

DE RE DE DICTO AMBIGUITIES IN TURKISH: DISAMBIGUATION BY LLMS
AND PROCESSING BY HUMANS

BÜŞRA MARŞAN

BOĞAZİÇİ UNIVERSITY

2023

DE RE DE DICTO AMBIGUITIES IN TURKISH: DISAMBIGUATION BY LLMS
AND PROCESSING BY HUMANS

Thesis submitted to the
Institute for Graduate Studies in Social Sciences
in partial fulfillment of the requirements for the degree of

Master of Arts
in
Linguistics

by
Büşra Marşan

Boğaziçi University
2023

De Re De Dicto Ambiguities in Turkish: Disambiguation by LLMs and Processing
by Humans

The thesis of Büşra Marşan

has been approved by:

Assist. Prof. Ümit Atlamaz
(Thesis Advisor)

Assist. Prof. Ömer Demirok
(Thesis Co-advisor)

Prof. Balkız Öztürk

Prof. Tunga Güngör

Prof. Olcay Taner Yıldız
(External Member)

June 2023

DECLARATION OF ORIGINALITY

I, Büşra Marşan, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisors and thesis committee at Boğaziçi University, including final revisions required by them.

Signature:

Date:

ABSTRACT

De Re De Dicto Ambiguities in Turkish: Disambiguation by LLMs and Processing by Humans

This study aims to investigate De Re De Dicto ambiguities in Turkish from different points of views: Formal semantics, computational linguistics and psycholinguistics.

From the formal semantics perspective, a survey of different accounts that aim to explain compositional semantics of De Re De Dicto ambiguities is provided. With free word order and optionally non-overt accusative marker interacting with information structure, Turkish poses an interesting case for De Re De Dicto ambiguities as their resolution is closely tied to many factors: Discourse and non-discourse elements in the conversational context, information structure, world knowledge, and belief states of the utterer.

From the computational linguistics perspective, a challenge data set in Turkish for natural language inference (NLI) is offered to test swiftly growing large language models' (LLMs) abilities in recognizing syntactic and morphological cues along with the world knowledge. For this purpose, a seed data of 743 sentence triplets were manually created and annotated by native speakers: One potentially De Re De Dicto ambiguous sentence, one De Re paraphrase and one De Dicto paraphrase. Then this seed data was augmented using a natural language augmentation algorithm specifically created for this task. An experiment was conducted to assess probing abilities of CNLI-TR and compare it with NLI-TR, the only other available NLI data set for Turkish.

From the psycholinguistics perspective, a self-paced reading experiment was conducted to investigate whether theoretical claims that De Re is "harder" to compositionally derive hold from a sentence processing perspective. Following disambiguating contexts, participants were shown De Re De Dicto ambiguous sentences and were asked comprehension questions targeting specific readings parallel to disambiguating contexts. Reading time of De Re De Dicto ambiguous sentences and response time of comprehension questions were compared across

disambiguating contexts to explore whether there was a significant and credible processing difference.



ÖZET

Türkçe’de De Re De Dicto Muğlaklıklarının Dil Modelleri Tarafından Çözümlemesi ve İnsanlar Tarafından Anlaşılması

Bu çalışma, Türkçedeki De Re De Dicto muğlaklıklarını farklı açılardan incelemeyi amaçlamaktadır: Biçimsel anlambilim, hesaplama dilbilim ve psikodilbilim.

Biçimsel anlambilim perspektifinden, De Re De Dicto belirsizliklerinin bileşimsel anlamlarını açıklamayı amaçlayan farklı açıklamaların bir araştırması sağlanmaktadır. Serbest kelime sırası ve isteğe bağlı olarak bilgi yapısıyla etkileşime giren açık olmayan belirteç belirteci ile Türkçe, De Re De Dicto belirsizlikleri için ilginç bir durum teşkil ediyor çünkü bunların çözümü birçok faktöre yakından bağlı: Konuşma bağlamında söylem ve söylem dışı unsurlar, bilgi yapısı , söyleyenin dünya bilgisi ve inanç halleri.

Hesaplama dilbilim açısından, hızla büyüyen büyük dil modellerinin (LLM’ler) dünya bilgisi ile birlikte sözdizimsel ve morfolojik ipuçlarını tanımadaki yeteneklerini test etmek için doğal dil çıkarımı (NLI) için Türkçe bir meydan okuma veri seti sunulmaktadır. Bu amaçla, 743 cümle üçlüsünün çekirdek verileri manuel olarak oluşturulmuş ve anadili İngilizce olan kişiler tarafından açıklanmıştır: Bir potansiyel olarak De Re De Dicto belirsiz cümle, bir De Re tefsiri ve bir De Dicto tefsiri. Daha sonra bu tohum verileri, bu görev için özel olarak oluşturulmuş bir doğal dil büyütme algoritması kullanılarak artırıldı. CNLI-TR’nin yoklama yeteneklerini değerlendirmek ve Türkçe için mevcut diğer tek NLI veri seti olan NLI-TR ile karşılaştırmak için bir deney yapılmıştır.

Psikolinguistik bakış açısından, De Re’nin bir cümle işleme perspektifinden bileşimsel olarak tutunmasının "daha zor" olduğuna dair teorik iddiaların olup olmadığını araştırmak için kendi hızınızda bir okuma deneyi yapıldı. Belirsizliği giderici bağlamların ardından, katılımcılara De Re De Dicto belirsiz cümleleri gösterildi ve belirsizliği giderici bağlamlara paralel olarak belirli okumaları hedefleyen anlama soruları soruldu. De Re De Dicto belirsiz cümlelerin okuma süresi

ve anlama sorularının yanıt süresi, anlamlı ve güvenilir bir işleme farkı olup olmadığını keşfetmek için belirsizliği giderici bağlamlarda karşılaştırıldı.



ACKNOWLEDGEMENTS

There is only one name written on the cover of each thesis, yet it is a team that stands behind every single one, offering unwavering support, invaluable guidance, and endless motivation to the author. This section is deeply personal to me as I strive to convey heartfelt appreciation to this incredible team. Ironically, despite my background as a linguist and translator, I'm at a loss for words that can truly capture the extent of my gratitude.

Each member of this team has made such incredible contributions to this thesis that I can't help but feel like I'm just the one pushing the buttons, responsible for typing out this massive study. This gargantuan study came to life thanks to the hard work and dedication of many people. Even though it might make the acknowledgments section longer than my actual thesis conclusion, I'm determined to personally thank every single person on the team as I just can't shake off the feeling that I'd be doing them a disservice if I didn't show my gratitude individually.

First and foremost, I want to express my heartfelt gratitude to my advisors, Dr. Ümit Atlamaz and Dr. Ömer Demirok, a million times over. Motivating someone, especially in the realm of academia where imposter syndrome looms large, is no small feat. They excelled not only in this aspect, but I consider myself incredibly fortunate that they mastered an even more challenging task that requires a lot of tact: The art of knowing precisely when to step in and hit the brakes. They skillfully prevented my perfectionist tendencies from pushing me into burnout territory.

Dear professors, thank you for being my pillars of support, lifting me up every time I descended into the abyss of self-doubt. Your gentle guidance was invaluable, knowing just when to intervene without stifling my potential for growth. I am truly grateful for your unwavering encouragement and your ability to strike the delicate balance between pushing me forward and safeguarding my well-being. Your unmatched patience and valuable guidance made this thesis possible. I will be forever grateful for your putting up with my meltdowns. Thank you for always being there for me as an academic advisor and thank you for your very valuable friendship. Your

wisdom and insight have been an unwavering source of inspiration, and I am truly fortunate to have had the opportunity to be your student.

During both my MA journey and the creation of this thesis, I consistently experienced the unwavering support of the Department of Linguistics and its esteemed professors. I am deeply thankful to all of them, and I want to extend a special appreciation to Balkız Öztürk, Pavel Logačev, Elena Guerzoni, and Kadir Gökgöz. Moreover, I would like to express my sincere gratitude to Tunga Güngör and Suzan Üsküdarlı from the Department of Computer Engineering, as well as Ceyda Arslan Kechriotis from Department of Turkish Language and Literature, for their invaluable contributions and support.

Another individual who dedicated a significant amount of time to this thesis and has been an incredible friend throughout the process is Utku Turk. I've reached a point where I'm tempted to devise a clever workaround to navigate the formatting restrictions set by the SBE just so I can sneak in an ASCII portrait of you somewhere within the manuscript, as a thank you note.

This research has been made possible through the generous grant of Boğaziçi University Dil Uygulama ve Araştırma Merkezi, and the unwavering dedication of my team of contributors, who generously shared their ideas, provided insightful annotations, and offered invaluable insights. I extend my heartfelt gratitude to Aslı Kuzgun, Ceren Oksal, Merve Doğan, and Serra Gök for their considerable time and effort invested in this endeavor.

I want to express my sincere thanks to the following individuals, each of whom has played a unique and invaluable role in my journey: Fatma Timuçin, who has been by my side every day since 2010, patiently listening to my enthusiastic ramblings about linguistics, and providing unwavering support. Arda Korkmaz, whose exceptional coding skills and emotional support have been instrumental in this endeavor. Ayda Abay, for her consistent encouragements throughout. Furkan Dikmen, not only for his friendship but also for his deep insights into semantics. Alp Ünal Ayhan and Aybüke Caner, who have provided uplifting pep talks at crucial

moments. Efe Güleler, for making his presence felt even across geographical distances. Olgun Dursun, who shared the painstaking journey of thesis writing alongside me, providing mutual support during challenging times. Last but not least, my kittens Mars and Venüs, who have brought joy and comfort as absolute babes and cutie pies.

Lastly, I wish to extend my deepest gratitude to my beloved family. Your unwavering patience, boundless support, and selfless sacrifices have propelled me to reach this point. I recognize how truly fortunate I am to have you all as my family. Thank you for being the foundation upon which I've built my journey. Annem, thank you for always embracing and celebrating the wonderfully unique and slightly mad individual that I am. You are behind everything that is good about me, and you are my favourite mum. Ayaya, your resolute determination is a testament to your remarkable strength. I'm grateful for inheriting your tenacity. Kedicim, my spirited and fervent supporter, your unwavering belief in me has been a driving force. Even when I doubted myself, you never did, and I am profoundly thankful for your enduring faith in my capabilities.

I dedicate this thesis to my oldest and dearest friend, Ahmet. Growing up by your side was one of the greatest gifts I have ever received. When I was writing this thesis, I was repeatedly told that things I wanted to cover would make at least two MA theses. Well, the joke is on them, that was the point all along. This one is for both of us! Thank you for always serving as my moral compass, even from afar. The void left by your absence will persist until I draw my very last breath.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION.....	1
1.1 De Re De Dicto ambiguities	2
1.2 Natural language inference	21
1.3 The scope of this thesis.....	30
1.4 Packages and technical details	30
1.5 Overview	31
1.6 Data availability	31
CHAPTER 2: DATA	32
2.1 Creating the seed data.....	33
2.2 Adversities and concerns	37
2.3 Statistics.....	46
2.4 CNLI-TR: What is the challenge?	48
2.5 Further improvements.....	49
CHAPTER 3: AUGMENTATION PIPELINE	51
3.1 Data augmentation strategies.....	51
3.2 Augmentation Algorithm	56
3.3 Validation and Post Editing	59
CHAPTER 4: COMPUTATIONAL EXPERIMENTS	61
4.1 LLM experiments	61
4.2 Random Forest and KNN experiments	68
4.3 Discussion and further research	69
CHAPTER 5: PROCESSING EXPERIMENT	71
5.1 Experiment design.....	72
5.2 Analysis and results.....	80
5.3 Discussion and further research	90
CHAPTER 6: CONCLUSION	93
APPENDIX A: Annotator Guidelines	95

APPENDIX B: Annotatoion Interfaces	103
APPENDIX C: Processing Experiment Materials	106
APPENDIX D: Processing Experiment Interface	112
REFERENCES	121



ABBREVIATIONS

3	third person
ABL	ablative
ACC	accusative
DAT	dative
GEN	genitive
GER	gerund
IMP	imperative
INF	infinitive
INS	instrumental
LOC	locative
NMLZ	nominalizer
NOM	nominative
PLU	plural
POSS	possessive
PROG	progressive
PRS	present
PST	past
SG	singular

LIST OF TABLES

Table 1. Specificity and Definiteness in Turkish as Described by kechriotis2009referentiality	16
Table 2. An illustration of different NLI relations	21
Table 3. A Comparison of Large NLI Resources	29
Table 4. Distribution of Gender and Etymological Origin of the Proper Nouns in the Data	39
Table 5. Inference Label Distribution of CNLI-TR	50
Table 6. Noise Creating Operations Proposed in EDA wei-zou-2019-eda	56
Table 7. Augmentation Operations Performed on Each Split	58
Table 8. Changes Made to the Inference Labels in Validation Process	59
Table 9. Inference Label Distribution in Train, Dev and Test Splits of CNLI-TR ..	61
Table 10. Precision, Recall and F1 Scores of Three Language Models Fine Tuned on CNLI-TR Dataset	62
Table 11. Comparison of Weighed Class LLM Experiments with Morphological Analysis	64
Table 12. KNN and RF Performance on NLI	69
Table 13. Comprehension Question Accuracy for Filler Items and Experiment Items	82
Table 14. T-Test Results of Comprehension Question Accuracy Across Conditions	83
Table 15. T-Test Results of RT Across Conditions	84
Table 16. ANOVA results	84
Table 17. Tukey's HSD Results	85
Table 18. Logistic Regression for Comprehension Question Accuracy	85

LIST OF FIGURES

Figure 1. Aristotle’s De Re De Dicto ambiguous sentences, cf. Keshet & Schwarz (2019)	3
Figure 2. A simplified syntactic tree of De Re reading according to Aquinas	4
Figure 3. A simplified syntactic tree of De Dicto reading according to Aquinas...	5
Figure 4. A sample De Re De Dicto ambiguous sentence set	7
Figure 5. Comparison of LF structures for De Re and De Dicto readings.....	8
Figure 6. An application of possible world pronoun in NP (adapted from percus2000constraints by keshet2019re)	10
Figure 7. Compositional implementation of scope theory with higher type traces keshet2019re	11
Figure 8. Composition at the top level.....	11
Figure 9. Compositional implementation of $\hat{}$ operator keshet2019re	12
Figure 10. $\hat{}$ operator rule	12
Figure 11. LF assumed by demirok2020pied.....	13
Figure 12. Truth conditions with lower type trace and higher type trace demirok2020pied.....	13
Figure 13. Results of the first experiment carried out by Zhang and Davidson	20
Figure 14. Results of the second and third experiments carried out by Zhang and Davidson	21
Figure 15. A sample annotated sentence triplet.....	32
Figure 16. Data augmentation for computer vision liu2020survey	52
Figure 17. Classification of text data augmentation liu2020survey	53
Figure 18. Sample annotation of seed data on BoAT interface	57
Figure 19. A sample augmented hypothesis pair	59
Figure 20. A sample augmented hypothesis triplet with proper noun mismatch ...	60
Figure 21. Precision scores of two subsplits of NLI-TR budur2020data.....	66
Figure 22. Morphological parsers’ effect on NLI performance	67
Figure 23. Baseline dialogue of the item #5.....	74

Figure 24. Comprehension question for the baseline dialogue of the item #5	75
Figure 25. Wide scope De Re dialogue of the item #5	75
Figure 26. Comprehension question for the wide scope De Re dialogue of the item #5	76
Figure 27. Narrow scope De Re dialogue of the item #5	76
Figure 28. Comprehension question for the narrow scope De Re dialogue of the item #5	76
Figure 29. De Dicto dialogue of the item #5	77
Figure 30. Comprehension question for the De Dicto dialogue of the item #5	77
Figure 31. Expected response accuracy per item before exclusions across conditions	81
Figure 32. Mean percentage of expected responses across conditions	86
Figure 33. Posterior plot of expected responses across conditions, compared to baseline	87
Figure 34. Posterior plot of response times across conditions, compared to baseline	88
Figure 35. Reading times of ambiguous sentences	89

CHAPTER 1

INTRODUCTION

This thesis aims to explore De Re De Dicto ambiguities in Turkish from three different perspectives: Formal semantics, computational linguistics, and psycholinguistics.

From the formal semantics perspective, a survey of different accounts that aim to explain compositional semantics of De Re De Dicto ambiguities is provided with a heavy focus on scope theory (Quine, 1956). Having free word order and optionally non-overt accusative marker interacting with information structure, Turkish poses an interesting case for De Re De Dicto ambiguities as their resolution is closely tied to many factors: Discourse and non-discourse elements in the conversational context, information structure, world knowledge, and belief states of the utterer. Scrambling and other word order changing operations along with the presence and absence of overt accusative marker in Turkish can make De Re or De Dicto readings inaccessible in otherwise ambiguous sentences.

From the computational linguistics perspective, a challenge data set in Turkish for natural language inference (NLI) is offered to test swiftly growing large language models' (LLMs) abilities in recognizing syntactic and morphological cues along with the world knowledge. For this purpose, a seed data of 743 sentence triplets were manually created and annotated by native speakers of Turkish: One potentially De Re De Dicto ambiguous sentence, one De Re paraphrase and one De Dicto paraphrase. For the annotation phase, the team of contributors marked inference relations between ambiguous sentences and their De Re and De Dicto readings. Then this seed data was augmented using a natural language augmentation algorithm specifically created for this task. Following the augmentation process, CNLI-TR, a challenge data set containing over 7300 annotated sentence triplets, was created. An experiment was conducted to assess probing abilities of CNLI-TR and compare it with NLI-TR, the only other available NLI data set for Turkish.

From the psycholinguistics perspective, a self-paced reading experiment was conducted to investigate whether theoretical claims that De Re is "harder" to compositionally derive hold from a sentence processing perspective. In the experiment, participants were first presented with a disambiguating context in the form of a dialogue. Then they were shown De Re De Dicto ambiguous sentences and asked comprehension questions targeting specific readings parallel to disambiguating contexts. Reading time of De Re De Dicto ambiguous sentences and response time of comprehension questions were compared across different disambiguating contexts to see whether there is a significant processing asymmetry between De Re and De Dicto readings.

1.1 De Re De Dicto ambiguities

Natural language allows us to travel to worlds beyond the world we live in. We can report others' beliefs or desires, reason about what would happen to kangaroos in worlds where they have no tails. Intensional operators are natural language expressions that make this travel and much more possible.

When noun phrases (NPs) are embedded under intensional operators, we arrive at De Re De Dicto ambiguities: Whether we interpret a noun phrase within the scope of an intensional operator or not determines which reading we get.

The fact that an intensional operator can change the "meaning" of certain phrases in a sentence was first attracted the attention of philosophers as early as 4th century BC (Novaes, 2003): In *Sophistikoi Elenchoi*, Aristotle discusses the issue with a heavy focus on modality¹ via a set of sentences (see Figure 1).

In both sentences, (a) interpretations are concerned with whether the men in question retain their ability to write or walk: Sitting does not cost an individual the ability to walk, nor not writing costs them the ability to write. They can always get up and walk, or grab a pen and write. However, we need to expand our horizons to some

¹Modern linguistic theory thoroughly discusses and proves that not only modals but all intensional operators (such as verbs like "want" or "wish" that express a desire) and adverbial quantifiers (Keshet & Schwarz, 2019) can affect how an NP in their scope is interpreted, but Aristotle's observations on De Re De Dicto ambiguities are limited to modals.

"It is possible for a sitting man to walk."

- a. A sitting man still possesses the ability to walk.
- b. It's possible for a man to simultaneously sit and walk.

"It is possible for a man who is not writing to write."

- a. A man who is not writing still possesses the ability to write.
- b. It's possible for a man to simultaneously write and not write.

Figure 1. Aristotle's De Re De Dicto ambiguous sentences, cf. Keshet & Schwarz (2019)

extent to jump the distance between (b) interpretations and base sentences as what these interpretations assert seem like paradoxes -Walking while sitting, or writing while not writing is not very possible if we are not advancing our arguments on a metaphorical level.

According to Aristotle, (a) and (b) are different readings of the same modal proposition. It is through "division and combination" (Novaes, 2003) that we arrive at different readings. In other words, the difference between (a) and (b) readings is the extent of the ability contained by the modal expression. For both readings, we accept the same logical structure but (a) reading describes the ability to walk or write through division while (b) reading describes the ability to walk and not walk at the same time, or write and not write at the same time through combination.

It is argued that Aristotle's approach to modal expressions seems "to raise a range of philosophical problems, especially concerning essentialism" (Novaes, 2003) instead of approximating a linguistic baseline. Yet, his assertions are still very valuable to both semantics and philosophy even after more than a millennia as he was the first to discuss how modals can affect the interpretation of certain elements in the sentence.

In 12th century, Peter Abelard revisits (see Kneale (1962)) Aristotle's discussion in *Sophistikoi Elenchoi* to make a clear distinction between two possible readings, and to coin their names:

- De Re² interpretation of a modal sentence makes an assertion about a "thing" (*re*),

²Lat. "of the thing"

- De Sensu³ interpretation, on the other hand, the modal predicates a "sense" (*sensus*).

Following Abelard's distinction, (a) interpretations above are De Re readings of the sentences as they talk about an individual who is sitting or not writing. Despite his current status, that man is able to walk, or write. Consequently, (b) interpretations are De Sensu readings in Abelard's terms: Instead of a man's abilities, they talk about the possibility of sitting while walking, or not writing while writing. It is worth highlighting that Abelard's distinction has its roots in Aristotle's arguments, hence he agrees that there are not different logical structures for De Re and De Sensu readings -Instead, he asserts that the distinction between two readings is based on a semantic ambiguity.

In the 13th century, several other scholars offered their contribution to the topic, including Thomas Aquinas (1886), Peter of Spain (1990), and William of Sherwood (1995). Amongst them, Aquinas makes an especially noteworthy contribution as he was the very first scholar to address the issue from a syntactic standpoint and use the term De Dicto⁴ instead of Abelard's De Sensu. Following Aristotle's term logic, Aquinas divides the sentence into two parts: The subject and the predicate. Through this division, he explains the difference between De Re and De Dicto readings: The subject is a "thing" in De Re readings (see Figure 2), while it is a whole clause in De Dicto readings (see Figure 3).

de re:

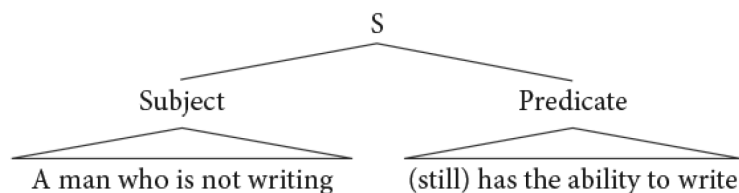


Figure 2. A simplified syntactic tree of De Re reading according to Aquinas

³Lat. "*of the sense*"

⁴Lat. "*of the meaning*"

de dicto:

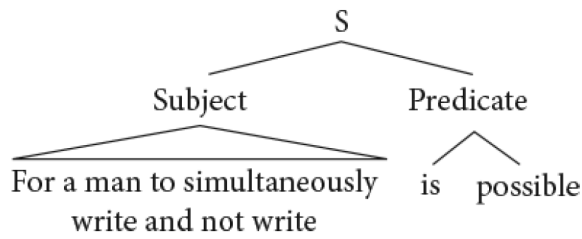


Figure 3. A simplified syntactic tree of De Dicto reading according to Aquinas

Peter of Spain also follows Aristotle's arguments and asserts that it is the modal expression that is ambiguous, not the clause embedded in it while William of Sherwood makes another attempt at introducing a syntactic criterion for the distinction by introducing a two-way categorization for modes. In William of Sherwood's categorization, modes in adverb form correspond to Aquinas's De Re reading while modes in noun form correspond to De Dicto reading.

In 20th century, logician George Henrik von Wright sheds a new light upon the issue by building a quantified modal logic system upon the C.I. Lewis's (1932) seminal work in modern logic titled *Symbolic Logic* (see Novaes (2003) for a detailed discussion). Using Aquinas's terms (1886), von Wright offers a detailed logical account of De Re De Dicto phenomenon (von Wright, 1951) . Prior (1952), Kneale (1962) and Hintikka (1962) later follows up with the issue through their unique arguments based on von Wright's.

Although very valuable, these historical contributions from Aristotle to Kneale (1962) and Hintikka (1962) are limiting for a perspective mainly concerned with the linguistic features and implications of De Re De Dicto ambiguities for two fundamental reasons:

- They focus on modal logic and modal expressions while disregarding other embedding operators (such as intensional operators like "want" or adverbial quantifiers like "probably") and possible world theory which allows a formal semantic analysis of intensionality (see Marcus (1961) and Plantinga (1976) for a thorough discussion of the relationship between intensionality and possible world theory).

- They attribute ambiguity to the sentence instead of noun phrases. In other words, they argue that the "meaning" of the entire sentence is ambiguous instead of the "meaning" of the NP which allows two interpretations.

As a result of these shortcomings of the philosophical perspectives, the modern linguistic discussion of De Re De Dicto ambiguities dates back to less than two centuries ago. With his principal work on modern analytic philosophy, Frege (1892) lays the foundations for a linguistic description of De Re De Dicto ambiguities through offering a comprehensive theory of sense and meaning. In his theory, Frege divorces concepts and objects from one another by introducing the distinction of sense (*Sinn*) and reference (*Bedeutung*). Through this distinction, it is possible to build a framework where different expressions can refer to the same object despite having different senses. For instance, let's assume that there is an individual named Susan whom we know of. She is a dedicated and a hardworking student, thus she is both the valedictorian and class president. The concepts of "valuedictorian" and "class president" have distinct and different meanings: One means the high achieving student that is tasked to give the farewell speech on behalf of her cohort, and the other means the leader of the student body. Yet both concepts have the same referent: Susan.

Equipped with the core concepts of sense and reference, Frege's theoretical framework is potent enough to allow the development of a more sophisticated theories of formal semantics. It is through this potency, Russell (1905) built his theory of definite descriptions under the light of Frege's observations.

Russell argues that definite descriptions behave like quantifiers that establish a unique description. In other words, definite descriptions are in contrast to names: They are logical equivalents of the conjunction of a set of properties that collectively refer to a single referent in the world, yet these properties do not presuppose the existence of their referent.

Let's explore this idea through an example: "the Ottoman sultan" is a definite description. The initial look at this expression seems to imply that this expression has

a unique, single entity that it refers to. Yet there is not an Ottoman sultan that rules over the geography known as Turkey as the country has been a republic since 1923. So, what does the expression "the Ottoman sultan" actually refer to? The theory put forward by Russell (1905) addresses this problem. He argues that definite descriptions like "the Ottoman sultan" do not have to have an actual referent object in the world. Instead, they are, in a sense, shorter and denser expressions of more complex descriptions. Thus, "the Ottoman sultan" actually is a description as follows: Someone who is of Ottoman descent, and someone who rules over the Ottoman geography. As a result, we don't need to presuppose existence of an entity who is "the Ottoman sultan" to analyse this phrase through the lens of logic.

It is through this approach to the issue of definite descriptions that Russell paves the way for a formal semantic definition of De Dicto meaning despite not using the term De Dicto in his work. In an ambiguous sentence like the one in Figure 15 ("Mary wants to marry a plumber."), De Dicto reading asserts that "a plumber" is not saturated by the existence of a single entity in the world whose profession is plumber. Instead, it is a description that shall be analyzed without an object that satisfies it.

Ambiguous sentence: Mary wants to marry a plumber. De Re reading: There is an individual who is a plumber, and Mary wants to marry them. De Dicto reading: Mary wants her future spouse to be a plumber.
--

Figure 4. A sample De Re De Dicto ambiguous sentence set

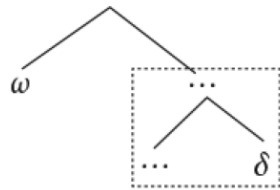
Following the path made accessible by Russell (1905), Quine (1956, 1971) provides a detailed discussion on propositional attitudes and draws a contrast between relational and notional readings of propositional attitudes. Without explicitly using the terms De Re and De Dicto, Quine borders on the ambiguities like Figure 15 and uses the notion of scope to distinguish different readings. He asserts the assumption that the extensions of predicates can vary across times and possible worlds.

