] gelecek. Bize [o]nun kitabı lazım. nom ←coref nom



CURRICULUM VITAE

FARUK BÜYÜKTEKİN

EDUCATION

Degree	Institution	Year of Graduation
Ph.D. in Cognitive Science	METU, Ankara	2025
M.Sc. in Cognitive Science	METU, Ankara	2018
B.A. in Foreign Language Education	Boğaziçi University, İstanbul	2007

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2025 – ongoing	METU, Ankara	Scientific Researcher
2023 – 2024	University of Cologne, Germany	Visiting Researcher
2019 Fall	University of Wolverhampton, UK	Visiting Student
2015 – 2017	Google NLU Project	Linguist
2014 – 2015	METU Turkish Discourse Bank	Annotator

PUBLICATIONS

- Büyüktekin, F., & Özge, U. (2024, August). A coreference corpus of Turkish situated dialogs. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)* (pp. 42-52).
- Büyüktekin, F., Çakır, M.P. & Acartürk, C. (2020). Analysis of Turkish referring expressions in a situated dialog context: The case of nominal demonstratives. In D. Zeyrek & U. Özge (Eds.), *Discourse Meaning: The View from Turkish* (pp. 103-136). Berlin, Boston: De Gruyter Mouton.