Extrapolating on the implications of possible world theory of modality, the vast majority of modern theoretical linguistic accounts of De Re De Dicto ambiguities build on Quine (1956) and refer to the notion of scope to semantically derive De Re and De Dicto readings from an ambiguous sentence.

It is essential to make a concrete definition of scope before moving on to the different theoretical accounts of De Re De Dicto ambiguities. Scope can be defined through the lens of syntax or semantics. From a syntactic point of view, the scope of an intensional operator is its c-command domain. Semantic definition of the scope, though, makes no reference to the structure: The scope of an operator contains the sentential elements whose intensional status is determined by the meaning of the operator.

The scope theory of De Re De Dicto marries these two definitions on the logical form (LF): An intensional operator's semantic scope is its syntactic scope on the LF. Thus, coming up with a simplistic and elegant derivation for De Re and De Dicto readings is possible through scope theory. When the ambiguous NP is in the c-command domain of the intensional operator in LF, we arrive at the De Dicto reading. By raising the NP in LF to extract it from the scope of its intensional operator, we are able to derive De Re reading.

a. *De dicto*:



b. *De re*:

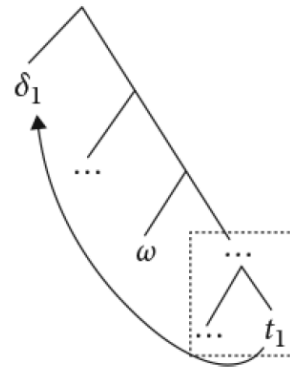


Figure 5. Comparison of LF structures for De Re and De Dicto readings

Assuming the LF structures depicted in Figure 5 for De Re and De Dicto readings, we are able to capture a wide range of phenomena, including quantification

over time, attitude verbs, and adverbial quantifiers that impact the interpretation of the NPs within their scope. However, the traditional scope theory of De Re De Dicto encounters certain limitations when confronted with data that goes beyond basic cases. These limitations include multiple embeddings, intermediate readings, scope islands, and non-specific De Re⁵ (Fodor, 1970) (for a detailed discussion of these cases, refer to Keshet & Schwarz (2019)).

To address these challenges, more refined approaches have been proposed, employing various tools offered by formal semantics to sever the link between a noun phrase's intensional status and its quantificational scope properties.

One such approach involves the use of intensional variables within the object language, departing from the traditional scope theory in how it represents the world of interpretation. Rather than encapsulating it as a parameter within the interpretation function, this approach overtly carries out quantification over worlds and times within the object language. Consequently, raising the NP from the scope of the intensional operator is no longer necessary to derive De Re readings.

The theoretical foundation of this approach draws upon previous empirical findings that highlight natural language's capacity for quantification over worlds (Cresswell, 2012), situations (Kratzer, 2007), and times (van Benthem, 1977; Kamp, 2013).

The incorporation of syntactically represented intensional variables within NPs to derive narrow scope De Re readings (referred to as Fodor's third reading) has gained prominence among various scholars. These accounts allow NPs to include phonologically and morphologically empty pronouns, specifically related to possible worlds or times, to fill the world or time arguments of predicates (Percus, 2000; Schwarz, 2012; Elbourne, 2013). By permitting different λ abstractors of intensional operators to bind these possible world pronouns, this approach divorces the

⁵Non-specific De Re, also known as Fodor's third reading, arises from the fact that the quantificational force of an indefinite NP can scope independently from its intensional status. In simpler terms, a De Re De Dicto ambiguous sentence like "Mary wants to marry a plumber" can yield a third reading where there are multiple individuals that satisfy the NP "a plumber." For instance, suppose there are three individuals named John, Jonathan, and James, all of whom are plumbers. Mary may want to marry any one of these gentlemen.

quantificational scope of an NP from its nominal predicate. Figure 6 provides an illustration of this approach, showcasing a De Re reading in a sentence that poses challenges, if not impossibilities, for traditional scope accounts of De Re De Dicto.

(S) If everyone in this room were outside, it would be empty.

De Re interpretation:

$\lambda w_0 [_{S1} \text{ if } [_{S2} \lambda w_1 [_{S3} \text{ everyone in this room in } w_0 \text{ were outside in } w_1 [\text{ it would be empty in } w_1]]]]$

For any world w , S is true in w iff for every accessible world w' such that everyone in this room in w is outside in w' , the room is empty in w' .

Figure 6. An application of possible world pronoun in NP (adapted from Percus (2000) by Keshet & Schwarz (2019))

In the De Re interpretation example shown in Figure 6, the world pronoun (w_0) associated with the NP, namely "everyone in this room," is interpreted relative to the world of evaluation assigned to the entire sentence. This allows us to derive the De Re reading. Conversely, if the world pronoun of the NP were w_1 instead of w_0 , it would be possible to derive a De Dicto reading. However, it is important to note that such a reading would result in a contradictory statement.

Introducing intensional variables in the object language significantly diverges from the traditional scope theory, yet there are other approaches that aim to sophisticate the traditional theory enough to account for cases like Fodor's third –Instead of completely ditching it.

One of such approaches, scope theory with higher type traces is based on Heim and Kratzerian assumptions regarding traces (Heim & Kratzer, 1998) and it disconnects scope and intensional status of the NPs to some extent to make narrow scope De Re readings possible.

By changing the semantic type of traces from e to $\langle et, t \rangle$, it is possible to interpret the NP relative to the world of evaluation for the entire sentence despite moving it out of the scope of the intensional operator (see Figure 7 and Figure 8).

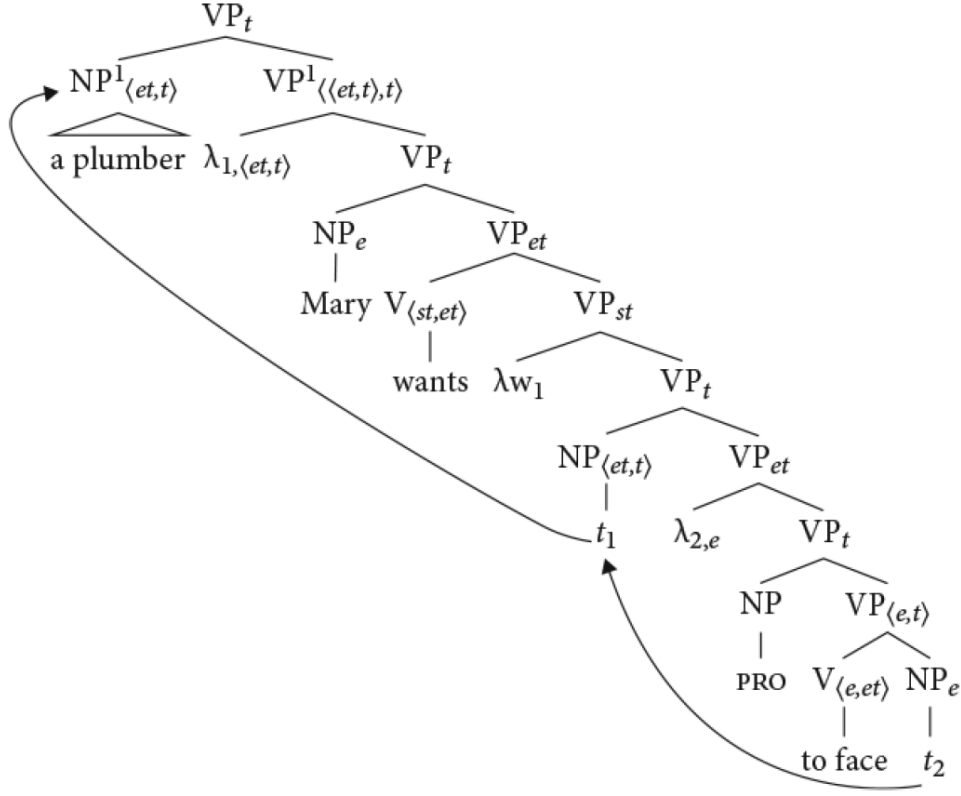


Figure 7. Compositional implementation of scope theory with higher type traces (Keshet & Schwarz, 2019)

$$\begin{aligned} \llbracket \text{VP}^1 \rrbracket^w (\llbracket \text{NP}^1 \rrbracket^w) &= \llbracket \text{wants} \rrbracket^w (\lambda w_1. [\lambda P_{et}. \exists y [y \text{ is a plumber in } w \text{ \& } P(y) = 1]]) \\ &\quad (\lambda x_e. \llbracket \text{face} \rrbracket^{w_1}(x) (\llbracket \text{PRO} \rrbracket^{w_1})) (\llbracket \text{Mary} \rrbracket^w) \end{aligned}$$

Figure 8. Composition at the top level

In an effort to provide scope theory with greater flexibility to derive narrow scope De Re readings, Keshet (2008, 2010) propose a novel approach based on a modified version of Intensional Functional Application (IFA) (Heim & Kratzer, 1998). Originally introduced as a tool for type repair, IFA enables the transition from extension to intension prior to Function Application. This operation becomes necessary when a function in the structure requires an intensional argument as input, while the syntax can only provide an extensional argument, resulting in a type mismatch.

A key characteristic of the approach proposed by Keshet (2010) is the introduction of an additional step in the derivation process, where type-shifting occurs using the $\hat{}$ operator and a novel rule (see Figure 12)

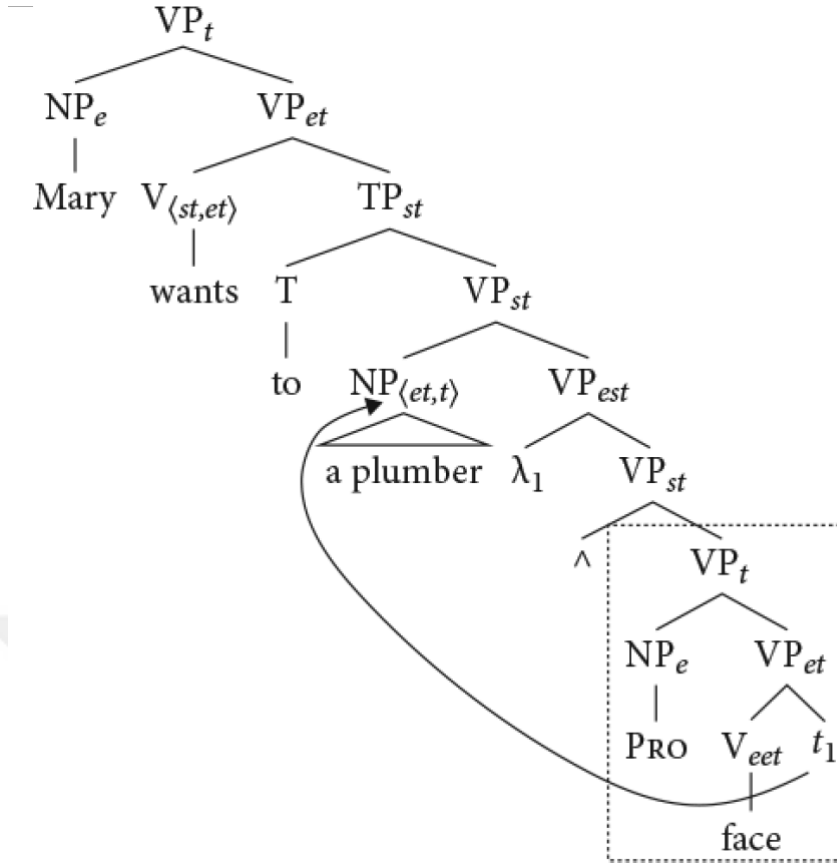


Figure 9. Compositional implementation of $\wedge$ operator (Keshet & Schwarz, 2019)

Intensional Abstraction

If α is a branching node and $\{\beta, \gamma\}$ is the set of its daughters, where β dominates only an $\wedge$ operator, then, for any situation s and variable assignment g , $\llbracket \alpha \rrbracket^{s,g} = \lambda s' \in D_s. \llbracket \gamma \rrbracket^{s',g}$.

(Modeled after Heim and Kratzer's Predicate Abstraction 1998: 186)

Figure 10. $\wedge$ operator rule

Unlike traditional approaches, there are no strict constraints on the placement of the $\wedge$ operator. Its insertion before the targeted function, which is in need of an argument, determines the success or failure of the derivation. To ensure that the quantificational scope of the NP does not impact its intensional status, the $\wedge$ operator must be inserted above the intensional operator, and the NP must be raised to a position between the two. This mechanism ensures that the quantificational scope and intensional properties of the NP can be handled independently.

An additional strategy worth mentioning in the context of modifying traditional scope theory is pied-piping. Demirok (2020) presents an exceptional

solution to exceptional De Re problems and other constructions that pose challenges for the traditional theory. In essence, the issue faced by the traditional theory in cases such as "If everyone in this room were outside, it would be empty" is that "everyone in this room" cannot scope out of the if-island, yet it must receive a De Re reading relative to the intensional operator.

Indeed, Demirok's proposed modifications (ibid), which encompass interpreting pied-piping at the level of LF, employing type-neutral λ abstraction, and utilizing higher type traces for scope reconstruction, significantly enhance the scope theory. These refinements allow the enriched scope theory to effectively tackle exceptional De Re cases and various other constructions in a unified manner. By incorporating these refinements, the theory becomes better equipped to handle complex scope phenomena, leading to more accurate and comprehensive analyses. These advancements contribute to a deeper understanding of scope interactions and provide valuable insights into the interpretation of natural language expressions.

$$\text{LF: } [\exists [\lambda_2 \text{ everyone in this room } \lambda_1 [\text{IDENT } t_2] [t_1 \text{ if } t_1 \text{ were outside}]]] \\ [\lambda_3 [\text{would } t_3 [\text{it be empty}]]]$$

Figure 11. LF assumed by Demirok (2020)

$$\begin{aligned} \times & \text{When } t_1 \text{ is of type } e: \llbracket (64) \rrbracket^w = 1 \text{ iff} \\ & \exists p \forall x: \llbracket \text{everyone in this room} \rrbracket^w(x) \rightarrow [p = \lambda w'. \llbracket \text{were outside} \rrbracket^{w'}(x)] \\ & \quad \& \llbracket \text{would} \rrbracket^w(p)(\lambda w'. \text{ it is empty in } w') \\ \checkmark & \text{When } t_1 \text{ is of type } \langle et, t \rangle: \llbracket (64) \rrbracket^w = 1 \text{ iff} \\ & \exists p: [p = \lambda w'. \llbracket \text{everyone in this room} \rrbracket^w(\llbracket \text{were outside} \rrbracket^{w'})] \\ & \quad \& \llbracket \text{would} \rrbracket^w(p)(\lambda w'. \text{ it is empty in } w') \end{aligned}$$

Figure 12. Truth conditions with lower type trace and higher type trace (Demirok, 2020)

Although aforementioned theoretical approaches differ from one another in terms of how to handle and derive De Re De Dicto ambiguities from a compositional

semantics perspective, they all lay on the same foundational values: Despite the early work on De Re De Dicto ambiguities by philosophers, the difference between De Re and De Dicto readings is not a mere issue of semantic ambiguity or difference in interpretation. De Re and De Dicto readings greatly differ from one another on logical structure. Across all frameworks discussed above, De Re is theoretically more complex than De Dicto: It either requires some covert movement in the logical form, or situation pronouns that are bound non-locally. The motivation behind this choice is evident: We need a more enriched context to arrive at the De Re reading. There must be at least one contextually salient individual who possesses the features or distinctions required by the sentence. Whereas De Dicto reading acts as a definition where the existence of a salient individual who saturates this definition is not necessary.

Instead of making novel theoretical claims, this thesis aims to investigate De Re De Dicto ambiguities in Turkish⁶ from different perspectives based on the aforementioned foundational values of the past theories that gained significant traction: From computational perspective, LLMs abilities to handle De Re De Dicto abilities is discussed. For this purpose, a natural language inference task is used as a proxy. From psycholinguistics perspective, human processing of De Re De Dicto ambiguities are examined. A self-paced reading task is used in the experiment design.

1.1.1 De Re De Dicto ambiguities in Turkish

The main distinction between De Re and De Dicto readings lies in the fact that De Re readings involve a specific individual (wide scope) or a group of individuals (narrow scope) while De Dicto readings do not refer to such specific individuals. Resolving De Re De Dicto ambiguities requires us, as humans, to utilize various sources of information: World knowledge, belief state of the utterer, linguistic and

⁶The reason why this thesis highlights De Re De Dicto ambiguities in Turkish is that with free word order, scope rigidity, and rich morphology, Turkish poses an interesting case for De Re De Dicto ambiguities. A thorough discussion on this is provided in the next subsection.

supralinguistic elements in the context, and in the case of some languages, linguistic cues.

Analytic languages with fixed word order like English often do not offer explicit linguistic cues to disambiguate potentially De Re De Dicto ambiguous sentences. In such cases, disambiguation heavily relies on the broader context and extra-linguistic factors.

- (1) Mary wants to call a plumber.
- (2) Marie veut appeler un plombier.
- (3) María quiere llamar a un plomero.

Sentences 1 through 3 satisfy all the necessary conditions to be De Re De Dicto ambiguous. They contain an intensional operator (to want) and an embedded NP (a plumber). However, in languages where no overt accusative morphology is present and scrambling is disallowed, disambiguating these sentences solely based on linguistic cues is impossible. In other words, it is possible to get both De Re and De Dicto reading from these sentences.

Without explicit linguistic cues available, disambiguating De Re De Dicto readings in such languages heavily relies on other contextual and supralinguistic factors. These may include the overall context, world knowledge, the beliefs and intentions of the speaker, and pragmatic considerations.

Turkish, as a language with free word order and rich case morphology, presents an interesting case for De Re De Dicto ambiguities. In Turkish, specific linguistic cues such as the overt accusative case marker, topicalization, scrambling, and noun incorporation can play a crucial role in disambiguating potentially De Re De Dicto ambiguous sentences by making one of these readings inaccessible.

In Turkish, objects that do not receive lexical case can be bare/caseless (Sentence 4), or carry the overt accusative case marker (Sentence 5). Moreover, the distinction between NP meaning and DP meaning in Turkish relies heavily on the presence or absence of overt accusative marker and indefinite article "bir" (Sentences 6 and 7) as Turkish has no overt determiner but overt case morphology (Ozturk,

Table 1. Specificity and Definiteness in Turkish as Described by Kechriotis (2009)

	Specificity	Definiteness
bir + NP	×	×
bir + NP + overt accusative	✓	×
NP + overt accusative	✓	✓

2004). The presence of the accusative marker and "bir" can indicate the specific interpretation of an NP as a definite or indefinite entity, providing important cues for De Re De Dicto disambiguation in Turkish.

Extensive research on Turkish case morphology, supported by empirical evidence, highlights a clear distinction between the accusative-marked and bare objects (Dede, 1986; Enç, 1991; Keleş, 2001). One of the key differences lies in the realm of referentiality, where the overt accusative case marker is associated with referential interpretations. Additionally, the presence of the indefinite article "bir" has the effect of rendering the NP indefinite, while its absence brings the NP closer to the domain of definite noun phrases due to the addition of definiteness (Kechriotis, 2009).

Table 1 provides a visual representation of this distinction, demonstrating how the presence of "bir" leads to indefiniteness, while the presence of the overt accusative case marker adds specificity to the NP.

(4) Meryem *bir* tesisatçı- $\emptyset$ çağırma*k* istiyor.
Meryem.NOM a plumber. $\emptyset$ call.NMLZ want.3.SG.PRS
‘Mary wants to call a plumber’

(5) Meryem *bir* tesisatçı-yı çağırma*k* istiyor.
Meryem.NOM a plumber.ACC call.NMLZ want.3.SG.PRS
‘Mary wants to call a plumber’

(6) Meryem tesisatçı-yı çağırma*k* istiyor.
Meryem.NOM plumber.ACC call.NMLZ want.3.SG.PRS
‘Mary wants to call [the] plumber’

(7) Meryem tesisatçı- $\emptyset$ çağırma*k* istiyor.
Meryem.NOM plumber. $\emptyset$ call.NMLZ want.3.SG.PRS
‘Mary wants to call a plumber’

Sentence (4) and (5) can be contrasted based on the presence of the overt accusative marker. In sentence (4), where there is no overt accusative case marking, the sentence is considered truly ambiguous. Without additional context or explicit clarification from the speaker, both De Re and De Dicto readings of the sentence are possible. However, in sentence (5), the presence of the overt accusative marker implies specificity, making the De Dicto reading inaccessible regardless of any additional disambiguating information provided in the conversational context.

Regarding sentences (5) and (6), the distinction lies in the presence of the indefinite article "bir." Since both sentences have the overt accusative case marker, neither allows for a De Dicto reading. However, the absence of "bir" in sentence (6) indicates that there is a plumber who is contextually salient and acknowledged by both the speaker and the listener.

Lastly, sentences (6) and (7) can be contrasted based on the presence or absence of "bir" and the overt accusative marker. In sentence (7), where both "bir" and the overt accusative marker are absent, noun incorporation is possible, rendering the De Re reading inaccessible in this sentence.

Apart from case morphology and the presence of "bir," word order changes such as topicalization or scrambling in Turkish can also influence the specificity of noun phrases (Taylan, 1984; İşsever, 2003). These word order changes interact with the information structure and take advantage of the scope rigidity characteristic of Turkish.

In Turkish, arguments presented on the left-hand side of the matrix verb are processed as new or novel information, while those presented on the right-hand side are treated as old or background information. Consequently, word order operations can make De Re or De Dicto readings inaccessible by altering the perceived information status of the noun phrases.

In summary, certain linguistic cues in Turkish, including case morphology, the presence of "bir," and word order changes, provide means to disambiguate De Re De Dicto ambiguities without necessarily relying on the belief state of the speaker

or world knowledge. Exploring the implications of De Re De Dicto ambiguities on the performance of Language Models (LLMs) and human processing opens up an interesting area of research.

1.1.2 Implications of De Re De Dicto ambiguities

As thoroughly discussed in the previous section, De Re De Dicto ambiguities have a certain significance in formal semantics. In addition, these ambiguities arise interesting questions for both NLP and psycholinguistics. Implications of De Re De Dicto ambiguities and these questions will be further discussed in the following subsections.

1.1.2.1 Large language models

De Re De Dicto ambiguities also pose an interesting case for language models. From the perspective of LLMs, two questions are very relevant: First, can LLMs recognize the presence of these ambiguities? Second, can they resolve these ambiguities?

In order for LLMs to recognize De Re De Dicto ambiguities, linguistic proficiency alone must be enough, yet the resolution of these ambiguities requires more information:

- World knowledge,
- Access to the belief state of the utterer,
- And in the case of Turkish, the ability to recognize certain linguistic cues.

There have been significant debate on whether LLMs have access to world knowledge, and whether carrying out tasks like NLI detection would require such knowledge (Zaenen et al., 2005; Condoravdi et al., 2003; Bowman et al., 2015; Chen et al., 2017). Moreover, there have been various studies to offer an external world knowledge base(Wang et al., 2019) and/or introduce knowledge graphs (Sharma et al., 2019; Wang et al., 2020) to improve the performance of LLMs in inference tasks. Yet certain questions on world knowledge and LLMs are still unanswered

-Such as: Do LLMs build a separate world knowledge module with the information presented, or can they separate linguistic knowledge and world knowledge when they have access to the latter.

Contrary to the situation regarding world knowledge, it is a certain fact that LLMs have no access to the belief state of the utterer or the source of the sentence unless such information is presented in a machine readable way to the LLMs.

The aim of this thesis is to probe into the case of linguistic information: To what extent language models are aware of it, and do they use this information while carrying out NLP tasks? To answer these questions, NLI task is used as a proxy to assess the abilities of LLMs, and a dataset (CNLI-TR) is built around De Re De Dicto ambiguities as the resolution of these ambiguities are possible through certain linguistic cues like the overt accusative marker, or word order.

1.1.2.2 Sentence processing and psycholinguistics

Although following different frameworks and methodologies to compositionally derive De Re and De Dicto readings, aforementioned theories agree upon one stipulation: Compositional derivation of De Re readings come with an additional complexity compared to De Dicto readings.

The theoretical approaches discussed in the previous sections either move NP out of the scope of the intensional operator, or introduce a morphologically null operator (such as a situation pronoun) above the NP to derive De Re readings. It is an intuitive assumption that LF movements and the use of null operators should cause additional processing load for the addressee of the utterance. As a result, there should be a difference between De Re and De Dicto readings regardless of the theoretical framework we adhere to in order to derive De Re and De Dicto meanings. Yet previous literature offers very limited examination of De Re De Dicto ambiguities through the lens of psycholinguistics.

One of the experimental studies worth mentioning is carried out by Zhang and Davidson (2021). In their experiment, Zhang and Davidson test the accessibility and

availability of De Re and De Dicto belief reports⁷ through three experiments where participants were asked to report truth value judgements.

In the first experiment, participants were presented with four scenarios where both De Re and De Dicto readings are available. After reading each scenario, participants were shown De Re and De Dicto target sentences containing definite DPs in possessive constructions and asked to pass their judgement on them (highly agree, uncertain, highly disagree). Participant judgements on De Dicto readings accumulated on the highly agree end while De Re judgements showed a bimodal pattern skewed towards highly agree end.

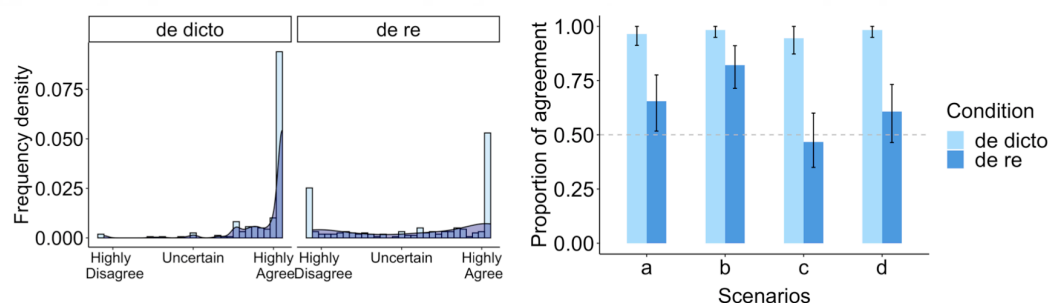


Figure 13. Results of the first experiment carried out by Zhang and Davidson

Focusing on bound De Re (Charlow & Sharvit, 2014), second and third experiment presented the participants 4 scenarios each, and asked judgement questions following each scenario. In contrast to the first experiment, scenarios in second and third experiment allowed either De Re or De Dicto reading and were accompanied by illustrations.

Results of the second and third experiment replicated the results of the first experiment: De Re readings showed bimodal agreement pattern and highlighted that the agreement rate of De Re readings were directly affected by the internal structure of the DP.

Yet neither of the psycholinguistics studies that have been carried out so far focuses directly on the processing cost of De Re and De Dicto readings. Through

⁷In their experiments, Zhang and Davidson only include wide scope De Re scenarios.

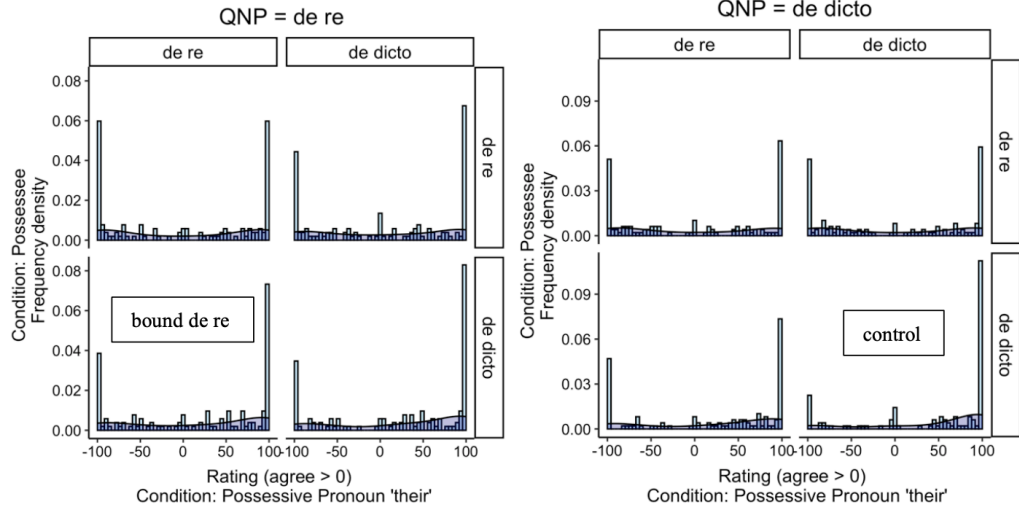


Figure 14. Results of the second and third experiments carried out by Zhang and Davidson

Premise	Hypothesis	Inference Relation
(P1) Venüs is a very fluffy and pretty cat.	(H1) Venüs is a mammal.	Entailment
(P2) Mars is playing with a ball of yarn at the moment.	(H2) Mars is sleeping now.	Contradiction
(P3) Two cats are playing ping pong on the third floor.	(H3) Two cats are wearing collars.	No relation

Table 2. An illustration of different NLI relations

a self-paced reading experiment, this thesis aims to investigate whether theoretical complexity of De Re and De Dicto readings translate to a processing complexity.

1.2 Natural language inference

Natural language inference (NLI) is one of the important yardsticks that allows us to gauge how much a model *actually* understands the language (Naik et al., 2018).

From a semantic standpoint, NLI involves identifying the logical relationship between two sentences: a hypothesis and a premise. Essentially, the task is to assess the relative truth value of the hypothesis given that the premise is TRUE.

When the hypothesis logically follows from the premise, an entailment relationship exists between them: If Venüs is a cat, then it is necessarily a mammal, since all felines are mammals. Thus, P1 entails H1 (see Table 2). On the other hand, when the hypothesis contradicts the premise, a contradiction relationship exists

between them: If Mars is playing with a ball of yarn, it cannot also be sleeping. As a result, P2 contradicts H2.

If it is impossible to determine the truth value of the hypothesis based on its premise, no relationship exists between the two sentences: Premise in P3 says nothing about collars, so it is not possible to establish a semantic relationship between P3 and H3.

Using NLI as a yardstick to assess the natural language understanding (NLU) capabilities of language models makes sense from both theoretical semantics and NLP perspectives.

From theoretical semantics perspective, understanding a sentence implies two abilities: Understanding its truth conditions and establishing how it relates to other sentences (Condoravdi et al., 2003). As a widely accepted framework of formal semantics, truth conditional semantics links the natural language to the world: When we understand a sentence, we don't merely unveil its linguistic meaning but also we understand its truth conditions (see Frege (1892) for the foundational argument behind truth conditional semantics, and refer to Devitt & Sterelny (1987), and Carston (2013, 2011) for the discussion of language relates to the world). For instance, to understand the sentence "It is snowing outside," one does not need to know today's weather or the actual truth value of the sentence. Instead, one needs to understand what the world should look like if the sentence is true.

Second tenet of understanding natural language from a theoretical semantics standpoint lies in the ability to "relate the text to other texts and ... [to] tell where texts agree or disagree in what they say" (Condoravdi et al., 2003). In other words, understanding natural language requires more than the ability to comprehend individual sentences in isolation. It also requires an understanding of how a sentence fits into a broader context, and connects to other sentences.

By design, NLI is an efficient task to assess NLP systems' abilities in the latter yardstick of understanding the meaning, namely, being able to tell the semantic

relationship between sentences (ibid). That is why it has tracked significant attention since very early days of NLP as a challenge task to measure NLU.

In his trailblazer work, Pearl (1988) highlights noteworthy potency of Bayesian networks in a wide array of tasks that require establishing a relationship between uncertainty and probabilistic dependencies, such as computer vision and NLP. Dynamic Bayesian networks he introduces were capable of carrying out tasks that were considered complex back then, and paved the way for significant developments in NLP systems. In a sense, Pearls work on Bayesian networks made NLI tasks possible but made no mention of NLI itself.

One of the first studies that brought inference and NLP together was carried out by Monz & de Rijke (2001). The authors aim to improve the performance of computational semantics systems through implementing a lightweight inference mechanism based on a set of heuristics. They test the efficiency of this mechanism on a large corpus of news articles and prove that their system performs with high precision and recall scores.

With a heavy focus on the role of intensionality, Condoravdi et al. (2003) offers a thorough and impactful discussion of NLU and its measurement in NLP systems. One of the biggest contributions of this study is the proposed framework for text understanding powered by the union of machine learning technologies, logical inference, and semantic theory. Through this framework, authors were able to successfully carry out various NLP tasks including NLI. Following this study, the idea of using NLI as a yardstick for measuring language understanding abilities of NLP systems gains more traction amongst scholars.

With the first Recognizing Textual Entailment (RTE-1) Challenge benchmark by PASCAL Network of Excellence in 2005, NLI solidifies its place in both literature and practice. In the proceedings of the first PASCAL challenge, Dagan et al. (2005) makes one of the most significant attempts to define the task at hand, recognizing textual inference, through the lens of logic:

Our applied notion of *textual entailment* is defined as a directional relationship between pairs of text expressions, denoted by T - the entailing "Text", and H - the entailed "Hypothesis". We say that T entails H if, typically, a human reading T would infer that H is most likely true.

Moreover, Dagan et al. (2005) go on to state that we, as humans, base such conclusions on common human understanding of language as well as common background knowledge." Yet a significant body of research on NLI takes an opposing stance. One such point of view worth mentioning here is laid out by Zaenen et al. (2005), interestingly enough, in a discussion of PASCAL initiative: They state that the PASCAL initiative provides a welcome endeavor" as it proposes "a competition testing NLP systems on their ability to understand language separate from the ability to cope with world knowledge". As Zaenen et al. (2005)'s definition of world knowledge borders the definition of truth conditional semantics, and is mainly based on veridicality which is, of course, part of our world knowledge as humans, it must be highlighted that world knowledge is a larger concept: It encompasses extralinguistic knowledge such as discourse and non discourse elements, and these elements play an important role in how we interpret linguistic elements like utterances.

Thus, the issues revolving around world knowledge remains prevalent even more than a decade after the RTE-1: Do language models use world knowledge? If so, to what extent? If it is actually the case that language models don't refer to the world knowledge even in the form of "background use" (Manning, 2006), or their abilities on drawing important cues from the world knowledge to successfully carry out crucial tasks like NLI or question answering are limited, our next step must involve abstracting and incorporating this knowledge in a way that will improve NLP systems.

With this motivation, one of the goals of this thesis is to assess the capabilities of language models: Can they learn and represent world knowledge? As a first step towards answering this question, our current study aims to probe into the issue through De Re De Dicto ambiguities, the resolution of which requires the belief state

of the utterer and, indubitably, world knowledge. The main question we ask here is whether language models are aware of De Re De Dicto ambiguity. For this purpose, we use an NLI task as a proxy.

1.2.1 Related work

One of the first benchmark efforts to assess the abilities of NLP systems was carried out by a group of Europe-based researchers led by Robin Cooper. Known as FraCaS (A Framework for Computational Semantics) Consortium, this project yielded valuable insights on computational semantic processing, and numerous articles between 1994 and 1996 (see Cooper et al. (1994; 1995; 1996c; 1996a; 1996b)). Two major contributions of FraCaS were the idea of using NLI to assess the linguistic competence of NLP systems and FraCaS test suite (Amblard et al., 2020).

FraCaS test suite⁸ (Cooper et al., 1996b) consists of 346 problem sets focusing on different linguistic phenomenon including ellipsis, generalized quantifiers, temporal reference and anaphora. Each problem contains one question, at least one premise, and one answer (Yes, No, Unknown, Undefined). FraCaS test suite was used by various NLP systems (MacCartney & Manning, 2007, 2008; Bernardy & Chatzikyriakidis, 2017; Pinkal & Koller, 2019), laid the foundations of many research projects (Lev, 2006; Ljunglöf & Siverbo, 2011; Siverbo, 2012; Bernardy & Chatzikyriakidis, 2017), and was translated into several languages (Kawazoe et al., 2017; Amblard et al., 2020), but one contribution of FraCaS is essential to mention here: Its proposition to use textual inference as a yardstick for assessing NLP systems greatly influenced PASCAL (Pattern Analysis, Statistical modelling and Computational Learning) Excellence Network which enriched the theory and practice of NLP with ample challenge benchmarks (Dagan et al., 2005, 2006; Vanderwende et al., 2006; Giampiccolo et al., 2007, 2008; Bentivogli et al., 2009, 2011).

⁸Data set can be accessed in Stanford NLP archive.

Following the RTE-1 (Recognizing Textual Entailment) Challenge benchmark offered by the PASCAL initiative (Dagan et al., 2005), numerous efforts have been made in creating NLI data sets.

Being one of the largest NLI datasets with 570,152 pairs of sentences, Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) is one of the most widely used benchmarks. Each pair in the dataset is annotated with entailment, contradiction or neutral tag through a crowdsourcing process. As a high quality dataset, SNLI covers relatively short⁹ sentences from various domains including fiction, news articles, conversational speech, and government reports. All sentence pairs in SNLI are gathered from Flickr30K corpus (Young et al., 2014). One of the shortcomings of SNLI is that it contains somewhat uniform sentences: There is little variation in register, style, and structure as all sentences are image captions. The authors state that the reason behind this design choice was to minimize the effects of variation on the interpretation of sentence pairs.

Introduced as an extension of SNLI dataset, Multi Genre Natural Language Inference Corpus (MultiNLI) (Williams et al., 2017) is designed to test the generalization abilities of LLMs trained on SNLI data. MultiNLI contains over 393.000 sentences in ten domains¹⁰ extracted from various sources: The American National Corpus (Reppen et al., 2005), and numerous freely available literary works of various authors including Isaac Asimov and Agatha Christie.

MultiNLI has two subsets: Matched and mismatched. Matched subset covers the same domains as SNLI while mismatched covers different domains. The idea behind this choice is to challenge LLMs trained on SNLI to see whether they perform well in novel domains and genres.

Containing sentence pairs in 15 languages including Turkish, Cross-lingual Natural Language Inference (XNLI) (Conneau et al., 2018) is another large extension

⁹The average length of sentences in SNLI is 20 tokens. Yet the average length varies across different domains and semantic tags. As per authors note, sentences with neutral tag are longer than entailment and contradiction ones.

¹⁰These domains include telephone dialogues, fiction, personal letters, travel guides, Wikipedia articles, and government documents.

of SNLI. The data in XNLI is gathered from the same source as MultiNLI. 750 sentences were selected from each domain to create 7500 pairs in English, then these sentences were translated into 14 target languages by professional translators. The main motivation of XNLI was to create cross-lingual dataset to improve LLMs abilities to perform tasks in multiple languages.

The Turkish portion of XNLI constitutes the very first NLI dataset in Turkish. 2 years after its creation, it was followed by the largest NLI dataset in Turkish: NLI-TR. In order to create a Turkish NLI dataset, Budur et al. (2020) opted to translate SNLI and MultiNLI datasets and combining them to offer NLI-TR.

Stating that the significant advancements in machine translation in the last few years result in high quality translation between hundreds of language pairs, Budur et al. (2020) used Amazon Translate to translate SNLI and MultiNLI datasets from English to Turkish. Through machine translation, two subsets were created: SNLI-TR and MultiNLI-TR. Semantic relations in both subsets were labelled as follows: Entailment, contradiction and neutral.

Using machine translation for an NLI dataset poses two threats towards obtaining high quality data: Low translation quality and loss of semantic relations between sentence pairs. In order to eliminate these threats, a translation quality assurance round was done by ten bilingual speakers.

For this purpose, 500 example sets were assessed by five bilinguals in terms of translation quality and intactness of semantic relations between pairs after translation. In order to ensure a standardized annotation, the contributors were given an annotation guide. For translation quality assessment, the contributors were asked to rate translation quality between 1 and 5 (5 being the highest score). Intra-Class Correlation was adopted to assess the translation quality and a score of 0.8426 was yielded. Using Krippendorff's alpha, inter-annotator agreement was calculated as 0.47.

For the semantic label consistency, labels in the translated data set were compared to the English data set to see if they match. This comparison yielded the

annotation-level analysis with a 0.78 agreement score. Moreover, semantic label annotations offered by the team of bilinguals were compared to one another to provide majority-level analysis with a 0.99 agreement score.

Following the translation quality validation tasks, Budur et al. (2020) conducts three experiments to test the role of morphological parsing in NLI tasks using pre-trained language models. In this study, these experiments and the performance of language models on NLI-TR were taken as a baseline for CNLI-TR data.

Other large NLI data sets and resources that are worth mentioning in this section are SICK (Sentences Involving Compositional Knowledge), SciTail, Social-IQ and PAWS (Paraphrase Adversaries from Word Scrambling).

SICK data set (Marelli et al., 2014; Bentivogli et al., 2016) contains image and video captions from 8K ImageFlickr (Hodosh et al., 2013) and SemEval-2012 STS MSR-Video Descriptions (Chen & Dolan, 2011). It was first used in SemEval 2014 shared task, and then translated into different languages (Thalenberg et al., 2018; Wijnholds & Moortgat, 2021; Ghasemi & Keyvanrad, 2021). What makes SICK noteworthy is that it contains a relatedness score in addition to 3-way inference labeling: Each hypothesis and premise pair is rated from 1 to 5¹¹. Both inference annotations and relatedness annotations were done by humans.

SciTail is a unique data set as it approaches entailment through a question answering perspective and consists of "naturally occurring" text and hypothesis pairs while a vast majority of NLI data sets consists of texts and their fabricated hypotheses for the entailment task. Text and hypothesis pairs in SciTail were yielded from multiple choice, school-level science questions (Khot et al., 2018). Potential answers to the questions were used as hypotheses. In addition, relevant sentences to potential answers were gathered from a large corpus to create new text and hypothesis pairs as well. This novel approach was then adapted by few other scholars to create NLI data sets in different languages (Amirkhani et al., 2020).

¹¹1 is least related and 5 is most related in terms of common words, similar syntactic structures and synonyms.

Table 3. A Comparison of Large NLI Resources

Dataset	Domain Ct	Language	Sentence Ct	Avg Sentence Length	Label Ct
SNLI	10	English	570,152 pairs	20 tokens	3
MultiNLI	5	English	392,702 pairs	22.3 tokens	3
XNLI	5	English and 14 more	392,702 pairs	21.3 tokens (EN) 16.8 tokens (TR)	3
NLI-TR	10	Turkish	1,003,000 pairs	9.1 tokens	3
FraCaS	N/A	English	346 problem sets	10 tokens	4
SICK	6	English	9,840 pairs	9.6 tokens	3
SciTail	1	English	27,000 sentences	9.7 tokens	3
Social-IQ	24	English	7,500 questions 52,500 answers	10.4 words	2
PAWS	N/A	English	108,463 pairs	17.3 tokens	2

As a multimodal data set, Social-IQ (Zadeh et al., 2019) derives inspiration from SciTail in terms of following question-answer data structure: It contains videos of social situations which serve as text, and questions about these videos and their answers which serve as hypotheses. For each question, there are correct and incorrect answers. Correct answers demonstrate **ENTAILMENT** relation while incorrect answers demonstrate **NO ENTAILMENT** relation.

Finally, PAWS data set takes the paraphrase route (Zhang et al., 2019): It aims to assess LLMs abilities to distinguish paraphrases from non-paraphrases through an NLI task. With the aim of using adversarial examples, PAWS contains data from Wikipedia and Quora. Texts and hypotheses (potential paraphrases) have "high bag-of-words overlap but different word order" (ibid), which makes PAWS a challenging data set.

A comparison of above mentioned NLI data sets can be found in Table 2. It is essential to note that with over a million sentence pairs, NLI-TR is one of the largest NLI data sets -both in Turkish and in general. A dominant trend is to use three-way distinction: **ENTAILMENT**, **NO RELATION**, and **CONTRADICTION**. Average sentence length, on the other hand, shows great variance inter and intra language: English data sets vary between 9.6 and 22.3 tokens, currently available Turkish data sets vary between 9.1 and 16.8 tokens.

1.3 The scope of this thesis

This thesis aims to investigate De Re De Dicto ambiguities in Turkish through the lens of computational linguistics and psycholinguistics. For this purpose, a De Re De Dicto inference data set is manually generated and annotated, then this seed data is augmented with a custom textual data augmentation algorithm. Combining the seed data and augmented data, a natural language inference challenge data set (CNLI-TR) is created.

Using CNLI-TR dataset, probing experiments were conducted on different language models and classification algorithms. The aim of these experiments was to see whether these models and algorithms can recognize morphosyntactic cues and patterns to successfully complete a natural language inference task.

In order to see whether there is a processing difference between De Re and De Dicto readings of ambiguous sentences, a psycholinguistics experiment was conducted with the participation of native speakers of Turkish. In this experiment, participants were presented with disambiguating contexts and De Re De Dicto ambiguous sentences in the form of a dialogue. Following each dialogue, they were asked a comprehension question to test whether they were able to get the intended reading from the disambiguating context. Reading times of ambiguous sentences and response times for comprehension questions were compared across different conditions.

Extensive body of theoretical work on De Re De Dicto ambiguities was taken as the basis for the discussion and assessment of the computational and psycholinguistics experiments.

1.4 Packages and technical details

The following list is all of the software and packages used in this thesis:

- Python libraries: Random, Types, Pandas, PyTorch (Imambi et al., 2021), Transformers (Wolf et al., 2019), NumPy, Requests, BeautifulSoup (Richardson, 2007)

- LLMs: BERT & multilingual BERT (Devlin et al., 2018), BERTurk (Schweter, 2020)
- R: magrittr (Bache, 2019), tidyverse (Wickham et al., 2019), feather, ggplot (Wickham, 2011)

1.5 Overview

This thesis is organized as follows: Second chapter of this thesis is dedicated to the CNLI-TR data set created as part of this study. Design choices, collection and annotation process of the seed data, important characteristics of the data set, and ethical concerns on the data collection are discussed in detail. Third chapter focuses on the data augmentation process needed to generate CNLI-TR data set from the seed data. Different approaches to data augmentation, challenges of textual data augmentation, augmentation pipeline used in this study and validation process are explained. Third chapter examines the computational experiments conducted as part of the natural language inference task. Experiment settings, performances of different LLMs in the inference task, and their implications are discussed. Fourth chapter addresses the issue of sentence processing and investigates whether there is a significant difference between De Re and De Dicto readings of ambiguous sentences through a self-paced reading experiment. Sixth chapter provides a thorough discussion of the De Re De Dicto phenomenon in Turkish based on the computational and processing experiments laid out in the previous chapters and concludes this study.

1.6 Data availability

All data and code used in this thesis including CNLI-TR data, augmentation pipeline, scraping and formatting scripts, models used for computational experiments, processing experiment materials, data collection environment and analysis scripts for the processing experiment are public on a GitHub repository dedicated to this research.

CHAPTER 2

DATA

Aligned with the aim of testing LLMs abilities in understanding De Re De Dicto ambiguities, a natural language inference challenge data set is created as part of this study: Challenge Natural Language Inference Dataset in Turkish, or CNLI-TR.

CNLI-TR focuses on De Re De Dicto ambiguities and consists of annotated sentence triplets: One sentence (*the text*), its De Re and De Dicto paraphrases (*the hypotheses*) (see Figure 15).

The text contains at least one DP in the form of "bir"¹ + NP under an intensional operator. The aim of this specific structure is to offer a text that is potentially De Re De Dicto ambiguous. Existence of overt accusative marker, lexical case, phrasal expressions, scrambling and other word order changing operations in some of the texts affect its ambiguous state, thus affecting the inference label of its De Re and De Dicto paraphrases.

Both De Re and De Dicto paraphrases contain only the information presented in the text, in other words, present no new information. De Re paraphrases strictly assert a wide scope De Re reading where there is only one unique entity that saturates "bir" + NP in the text while De Dicto paraphrases assert that there is no single, specific entity that saturates "bir" + NP.

Text: Ender, Ayşe gibi bir kızla evlenmek istiyor. <i>"Ender wants to marry a girl like Ayşe."</i>
De Re paraphrase: Ender'in evlenmek istediği kız Ayşe gibi. <i>"The girl Ender wants to marry is like Ayşe."</i>
De Re label: 1
De Dicto paraphrase: Ender evleneceği kız Ayşe gibi birisi olsun istiyor. <i>"Ender wants the girl he'll marry to be someone like Ayşe."</i>
De Dicto label: 1

Figure 15. A sample annotated sentence triplet

¹"bir" in Turkish corresponds to "a" or "one" depending on the context and surrounding structure.

In the sample text in Figure 15, there is an intensional operator, "istemek" (*to want*), and there is an NP, "Ayşe gibi bir kız" (*a girl like Ayşe*), embedded in it. Thus, it meets the necessary criteria for being potentially De Re De Dicto ambiguous.

- (1) Ender, Ayşe gibi bir kızla evlenmek istiyor.
Ender-NOM Ayşe-NOM like a girl-INS marry-NMLZ want-PROG-3SG.
"Ender wants to marry a girl like Ayşe."

De Re paraphrase asserts that there is a specific girl whom Ender wants to marry, and she is like Ayşe. De Dicto paraphrase asserts that Ender has a criterion for his future spouse: She should be a girl like Ayşe. Both De Re and De Dicto paraphrases have their own label which states the inference relationship between them and the text. All sentence triplets in the CNLI-TR follow this data structure.

CNLI-TR data set contains of two subsets in the above explained format: The seed data and augmented data. Seed data consists of 743 sentence triplets and is manually generated and annotated by trained domain experts. Augmented data is artificially generated using the seed data and an augmentation pipeline created for the purposes of this study. After being generated, augmented data is revised and validated by the domain experts to ensure data quality and semantic label accuracy.

Details of the seed data generation, ethical concerns and statistics regarding the data set is discussed in the following sections of this chapter. Augmentation and validation processes are discussed in a dedicated chapter (see Chapter 3).

2.1 Creating the seed data

The creation of the CNLI-TR data set involved two main steps: the generation of the seed data and the augmentation process. To ensure the accuracy and quality of the data, a team of linguists who were native speakers of Turkish actively contributed to the study.

The recruitment of contributors was conducted through an announcement that was shared with the graduate students of the Linguistics Department at Boaziçi University. From the pool of applicants, four MA students who had successfully completed their remedial year and possessed advanced knowledge of semantics were

selected to participate in the project. Prior to joining, the chosen contributors were provided with detailed information about the project, ensuring they were well-informed about its objectives and requirements. Monetary compensation was provided to acknowledge the time and effort they dedicated to the project.

Once the contributors joined the project, a briefing session was conducted by the author to provide them with a clear understanding of the project’s roadmap, timeline, and targets.

The generation of the seed data involved two distinct steps. In the first step, the contributors were tasked with writing 50 novel sentences that had the potential to be De Re De Dicto ambiguous. In the second step, they were assigned 145 potentially ambiguous sentences for which they were required to write De Re and De Dicto paraphrases.

After the completion of the generation tasks, subsets of premise and hypothesis pairs were created and assigned to the contributors for inference annotation.

To ensure consistency and adherence to guidelines, contributors were provided with task-specific guidelines² and sample data sets. Additionally, they attended a workshop conducted by a domain expert (the author of the thesis) to enhance their skills and understanding of the task. Prior to finalizing their contributions, a pilot round was conducted, and the data produced by the contributors was reviewed by the domain expert to address any inconsistencies or issues. Feedback on the pilot task was provided to the contributors, allowing them to clarify any misunderstandings or seek further guidance from the expert.

By following these meticulous steps, including guidelines, workshops, pilot rounds, and continuous communication with the domain expert, the CNLI-TR data set was carefully constructed to ensure accuracy, consistency, and reliability.

During the first round of data generation, the team of contributors received instructions that encouraged them to introduce word order variations, as well as

²See Appendix A for the guidelines, and Appendix B for the annotation interfaces.

lexical and semantic case variations, in order to avoid creating data that was overly uniform. The aim was to ensure diversity and richness in the data set. Additionally, the contributors were encouraged to incorporate phrasal expressions, sayings, and idioms into their sentences, adding further complexity and nuance.

The ambiguous sentences generated by the contributors underwent careful examination by the domain expert. This step was crucial to ensure the quality and accuracy of the data. The expert reviewed the sentences and made revisions whenever necessary, addressing any potential issues or inconsistencies. This iterative process of review and revision helped refine the data and maintain its high quality standards.

By providing the contributors with the freedom to introduce word order variations and incorporate various linguistic elements, and by subjecting the generated sentences to expert scrutiny, the CNLI-TR data set achieved a balance between diversity and quality, making it a valuable resource for studying De Re De Dicto ambiguities in Turkish.

For the second data generation round, manually generated ambiguous sentences were labelled and sorted by the expert following these criteria:

- Word order variance. As a free word order language Hoffman (1995), Turkish allows different word order configurations in line with information structure (see Taylan (1984) and İşsever (2003) for a detailed discussion).
- Structural, semantic and lexical case variation. Some verbs in Turkish require the use of lexical case on their object. For instance, "hoşlanmak" (*to like*) requires ablative case on its object instead of accusative (see Marşan et al. (2021) and Kara et al. (2020) for a detailed discussion on particularities of lexical case in Turkish).
- Presence of modifiers before the ambiguous noun. In order to make the seed data more diverse, various modifiers (such as "başka" (*other; another*) and adjectives like "güzel" (*pretty*)) were used to modify object NPs.

- Author of the ambiguous sentences. Both the domain expert and collaborator team created gold sentences, thus each ambiguous sentence has an author."

gll

In the second data generation task, subsets of 145 potentially ambiguous sentences were created based on the categorization explained earlier. Each subset contained an equal amount of data from each category and was assigned to a contributor for the creation of paraphrases. To ensure objectivity and avoid biases, contributors were not assigned the potentially ambiguous sentences they had authored in the previous round. The order of the sentences within each subset was randomized to prevent any repetitive patterns or biases in the paraphrases.

Once the contributors completed the task of creating paraphrases, each paraphrase underwent a thorough review by the domain expert. Corrections and revisions were made as necessary to maintain the accuracy and quality of the data. This review process ensured that the paraphrases captured the intended meanings and accurately reflected the De Re and De Dicto interpretations of the original potentially ambiguous sentences.

The third task involved inference annotation (0 for no inference, 1 for inference), where each subset of premise and hypothesis pairs was assigned to contributors for annotation. The contributors annotated either the De Re or De Dicto paraphrase of the potentially ambiguous sentences. Each pair was annotated by two contributors. It was found that 89% of the inference annotations were consistent between the two contributors. For the remaining 11% of the annotations, a third contributor was assigned to determine the correct inference label and resolve any disagreements.

After the completion of the inference annotation task, the data was prepared and ready for the augmentation process, which would further enhance the dataset for studying De Re De Dicto ambiguities in Turkish.

2.2 Adversities and concerns

Data generation subtask of creating CNLI-TR data set faced some adversities due to the nature of the task, typological features of the Turkish language, and various limitations regarding the resources. In order to overcome these adversities, various new strategies were adopted. This section will take a closer look at these adversities and strategies while also discussing the concerns shaping the data generation task.

2.2.1 Ethical data collection

As the demand for extensive and reliable data in the field of NLP continues to surge, ethical concerns regarding data collection and generation have garnered significant attention. The aim is to ensure that the vast amounts of data employed to train LLMs, foundational models, and other NLP systems are sourced ethically and are free from biases.

The burgeoning discourse on ethics in NLP and data collection revolves around several key focal points (Larson, 2017; Leidner & Plachouras, 2017; Prabhumoye et al., 2020; Shmueli et al., 2021; Rogers et al., 2021):

- **Ethical Awareness:** It is imperative for researchers to uphold ethical considerations throughout every stage of their work. From the initial design to the final implementation, ethical concerns must serve as guiding principles. Researchers should remain vigilant in identifying potential biases within the data and carefully contemplate the societal impacts of their research.
- **Data Quality and Integrity:** The efficacy of NLP systems and LLMs is heavily reliant on the quality and integrity of the data used for training. Therefore, it is crucial to meticulously assess the reliability of data sources, rigorously verify annotations, and proactively address potential challenges such as noise and misinformation.
- **Data Bias and Fairness:** As previously mentioned, the nature of the data significantly influences the behavior and performance of NLP systems and

LLMs. When models are trained on biased data, they inherently perpetuate and amplify biases. Such biased models and systems propagate harmful stereotypes, reinforce problematic ideologies such as racism and misogyny, and perpetuate bias within technology, academia, and society. To combat this, researchers must exercise vigilance during the processes of data collection, data generation, and annotation to mitigate bias effectively.

- **Transparency and Explainability:** NLP systems must prioritize transparency to facilitate fair assessment and improvement. Transparent systems not only enhance the research environment but also offer clearer and more predictable impacts on society and technology. Researchers must uphold accountability by diligently documenting their research and systems, providing detailed explanations that are easily accessible to users and fellow researchers.

Throughout the execution of this study, meticulous care was devoted to adhering to the aforementioned key points, while upholding the principles of research ethics concerning the rights of the contributors and the data collection processes.

All contributors were informed of the nature, workload and deadlines of the tasks *prior* to joining the study, and they were compensated at a fair hourly rate above the minimum wage.

While discussing certain criteria that benchmark data sets should meet, Bowman & Dahl (2021) state that testing models and data sets for "plausible and harmful social bias" is, regardless of the intentions or lack thereof, contributes to the problem of developing significantly biased and harmful models. Aligned with this perspective, we executed our data set generation efforts while bearing potential biases in mind.

During both seed data generation and data augmentation process, the use of words and phrases with sexist, classist, or racist connotations is carefully and deliberately avoided. While generating synset subsets, these words were removed.

An example of such an instance was posed by the synset of the word "kadın" (*woman*) which contains the following words: "hanım" (*lady*), "hanımefendi"

Table 4. Distribution of Gender and Etymological Origin of the Proper Nouns in the Data

		Count	Percentage
Gender	Feminine	1168	45.8%
	Masculine	1264	42.3%
	Genderless	325	11.7%
Origin	Turkish	1420	51.5%
	Arabic	524	19%
	Persian	229	8.3%
	Hebrew	7	0.25%
	Multiple origin	144	5.2%
	Unknown	428	15.5%

(*madam*), "bayan" (*miss, mistress*), "karı" (*vulg. broad, woman*). While creating the subset, "karı" was removed from the list as it is often used as a pejorative term.

Moreover, gender and etymological origins of the proper nouns were considered while SR operation was being carried out. An equal distribution of feminine and masculine names was aimed. In addition, names of different origin were included in SR operations. As a result, a balanced gender distribution and etymological origin variety were achieved (see Table 4).

To uphold transparency and accountability, every step of the data generation process was diligently documented. This comprehensive documentation, accompanied by clear contributor guidelines, was openly shared and made accessible to the public. It was made available on GitHub, alongside the data itself, facilitating transparency and encouraging further scrutiny and collaboration.

2.2.2 Representational abilities and linguistic variance

Representational abilities and linguistic variance are fundamental features of a truly challenging NLI data set, as well as essential for creating high-quality NLP resources. Firstly, NLP models are expected to handle different uses of language -such as dialectal variations, formal and informal registers, word plays, or artistic and metaphoric uses of language. To achieve this, the training data must encompass a

broad spectrum of linguistic variation, including diverse language styles, extensive lexical variations, and various linguistic structures and phenomena.

Secondly, NLP models are expected to have consistently high performance across different tasks and domains. Emphasizing linguistic variance in NLP resources enables exposure to a wide array of linguistic phenomena, empowering models to handle the intricacies and subtleties of natural language while adapting effectively to different domains and challenges.

With the primary objective of constructing CNLI-TR as a demanding data set that thoroughly assesses the language understanding abilities of LLMs, ensuring linguistic variance stood as a paramount concern throughout the data generation processes. To fulfill this objective, multiple contributors were hired to avoid uniformity, and they were strongly encouraged to employ different syntactic structures, vary their register across different sentence sets, utilize idioms and phrasal expressions, experiment with diverse word orders, and employ various strategies to create both De Re and De Dicto paraphrases.

The conscientious approach taken during the development of CNLI-TR has yielded remarkable linguistic variance, despite being centered around a single linguistic phenomenon. By incorporating voice alternations, word order modifications, lexical variations, and the inclusion of idioms, a substantial level of diversity has been achieved across the De Re and De Dicto paraphrases within the data set. This deliberate attention to detail has successfully resulted in a rich and varied set of linguistic expressions, enhancing the overall quality and challenge presented by CNLI-TR.

To maintain that De Re and De Dicto paraphrases don't look too much alike, deliberate efforts were made to introduce word order variations. For instance, in sentence set 2, the De Re paraphrase employs topicalization by reordering the constituents, specifically highlighting "yorucu bir görev" (*a tiring task*), thereby diverging from the default word order in Turkish. On the other hand, the De Dicto paraphrase within the same set incorporates a relatively uncommon and outdated

structure, namely "istiyor ki" (*wants such that*), introducing syntactic variance. These intentional choices were part of the attempt to ensure differentiation and enhance the linguistic diversity between the two paraphrase types.

- (2) a. Miray yorucu bir görevi tamamlamak istiyor.
Miray.NOM tiring a task.ACC complete.NMLZ want.3.SG.PROG
'Miray wants to complete a tiring task'
b. Yorucu bir görev var Miray'ın tamamlamak istediği.
tiring a task exist.3.SG.PRS Miray.GEN complete.NMLZ want.NMLZ.POSS

'There is a tiring task that Miray wants to complete' (De Re)

- c. Miray istiyor ki tamamlayacağı görev yorucu olsun.
Miray want.3.SG.PROG ki complete.NMLZ.POSS task tiring be.IMP.3.SG
'Miray wants such that the task she will complete shall be tiring' (De Dicto)

In sentence set 3, the De Re paraphrase of the ambiguous sentence incorporates the idiomatic expression "gözüne kestirmek" (*to lay sb's eyes on sth*). During the data generation process, idioms such as this one, along with a variety of phrasal expressions, were intentionally employed. The inclusion of these idioms and expressions aimed to provide a challenge for LLMs, not only through the De Re and De Dicto ambiguities but also by presenting rich and often metaphorical language constructs. By incorporating such idiomatic and figurative expressions, CNLI-TR offers a comprehensive and diverse range of linguistic challenges for language models to navigate and comprehend.

- (3) a. Nur hediyelik eşya dükkanından bir bileklik almak
Nur.NOM gifting object shop.POSS.ABL a bracelet buy.NMLZ
istiyor.
want.3.SG.PROG
'Nur wants to buy a bracelet from the gift shop'
b. Nur'un hediyelik eşya dükkanında gözüne kestirdiği nesne bir
Nur.GEN gifting object shop.POSS.LOC eye.POSS.DAT cut.NMLZ object a
bileklik.
bracelet
'The object that Nur has her eyes on at the gift shop is a bracelet' (De Re)

- c. Nur hediyelik eşya dükkanından alacağı nesnenin bileklik
Nur gifting object shop.POSS.ABL buy.NMLZ.POSS object.GEN bracelet
olmasını istiyor.
be.NMLZ.POSS.ACC want.3.SG.PROG
‘Nur wants the object she will buy from the gift shop to be a bracelet’ (De Dicto)

To further enhance diversity both within and across sentence sets, the strategy of utilizing synonyms and quantifiers was employed. In sentence set 4, the ambiguous sentence employs the word "konu" to convey the meaning of *topic*. However, in the corresponding De Re paraphrase, a synonym, "mevzu", is used to maintain semantic variation. Similarly, the De Dicto paraphrase employs a quantifier, "herhangi" (*any*), to indicate that there is no specific or unique topic that Janset desires to discuss. By incorporating synonyms and quantifiers in this manner, CNLI-TR achieves a broader range of linguistic expressions, adding depth and variation to the data set while challenging language models to accurately interpret and differentiate between subtle nuances of meaning.

- (4) a. Janset başka bir konudan bahsetmek istiyor.
Janset.NOM another a topic talk.NMLZ want.3.SG.PROG
‘Janset wants to talk about another topic’
b. Janset’in bahsetmek istediği mevzu şu an konuşulandan
Janset.GEN talk.NMLZ want.SG.POSS topic this moment talk.GER.ABL
farklı.
different
‘The topic that Janset wants to talk about is different from the one that is being talked about right now’ (De Re)
c. Janset şimdiki konunun dışında herhangi bir konudan bahsetmek
Janset present topic.GEN out.POSS.LOC any a topic.ABL talk.NMLZ
istiyor.
want.3.SG.PROG
‘Janset wants to talk about any topic other than the one that is being talked about right now’ (De Dicto)

In addition to the previously mentioned strategies, several other techniques were employed to further diversify the data within CNLI-TR. These techniques include passivization, negation, conjunction, modification, and various other linguistic strategies.

2.2.3 Bias and annotation artifacts

Disagreements, bias and annotation artifacts are significant threats against data quality.

Disagreements occur when different annotators have varying judgments on the same data, which can be attributed to factors like task definition ambiguity, inadequate annotator training, or the differences between how humans representation of linguistic meaning varies (Pavlick & Kwiatkowski, 2019). In cases where disagreements stem from poor task definition or insufficient training, they should be treated as noise that undermines data quality and reliability. Such inconsistencies make it challenging for LLMs and NLP systems to determine the correct label, hindering their ability to perform the given task effectively.

However, it is important to note that disagreements arising from variances in linguistic meaning representation should be viewed as inherent disagreements rather than mere noise. These discrepancies can persist even when annotators receive the same guidelines and extensive training. Notably, even large and high-quality NLI datasets like SNLI and MultiNLI have been found to exhibit inconsistencies in annotation, as discovered by Pavlick & Kwiatkowski (2019).

Eliminating all occurrences of inherent disagreements is neither feasible nor straightforward, as they can originate from variations in interpersonal grammar, divergent world interpretations, or differing domain knowledge levels among annotators. Instead, a more effective approach is to equip NLP models to learn from and leverage these inherent disagreements to achieve a more robust performance.

To address noisy disagreements, comprehensive and extensive training should be provided to annotators, ensuring they have a solid understanding of the task. Additionally, each data point should be annotated by multiple annotators to obtain a consensus and mitigate the impact of individual annotator biases or errors.

Bias in annotation can arise from various sources, including personal beliefs, cultural backgrounds, individual tastes, and preferences. These biases, if present in the data, can lead to imbalances, unfairness, or even render the data useless for

certain applications. Furthermore, the existence of social biases within the data can propagate through models trained on such data, resulting in problematic, unfair, and unjust models. This can have detrimental effects on society by reinforcing biases, perpetuating discrimination, and negatively impacting minority groups.

It is crucial to address and mitigate bias in annotation to ensure that the resulting models and systems are fair, unbiased, and aligned with ethical standards. This involves taking proactive measures during the data collection and annotation processes to identify and minimize biases. Annotators should receive explicit guidelines and training to mitigate their own biases and adhere to ethical considerations. Additionally, careful monitoring and evaluation of the annotated data should be carried out to detect and rectify any potential biases.

And the final point of this subsection, annotation artifacts have a direct impact on the generalization abilities of LLMs, influencing their performance and potentially misleading assessments of their true natural language understanding capabilities. In their work, Gururangan et al. (2018) identify three types of annotation artifacts:

- Certain words or expressions that strongly imply a specific label: Adverbial modifiers like "always" or "never," negation, or quantifiers can serve as cues for inferring a particular label. These linguistic elements can inadvertently bias the annotations.
- Annotator heuristics: To expedite the annotation process and conserve cognitive resources, annotators may develop heuristics based on patterns observed in the data. For example, after encountering a substantial amount of data, annotators may generalize their own rules, such as "If a paraphrase contains negative polarity items (NPIs), it should be assigned an inference label of 0." These heuristics can introduce biases and artifacts into the annotations.
- Lexical and syntactic biases in the data: Ambiguities, complex sentence structures, infrequent words or expressions, or other biases in the data can lead to annotation artifacts. These artifacts may arise due to difficulties in

processing certain linguistic patterns, resulting in inconsistencies or biased labels.

The issue with annotation artifacts is that LLMs can exploit them without truly understanding the underlying semantics. Models can "memorize" the artifacts present in the training data and achieve high performance, yet their success may be based on exploiting these artifacts rather than genuine comprehension of the data. Consequently, a high performance on data with significant annotation artifacts does not necessarily indicate true NLU capabilities.

To address the potential pitfalls of data generation and ensure the quality of CNLI-TR, several precautions were taken. These precautions aimed to mitigate noisy disagreements, bias, and annotation artifacts, thus enhancing the reliability and usefulness of the dataset.

To prevent noisy disagreements, the team of contributors underwent extensive training before each task. They were provided with accessible guidelines to refer to whenever they encountered confusion or ambiguity. Prior to each task, a pilot run was conducted to ensure that all contributors had a clear understanding of the requirements and expectations. Additionally, a group chat was established for the contributors, enabling them to ask questions and seek clarifications from both their peers and the domain expert. These measures helped align the contributors' judgments and minimize disagreements.

To address bias and annotation artifacts, a careful approach was taken. Contributors did not label the data they created to avoid any potential bias. The process involved assigning the ambiguous sentences generated by one contributor in the first task to other contributors for the second task, where they were responsible for creating De Re and De Dicto paraphrases. In the third task, contributors did not label any ambiguous sentences or paraphrases they had written themselves. This approach reduced the potential biases introduced by individual contributors during the annotation process.

Furthermore, to mitigate the impact of annotator heuristics and the presence of lexical and syntactic biases in the data, contributors were encouraged to generate data with high linguistic and stylistic variation in the first two tasks. This diversity in language styles and structures helped minimize the influence of specific annotator heuristics. In the third task, sentence triplets were distributed evenly among the contributors, ensuring that each contributor encountered a wide range of linguistic patterns, such as topicalized sentences, negated paraphrases, and lexical variations. This approach helped address the issue of biases stemming from specific linguistic patterns or structures.

By implementing these precautions, we aimed to produce a high-quality data set with reduced disagreements, bias, and annotation artifacts. The careful training, guidelines, peer interactions, multiple annotations, and emphasis on linguistic variation collectively contributed to enhancing the reliability and integrity of the data.

2.3 Statistics

As highlighted in the previous sections, achieving diversity within the CNLI-TR dataset was a key objective. This diversity encompassed various linguistic aspects such as case markers, word order, lexical variations, and phrasal expressions, while maintaining the specific syntactic and semantic structures that give rise to De Re and De Dicto ambiguities.

By incorporating such variance in case markers, CNLI-TR offers a rich and comprehensive range of linguistic patterns. This diversity contributes to the robustness and versatility of the data set, enabling NLP models to effectively handle and understand the complexities and nuances of natural language.

In order to illustrate the relationship between word order and information structure in Turkish, different word orders were opted for in both potentially ambiguous sentences and their paraphrases. Word order variance in the potentially ambiguous sentences served both for disambiguation purposes and making the data set less unified while in De Re and De Dicto paraphrases, word order variance

reduced similarity within the paraphrases, and between ambiguous sentences and paraphrases to make the data set more challenging.

Finally, phrasal expressions were incorporated to both test the world knowledge abilities and generalization abilities of LLMs. Sentences like (5) check all the boxes to have De Re De Dicto ambiguity: An intensional operator and an NP without overt accusative marker in its scope. Despite there is nothing structural that makes De Re reading inaccessible, such sentences only allow De Dicto reading as they contain idiomatic expressions. Sentence (5) does not talk about an actual blind person or the actual event of Allah giving them a pair of eyes. Instead, it conveys a meaning where it is highlighted that the needs of people in dire situations are answered by a metaphysical force/Allah with abundance.

The inclusion of phrasal expressions within CNLI-TR serves multiple purposes, including testing the world knowledge abilities and generalization capabilities of LLMs. Examples such as sentence (5) exemplify the incorporation of phrasal expressions to elicit De Re and De Dicto ambiguity. These sentences are designed to feature an intensional operator and an NP without an overt accusative marker within its scope.

Despite the absence of any structural constraints that would render the De Re reading inaccessible, sentences like (5) exclusively allow for a De Dicto interpretation due to the presence of idiomatic expressions. In the context of sentence (5), the intention is not to discuss an actual blind person or the literal act of Allah providing them with a pair of eyes. Instead, the sentence conveys a metaphorical meaning wherein it emphasizes that people in dire situations have their needs abundantly met by a metaphysical force or Allah.

By incorporating such phrasal expressions and idiomatic language, CNLI-TR aims to assess the LLMs' understanding of metaphors and ability to generalize beyond literal interpretations. These examples challenge the models to grasp the deeper meaning and metaphorical nuances embedded within the language. This

further enhances the dataset’s capacity to evaluate the NLP models’ comprehension of complex linguistic phenomena and their real-world implications.

(5) K r istedi bir g z, Allah verdi iki g z.
blind.NOM want.3.SG.PST an eye Allah.NOM give.3.SG.PST two eye
Lit. ‘The blind wanted [one] eye but Allah bestowed them two eyes’

(6)  m  diledi bir burun, Allah bahsetti iki burun.
blind.NOM want.3.SG.PST a nose Allah.NOM give.3.SG.PST two nose
Lit. ‘The blind wished for [one] nose but Allah granted them two noses’

Following the augmentation process in CNLI-TR, it is indeed observed that the idiomatic sentences in the seed data lose their idiomatic power and the blocking effects on the De Re reading. This is primarily due to the replacement of certain elements of the sentence with synonyms or semantically related words, which alters the original idiomatic expression.

As a consequence, sentences like (5) in the augmented portion of CNLI-TR no longer maintain their idiomatic nature and restrict the interpretation solely to the De Dicto reading. On the other hand, sentence (6) exhibits the same structure but allows for both De Re and De Dicto readings, albeit sounding somewhat unnatural.

As a result of the augmentation strategy we followed, augmented portion of the CNLI-TR contains no phrasal/idiomatic expressions.

2.4 CNLI-TR: What is the challenge?

CNLI-TR, constructed around the De Re De Dicto ambiguities, serves as a platform for assessing the linguistic comprehension capabilities of LLMs, particularly in relation to recognizing various linguistic cues that influence the inference status of sentence triplets. These cues encompass elements such as word order, case markers, voice, the presence of quantifiers and modifiers, as well as idiomatic expressions.

The objective of CNLI-TR is not solely to identify instances of ambiguity but rather to capture the subtle details that resolve the ambiguity. By incorporating a wide range of linguistic features, the dataset challenges LLMs to discern and effectively

utilize morphological and information structural aspects to infer the correct interpretation.

Most of the existing research on NLI capabilities of LLMs has primarily focused on high resource languages, including English, French, Spanish, and German. These languages often exhibit fixed word order and possess limited morphological complexity due to their analytic or isolating nature. In contrast, languages like Turkish, with rich morphology and flexible word order, have received relatively less attention in exploring how well LLMs can grasp and leverage morphological information in NLI tasks. Consequently, our current understanding of LLMs’ performance in this aspect remains limited and somewhat unreliable.

Given this context, CNLI-TR presents a unique opportunity for both LLMs and researchers to delve deeper into the challenges posed by morphology. By emphasizing the significance of morphology and information structure, the dataset encourages a more comprehensive examination of the morphological intricacies within the Turkish language and similar morphologically rich languages. This, in turn, fosters a deeper understanding of how LLMs can effectively capture and utilize morphological cues for enhanced language understanding.

2.5 Further improvements

Currently, an overwhelming majority of the sentence triplets in CNLI-TR data set is truly ambiguous —Almost 80% of the texts allow inferring both hypotheses, which makes CNLI-TR a very unbalanced data set.

It is important to address the imbalance in the CNLI-TR data set, where a significant majority of sentence triplets (almost 80%) allow for both hypotheses. This lack of balance can affect the performance and generalizability of NLP models trained on the data set. To improve CNLI-TR, efforts can be made to create new sentence triplets that introduce more diversity and balance in terms of the inference status.

In addition, incorporating naturally occurring data can enhance the authenticity and expressiveness of CNLI-TR. This can involve collecting and annotating sentences from various sources such as books, articles, or online texts, ensuring a broader representation of language use and reducing the reliance on fabricated or artificial data.

Table 5. Inference Label Distribution of CNLI-TR

	Train	Dev	Test
Inference	80%	79%	81%
No inference	20%	21%	19%

Introducing contextual information, such as preceding or following sentences, can indeed provide additional cues to resolve ambiguity and enhance the understanding of the inference status. By considering the surrounding context, models can benefit from a wider range of linguistic cues and make more accurate predictions.

Expanding the dataset and incorporating these improvements will not only address the issues related to imbalance and artificiality but also contribute to the overall quality and usefulness of CNLI-TR as a resource for evaluating and advancing LLMs.

CHAPTER 3

AUGMENTATION PIPELINE

Current developments in transformers and neural networks allow language models to introduce significant amounts of parameters to improve predictions in classification tasks. Hence the size of the data plays an important role in the performance of language models (Tong & Koller, 2001; Tang et al., 2015; Liu et al., 2020) as data sparsity leads to issues like overfitting (Hawkins, 2004).

Hence, the primary objective of this study was to provide a high-quality, extensive data set. Generating a sufficiently large data set manually is both time-consuming and expensive. Moreover, identifying De Re De Dicto ambiguous sentences within a corpus or through web scraping proves to be exceedingly challenging due to their specific structural requirements. Furthermore, the process of locating De Re De Dicto ambiguous sentences necessitates the additional task of creating their De Re and De Dicto paraphrases. Consequently, augmenting the existing seed data emerged as the more feasible approach for CNLI-TR.

3.1 Data augmentation strategies

Neither sparse data issue (especially when the data is labelled and of high quality) nor data augmentation is a new topic of discussion. In fact, data augmentation has been first used in computer vision tasks (Simard et al., 1998) and are now a common practice in both computer vision and speech recognition (Krizhevsky et al., 2017; Ko et al., 2015). Considering that augmentation techniques like changing the brightness, color, or orientation of an image or adjusting the pitch, volume and speed of an audio don't necessarily change the "semantics of the input (Liu et al., 2020), augmentation is not a very challenging task for computer vision and speech recognition.

Simply changing the rotation of an image, or getting the reflection of it allows creating multiple "augmented" versions without tampering with the quality of contents of the image (see Figure 16 for a sample augmented data for computer vision). Yet the same cannot be said for text data since preserving semantic labels

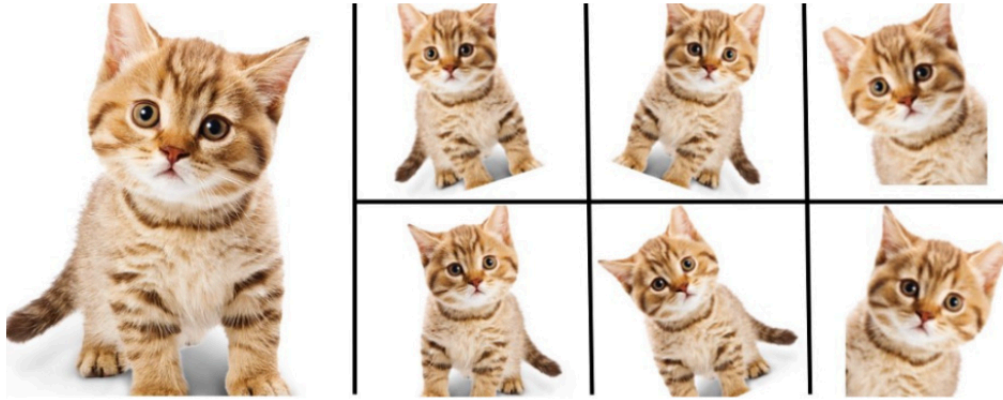


Figure 16. Data augmentation for computer vision Liu et al. (2020)

through augmentation process is rather strenuous. As a result, data augmentation for NLP is still a challenging and relatively new practice in its early stages.

Despite its challenging nature, recent studies clearly illustrate that text augmentation improves model performance especially in classification tasks (Wang & Yang, 2015; Wei & Zou, 2019) and highlight the need for it.

In order to address the issue of augmenting text without compromising the quality or creating semantically irrelevant data, different strategies have been followed. Thoroughly discussing these strategies, Liu et al. (2020) offer a classification (see Figure 17):

- **Unconditional data augmentation:** Using unlabelled data, unconditional augmentation offers two paths: Back-translation and lexical substitution. Back translation is performed by translating the text into a different language and re-translating it back into the source language. Lexical substitution is based on creating noise by randomly adding new words, replacing or removing words as well as changing the word order.
- **Conditional data augmentation:** Conditional augmentation requires the use of labelled data and a language model. Yet it has a higher implementation cost and requires a relatively higher amount of data than unconditional augmentation, so the language models can perform well.

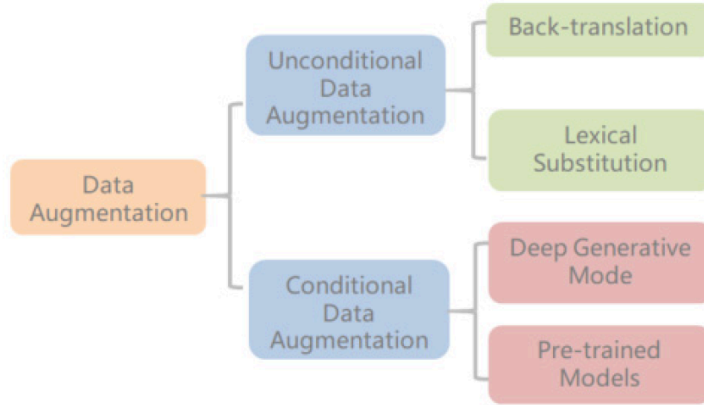


Figure 17. Classification of text data augmentation Liu et al. (2020)

To identify the most suitable strategy for our specific objectives and the distinctive characteristics of our data, we thoroughly evaluated both conditional data augmentation and unconditional data augmentation approaches.

Considering conditional data augmentation initially seemed counterintuitive due to the limited size of our seed data, which consists of only 743 sentence triplets. This data set size is considerably small to effectively implement even the most advanced deep generative models. Additionally, pre-trained models also exhibit subpar performance when trained on such a small data set.

Similar to conditional data augmentation techniques, unconditional data augmentation operations pose some serious setbacks as well mainly because of the characterising property of CNLI-TR data set: Potential ambiguity of premise sentences. This ambiguity is caused by a very specific syntactic structure: All premise sentences have an intensional operator and a NP c-commanded by this intensional operator. The randomness of unconditional data augmentation operations risks the integrity of this structure and as a result, the potential ambiguity.

The first method of unconditional data augmentation is back translation: The seed data is translated to a mediator language, then this target sentence is re-translated to the source language to create augmented text data. Due to budgetary restrictions, working with human translators was not possible for our study. Hence the only viable option for back translation was using machine translation tools yet

this method compromised the syntactic integrity and case morphology vital to preserving the potential ambiguity in the text.

Despite the considerable progress made in neural machine translation and the subsequent enhancements in the quality of machine translation (Sennrich et al., 2015), the effectiveness of back translation is directly influenced by typological constraints. When the mediator and source language pairs share a significant degree of typological similarity, or are even linguistically related, such as English and French, back translation tends to yield favorable results by providing augmented text that preserves the semantic profile to a large extent. This phenomenon is clearly illustrated in sentence set (1), where sentence (1a) is translated into French using Google Translate, resulting in sentence (1b). Subsequently, sentence (1b) is translated back into English, resulting in sentence (1c). Through this process, neither crucial information was lost nor an ungrammatical sentence was generated.

- (1) a. Tijen wants to eat a very delicious and *hearty* dessert like the ones in the restaurant.
- b. Tijen veut manger un dessert très savoureux et *copieux*
Tijen.NOM want.3.SG.PRS eat.INF a dessert very tasty and copious
comme ceux du restaurant.
like those of restaurant
'Tijen wants to eat a very tasty and copious dessert like the ones in the restaurant.'
- c. Tijen wants to eat a very tasty and *filling* dessert like the ones in the restaurant.

When the source and mediator language pairs exhibit significant typological differences, such as Turkish and English, back translation tends to lead to the loss of crucial information, including morphological cues that can alter the inference status and labels of the hypotheses. An example of this can be observed in sentence set (2), where sentence (2a) is translated into English using Google Translate. The resulting English sentence, (2b), is then translated back into Turkish, resulting in (2c). While (2b) represents a valid translation, several issues arise when returning to Turkish in (2c). Firstly, *hearty* is translated into Turkish as "içten", which actually means

sincere. Secondly, the explicit accusative marker on "tatlı" (*dessert*) in (2a) is lost in (2c).

- (2) a. Tijen restorandakiler gibi lezzetli ve doyurucu bir tatlı-yı
Tijen.NOM restaurant.LOC.ki.PLU like tasty and filling a dessert.ACC
yemek istiyor.
eat.INF want.3.SG.PRS
'Tijen wants to eat *the* tasty and filling dessert like the ones in the restaurant'
- b. Tijen wants to eat a tasty and *heartly* dessert like the ones in the restaurant.
- c. Tijen, restorandakiler gibi lezzetli ve içten bir tatlı-∅ yemek
Tijen.NOM restaurant.LOC.ki.PLU like tasty and sincere a dessert-∅ eat.INF
istiyor.
want.3.SG.PRS
'Tijen wants to eat *a* tasty and *sincere* dessert like the ones in the restaurant'

The embedded object in sentence (2a) bears overt accusative case marking which indicates specificity (Enç, 1991; Kechriotis, 2009). Consequently, the De Re paraphrase of (2a) must be labeled as 1 (INFERENCE), while its De Dicto paraphrase must be labeled as 0 (NO INFERENCE). However, in (2c), the embedded object lacks the overt accusative case marking. As a result, both the De Re and De Dicto paraphrases of (2c) should be labeled as 1 (INFERENCE).

A rather popular method of unconditional data augmentation is creating noise in the seed data. For this purpose, Wei & Zou (2019) proposes a language-independent framework called EDA (Easy Data Augmentation) which consists of four operations (see Table 6): Synonym replacement (SR), random insertion (RI), random swap (RS), and random deletion (RD).

The most important quality of EDA is the fact that it doesn't require a significant amount of data to be successfully carried out, yet it has its own pitfall as an unconditional augmentation technique: Proposed operations are carried out randomly. Hence the risk of compromising semantic labels and syntactic structure is still prevalent, especially with random swap and random deletion as these two operations can remove core elements like the intensional operator and DP, or move DP out of the c-command domain of intensional operator.

Table 6. Noise Creating Operations Proposed in EDA (Wei & Zou, 2019)

Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
Synonym Replacement (SR)	A <i>lamentable</i> , superior human comedy played out on the backward road of life.
Random Insertion (RI)	A sad, superior human comedy played out on <i>funniness</i> the back roads of life.
Random Swap (RS)	A sad, superior human comedy played out on <i>roads</i> back <i>the</i> of life.
Random Deletion (RD)	A sad, superior human out on the roads of life.

After much deliberation, we decided that the best route for us was following a hybrid approach in order to successfully augment our relatively smaller seed data set while preserving the inference labels. Thus, we decided to carry out lexical substitution, more specifically synonym replacement operation, on labelled data.

The details of the augmentation algorithm and lexical substitution process is thoroughly discussed in the following section.

3.2 Augmentation Algorithm

An augmentation algorithm that performs non-random lexical substitution operation was created for this study. The main idea behind the algorithm is to deliberately replace certain elements in each sentence with their synonyms, hyponyms, or hypernyms. In order to avoid random replacements that would change the semantic profile of the sentence, the algorithm refers to manually generated labels in the data.

Before being augmented, the seed data was first divided into 6 splits for the ease of data handling. Then these splits were converted into CoNLL-U format using a conversion algorithm. Using BoAT (Akkurt et al., 2022), the data was semi-automatically annotated in UD-style¹ by a single annotator who is familiar with the framework. The reasoning behind working with a single annotator for this task was to ensure coherency and uniformity.

¹Universal Dependencies (UD) is a morphosyntactic annotation framework and a collaborative treebank project. See De Marneffe et al. (2021) for details.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
+ 1	Deniz	Deniz	PROPN	-	Case=Nom Number=Sing Person=3	-	-	-	-
+ 2	bir	bir	DET	Indef	-	-	det	-	-
+ 3	dili	dil	NOUN	-	Case=Acc Number=Sing Person=3	-	-	-	change=yes
+ 4	konuşmak	konuş	VNOUN	-	Case=Nom Number=Sing Person=3 Polarity=Pos VerbForm=Vnoun	-	ccomp	-	-
+ 5	isterse	-	-	-	-	-	-	-	-
+ 6	kesinlikle	kesinlikle	ADV	-	-	-	advmod	-	change=yes
+ 7	hemen	hemen	ADV	-	-	-	advmod	-	change=yes
+ 8	öğrenir	öğren	VERB	-	Aspect=Hab Number=Sing Person=3 Polarity=Pos Tense=Pres	0	root	-	-
+ 9	.	.	PUNCT	Stop	-	-	punct	-	SpacesAfter=\n

Figure 18. Sample annotation of seed data on BoAT interface

The annotation process only addressed word forms, UPOS tags and morphological features of words that are relevant to the augmentation process (see Figure 21). No dependency relation was annotated, moreover no annotation was done for words that will not undergo lexical substitution process. In order to guide the algorithm throughout the augmentation process, words to be replaced were marked with `change=yes` function in their MISC column.

After being annotated, the seed data was reviewed once by the annotator to ensure consistency and accuracy. Then it was fed into the augmentation algorithm for SR which strictly follows the constraints laid out below:

- Proper nouns are replaced with proper nouns.
- Intentional operators are replaced with intentional operators.
- Adjectives are replaced with synonymous adjectives.
- Object NPs are replaced with synonymous NPs.

For proper nouns, a database was generated using a web scraper. This database contains names, their origin and gender. Intensional operators are stored in a manually generated list.

For nouns, adverbs and adjectives, relevant synsets were extracted from Turkish WordNet (Bakay et al., 2021). Then these synsets were manually revised to exclude instances that would compromise the meaning or semantic integrity of the

seed sentences. When synonyms were not available, hyponym and hypernyms were used. After the manual revision process, a subset of WordNet synsets were created.

Following the predefined SR constraints, the augmentation algorithm refers to the proper noun database, intensional operator list, and synset list. Since seed data only consisted of 743 sentences, several rounds of augmentation operations were carried out on each split to create a large data set. For this purpose, three different SR operations were defined:

- Operation 1: Replace proper nouns.
- Operation 2: Replace substantives.
- Operation 3: Replace proper nouns and substantives.

These operations were sequentially applied to each split, following a predetermined order as outlined in Table 21. To introduce sufficient variation, each split underwent two different operations in different orders. As a result, nine versions of each split were created, with each version experiencing a distinct combination of operations in varying sequences.

Table 7. Augmentation Operations Performed on Each Split

Input	Operation	Output
Split n	Operation 1	V1
Split n	Operation 2	V2
Split n	Operation 3	V3
V1	Operation 2	V4
V1	Operation 3	V5
V2	Operation 1	V6
V2	Operation 3	V7
V3	Operation 1	V8
V3	Operation 2	V9

Following the augmentation process, a total of 8,800 sentence triplets were generated. However, due to the absence of proper nouns in some sentences from the seed data, duplicate sentences emerged during the augmentation process. These

Table 8. Changes Made to the Inference Labels in Validation Process

	De Re	De Dicto
1 to 0	854	788
0 to 1	791	788
Total Changes	1645	1576

duplicates were subsequently removed, resulting in a final data set comprising 7,420 unique sentence triplets and a total of 22,260 individual sentences.

3.3 Validation and Post Editing

After the augmentation process, all versions of the dataset underwent a thorough revision by a domain expert to ensure the quality of the augmentations. During this review, any case morphology errors and typos were corrected.

Following the initial revision round, specific attention was given to the versions in which the substantives underwent the SR operation (V2, V3, V4, V5, and V7). In these versions, additional revisions were conducted by the contributors, and any necessary corrections were made to the inference labels.

Following the removal of duplicates, the augmented dataset comprised a total of 20,031 unique sentences, accompanied by 13,354 inference annotations. A domain expert conducted case morphology and typo fixes for 9,052 sentences. Additionally, after the augmentation process, contributors made changes to the inference labels, amounting to only 24.12

Out of the 3,221 instances where the inference labels were revised, 2,423 of them were attributed to nuances in the meaning of adjectives. These nuances necessitated adjustments to ensure the accurate representation of the intended meaning.

Text: Ergalip *değişik* bir konudan bahsetmek istiyor. (nto_24289)
"Ergalip wants to talk about a different topic."

De Re paraphrase: Ergalip'in bahsetmeyi istediği *özgün* bir konu var. (ndm_96578)
"There is an original topic that Ergalip wants to talk about."

Inference: 0

Figure 19. A sample augmented hypothesis pair

Text: Cankut bir tabibe görünmek istiyor. (<i>lbo_12055</i>)
"Cankut wants to see a doctor."
De Dicto paraphrase: Suat'n görünmek istediği şahısın mesleği tabiplik. (<i>tke_99288</i>)
'The profession of the person that Suat wants to see is medicine.
Inference: 0
De Dicto paraphrase: Suat'ın görünmek istediği herhangi bir tabip. (<i>wwq_33595</i>)
"Suat wants to see any doctor."
Inference: 0

Figure 20. A sample augmented hypothesis triplet with proper noun mismatch

One of the label change instances is illustrated in Figure 19. Here the object NP in the text is modified by "değişik" (*different*) while the same NP is modified by another adjective with a similar but different meaning: "özgün" (*original*). Hence, the inference relation of this augmented pair was annotated as 0 (NO INFERENCE) while the corresponding sentence in the seed data had 1 (INFERENCE) label.

After the label changes, both hypotheses of 399 sentence triplets were annotated with 0 (NO INFERENCE). There were two reasons behind this revision. The first reason was the mismatch in proper names as a result of Operation 1 and/or Operation 3 (see Figure 20).

Another important aspect to consider was the presence of idioms, phrasal expressions, and proverbs in the seed texts. During the augmentation process, these sentences underwent transformations that caused them to lose their original intended meanings, leading to a shift in their inference labels towards the negative direction. In the case of idioms and phrasal expressions, the De Dicto readings were no longer preserved, while for proverbs, both readings were lost.

Despite these changes, the entirety of these triplets was retained within the dataset. This decision was made in order to maintain a diverse range of examples and to serve as a benchmark for evaluating the ability of language models to detect the absence of inference relations in instances that do not rely on external world knowledge.

CHAPTER 4

COMPUTATIONAL EXPERIMENTS

For the experiment phase, CNLI-TR was split into three sets for a 95% confidence rate. Train set consists of 10,726 sentence pairs, development set consists of 2,094 sentence pairs, and test set consists of 1,619 pairs. While splitting the data, a balanced distribution of inference (1) and no inference (0) labels across the data sets was ensured.

Table 9. Inference Label Distribution in Train, Dev and Test Splits of CNLI-TR

	Train	Dev	Test
Inference	80%	79%	81%
No inference	20%	21%	19%

In order to see model performance on CNLI-TR data set, two sets of experiments were conducted. In the first set, we compared LLMs performance on CNLI-TR and NLI-TR by replicating Budur et al. (2020)’s experiments on BERT, BERTurk and multiBERT. In the second set, we investigated the performance of less sophisticated classification algorithms compared to LLMs: Random Forest and K-Nearest Neighbors (KNN) algorithm.

The following subsections will present and discuss the results of these experiments.

4.1 LLM experiments

To gauge the performance of the LLMs on CNLI-TR data, we undertook a dual-phase experimentation process: During the first phase, our objective was establishing a foundational baseline. This served as a prerequisite for facilitating a comprehensive juxtaposition between NLI-TR and CNLI-TR. Subsequently, in the second phase, our focus shifted to exploring the impact of morphological analysis on NLI tasks.

The research conducted by Budur et al. (2020) on NLI-TR represents a comprehensive study on both NLI task and NLI data in Turkish. In our work, we

adopted their experiments and experiment settings as the baseline and replicated them using CNLI-TR data.

Following the methodology of Budur et al. (2020), we conducted our LLM experiments on BERT-En (Devlin et al., 2018), BERT-Multi, and BERTurk (Schweter, 2020). Each model was individually fine-tuned for three epochs, employing a maximum sequence length of 128, a batch size of 8, and a learning rate of 2×10^{-5} . We utilized the HuggingFace Transformers library (Wolf et al., 2020) for implementation, adhering to the exact procedures followed in the NLI-TR research.

4.1.1 Setting the foundational baseline

For the first round of LLM experiments, all models underwent fine-tuning on the train split of CNLI-TR and were subsequently evaluated on the test set. However, due to the unbalanced distribution of inference labels in the dataset (see Figure 9), the predictions made by all three models were heavily skewed towards the inference class (1), as indicated in Table 10. Consequently, these models failed to identify sentence triplets where both paraphrases lacked any inference, and consistently predicted 1 even for unambiguous sentences. This included sentences where the object NP possessed an overt accusative marker, or non-canonical word order, rendering the De Dicto paraphrase inaccessible.

Table 10. Precision, Recall and F1 Scores of Three Language Models Fine Tuned on CNLI-TR Dataset

		BERT-En	BERT-Multi	BERTurk
Non-weighed	Precision	0.81	0.83	0.81
	Recall	0.91	0.92	0.90
	F1	0.85	0.87	0.85
Weighed	Precision	0.68	0.73	0.79
	Recall	0.71	0.74	0.75
	F1	0.69	0.73	0.76

In an effort to address the data imbalance and obtain a more comprehensive evaluation of the language models' capabilities, we conducted an additional experiment. This time, we utilized a cross-entropy loss function with adjusted class

weights. The weight assigned to the inference class (1) was reduced, while the weight for the no inference class (0) was increased. As a result, although the prediction recall was heavily impacted, the models exhibited improved performance in detecting instances of no inference.

Yet even with adjusted class weights, all three models failed to recognize the instances of idiomatic expressions, phrasal verbs, and similar metaphoric uses of language that changed the inference label status of the hypotheses.

4.1.2 Investigating the effects of morphological analysis

After compensating for the label imbalance in the data set, we investigated the effect of morphological analysis in our NLI task. For this purpose, Budur et al. (2020) added a morphological analysis layer with regular tokenization on top of the LLMs and compare the performance of 3 different parsers (Akin & Akin, 2007; Sak et al., 2011; Öztürel et al., 2019) and showed that in scenarios where sparse data is an issue, morphological analysis provides supplemental information which improves LLM performance. As the size of dataset increases, the presence of a morphological analysis layer in the LLM did not result in a significant improvement, particularly after 20,000 intervals (depicted in Figure 22).

As CNLI-TR was relatively smaller than NLI-TR, exploring the effects of morphological analysis was promising. For this purpose, we carried out another round of LLM experiments where we incorporated a layer of morphological analysis.

Maintaining uniformity with our prior experimental parameters and class weights, we retained the same configurationsave for one crucial alteration: the integration of the Stanza morphological parser (Qi et al., 2020). This augmentation facilitated a more granular understanding of the interplay between morphology and the underlying language tasks.

During this phase, we carried out three distinct experiments. In the first experiment, we first did a morphological analysis of sentence triplets and yielded

features for each lemma. Then we concatenated these features to each sentence. This resulted in a modified input structure as follows:

Ali bir kediyi gördü. | PROPN | Case=Nom | Number=Sing | Person=3 |
DET | Case=Nom | Number=Sing | Person=3 | NOUN | Case=Acc | Number=Sing |
Person=3 | VERB | Number=Sing | Person=3 | Tense=Past

For the second experiment, we again did a morphological analysis of the sentence triplets. However, in this iteration, we took a different route by appending the morphological attributes of each word to the word itself:

Ali .PROPN | Case=Nom | Number=Sing | Person=3 bir .DET | Case=Nom |
Number=Sing | Person=3 kedi .NOUN | Case=Acc | Person=3
gördü .VERB | Number=Sing | Person=3 | Tense=Past

In our last experiment, we took a different approach by providing only the morphological features to the LLMs. This resulted in a simplified input structure, as depicted below:

PROPN | Case=Nom | Number=Sing | Person=3 | DET | Case=Nom | NOUN |
Case=Acc | Person=3 | VERB | Number=Sing | Person=3 | Tense=Past

Compared to the weighed class experiment setting in the first round, morphological analysis layer had a noticable impact on the LLM performance (see Table 11).

Table 11. Comparison of Weighed Class LLM Experiments with Morphological Analysis

		BERT-En	BERT-Multi	BERTurk
Morphological Features Concatenated to the SENTENCE	Precision	69	77	79
	Recall	71	74	75
Morphological Features Concatenated to the LEMMA	Precision	75	80	83
	Recall	72	78	81
Morphological Features ONLY	Precision	67	77	78
	Recall	73	76	78

The approach where morphological features were concatenated to the lemma showed the most substantial increase in precision. On the other hand, when only morphological features were fed into the LLMs, the results were actually worse than the baseline values we established in the previous round of experiments.

The approach of concatenating morphological features to the sentences yielded subpar results. This drop in performance can be attributed to the limitations of LLMs in effectively parsing and correlating morphological features with their corresponding lemmas.

Despite the consistent improvement in recall scores achieved by introducing additional morphological information, all three LLMs still struggled to accurately identify unambiguous instances, failing to attain high performance levels.

Furthermore, the integration of morphological information proved ineffective in recognizing idiomatic expressions and the figurative use of language.

4.1.3 Overview of the results

Both rounds of experiments yielded results that demonstrated the language models' ability to successfully identify true ambiguities, where both the De Re and De Dicto hypotheses were inferable. However, these models struggled in recognizing the structural and linguistic elements that indicate the absence of De Dicto inferences. Furthermore, they faced challenges in detecting instances where neither the De Re nor the De Dicto inference was available. Consequently, these experiments show that language models fail to incorporate morphological cues and word-order based cues regarding the information structure while making decisions which allow disambiguation of De Re De Dicto ambiguities.

In comparison to the performance of NLI-TR (depicted in Figure ??), the precision achieved in the CNLI-TR experiments is notably lower. This indicates that our initial objective of challenging the abilities of LLMs has been accomplished using the CNLI-TR data set. The lower precision values suggest that the LLMs faced

Model Name	SNLI-TR		MultiNLI-TR	
	Dev	Test	Matched Dev	Mismatched Dev
BERT-En	81.83%	82.09%	69.98%	70.56%
BERT-Multi	85.37%	85.12%	75.97%	76.34%
BERTurk	87.28%	87.04%	79.58%	80.87%

Figure 21. Precision scores of two subsplits of NLI-TR (Budur et al., 2020)

greater difficulty in accurately determining the correct inferences in the CNLI-TR task.

An essential question that arises at this juncture is: What exactly do LLMs learn about the structure of language? The CNLI-TR experiments illustrate that LLMs do not pay significant attention to morphosyntactic cues, at least in the case of Turkish. In their NLI-TR experiments, Budur et al. (2020) investigate the impact of morphological parsers on the NLI performance of the models in a separate experiment. The parameters used in this experiment were the same as those in the baseline experiments, which served as the foundation for our own experiments. The models were trained for 1 epoch, and the accuracy in the evaluation sets was measured at intervals of 1,000 training steps. Interestingly, the presence of an additional morphological parser layer in the LLM did not result in a significant improvement, particularly after 20,000 intervals.

This pattern highlights the importance our question and implies that LLMs might utilize morphosyntactic information in a less efficient manner than initially anticipated. The reason for this could be attributed to the architectural characteristics of LLMs, the pretraining data they were exposed to, or the fact that BERT-En and BERT-Multi were not specifically pretrained for Turkish, and BERTurk only utilized 30 GB of data for pretraining. In any case, we must carefully consider our expectations from LLMs and the manner in which we pretrain them.

Another noteworthy observation is that neither of the LLMs tested in this stage demonstrated the ability to recognize idioms, phrasal expressions, and proverbs within the dataset. Expressions such as "Kör istedi bir göz, Allah verdi iki göz" were treated on par with other sentence triplets in the data. Consequently, the models failed

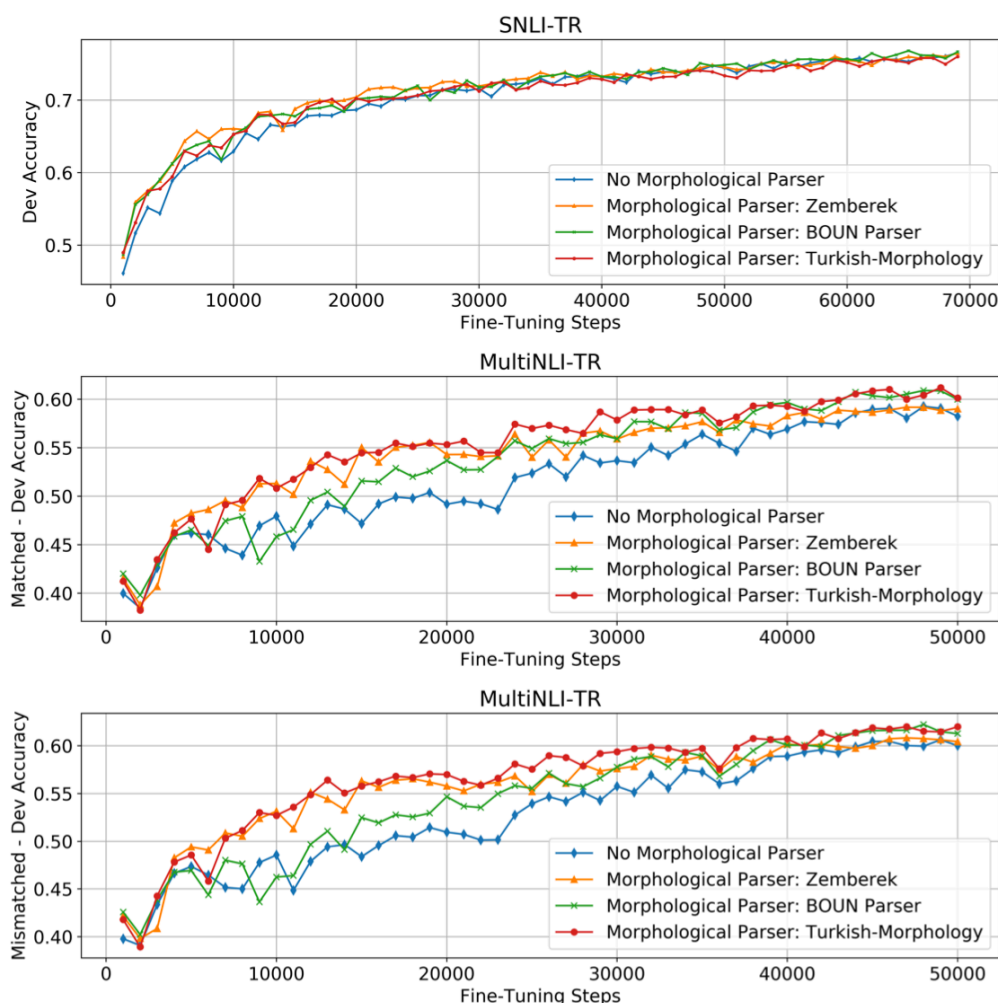


Figure 22. Morphological parsers' effect on NLI performance

to detect the presence of metaphorical language usage and the absence of De Re readings in such cases. This highlights a limitation in the models' understanding of figurative language and their ability to discern nuanced linguistic elements.

Lastly, our experiments underscore the necessity to redefine our interpretation of "language comprehension" in the context of LLMs. Even when provided with morphological analysis, these models tend to either disregard or fail to link this supplementary information to the primary task at hand, which involves discerning the Natural Language Inference (NLI) relationships within sentence triplets.

Furthermore, these models do not establish the expected connection between word order and the underlying information structure, a capability we humans rely

on to draw inferences between sentences and identify situations where De Re or De Dicto paraphrases are not applicable.

4.2 Random Forest and KNN experiments

As part of our computational experiments, we evaluated the NLI capabilities of two learning algorithms: Random Forest (RF) (Liu et al., 2012) and K-Nearest Neighbor (KNN) (Soucy & Mineau, 2001).

To conduct these experiments, we merged the train and dev splits of CNLI-TR and utilized them for training purposes. The test split was reserved for evaluating the performance of the models. Given the significant impact of the imbalanced nature of the data set on the model performance in LLM experiments, we addressed this issue by introducing class weights in both the RF and KNN experiments. This allowed us to mitigate the effects of data imbalance and enhance the overall performance of the models.

Two experiments were performed for both algorithms. The first experiment involved using lemmatization, while the second experiment was conducted without lemmatization. The purpose of this comparison was to determine whether lemmatization had a positive impact on the performance of the algorithms.

For the RF algorithm, we calibrated it with 100 n estimators and 5 minimum sample leaves. To balance class weights, we utilized the `balanced_subsample` object in the RF library. This approach helped address the imbalanced nature of the data set and improve the overall performance of the Random Forest model.

Regarding the KNN algorithm, we calibrated it with 200 n neighbors and employed the distance weight option. This option assigned weights to points based on the inverse of their distance, enabling the model to consider the relevance of neighboring points when making predictions.

Interestingly, the introduction of lemmatization had a negative impact on recall but a positive effect on precision for both inference and no inference labels in both the KNN and RF algorithms (see Table 12).

Despite their simpler nature compared to the more sophisticated LLMs discussed in the previous section, both KNN and RF demonstrated performance that was comparable to those advanced models. This finding highlights the effectiveness of these traditional machine learning algorithms in tackling the NLI task, even in comparison to more complex language models.

Table 12. KNN and RF Performance on NLI

		KNN		RF	
		Precision	Recall	Precision	Recall
No	Label 1	83	78	80	78
Lemmatization	Label 0	29	40	28	40
With	Label 1	85	35	84	98
Lemmatization	Label 0	97	35	83	36

Again, similaar with LLMs, both models failed to detect a very large portion of no inference paraphrases.

4.3 Discussion and further research

Assessing the abilities of LLMs should not be considered as a simple process that measures LLM performance on a set of tasks. The linguistic awareness and generalization abilities of LLMs should be considered in addition to task performance.

With a challenging data set like CNLI-TR, even state of the art LLMs fail to detect nuanced linguistic cues to detect no inference instances, or idiomatic expressions which require both world knowledge and linguistic knowledge.

What CNLI-TR has shown us is that currently available LLMs have a subpar ability to recognize the interaction between word order, morphosyntax, and information structure. Moreover, they seem to not prioritize information provided by the morphology while making NLI predictions. This is an important takeout that should be considered while advancing LLMs and foundational modals.

In further research projects, we aim to expand CNLI-TR data set by introducing more data to balance out the inference labels, test other models, and investigate the contribution of other morphological parsers to the LLM performance.



CHAPTER 5

PROCESSING EXPERIMENT

As extensively discussed in Chapter 1, all established formal linguistic analyses of De Re De Dicto ambiguities predominantly view the De Dicto reading as the default interpretation: It requires no adjustments, movement, or operation in LF. Additional operations, such as movement or quantificational scope changing, are only referred to for deriving the De Re readings. With this context in mind, our primary research question, which this experimental study seeks to address, is as follows: Does the inherent complexity asymmetry observed in theoretical accounts of De Re De Dicto ambiguities translate into the processing of these readings? In other words, we aim to investigate whether there are any discernible differences in the online or offline processing of De Re and De Dicto readings.

The existing body of experimental research on De Re De Dicto ambiguities is relatively limited; however, the available results are intriguing and thought-provoking. As mentioned in the study by Tiskin (2021), two experiments conducted by Zhang & Davidson (2021), Drozdovskaia (2021) demonstrate that in English and Russian, the De Re interpretation is generally more challenging to comprehend compared to the De Dicto interpretation. However, it is important to note that this pattern may not hold true for other languages.

In this experiment, our objective is to investigate De Re De Dicto ambiguities in Turkish from a psycholinguistic perspective, employing an approach that has not been previously utilized in experimental research on these ambiguities. We aim to compare the online and offline processing costs of De Re and De Dicto readings by presenting these ambiguities in the context of natural language dialogues.

Additionally, we seek to explore both narrow scope and wide scope interpretations of De Re readings. These specific aspects have not yet been thoroughly examined by previous scholars investigating the processing of De Re De Dicto ambiguities.

For this purpose, a self-paced reading experiment was designed and shared with 186 participants mainly from Boğaziçi University through PCIBexFarm. In the

following sections, the experiment design and results will be discussed in detail. All experiment materials and fillers are available on a GitHub repository, and in Appendix C. A demo run of the experiment can be accessed on PCIBex Farm.

5.1 Experiment design

To investigate the claims regarding complexity made by compositional semantics accounts, a self-paced reading experiment was conducted. The experiment items were presented to participants in the form of a dialogue between two individuals, aiming to closely resemble naturally occurring instances of De Re De Dicto ambiguity.

After completing the self-paced dialogue reading task, participants were then presented with comprehension questions. These questions required participants to response Yes/No questions based on their understanding of the dialogue they had just read. The primary objective of these comprehension questions was twofold: first, to assess whether participants were able to access the intended disambiguation of the De Re De Dicto ambiguous sentence presented in the dialogue; and second, to record response times as a measure of the offline processing burden associated with the interpretations of De Re and De Dicto readings.

In the subsequent subsections, a comprehensive overview of the experiment design, including its various components and details will be provided.

5.1.1 Participants

Prior to conducting the experiment, ethical approval was obtained from the Boğaziçi University Ethics Committee (refer to Appendix ?? for the corresponding document). The experiment was then administered to the students of the Boğaziçi University Linguistics Department. Participants who completed the experiment were granted extra credit in a Linguistics course of their choice. Furthermore, the experiment was also shared with the students of the Computer Engineering Department and the author's social circle.

A total of 186 participants, all of whom were native speakers of Turkish, took part in the experiment. The age range of the participants was 17 to 61, with a median age of 19.

For the purpose of analysis, participants whose reading times significantly deviated from the average were excluded from the data set¹.

5.1.2 Experiment materials

The experiment materials were designed to include four different conditions: Baseline, narrow scope De Re, wide scope De Re, and De Dicto. Each experimental item had four variations, corresponding to these conditions. Within each condition, a disambiguating dialogue was provided in order to favor either the narrow scope De Re interpretation, wide scope De Re interpretation, De Dicto interpretation, or the baseline where no explicit disambiguation occurred.

Following the dialogue, participants were presented with the De Re De Dicto ambiguous sentence as part of a self-paced reading task. It is important to note that the ambiguous sentence remained the same across all conditions for each experimental item. After reading the ambiguous sentence, participants were asked a Yes/No comprehension question that specifically targeted the intended disambiguation based on the given condition. The purpose of these comprehension questions was to assess participants' understanding and interpretation of the disambiguated meaning in each condition.

If participants provided the expected response to the comprehension questions, it indicated that they were able to successfully access and comprehend the intended disambiguation of the ambiguous sentence. In other words, their response aligned with the disambiguated meaning targeted in each specific condition.

The decision to present the disambiguating dialogue first and then the ambiguous sentence was motivated by the aim to minimize any potential confounding effects. All disambiguating dialogues were structured in a consistent manner, with

¹RTs shorter than 50 ms and longer than 2000ms were excluded.

Person A initiating the dialogue by mentioning a common friend, and Person B providing additional information that allowed for disambiguation. Although the lengths and perceived difficulties of the dialogues varied across conditions, the ambiguous sentence itself had the same length, perceived difficulty, and lexical items across different conditions. This approach enabled the experiment to be conducted in a more controlled manner.

By presenting the disambiguating context first and not collecting reading time for that part of the dialogue, the focus was primarily placed on the reading time of the ambiguous sentence itself. This allowed for a more direct comparison of reading times across conditions, as the participants encountered the same ambiguous sentence in each condition. By maintaining consistency in the length, difficulty, and lexical items of the ambiguous sentence, the experiment aimed to isolate the effects of the different disambiguating contexts on participants' reading times.

A: Hakan ödül töreninde ne giyeceğine karar verebilmiş mi, haberin var mı?
"Has Hakan decided what he will wear at the awards ceremony? Do you know anything about it?"

B: Geçen gün konuştuğumuzda heyecanlıydı, bana bir gömlek giymek istediğini söyledi.
"He was excited when we spoke the other day, he told me that he wanted to wear a shirt."

A: Bizim Hakan bir gömlek giymek istiyor demek. (SPR)
"So our Hakan wants to wear a shirt, huh."

Figure 23. Baseline dialogue of the item #5

The baseline dialogue of an item from the experiment is presented in Figure 23. In this condition, no disambiguation occurs within the dialogue. The reading time of the ambiguous sentence in this condition serves as a baseline against which the De Re and De Dicto conditions can be compared.

The first part of the dialogue, consisting of Person A's first sentence and Person B's reply, is presented all at once. When the participant presses the space button, the dialogue disappears, and the ambiguous sentence in bold appears in

Self-Paced Reading (SPR) format. This format of presenting dialogues is consistent across all conditions and items.

After reading the ambiguous sentence, participants are presented with a comprehension question. They are asked to indicate "Evet" (*Yes*) if the information presented can be inferred from the dialogue, and "Hayır" (*No*) if it cannot be inferred.

Hakan ödül töreninde şık bir pijama giymek istiyor.
"Hakan wants to wear a chic pajama at the awards ceremony."
Intended response: No

Figure 24. Comprehension question for the baseline dialogue of the item #5

In the baseline condition, where no disambiguation occurs in the dialogue, the comprehension questions do not target a specific reading. Instead, they aim to assess the participants' understanding of the dialogue. For example, in baseline condition of the item #5 in Figure 24, the comprehension question asks whether Hakan wants to wear pajamas. The intended response must be "no" since both the preceding dialogue and the ambiguous sentence mentioned that Hakan wants to wear a shirt, not a pajama.

A: Bu sene ödül törenini Hakan sunacakmış, duydun mu? Hatta kıyafetini bile seçmiş.
"Have you heard that Hakan is going to present the awards ceremony this year? He has even chosen what he will wear."
B: Evet, kendisi söyledi. çok da heyecanlı! Bir gömlek beğenmiş, çok şık. Törene onunla gitmek istiyor.
"Yes, he told me himself. He is very excited! He found a shirt he liked, very chic. He wants to wear it to the ceremony."
A: Bizim Hakan bir gömlek giymek istiyor demek. (SPR)
"So our Hakan wants to wear a shirt, huh."

Figure 25. Wide scope De Re dialogue of the item #5

The wide scope De Re dialogue for the same item is presented in Figure 25. In this dialogue, it is asserted that there is a specific, unique shirt that Hakan wants

to wear. After the dialogue, the same ambiguous sentence (written in bold) as in the baseline dialogue is presented.

Hakan ödül töreninde herhangi bir gömlek giyecek.

"Hakan will wear just any shirt at the ceremony."

Intended response: No

Figure 26. Comprehension question for the wide scope De Re dialogue of the item #5

The comprehension question for this dialogue is presented in Figure 26. As a disambiguation occurs in the preceding dialogue, the question specifically targets the narrow scope De Re reading.

A: Ödül töreni sunacağı için Hakan çok heyecanlı. Ne giyeceğini seçmeye çalışıyordu.

"Hakan is very excited because he will present the awards ceremony. He was trying to choose what to wear."

B: Ortama uygun kıyafetlerini bir kenara ayırmış, hepsi de gömlek! Onlardan birini giyecekmiş.

"He set aside his clothes that suit the event, all are shirts! He is going to wear one of them."

A: Bizim Hakan bir gömlek giymek istiyor demek. (SPR)

"So our Hakan wants to wear a shirt, huh."

Figure 27. Narrow scope De Re dialogue of the item #5

For the narrow scope De Re condition of the same item, Person B asserts that there are several shirts that Hakan deemed suitable for the occasion and he is going to wear one of them. Please refer to Figure 27 for the dialogue.

Hakan ödül töreni için seçtiği gömleklerden birini giymek istiyor.

"Hakan wants to wear one of the shirts that he has chosen for the awards ceremony."

Intended response: Yes

Figure 28. Comprehension question for the narrow scope De Re dialogue of the item #5

The comprehension question in the narrow scope De Re condition (Figure 28) specifically targets the disambiguation introduced in the dialogue. Participants are asked whether Hakan wants to wear one of the shirts he has chosen for the event. If participants have successfully understood the disambiguating information provided in the dialogue, their response should align with the intended response.

A: Hakan çok heyecanlı, ödül törenini sunmaya hazırlanıyor.
"Hakan is very excited, he is preparing for presenting the awards ceremony."

B: Biliyorum, kıyafet konusunda benimle konuştu. Ortama en uygun giysi gömlek olur diye düşünmüş, gömlek alacak.
"I know, he talked to me about the outfit. He thought that the best thing to wear would be a shirt so he will buy one."

A: Bizim Hakan bir gömlek giymek istiyor demek. (SPR)
"So our Hakan wants to wear a shirt, huh."

Figure 29. De Dicto dialogue of the item #5

And in the De Dicto condition of the same item (Figure 29), it is asserted that the clothing item Hakan wants to wear is a shirt, but there is no specific shirt that he wants to wear yet —He intends to buy one.

Hakan ödül törenini sunarken pijama giymek istiyor.
"Hakan wants to wear pajamas at the awards ceremony."

Intended response: No

Figure 30. Comprehension question for the De Dicto dialogue of the item #5

For the comprehension question of De Re condition (Figure 30), the target is, again, to check if the participant was able to access the De Dicto reading.

A total of 24 experiment items were prepared, divided into 4 conditions, resulting in a total of 96 dialogues. The comprehension questions were carefully balanced across items and conditions to ensure an equal distribution of Yes and No responses. Each ambiguous sentence followed a consistent format, including a proper name, "bir" + NP, and the intensional operator "istemek" (*to want*).

The length of each ambiguous sentence ranged from 42 to 52 characters. The disambiguating dialogues had a length range of 135 to 185 characters, with a maximum difference of 25 characters between conditions of the same item. To minimize variance, the use of rare lexical items or challenging syntactic structures was avoided in both the disambiguating dialogues and the ambiguous sentences. To ensure this, the bigram and unigram probabilities of each condition for each item were calculated against the Turkish National Corpus (Aksan & Aksan, 2009). The bigram and unigram probabilities of the ambiguous sentences were compared to each other, as were the probabilities of the disambiguating dialogues. Edge cases in dialogues and ambiguous sentences were revised.

5.1.3 Fillers

A total of 21 filler dialogues were also prepared for the experiment, following the same constraints and procedures as the main items. The comprehension questions for the fillers were balanced across items to ensure an equal distribution of Yes and No responses. The length of the filler dialogues ranged from 154 to 181 characters, while the SPR sentences shown after the dialogues had a length range of 43 to 76 characters.

Similar to the experiment items, the bigram and unigram probabilities of both the filler dialogues and the following sentences were calculated using the same procedure. This allowed for a comparison of the probabilities to ensure consistency in the linguistic properties of the fillers.

5.1.4 Procedure

The experiment began with a welcome page where participants were provided with a detailed explanation of the experiment procedure. This ensured that they had a clear understanding of what was expected from them. Following the welcome page, participants proceeded to a trial run where they were shown 5 items in the same format as the main experiment.

During the trial run, the comprehension questions targeted only comprehension as the SPR sentences were not ambiguous. If a participant gave an incorrect response, they were presented with a warning message stating "Daha dikkatli okumalısın!" (*You must read more carefully!*). This served as a reminder for participants to pay closer attention to the task. Conversely, if a participant gave the correct response, they were presented with a congratulatory message saying "Bravo, bu diyalogu anladın!" (*Congratulations, you understood this dialogue correctly!*). These messages were implemented to ensure that participants understood the nature of the task and to encourage careful reading.

After the trial run, participants were notified that they were about to start the actual experiment. In the actual experiment, participants received no feedback on their responses, as there were no "correct" responses per se for the experiment items. The comprehension questions were designed to assess whether the intended disambiguation was accessible to the participant, so there were no objectively correct responses but rather expected responses based on the intended disambiguation.

To ensure that each participant was exposed to every condition of the experiment items, a Latin Square design (Grant, 1948) was employed. This design allowed for a balanced distribution of conditions among the participants. Additionally, all participants were shown the same set of filler dialogues to maintain consistency across the experiment.

The order of both the fillers and experiment items was randomized to minimize any potential order effects. This randomized presentation ensured that each participant encountered a unique sequence of stimuli.

Reading times of the ambiguous sentences were recorded to capture participants' online processing of the sentences. By examining these reading times, researchers could gain insights into how participants processed and understood the ambiguous constructions in real-time.

Furthermore, response times for the comprehension questions were recorded to assess participants' offline processing. These response times provided information

about the cognitive effort required to comprehend and response the questions related to the dialogues.

By collecting both reading times and response times, we were able to gather data on both the online and offline processing aspects of the participants' comprehension of the experiment materials.

5.2 Analysis and results

After the data collection period, the participant responses underwent a careful revision and filtering process. Several criteria were applied to determine which data points should be excluded from the analysis. The following groups of participants were excluded:

- Non-native speakers of Turkish: Participants who were not native speakers of Turkish were excluded from the analysis. This decision was made to ensure that the data accurately represented the processing patterns of native Turkish speakers.
- Participants with low comprehension performance: Participants who failed to provide correct responses to more than 50% of the comprehension questions in both the baseline conditions of the experiment items and the filler items were excluded. This criterion aimed to ensure that only participants who demonstrated a satisfactory level of comprehension were included in the analysis.
- Participants with extreme reading times: Participants whose reading times for the SPR sentences fell below 50ms or exceeded 2000ms were excluded. These extreme reading times could indicate potential data quality issues or outliers that might skew the results.
- Participants with uniform response patterns: Participants who consistently responded with either all "Evet" or all "Hayır" to all comprehension questions were excluded. This criterion aimed to eliminate participants who might have

provided indiscriminate or biased responses, possibly due to inattentiveness or lack of understanding.

By applying these exclusion criteria, the data set was refined to include data from participants who met the necessary criteria for reliable analysis. This process ensured the quality and integrity of the data used in the subsequent stages of the study.

During the pre-analysis stage, it was observed that four specific experiment items (Items #1, #9, #16, and #22) displayed significant divergence in the reading times of the SPR sentences across all conditions. Upon closer examination, a recurring pattern was identified: all of these items contained noun compounds in the NP slot of the ambiguous sentence. Specifically, Item #1 had "bir Avrupa ülkesi" (*a European country*), Item #9 had "bir Yunan adası" (*a Greek island*), Item #16 had "bir müzik öğretmeni" (*a music teacher*), and Item #22 had "bir bitki çayı" (*an herbal tea*).

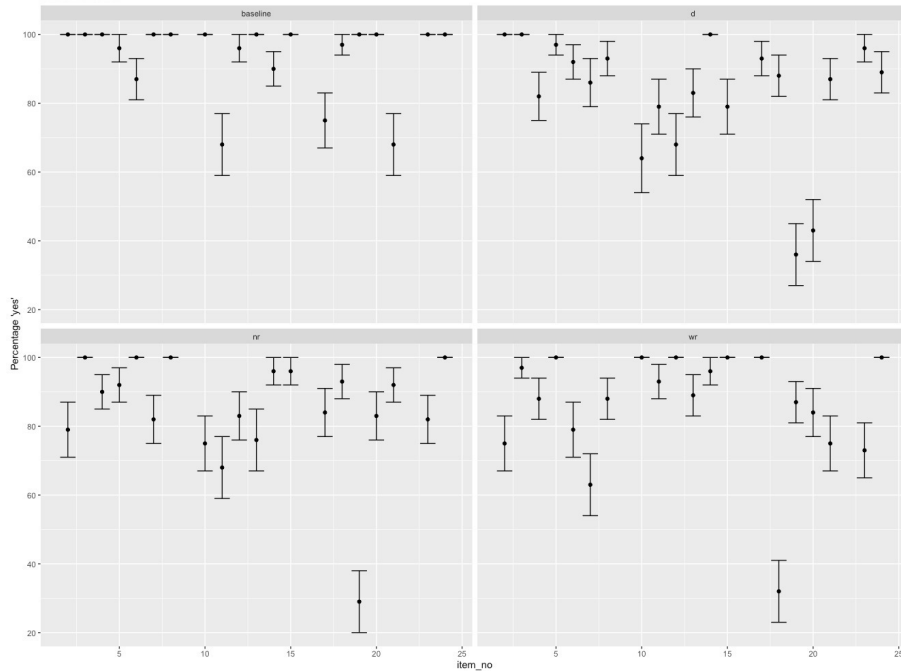


Figure 31. Expected response accuracy per item before exclusions across conditions

Due to the noticeable deviation and the common characteristic of these items, a decision was made to exclude them from the analysis. This step aimed to ensure data quality and avoid potential confounding effects that could arise from the specific linguistic structures present in these items. By excluding these items, the subsequent analysis would focus on the remaining data set, providing a more reliable assessment of the processing patterns for the De Re De Dicto ambiguities under investigation.

Table 13. Comprehension Question Accuracy for Filler Items and Experiment Items

		Accuracy
Fillers		0.95
Experiment Items	Total	0.91
	Baseline	0.99
	De Dicto	0.91
	Wide Scope De Re	0.84
	Narrow Scope De Re	0.93

The data analysis was performed in R, and confidence intervals were calculated following the methods proposed by Morey et al. (2008). The mean filler accuracy was found to be 0.951 (see Table 13). Mean filler accuracy was computed by comparing participant responses with the correct responses to the comprehension questions.

This high accuracy suggests that the feedback provided during the trial run and the detailed explanation of the task in the welcome screen were effective in ensuring that participants understood the requirements of the experiment and performed the comprehension questions accurately.

To assess the accuracy of the experiment items, the participants' responses were compared to the expected responses for each item (see Figure 13 and Fi). The accuracy of the baseline condition was found to be the same as the filler accuracy, indicating that the baseline dialogues were designed accurately and did not introduce any confusing elements for the participants.

However, there was a significant drop in mean accuracy in both the De Dicto and De Re conditions compared to the baseline and filler accuracy. This suggests that

the disambiguating contexts had an effect on the participants' accuracy in responding to the comprehension questions.

In order to further examine the experiment results and determine whether there is a significant online or offline processing difference between De Dicto and De Re readings, we carried out several analyses on comprehension question accuracy, response time, and reading time. Each of these analyses will be discussed in the following subsections.

5.2.1 T-tests

We began by carrying out one-sample t-tests, comparing the baseline with the De Dicto condition, the baseline with the narrow scope De Re condition, and the baseline with the wide scope De Re condition. These t-tests were instrumental in helping us determine whether there existed a statistically significant distinction in reading times and comprehension question accuracy among the De Dicto and De Re scenarios, as well as the baseline.

Table 14. T-Test Results of Comprehension Question Accuracy Across Conditions

	t	df	p value
De Dicto	12.33	567	<2.2e-16
Wide Scope De Re	8.3278	567	<6.205e-16
Narrow Scope De Re	8.9735	567	<2.2e-16

The results of the one sample t-tests indicate a noteworthy variation in comprehension question accuracy compared to the baseline across different conditions with 95% confidence interval (see Table 14). Very small p-value asserts that there is strong evidence against the null hypothesis. This underscores the observation that, despite all readings being accessible to the participants, they exhibited reduced accessibility in comparison to the baseline condition.

Similar to comprehension question accuracy, one sample t-test results of comprehension question RTs show with 95% confidence interval that there is

Table 15. T-Test Results of RT Across Conditions

	t	df	p value
De Dicto	-11.127	567	<2.2e-16
Wide Scope De Re	-8.7954	567	<2.2e-16
Narrow Scope De Re	-15.694	567	<2.2e-16

statistically significant difference between baseline and experiment conditions (see Table 15). Consistently small p-values across t-tests strongly affirm the evidence against null hypothesis.

5.2.2 ANOVA and logistic regression

Once we established the presence of a statistically significant impact on response time (RT) and comprehension question accuracy across various conditions, we pursued a more detailed analysis of this effect. To achieve this, we employed logistic regression analysis for comprehension question accuracy and self-paced reading sentences, and utilized analysis of variance (ANOVA) for response time, considering it as continuous data with categorical predictors.

Table 16. ANOVA results

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Condition	3	6318325	2106108	8.604	1.17e-05
Residuals	1378	337311699	244784		

The ANOVA outcomes reveal a robust effect supported by a significantly low p-value (see Figure 16). This led us to embark on a post-hoc test, aiming to delve deeper into the specific variations between distinct conditions. To achieve this, we conducted a Tukey's Honestly Significant Difference (HSD) test (see Figure 17).

Based on the Tukey multiple comparisons of means, the results indicate that the mean response time (RT) in the De Dicto condition is 77.43 units greater than the baseline. In contrast, this difference extends to 155.23 units in the case of the wide scope De Re condition. When considering their corresponding confidence intervals and p-values, these comparisons demonstrate a statistically significant divergence.

Table 17. Tukey's HSD Results

	Difference	Lwr	Upr	P Adj
De Dicto & Baseline	77.4317703	-17.40363	172.2672	0.1535268
Narrow Scope De Re & Baseline	154.3069523	57.48713	251.1268	0.0002560
Wide Scope De Re & Baseline	155.2286359	64.49814	245.9591	0.0000686
Narrow Scope De Re & De Dicto	76.8751820	-28.50743	182.2578	0.2386981
Wide Scope De Re & De Dicto	77.7968656	-22.02017	177.6139	0.1866597
Wide Scope De Re & Narrow Scope De Re	0.9216836	-100.78262	102.6260	0.9999954

In essence, the findings suggest a notable variation in RT between De Re readings and the baseline condition, whereas no such observable effect is apparent between De Dicto and the baseline, as well as among the other experimental conditions.

Table 18. Logistic Regression for Comprehension Question Accuracy Residuals

Min	1Q	Median	3Q	Max
-1514.87	-356.56	25.01	380.87	975.44

Coefficients				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2022.56	23.43	86.333	<2e-16
De Dicto	77.43	36.87	2.100	0.0359
Narrow Scope De Re	154.31	37.64	4.099	4.38e-05
Wide Scope De Re-	155.23	35.27	4.401	1.16e-05

The results of the linear regression analysis highlight the significance of the predictor variables (De Dicto, wide and narrow scope De Re) in predicting the response variable, RT. The coefficients corresponding to these predictor variables indicate the projected alteration in the response variable linked to each respective condition, in comparison to the baseline condition.

5.2.3 Bayesian analysis

Following linear regression and ANOVA, we conducted a Bayesian analysis to explore expected response accuracy and reading time of SPR sentences.

There was no credible difference in the expected response accuracy between the De Re and De Dicto conditions which is illustrated in Figure 32. As credible intervals of experience items overlap, it is not possible to assert a processing difference between different conditions based on the current data. Aligned with the findings of our previous analyses, though, there is a significant difference between baseline and the experiment conditions.

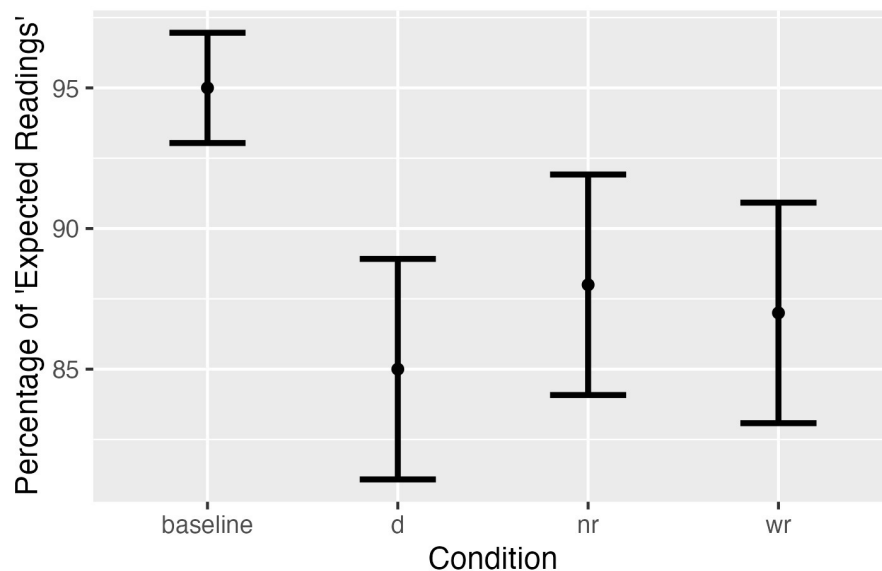


Figure 32. Mean percentage of expected responses across conditions

Examining the Bayesian analysis of anticipated responses among various conditions, as depicted in Figure 33, yields additional perspective on the distinctions between these conditions. In this analysis, we contrast the expected responses to the comprehension questions of the experimental items with the baseline.

The average anticipated response accuracies across all conditions exhibit a certain degree of similarity, as indicated by the credible intervals that reflect the uncertainty surrounding these estimates.

Notably, the credible interval line for the wide scope De Re condition crosses the 0 point, indicating that the expected response accuracy in this condition may not significantly differ from the baseline condition. However, it is important to consider the overlapping credible intervals and the overall similarity of the mean accuracies across conditions. This suggests that, while there may be some variability, there is no clear evidence to support substantial differences in expected response accuracy between the conditions.

It is worth noting that the inclusion of random slopes and intercepts for item and subjects in the analysis helps account for individual variation and item-specific effects, further enhancing the validity of the results.

This analysis highlights a discernible "accessibility" contrast between the baseline and experimental conditions. However, drawing conclusions regarding the relative accessibility among the experiment conditions themselves remains challenging.

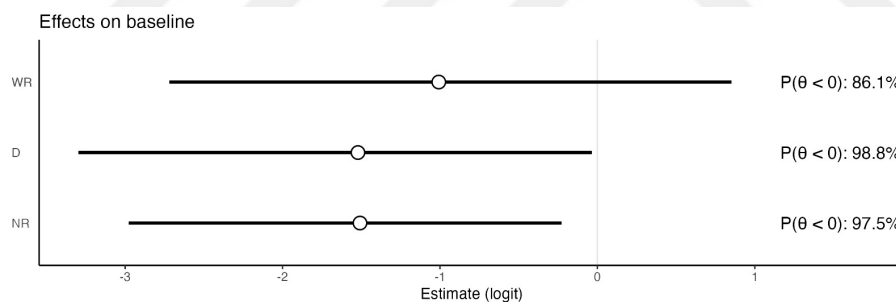


Figure 33. Posterior plot of expected responses across conditions, compared to baseline

The Bayesian analysis of response times across conditions, as depicted in Figure 34, provides insights into the differences in processing times for comprehension questions. The mean response times for all conditions are relatively similar, as indicated by the overlapping credible intervals.

It is noteworthy that the credible interval lines for both the De Dicto and wide scope De Re conditions cross the 0 point, suggesting that there is no substantial evidence of significant differences in response times compared to the baseline

condition. This indicates that participants took a similar amount of time to respond to comprehension questions across all conditions.

Considering the overlapping credible intervals and the similarity of mean response times, it can be concluded that there is not a significant difference in response times between the baseline, De Dicto, and wide scope De Re conditions. These results suggest that the processing burden associated with these conditions is comparable, at least based on the current data.

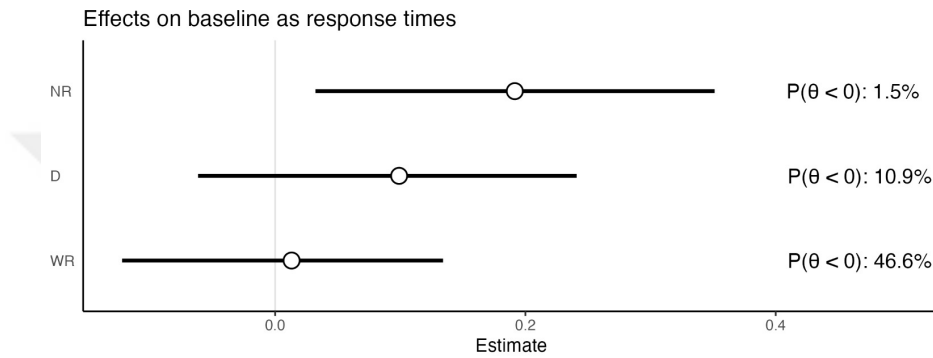


Figure 34. Posterior plot of response times across conditions, compared to baseline

To examine the online processing of the ambiguous sentences, the reading times of specific elements within the sentences were analyzed. Specifically, the reading times of the determiner and two preceding words (det-2, det-1), the determiner (det), the noun, and the following word (noun+1) were considered.

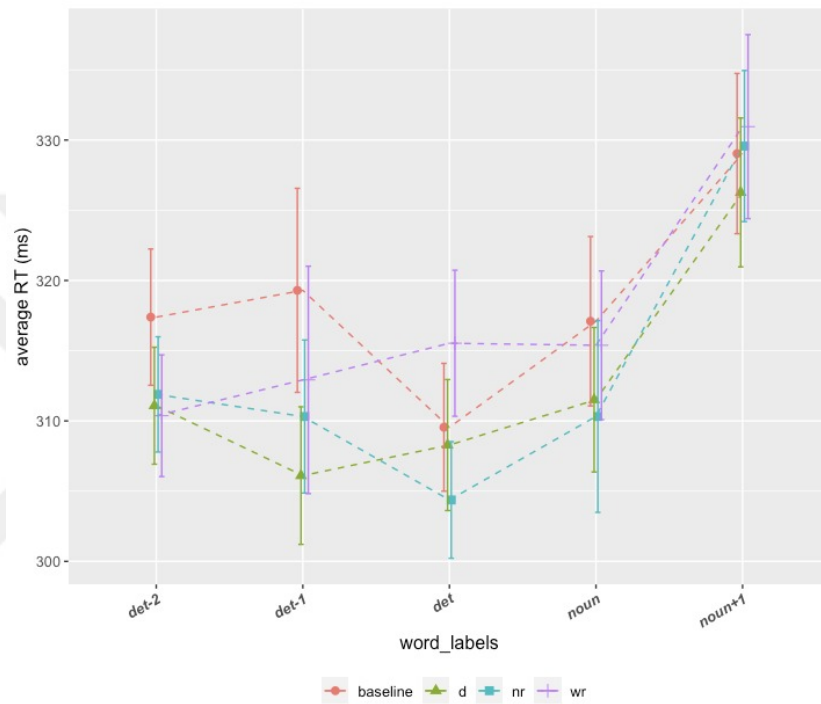
As an example, let's take the ambiguous sentence from item #5: "Bizim Hakan bir gömlek giymek istiyor demek." The reading times of "bizim" (det-2), "Hakan" (det-1), "bir" (det), "gömlek" (noun), and "giymek" (noun+1) were analyzed (see (1) below).

- (1) Bizim Hakan bir gömlek giymek istiyor demek.
 our Hakan.NOM a shirt wear.INF want.3.SG.PRS say.INF
 'So our Hakan wants to wear a shirt, huh.'

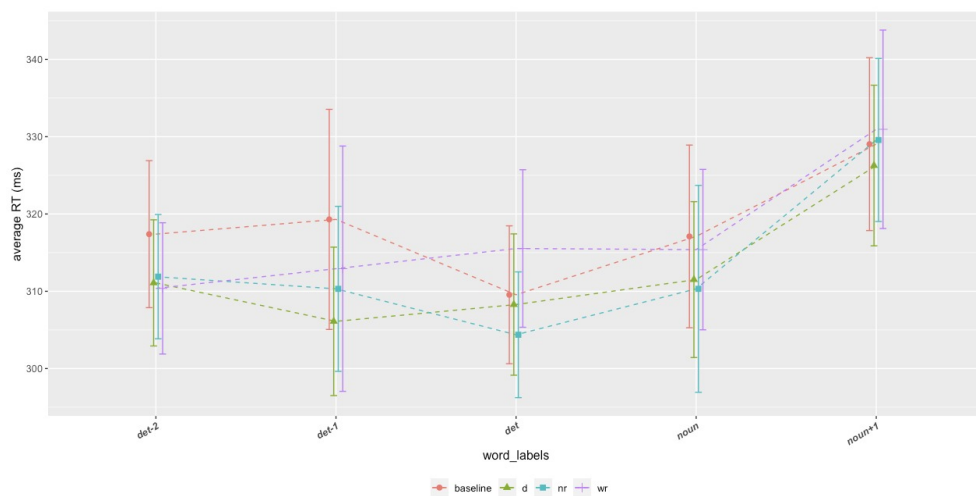
The analysis of reading times for the ambiguous sentences across conditions, incorporating credible intervals, indicates that there is no significant effect. This is illustrated in Figure 35. The overlapping credible intervals suggest that the reading

times for these specific elements within the sentences did not differ significantly between the conditions.

Based on the current data, there is no strong evidence to suggest that the disambiguating contexts in the experiment had a significant impact on the online processing of the ambiguous sentences, as reflected in the reading times of the analyzed elements.



(a) Without credible interval



(b) With credible interval

Figure 35. Reading times of ambiguous sentences

5.3 Discussion and further research

The aim of the study was to investigate the online and offline processing of De Re and De Dicto ambiguities in Turkish. For the online processing, reading times of the disambiguated ambiguous sentences were recorded. However, due to the high overlap in reading times across conditions (Figure 35), it is not possible to infer a significant difference even without considering credible intervals. The only potentially promising point for further exploration is the difference between the noun position in the De Dicto condition and the baseline condition. To delve deeper into this point, a mixture model analysis could be employed.

Regarding the offline processing assessment, response times of the comprehension questions were recorded across conditions, revealing an unexpected pattern. The response times for the De Dicto and wide scope De Re conditions were shorter compared to the narrow scope De Re condition. This finding is significant as it goes against the notion that narrow scope De Re interpretations impose a heavier derivation burden, as suggested by formal linguistic accounts (Tiskin, 2021; Zhang et al., 2019; Drozdovskaia, 2021). Furthermore, there was no significant difference in response times between the De Dicto and wide scope De Re readings. This challenges the common argument among scholars that "De Re is harder" even in processing (Tiskin, 2021; Zhang et al., 2019; Drozdovskaia, 2021).

These findings suggest that the processing patterns of De Re and De Dicto ambiguities in Turkish may differ from the expectations based on formal linguistic accounts, or there can be variation across languages. Further analysis and exploration are needed to better understand the underlying factors contributing to these processing patterns.

In addition to response times, the study also examined the acceptability of De Re and De Dicto readings through comprehension questions. The mean expected response rating was over 85% across all conditions, indicating a high level of acceptability. Furthermore, the distribution of expected responses by participants was skewed towards higher ratings. This contrasts with the findings of Zhang & Davidson

(2021), who reported a bimodal pattern in acceptability judgments for De Re readings and consistently higher judgments for De Dicto readings. The divergent findings in our study compared to previous research highlight the need for further and more in-depth investigation of De Re and De Dicto ambiguities in Turkish.

To summarize, our experimental research aimed to address a gap in the current literature by exploring De Re and De Dicto ambiguities from a psycholinguistic perspective. This study is the first of its kind in Turkish, setting a baseline for future research in this language. Furthermore, previous experimental work on English has been limited and does not include the narrow scope De Re interpretation. Considering that narrow scope De Re has posed challenges for traditional scope theories, comparing it with wide scope De Re and De Dicto interpretations is important to provide a foundation and potential inspiration for theoretical research.

Future studies in this area should further investigate the online processing of De Re and De Dicto ambiguities. In our study, the ambiguous sentences had slight variation in the distance between the noun and the verb. It would be valuable to conduct further research by controlling this distance and investigating the precise location of integration during online processing.

Our research has identified two primary limitations that warrant discussion. Firstly, our experiment design involved disambiguating the context before presenting potentially ambiguous sentences. This design was chosen to minimize complicating factors, as controlling disambiguating contexts in terms of length and difficulty posed challenges. However, this choice introduces an intriguing question: Given that participants were already aware of the disambiguating context when encountering the "ambiguous" sentence, can we still consider that sentence as truly ambiguous? Could the pre-existing understanding of the correct LF structure influence the syntactic structure participants attributed to the "ambiguous" sentence? Future investigations in this area should consider alternative experiment designs to delve deeper into these

inquiries. One possible avenue might involve adopting an experiment design aligned with surprisal accounts, moving away from the self-paced reading task.

Secondly, our experiment materials intentionally exclude references to the belief states of the speaker, listener, or the third party under discussion. This decision was made to maintain consistency across disambiguating contexts and ambiguous SPR sentences. Consequently, our exploration of the narrow scope De Re is confined to scope properties, excluding the interplay between the noun phrase (NP) and the belief states of the speaker and the individual in question. Future research should address this intricate interaction to provide a more comprehensive understanding.

Our study underscores the need for alternative experiment designs to address the potential impact of pre-existing context disambiguation and emphasizes the importance of exploring the interrelation between NP scope and belief states in subsequent research endeavors.

CHAPTER 6

CONCLUSION

This study offers a high quality challenge NLI dataset with De Re De Dicto ambiguities. De Re De Dicto ambiguities present a challenge because while in certain cases linguistic structure can help resolving the ambiguity, in others world knowledge or the belief state of the speaker is required.

The CNLI-TR data set aims to test if the language models can identify De Re De Dicto ambiguities, and resolve these ambiguities when linguistic cues like case morphology and changes to the word order are available. In doing so we explored a hybrid approach to text data augmentation and offered high quality, manually revised data set for Turkish, a low resource language.

Considering that our augmentation strategy managed to preserve 74.88% of the initial inference labels and successfully augmented our seed data around 9 times of its initial size, we believe that we reached our goal in this area. Our aim for the near future is expanding this data set by incorporating more instances and balancing the inference label distribution. Our future goal is to enhance the data by adding linguistic and contextual cues that will provide enough world knowledge or belief states to help resolve the truly ambiguous sentences.

To assess the LLM performance in resolving De Re De Dicto ambiguities within an NLI context, we conducted computational experiments involving BERT-En, BERTurk, and BERTMulti. These experiments were conducted both with and without the inclusion of a morphological analysis layer. Notably, the introduction of morphological analysis yielded a marked improvement in the language model’s performance, resulting in gains of up to 3 points. Among the models tested, BERTurk emerged as the most proficient performer.

However, across both experimental conditions, namely with and without the morphological analysis layer, the LLMs demonstrated a consistent proficiency in recognizing the presence of De Re De Dicto ambiguities. Strikingly, they struggled when faced with instances where one or more interpretations were inaccessible. This

observation prompted a reassessment of the role played by morphological information in the decision-making processes of LLMs within tasks like NLI.

Based on our experimental outcomes, it becomes apparent that LLMs possess a lesser grasp or utilization of morphological cues than what existing literature estimates suggest. This revelation might be attributed, in part, to the constrained pretraining data used in Turkish-fluent models like BERTurk, which drew from a relatively limited pool of around 30 GBs of pretraining data. Another plausible explanation lies in the dataset used, CNLI-TR, which demonstrates an imbalance favoring sentences with genuine ambiguity over those that are unambiguous. This skewed dataset composition potentially influences the LLMs' performance in this specific context.

To evaluate human capacity in resolving De Re De Dicto ambiguities, we undertook a psycholinguistics experiment. Our methodology encompassed a disambiguating dialogue, succeeded by a SPR-format De Re De Dicto ambiguous sentence, culminating in a comprehension question targeting the intended interpretation. By analyzing both online and offline sentence processing through RTs and question accuracy, we aimed to shed light on the intricacies of this cognitive process.

Our findings were subjected to rigorous analysis involving t-tests, ANOVA examination, and Bayesian analysis. These analytical methods were instrumental in corroborating that De Re and De Dicto ambiguities manifest distinct patterns in offline processing compared to the baseline condition.

For a comprehensive understanding of the real-time processing disparities and the subtleties delineating De Dicto from De Re readings, we acknowledge the necessity for further experimental endeavors. The exploration of online processing discrepancies warrants additional investigations to unravel the finer nuances inherent in these linguistic phenomena.

APPENDIX A

Annotator Guidelines

CNLI-TR Project - Handout 1

Büşra Marşan

Abstract

This project aims to explore De Re/De Dicto phenomenon from various aspects including its significance in computational linguistics and Natural Language Processing. Our primary goal is to create a De Re/De Dicto ambiguity data set to explore a language model's capabilities of "understanding" this ambiguity. Within the framework of this project, we define "understanding" a sentence as knowing what can and cannot be entailed from that. In this handout, we will get familiar with the De Re/De Dicto phenomenon, explore the NLP task we aim to do, and discuss your first task as the contributors.

1 An Introduction to De Re/De Dicto Phenomenon

An ambiguous sentence lies at the core of De Re/De Dicto phenomenon. The interaction between noun phrases and intensional operators trigger the ambiguity (see Keshet and Schwarz (2019) for a detailed discussion). As a result, it is possible to get at least two readings from the ambiguous sentence: De Re and De Dicto.

Let's say that we have a common friend, Ayşe. I learned something about her and I am sharing it with you:

- (1) Ayşe bir tesisatçı ile evlenmek istiyor.
"Ayşe wants to marry a plumber."

(1) contains both elements that are required for a De Re/De Dicto ambiguous sentence: A noun phrase (bir tesisatçı "a plumber") and an intensional operator (istemek "to want"). Hence, it is possible to get the following interpretations:

Specific De Re: There is an individual who is a plumber, let's call him Ali. Ayşe wants to marry this person. She may or may not know that he is a plumber. In this interpretation, "a plumber" refers to an actual person¹. Hence, we have a De Re reading.

Non-Specific De Re: There is another individual who is a plumber, let's call him Veli. Ayşe wants to marry either Ali or Veli. She hasn't decided which one yet but it is certain that she wants to marry one of them. In this interpretation, "a plumber" refers to actual people. Hence, we have another De Re reading.

De Dicto: Ayşe has very specific desires: She wants to get married and she wants to marry a person who is a plumber. She hasn't met anyone special yet, there is not a specific person in her mind. She just has this affection towards plumbers and she can marry any plumber.

Another interesting thing to note about De Re/De Dicto ambiguity is that strong quantifiers and operators that shift the time of evaluation can cause De Re/De Dicto Ambiguity:

- (2) 2050 senesinde Dilbilim bölümündeki profesörlerin çoğu emekli olacak.
"By 2050, most professors in Linguistics department will retire."

Here our noun phrase is "Dilbilim bölümündeki profesörlerin çoğu" and our operator is "emekli olacak." Hence, it is possible to get the following interpretations:

De Re: The majority of the professors who are working at the linguistics department **NOW** will be retired by 2050. In this interpretation, the noun phrase ("Dilbilim bölümündeki profesörlerin çoğu") is

¹The terms De Re and De Dicto can be confusing but learning their literal meaning can help you distinguish them: "De Re" means "of thing", while "De Dicto" means "of meaning/concept." Note that "dicto" is "saying, expression" in Latin and "dictionary" is derived from this root.

anchored in the utterance context with respect to the time and world of evaluation.

De Dicto: The majority of the professors who **WILL BE** working at the linguistics department **in 2050** will retire in 2050. Here, the noun phrase (“Dilbilim bölümündeki profesörlerin çoğu”) is evaluated relative to the time and world at which the main proposition expressed by the sentence is evaluated (year 2050).

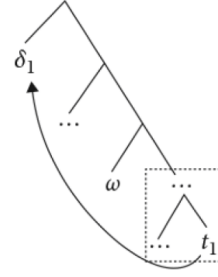
1.1 Semantics of De Re/De Dicto

I don't think we need to go into the intricate semantic details of the issue since there is a lot to cover and quite frankly, they are very peripheral to our task. I just want to point out that the widely accepted consensus is that De Re is interpreted above the intensional operator while De Dicto is interpreted in situ. Below you can find a simple schema where w is the intensional operator and d is the noun phrase.

a. *De dicto*:



b. *De re*:



2 The Goal(s) of The Project

What we want to achieve through CNLI-TR Project is to see if language models like BERT are able to detect De Re/De Dicto ambiguities. In order to do that, we aim to design an inference task where the language model is presented with the ambiguous sentence and one of the possible readings (De Re or De Dicto). If the model is able to “see” the ambiguity, it must be able to say that both De Re and De Dicto readings can be inferred from the ambiguous sentence.

In order to train the model, we need large quantities of data. This is where you as contributors come into play!

For this purpose, 500+ novel De Re/De Dicto ambiguous sentences were written by me (Büşra) and we need your support to expand the data set. After your contributions, we will feed the seed data (manually generated sentences) into a language augmentation pipeline to create similar sentences.

3 Our First Task: Sentence Generation

You can find the sentences written by me on this link: <https://docs.google.com/spreadsheets/d/1tEP0-4B9Z9Q0u1KEBB-Y6CwwZxG0vWF1mDJTxDTRFY/edit?usp=sharing>

First of all we would like you to review them, get familiar with them. Then we would like you to write 50 novel sentences each. In order to make sure that we are on the same page, we should start with 10 sentences each and review them. After the review session, you can write 40 more.

You can get as creative as you want with the sentences. All we ask is that they should contain an intensional operator, a proper name, and a noun phrase. You can try writing conditional sentences, questions etc. or play with the word order.

Following this task, we will write De Re and De Dicto readings for these ambiguous sentences. Then the augmentation task will be done and we will ask for your help for your final task: Reviewing the augmented data. Theoretically, the augmentation pipeline must provide us with meaningful sentences but in practice, we probably will come across some nonsensical and/or freak of nature sentences. We need to eliminate such sentences to ensure that our data is not noisy and the language model can properly learn from it.

References

Keshet, E., & Schwarz, F. (2019). De re/de dicto. *The Oxford handbook of reference*, 167–202.

CNLI-TR Project - Handout 2

Büşra Marşan

Dear Contributors,

We have completed the first task! Thank you for your time, patience, and diligence. Now we have somewhere around 600 ambiguous sentences. For our next task, we will write De Re and De Dicto paraphrases of these sentences. In this handout, you will be offered a detailed guide and discussion so that the task and our purposes are clear. At the very end you can also find a brief summary.

1 Task 2: Writing De Re & De Dicto Sentences

For our second task, we will write De Re and De Dicto paraphrases for each ambiguous sentence.

It is only possible to get either De Re or De Dicto paraphrase from some sentences. In such cases, you should write a one De Re and one De Dicto sentence following the strategies discussed below. Then you should put a 0 (zero) in the column next to the “impossible” sentence -the De Re or De Dicto paraphrase.

Figure 1: Sample annotation ¹

Sentence	De Re Reading (unique entity)	RJ	De Dicto Reading (various referents)	DJ
Gamze bir doktorla evlenmek istiyor.	Gamze'nin evlenmek istediği kişi bir doktor.		Gamze evleneceği kişinin mesleğinin doktorluk olmasını istiyor.	
Selin gruptan birinin hasta olduğunu düşünüyor.	Selin'in hasta olduğunu düşündüğü belli bir grup üyesi var.		Selin hasta olan kişinin bir grup üyesi olduğunu düşünüyor.	
Jüri, yarışmacının hile yaptığını sanıyor.				
Güvenlik görevlileri birinin kapıyı zorladığına inanıyor.				
IT departmanı birinin siteyi hacklemeye çalıştığını düşünüyor.				

¹ RJ = De Re Judgement

DJ = De Dicto Judgement

Similar to the previous task, we are after variance. We need word order variance, different structures, and different strategies so that our De Re sentences or De Dicto sentences aren't too similar to one another. If our sentences are too similar, the algorithm will just memorize them instead of actually “learning” patterns etc. That is why we want our sentences to vary, to look different from one another.

Another point to consider is that we shouldn't introduce contentful information that is not in the original sentence. Let's say our ambiguous sentence is as follows:

(A) Orkun başka bir kravat takmak istiyor.

“Orkun wants to wear a different tie.”

We can come up with many De Re paraphrases, such as:

(1) Orkunun takmak istediği farklı bir kravat var.

“There is a different tie that Orkun wants to wear.”

(2) Orkun'un takmak istediği kravat başka.

“The tie Orkun wants to wear is different.”

(3) Orkun bundan başka bir kravat takmak istiyor.

“Orkun wants to wear another tie.”

(4) Orkun kırmızı kravatı değil lacivert kravatı takmak istiyor.

“Orkun wants to wear the red tie.”

In sentences (1), (2), and (3), we don't offer additional information -information that cannot be found in sentence (A). That is why they are all acceptable for this task. Sentence (4), on the other hand, offers additional information (the colour of the necktie). That is why it is unacceptable.

It is possible to come up with several De Re sentences (and De Dicto sentences) for a single ambiguous sentence. You can follow the strategies I discuss below to write De Re and De Dicto sentences or you can come up with your own.

2 De Re Sentences

As we already know, De Re sentences make reference to a **specific** entity in the **real world**. In De Re paraphrases, we aim to convey that meaning. For this purpose, we can use the strategies & structures below:

(A) Gamze bir doktorla evlenmek istiyor.

"Gamze wants to marry a doctor."

(1) Gamze'nin evlenmek istediği kişinin mesleği doktorluk.

"The profession of the person Gamze wants to marry is doctor."

(2) Gamze'nin evleneceği kişi bir doktor.

"The person Gamze will marry is a doctor."

(3) Gamze'nin evlenmek istediği kişi bir doktor.

"The person Gamze wants to marry is a doctor."

(If you can come up with other strategies and/or structures, please use them!)

Please note that our De Re sentences should not make reference to a set of entities. They should make reference to a unique and single entity.

3 De Dicto Sentences

As we discussed before, De Dicto sentences don't make reference to an entity. In De Dicto paraphrases of the ambiguous sentences, we aim to convey this meaning. For this purpose, we can follow the strategies/replicate the structures below:

(A) Gamze bir doktorla evlenmek istiyor.

"Gamze wants to marry a doctor."

(1) Gamze eşinin bir doktor olmasını istiyor.

"Gamze wants her (future) spouse to be a doctor."

(2) Gamze evleneceği kişinin mesleğinin doktorluk olmasını istiyor.

"Gamze wants the profession of the person she will marry to be a doctor."

(If you can come up with other strategies and/or structures, please use them!)

4 Summary

After you read this handout, we should have a Zoom meeting together to see if we are on the same page. After that, we can divide the data set and start writing De Re and De Dicto sentences.

Our Task: We will write De Re and De Dicto paraphrases for ambiguous sentences.

What to Bear in Mind

- Some "ambiguous" sentences don't have both paraphrases. In such cases, please write both De Re and De Dicto sentences. Then put a 0 (zero) in the column where a certain paraphrase is not possible in your opinion (see Figure 1).
- Our De Re sentences must make reference to a unique, single individual in the real world. Our De Dicto sentences must have variance in their referents.
- Variation is important & needed. Word order variance, different structures, different strategies are all welcome!
- We shouldn't offer additional information in De Re and De Dicto sentences.
- If you are not sure about a sentence, please write to our WhatsApp group so that we can discuss together.

CNLI-TR Project - Handout 3

Büşra Marşan

Dear Contributors,
We have completed our first two tasks so now we have around 30,000 sentences (ambiguous sentences, de re paraphrases, and de dicto paraphrases).
Our last task is to give judgements on these sentences. The roadmap for this task will be explained in this handout.

1 Task 3: Well Formedness Check & Inference Check

For our first task, we wrote 600+ ambiguous sentences. For our second task, we wrote 600+ De Re and 600+ De Dicto paraphrases for the ambiguous sentences. After that, I augmented our manually generated dataset using an augmentation pipeline.

The augmentation pipeline took manually generated sentences (the sentences we wrote in our first two tasks) as its input and applied a set of operations to produce similar sentences. Thus, we have a dataset that contains around 30,000 sentences now: 10,000 ambiguous sentences, 10,000 De Re paraphrases, and 10,000 De Dicto paraphrases.

For our third and final task, we have two goals:

- Ensuring that all sentences generated by the augmentation pipeline are grammatical and convey the meaning we want them to convey,
- Marking the inference relation or lack thereof between the ambiguous sentences and their De Re & De Dicto paraphrases.

In the following subsections, I explain the details of these two subtasks. You can do them separately (you can check the grammaticality & intended meanings of the sentences in your first run, then you can mark the inference relationship between them in your second run) or you can do them simultaneously (you can go through the sentences once to check grammaticality and intended meanings, and also mark the inference relationship).

1.1 Task 3a: Well Formedness Check

For this subtask, we need to check the generated sentences. They must:

- **Make sense.**
Our sentences must be well formed. Following sentences fall short of this criterion:
Leyla bir odayı okumak istiyor.
Canan başka birinin bir topuklu kapıyı yazdığından şüpheleniyor.
- **Be grammatical.**
Our sentences must be grammatical. The following sentence falls short of this criterion:
Ayşe'nin bir kediyi evlat edinmeleri düşünecek.
- **Have an intensional operator.**
Our sentences must have an intensional operator. Following sentences fall short of this criterion:
Venus bir bardağı kırdı.
Ayşe'nin bir kızı var.
- **Have a DP.**
Our sentences must have a DP **C-commanded** by the intensional operator. Following sentences fall short of this criterion:

Asude yazı yazmak istiyor.
Kamuran hırsızlıktan şüpheleniyor.

- **Have proper case morphology and orthography.**
 Our sentences must have proper case morphology and orthography.
Ceyda bir kitaba okumak istiyor.
Ceyda bir kitabı okumak istiyor.
Hakan başka birgomlek almayı arzuluyor..
Hakan başka bir gömlek almayı arzuluyor..

Ideally, our sentences should meet all of the criteria above yet there are three possible cases which I will explain below.

Case 1: All is Well

We must first examine the ambiguous sentences (see the column A (Sentence) in Figure 1.1). They must meet **all** of the above criteria. If they do, we must put 1 in column B (CHECK) (see highlighted column in Figure 1.1).

Then we must proceed to check De Re and De Dicto paraphrases. They must also make sense, be grammatical, and have proper case and orthography. If that's the case, we are done with the grammaticality judgement subtask.

Figure 1: Sample annotation¹

E12	A	B	C	D	E	F
	Sentence	CHECK	De Re Reading (unique entity)	RJ	De Dicto Reading (various referents)	DJ
1	Gamze bir doktorla evlenmek istiyor.	1	Gamze'nin evlenmek istediği kişi bir doktor.	1	Gamze evleneceği kişinin mesleğinin doktorluk olmasını istiyor.	1
2	Başka bir tabloda boyamak sanıyor Abidin.	0	Abidin'in boyamak sandığı başka bir tablo var.		Abidin bunun dışında bir tabloda boyamak sanıyor.	
3	Selin gruptan birinin hasta olduğunu düşünüyor.	1	Selin'in hasta olduğunu düşündüğü belli bir grup üyesi var.	1	Selin hasta olan kişinin bir grup üyesi olduğunu düşünüyor.	0
4	Jüri, bir yarışmacının hile yaptığını sanıyor.	1	Jürinin hile yaptığını sandığı bir yarışmacı var.		Jüri bir yarışmacının hile yaptığını sanıyor ama hangi yarışmacı olduğunu bilmiyor.	
5	Serra başka bir odayı süpürmek istiyor.	1	Serra'nın süpürmek istediği oda bundan başka bir oda.		Serra bu odayı süpürmek istemiyor.	

¹ RJ = De Re Judgement
 DJ = De Dicto Judgement

Case 2: Case and/or Orthography Issues

If the ambiguous sentence (column A (Sentence) in Figure 1.1) and De Re & De Dicto paraphrases have only case and/or orthography issues, we should fix them. Then we should put 1 in column B (CHECK) (see highlighted column in Figure 1.1). After that, we are done with the grammaticality judgement subtask.

Case 3: Major Issues with Well Formedness

There can be major issues in ambiguous sentences, De Re paraphrases or De Dicto paraphrases.

We should start with checking the ambiguous sentence (column A). If it fails to comply with any of the criteria stated above (other than proper case and orthography), we must put a 0 in column B (see line 3 Figure 1.1 for an example). When the ambiguous sentence has major issues, we don't need to check De Re and De Dicto paraphrases.

If there are no major issues in the ambiguous sentence, we should proceed to check De Re and De Dicto paraphrases. As discussed in Case 2, we should fix minor issues like case and orthography.

If the De Re or De Dicto paraphrases of the ambiguous sentences fail to comply with any other criteria, we need to mark them by filling their cell with red (see Figure 1.1).

Figure 2: Sample issue ²

C5 Jürinin hile sandığı bir yarışmacı var.						
	A	B	C	D	E	F
1	Sentence	CHECK	De Re Reading (unique entity)	RJ	De Dicto Reading (various referents)	DJ
2	Gamze bir doktorla evlenmek istiyor.	1	Gamze'nin evlenmek istediği kişi bir doktor.	1	Gamze evleneceği kişinin mesleğinin doktorluk olmasını istiyor.	1
3	Başka bir tablodan boyamak sanıyor Abidin.	0	Abidin'in boyamak sandığı başka bir tablo var.		Abidin bunun dışında bir tablodan boyamak sanıyor.	
4	Selin gruptan birinin hasta olduğunu düşünüyor.	1	Selin'in hasta olduğunu düşündüğü belli bir grup üyesi var.	1	Selin hasta olan kişinin bir grup üyesi olduğunu düşünüyor.	0
5	Jüri, bir yarışmacının hile yaptığını sanıyor.	1	Jürinin hile sandığı bir yarışmacı var.		Jüri bir yarışmacının hile yaptığını sanıyor ama hangi yarışmacı olduğunu bilmiyor.	
6	Serra başka bir odayı süpürmek istiyor.	1	Serra'nın süpürmek istediği oda bundan başka bir oda.		Serra bu odayı süpürmek istemiyor.	
7						

² RJ = De Re Judgement

DJ = De Dicto Judgement

1.2 Task 3b: Inference Check

For this subtask, we need to check the inference relationship between the ambiguous sentence and its different paraphrases.

We will only consider ambiguous sentences that pass our grammaticality test. If the value in CHECK column is 0, we must pass that line. We don't need to provide a judgement on De Re or De Dicto paraphrase.

If the value in CHECK column is 1, we need to assess if we can infer De Re and De Dicto paraphrases from the ambiguous sentence. There will be instances where we can infer both paraphrases, only one paraphrase or neither De Re nor De Dicto paraphrases.

If we can infer De Re paraphrase from the ambiguous sentence, we must put 1 in column D (RJ) (see Figure 1.2). If we cannot infer the paraphrase, we must put 0.

If we can infer De Re paraphrase from the ambiguous sentence, we must put 1 in column F (RJ) (see Figure 1.2). If we cannot infer the paraphrase, we must put 0.

Figure 3: Sample issue ³

E12	fx					
	A	B	C	D	E	F
1	Sentence	CHECK	De Re Reading (unique entity)	RJ	De Dicto Reading (various referents)	DJ
2	Gamze bir doktorla evlenmek istiyor.	1	Gamze'nin evlenmek istediği kişi bir doktor.	1	Gamze evleneceği kişinin mesleğinin doktorluk olmasını istiyor.	1
3	Başka bir tablodan boyamak sanıyor Abidin.	0	Abidin'in boyamak sandığı başka bir tablo var.		Abidin bunun dışında bir tablodan boyamak sanıyor.	
4	Selin gruptan birinin hasta olduğunu düşünüyor.	1	Selin'in hasta olduğunu düşündüğü belli bir grup üyesi var.	1	Selin hasta olan kişinin bir grup üyesi olduğunu düşünüyor.	0
5	Jüri, bir yarışmacının hile yaptığını sanıyor.	1	Jürinin hile yaptığını sandığı bir yarışmacı var.		Jüri bir yarışmacının hile yaptığını sanıyor ama hangi yarışmacı olduğunu bilmiyor.	
6	Serra başka bir odayı süpürmek istiyor.	1	Serra'nın süpürmek istediği oda bundan başka bir oda.		Serra bu odayı süpürmek istemiyor.	
7						

³ RJ = De Re Judgement

DJ = De Dicto Judgement

2 Notes

Before we start our final task, we must meet on Zoom to make sure that we are all on the same page. I will text you to find a day that works for all of us. After our Zoom meeting, you can start the task.

If you are not sure about the grammaticality of a sentence, confused by the inference check, or have any questions, please write to our WhatsApp group so we can discuss the issue together.

APPENDIX B

Annotatoion Interfaces

	A	I
1	Potentially Ambiguous Sentence	Notes
731	Akşam için pişirilen bir kurabiyenin kaç derecede pişirildiğini Abraham bilmiyordu.	
732	Astrid misafirlerinin eve gelirken bir hediye getirip getirmeyeceğinden şüpheliydi.	
733	Her katılımcı bir masaya oturmak gerektiğini düşündü.	
734	Kimileri bir oduncunun tüm odunları keseceğine inanmıştı.	
735	Sağduyu sahibi bir satıcı herkesin ödeyebileceği fiyatlar belirlemesi gerektiğini bilirdi.	
736	Gaddar adam fidanlardan birini sökmeyi arzuluyordu.	
737	3A sınıfından Emre babasıyla okuyabileceği bir şiiri ezberlemek istiyordu.	
738		
739		
740		
741		
742		
743		
744		
745		
746		
747		

Annotation interface for Task 1 where the contributors were asked to write De Re De

Dicto ambiguous sentences.

	A	B	C	D	E
1	Sentence	De Re Paraphrase (unique entity)	RJ	De Dicto Paraphrase (various referents)	DJ
2	Gamze bir doktorla evlenmek istiyor.	Gamze'nin evlenmek istediği kişi bir doktor.		Gamze evleneceği kişinin mesleğinin doktorluk olmasını istiyor.	
3	Selin gruptan birinin hasta olduğunu düşünüyor.	Selin'in hasta olduğunu düşündüğü belli bir grup üyesi var.		Selin hasta olan kişinin bir grup üyesi olduğunu düşünüyor.	
4	Jüri, yarışmacının hile yaptığını sanıyor.				
5	Güvenlik görevlileri birinin kapıyı zorladığına inanıyor.				
6	IT departmanı birinin siteyi hacklemeye çalıştığını düşünüyor.				
7					
8					

Annotation interface for Task 2 where the contributors were asked to write De Re and De Dicto paraphrases, and write inference labels for the paraphrases.



	A	B	C	D
1	Sentence	CHECK	De Re Paraphrase (unique entity)	De Dicto Paraphrase (various referents)
2	Gamze bir doktorla evlenmek istiyor.	1	Gamze'nin evlenmek istediği kişi bir doktor.	Gamze evleneceği kişinin mesleğinin doktorluk olmasını istiyor.
3	Başka bir tabloda boyamak sanıyor Abidin.	0	Abidin'in boyamak sandığı başka bir tablo var.	Abidin bunun dışında bir tabloda boyamak sanıyor.
4	Selin gruptan birinin hasta olduğunu düşünüyor.	1	Selin'in hasta olduğunu düşündüğü belli bir grup üyesi var.	Selin hasta olan kişinin bir grup üyesi olduğunu düşünüyor.
5	Jüri, bir yarışmacının hile yaptığını sanıyor.	1	Jürinin hile sandığı bir yarışmacı var.	Jüri bir yarışmacının hile yaptığını sanıyor ama hangi yarışmacı olduğunu bilmiyor.
6	Serra başka bir odayı süpürmek istiyor.	1	Serra'nın süpürmek istediği oda bundan başka bir oda.	Serra bu odayı süpürmek istemiyor.
7				
8				

	A	B	C	D	E	F
1	Sentence	CHECK	De Re Paraphrase (unique entity)	RJ	De Dicto Paraphrase (various referents)	DJ
2	Gamze bir doktorla evlenmek istiyor.	1	Gamze'nin evlenmek istediği kişi bir doktor.	1	Gamze evleneceği kişinin mesleğinin doktorluk olmasını istiyor.	1
3	Başka bir tabloda boyamak sanıyor Abidin.	0	Abidin'in boyamak sandığı başka bir tablo var.		Abidin bunun dışında bir tabloda boyamak sanıyor.	
4	Selin gruptan birinin hasta olduğunu düşünüyor.	1	Selin'in hasta olduğunu düşündüğü belli bir grup üyesi var.	1	Selin hasta olan kişinin bir grup üyesi olduğunu düşünüyor.	0
5	Serra başka bir odayı süpürmek istiyor.	1	Serra'nın süpürmek istediği oda bundan başka bir oda.		Serra bu odayı süpürmek istemiyor.	
6						

Annotation interface for Task 3 where the contributors were asked to go over augmented data for validation.

Processing Experiment Materials

[illegible]

107

Fillers					
Dialogue	Char	Sentence	Char	Question	Ans
A: Feriha'nın tırnağı kırılmış, ofiste ortalgı birbirine kattı! Canı çok acıyormuş, tadı çok kaçmış. \$ B: Yok artık, bir de bayılsaymış! Demek o yüzden benden bir törpü istedi.	177	A: Aaa, Feriha senden törpü mü istedi? Hiç bilmiyordum!	55	Feriha tırnağı kırıldığı için bayıldı.	H
A: Bizim Hüseyin çaycılığa başlayacakmış diye duydum, hiç haberin var mı? \$ B: Evet evet, haberim var! İşlek bir caddede kahvehane açmak gibi bir niyeti var hatta.	163	A: Desene bizim Hüseyin artık işyeri sahibi olacak!	51	Çaycı Hüseyin bir restoran açmak istiyor.	H
A: Canan plajda güneşlenirken uyuymuş. Her yeri yanmış! Hem kıpkırmızı, hem de canı acıyor. \$ B: Vah vah, geçmiş olsun. Yanık yerlerine bir krem sürse çok iyi gelir.	169	A: Bir krem bütün o yanıkları iyileştirir mi diyorsun?	54	Canan balkonda güneşlenirken yanmış.	H
A: Müjdemî isterim, Nihal üniversite sınavında derece yapmış! \$ B: Niyahet geçiyor şu sınavı! Geçen sene de sınava girdi ama sınavda fenalaştığı için hiçbir yeri kazanamadı.	174	A: Demek Nihal daha önce de üniversite sınavına girdi.	54	Nihal üniversite sınavında iki kere derece yapmış.	H
A: Recep ile dağ yürüyüşü yapacaktık ama rotayı öğrendikten sonra beni ekti! \$ B: Recep'in köpek fobisi var. Eğer yol üzerinde köpek olabileceğini öğreniyse ondan ekmiştir seni.	178	A: Demek Recep küçücük bir köpektен korkuyor ha!	48	Recep'in yükseklik fobisi olduğu için yürüyüşe katılmayacak.	H
A: Hilal çok fantastik bir rüya görmüş dün. Tam anlatıyordu telefon geldi, dinleyemedim. \$ B: Bana anlattı, çok acayipti. Bir ejderhanın sırtında tüm dünyayı gezmiş rüyasında.	175	A: Bizim Hilal bir ejderha avcısı mı olmuş rüyasında?	53	Hilal rüyasında bir gergedanın sırtında dünyayı gezmiş.	H

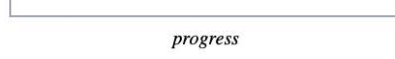
A: Ceyhun evde çok sıkılmış, bir yerlere gitmek istiyor. Ne dersin? \$ B: Ben evden çıkmak isterim ama Ceyhun pikniğe gitmek istiyor. Bahar alerjisi yüzünden ona gelemem. A: Ercan yeni iş bulmuş, duydu mu? Ben duyunca çok mutlu oldum doğusu! \$ B: Duydum, duymaz mıyım! Nefis bir haber gerçekten de. Havaalanında güvenlik görevlisi olarak çalışacakmış.	168	A: Ceyhun bir yerde piknik yapmak istiyor demek.	48	Ceyhun dışarı çıkmak istemiyor, evde kalmak istiyor.	H
A: Salih okula geri dönmek istiyormuş. Dün markette karşılaştık, o zaman bahsetti. \$ B: Bana da anlattı biraz. Sınavlara tekrar girip üniversite öğrencisi olmak niyetindeymiş.	181	A: Demek Ercan havaalanında çalışmayı planlıyor.	48	Ercan karakolda güvenlik görevlisi olarak çalışacak.	H
A: Leyla işinden hiç ama hiç memnun değil. Ne kendini geliştirebiliyor ne de önünü görebiliyor. \$ B: Maalesef... Tam da bu yüzden yeni bir iş arıyor kendine bir süredir.	175	A: Salih tekrar üniversiteli olacak desene.	43	Salih sınava girip liseyi tekrar okumak istiyor.	H
A: Mikail'in seracılığa başladığını duydum. Hep böyle ilginç şeyler peşinde doğrusu! \$ B: Toprakla uğraşmayı çok seviyor. Şimdi serasında avokado yetiştirmek için uğraşıyor.	169	A: Aaa, Leyla yeni bir iş arıyor demek kendine.	47	Leyla işinden çok memnun ama yine de yeni bir iş arıyor.	H
A: Bedri'nin sana bir sürprizi varmış! Doğum günün yaklaştığı için sana bir jest yapmak istemiş. \$ B: Evet, söyledi. Çok ince düşünmüş, benim bir portremini çizicekmiş.	173	A: Mikail avokado yetiştirecek ha, bir yaşına daha girdim!	58	Mikail seracılığa ananas yetiştirmek için başlamış.	H
	166	A: Demek Bedri senin bir portreni çizicek. Gerçekten iyi düşünmüş.	66	Bedri arkadaşının doğum günü için bir duvar resmi çizmeyi planlıyor.	H

A: Meral'i gördüm koridorda. Başhekime kız kardeşi ile çalışmak istemediğini söylüyordu. Ne alaka? \$ B: Meral'in kız kardeşi de doktor, hem de bizim hastanede. Alakası budur.	174	A: Meral'in birlikte çalışmak istemeyeceği bir doktor olacağını düşünmezdim.	76	Meral, doktor olan kız kardeşi ile beraber çalışmak istemiyor.	E
A: Merve çok çalışmaktan şikayetçi. Oğluna hiç vakit ayıramadığı için epey mutsuzdu. \$ B: Geçenlerde yeni iş arıyordu. Demek iş değiştirmek istemesi bu yüzden mi?	162	A: Desene sen Merve kendine yeni bir iş bulmak istiyor diye!	60	Merve oğluna daha fazla vakit ayırmak için yeni iş arıyor.	E
A: Baloya Şeyda da bizimle gelecek. Haberin var değil mi? \$ B: Yoktu doğrusu! Söylediğin için teşekkürler. Geçen hafta kendine yeni ayakkabı bakıyordu, bundanmış demek.	168	A: Şeyda baloda giymek için ayakkabı mı almak istiyor yoksa?	60	Baloya gitmeyi planlayan Şeyda kendine yeni ayakkabı bakıyor.	E
A: Nurdan bu odaya yeni bir perde asmak istiyor. Şu anki perdelerin odayı kasvetli gösterdiğini düşünüyor. \$ B: Haklı doğrusu! Buraya şöyle açık renkli perdeler asmak lazım.	173	A: Nurdan buraya açık renkli perde mi asmak istiyor yani?	57	Nurdan odanın kasvetli göründüğünü düşünüyor.	E
A: Yakışıklı bir adam Feride'ye ismini sormuş. Sonra da ona bir pastahanede çay ısmarlamış. \$ B: Feride de öyle her yakışıklıya adını söylüyor muymuş? Olacak iş değil!	167	A: Şu şehirde Feride'nin adını bilmeyen bir yakışıklı kalmadı ki.	65	Yakışıklı bir adam Feride'ye çay ısmarlamış.	E
A: Candan bir opera bestelese ne güzel olur. Sesinin ihtişamına öyle yakışır ki! \$ B: Böyle düşünüyorsan sana çok güzel haberlerim var. Candan bir müzik kayıt şirketi ile anlaşmış.	180	A: Yoksa bir opera albümü mü kaydedecek o şirket ile?	53	Candan bir müzik şirketi ile anlaşma yapmış.	E

A: Cemil'in bu yarış kazanabileceğini hiç ama hiç sanmıyorum! Onu hiç antrenmanda görmüyorum. Ş B: Ben de seninle aynı fikirdeyim. Cemil yarışa hazırlanmak dışında her şeyi yapıyor!	182	A: Bu gidişle Cemil sadece bu yarış değil hiçbir yarış kazanamaz!	67	Cemil gireceği yarışa gerektiği kadar iyi hazırlanmıyor.	E
A: Magazin haberlerinde gördüm, Dost Oyeski bu ara varını yoğunu ganyana yatırıyor. Ş B: Yolu yol değil onun! Çok fena bir kumarbaza dönüştü gerçekten.	154	A: Dost Oyeski'nin bir kumarbaz olacağı kırk yıl düşünsem aklıma gelmezdi.	74	Dost son zamanlarda kumar oynamaya çok vakit ve para harcıyor.	E
A: Hülya hamurışları konusunda kendini çok geliştirmiş. Geçenlerde yaptığı kekt nefisti! Ş B: Bir sürü tarif öğrenmiş, nefis tatlılar yapıyor. Hayali bir pastane açmakmış.	171	A: Demek Hülya'nın hayallerinde pastacılık var!	47	Hülya bir pastacı açmak istediği için bir sürü yeni tarif öğreniyor.	E

APPENDIX D

Processing Experiment Interface



Merhaba!

Türkçe tümcelerın nasıl işlendiđi üzerine yaptığımız bu deneye katılmaya sizi davet ediyoruz. Araştırmaya katılmayı kabul ettiđiniz taktirde sizden bazı cümleleri okumanız ve bu cümlelere dair sorulacak sorulara cevap vermeniz beklenmektedir. Ayrıca bir sonraki ekranda istenilen bilgileri de sağlamanızı rica ediyoruz.

İsminiz ve kişisel bilgileriniz tamamen gizli tutulacaktır. Çalışmaya katılmanız tamamen isteđe bađlıdır. Sizden ücret talep etmiyoruz ve size herhangi bir ödeme yapmayacağız. Sizden alınan dil verisi ileride başka çalışmalar için de kullanılabilir.

İstediđiniz zaman çalışmaya katılmaktan vazgeçebilirsiniz. Bu durumda sizden almış olduđumuz veriler imha edilecektir. Yaptık istediđimiz araştırmanın size herhangi bir risk getirmesi beklenmemektedir.

Deney yaklaşık 15 dakika sürecektir.

Deneye yalnızca **bilgisayar üzerinden** katılabilirsiniz, **MOBİL CİHAZLARDAN KATILMANIZ MÜMKÜN DEĞİLDİR**. Dış etkenlerden etkilenmememiz adına deneye yalnızken katılmanızı öneririz.

Çalışmayla ilgili sorularınız var ise araştırmanın yürütücöleri ile iletişime geçebilirsiniz.

Şimdiden vaktiniz ve katkınız için çok teşekkürler!

Araştırmanın yürütücöleri:

Dr. Ümit Atlamaz

Dr. Ömer Demirok

Büşra Marşan

Boğaziçi Üniversitesi, Dilbilim Bölümü

Utku Türk

University of Maryland, Department of Linguistics

☐ Yukarıda yazılanları anladım. Çalışmaya katılmayı kabul ediyorum.

Devam etmek için tıklayın

Participants were first shown a welcome screen where they found general information about the experiment: Duration, purpose etc.



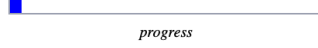
progress

Anadiliniz nedir? (Birden fazla anadiliniz varsa lütfen hepsini yazınız.)

Kaç yaşındasınız?

Devam

Participants were asked about their native language(s) and age.



progress

Deneyimize hoşgeldiniz!

Bu deneyde sizden verilen diyalogları okumanızı, sonrasında okuduğunuz diyaloglarla ilgili cümlelere cevap vermenizi istiyoruz. Diyaloglar ve ilgili cümleler size **parça parça** gösterilecektir. İlk olarak ekranın ortasında bir **yıldız (*)** göreceksiniz. Ardından diyalogun ilk kısmını göreceksiniz:

A: Ayşe kendisine yeni bir yazlık yaptırmıyormuş.

B: Hem de Bodrum'da yaptırmıyormuş!

Boşluk tuşuna bastığınızda diyalogun ilk kısmı kaybolacak ve ikinci kısmını göreceksiniz. İlk başta sadece **cümlelerin ne uzunlukta olduğunu** gösteren kısa çizgiler göreceksiniz:

Boşluk tuşuna tekrar bastığınızda cümlelerin bir kısmı görünür ve önceki kısmı kaybolur. Cümle bitene kadar siz boşluk tuşuna bastıkça cümlelerin yeni bir kısmı görünür hale gelir.

----- Ayşe -----
----- kendisine -----
----- Bodrum'da -----
----- bir -----
----- yazlık -----
----- yaptırmıyor!

İçeriği anlamak için cümleleri kendi **normal okuma hızınızda** okuyun. Okurken lütfen bir parmağınızı **BOŞLUK TUŞU** üzerinde tutun. Bir cümleyi okuduktan sonra bir sonrakinin görünmesi için her zaman boşluk tuşuna basın.

Okuduğunuz diyalogları **anlama soruları** takip edecek. Bu sorularda size **diyaloga dair bir cümle** gösterilecek ve bu cümlelerin verilen diyalogdan çıkarılıp çıkarılmadığı sorulacaktır. Eğer cümlede verilen bilgi diyalog ile **örtüşüyorsa Evet**, diyalog ile **örtüşmüyorsa ya da çelişiyorsa Hayır** cevabını vermelisiniz.

Lütfen bu soruyu uygun cevaba tıklayarak veya **F (=EVET, ÖRTÜŞÜYÖR)** ya da **J (=HAYIR, ÖRTÜŞMÜYÖR)** rakamına basarak cevaplayın.

Cevap verdikten sonra yanıtınızı gözden geçiremez ya da değiştiremezsiniz. Cevap verdikten sonra tekrar ekranın ortasında bir **yıldız (*)** göreceksiniz. Bu yeni bir cümle görmeye başlayacağınızın işareti. Lütfen boşluk tuşuna basarak ilerleyiniz.

Bir anlama sorusu şöyle görünebilir:

Ayşe kendisine çanak-kale'de bir çiftlik yaptırmıyor.

F (Evet) J (Hayır)

İlk önce 5 örnek diyalog ile bir alıştırmaya yapacaksınız. Bu alıştırmalarda size cevaplarınızın doğruluğu ile ilgili geridönüt verilecektir. Fakat deney başladıktan sonra artık geridönüt almayacaksınız. Alıştırmaları bitirdikten sonra gerçek deneyi başlayacaksınız.

Deneyle ilgili bir sorunuz varsa lütfen araştırmacı ile iletişime geçin:

Büşra Marşan, busra.marsan@boun.edu.tr

Teşekkürler!



progress

A: Çiçek öyle bir şiir okudu ki hepimizi hüngür hüngür ağlattı. Ne kadar yetenekli bir çocuk!

B: Öyle gerçekten. Haftaya da Shakespeare'in sonelerinden birini okumayı planlıyor.

Disambiguating context in the form of a dialogue.



progress

Demek _____

Ambiguous sentence, shown in the self paced reading format. Each time the participant hit the space button, a new word of the sentence was shown.



progress

Çiçek Shakespeare'in sonelerini okumak istiyor.

F (Evet) J (Hayır)

Sample comprehension question.



progress

Mahide bir muhtarın portresini çizecek.

F (Evet) J (Hayır)

Daha dikkatli okumalısın!



progress

Yulaf kavrulmuş balık yemek istediği için mutfakta.

F (Evet) J (Hayır)

Bu diyalogu anladın, bravo!

Participants received feedback to their answers in the practice round: "You need to read more carefully!" or "You understood this dialogue, bravo!"



progress

Alıştırma turunu tamamladınız. Deneye başlayabilirsiniz.

Artık geridönüt almayacaksınız.

Deneye başlamadan önce parmaklarınızı 'boşluk', 'F' ve 'J' tuşlarına yerleştirmenizi, telefonunuzu sessize almanızı ve deneyi tam ekran yapmanızı öneriyoruz. Deneye başlamak için aşağıdaki butona basınız.

Deneyi başlat

Participants were notified after they finished the practice round and before they proceeded to the actual experiment.

progress

Deneyi tamamladınız.

Cevaplarınız başarıyla kaydedildi.

Vaktiniz ve katılımınız için çok teşekkürler!

Dilerseniz, aşağıda çalışma ile ilgili detayları okuyabilirsiniz. Bu detayları henüz deneyi yapmamış arkadaşlarınızla paylaşmamanızı rica ederiz.

Ekstra kredi almak istiyorsanız, aşağıdaki butona basıp bir sonraki sayfaya geçebilirsiniz.

Çalışmanın Amacı

Bu çalışmada özel bir muğlaklık türü olan De Re De Dicto'nun insanlar tarafından nasıl yorumlandığını inceliyoruz. Çalışmaya dair bir sorunuz varsa araştırmacı ile aşağıdaki mail adresi üzerinden iletişime geçebilirsiniz:

Büşra Marşan busra.marsan@boun.edu.tr

Araştırmayla ilgili haklarınız konusunda Boğaziçi Üniversitesi Sosyal ve Beşeri Bilimler Yüksek Lisans ve Doktora Tezleri Etik İnceleme Komisyonu'na (SOBETİK) (sbe-ethics@boun.edu.tr) danışabilirsiniz

Ekstra Kredi

Boğaziçi Üniversitesi öğrencisiyseniz ve Ling kodlu derslerden birinden **ekstra kredi almak istiyorsanız** lütfen aşağıdaki butona basarak ilgili forma gidiniz.

Ekstra Kredi

Participants were informed about the purpose and topic of the study after they completed the experiment. They were also presented with a link that leads to a form for extra credit.

Deney sona erdi. Yanıtlarınız otomatik olarak kaydedildi. Katılımınız için çok teşekkür ederiz!

Deneye katılımınız karşılığında kredi istiyorsanız hangi ders için kredi istediğinizi lütfen belirtiniz (birden fazla şubeli dersler için şubenizi de belirtiniz). Eğer belirttiğiniz dersin hocası deneyin duyurusunu yapmadıysa, o ders için ekstra kredi isteyemezsiniz.

Adınız:

Öğrenci numaranız:

Ders kodu:

Dersi veren hocanızın adı:

(Bu bilgiler deney yanıtlarından ayrı olarak kaydedilecektir ve herhangi bir şekilde deney cevaplarınızla eşleştirilemez.)

[→ Ekstra Kredi için bilgilerimi gönder!](#)

Extra credit form.

REFERENCES

- Akın, A. A. & Akın, M. D. (2007). Zemberek, an open source nlp framework for turkic languages. *Structure*, 10(2007), 1–5.
- Akkurt, S. F., Marşan, B., & Uskudarli, S. (2022). Boat v2—a web-based dependency annotation tool with focus on agglutinative languages. *arXiv preprint arXiv:2207.01327*.
- Aksan, Y. & Aksan, M. (2009). Building a national corpus of turkish: Design and implementation. *WorNing Papers in Corpus-based Linguistics and Language Education. ToNyo: TUFs*, (3), 299–310.
- Amblard, M., Beysson, C., de Groote, P., Guillaume, B., & Pogodalla, S. (2020). A French Version of the FraCaS Test Suite. In *LREC 2020 - Language Resources and Evaluation Conference* (p.9). Marseille, France.
- Amirkhani, H., AzariJafari, M., Pourjafari, Z., Faridan-Jahromi, S., Kouhkan, Z., & Amirak, A. (2020). Farstail: A persian natural language inference dataset. *arXiv preprint arXiv:2009.08820*.
- Aquinas, T. (1886). *De propositionibus modalibus*. Sancti Thomae Aquinatis Opuscula philosophica et theologica. S. Lapi.
- Bache, S. M. (2019). Wickham h. magrittr: a forward-pipe operator for r. 2014. *URL https://CRAN.R-project.org/package=magrittr*. R package version, 1.
- Bakay, Ö., Ergelen, Ö., Sarıms, E., Yıldırım, S., Arıcan, B. N., Kocabalcıoğlu, A., Özçelik, M., Saniyar, E., Kuyrukçu, O., Avar, B., et al. (2021). Turkish wordnet kenet. In *Proceedings of the 11th Global Wordnet Conference* (pp. 166–174).
- Bentivogli, L., Bernardi, R., Marelli, M., Menini, S., Baroni, M., & Zamparelli, R. (2016). Sick through the semeval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Language Resources and Evaluation*, 50, 95–124.
- Bentivogli, L., Clark, P., Dagan, I., & Giampiccolo, D. (2009). The fifth pascal recognizing textual entailment challenge. In *TAC: Citeseer*.
- Bentivogli, L., Clark, P., Dagan, I., & Giampiccolo, D. (2011). The seventh pascal recognizing textual entailment challenge. In *TAC: Citeseer*.
- Bernardy, J.-P. & Chatzikyriakidis, S. (2017). A type-theoretical system for the fracas test suite: Grammatical framework meets coq. In *IWCS 2017-12th International Conference on Computational Semantics-Long papers*.
- Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- Bowman, S. R. & Dahl, G. E. (2021). What will it take to fix benchmarking in natural language understanding? *arXiv preprint arXiv:2104.02145*.

- Budur, E., Özçelik, R., Güngör, T., & Potts, C. (2020). Data and representation for Turkish natural language inference. *arXiv preprint arXiv:2004.14963*.
- Carston, R. (2011). Truth-conditional semantics. *Philosophical Perspectives for Pragmatics*, 10, 280.
- Carston, R. (2013). 12. implicature, explicature, and truth-theoretic semantics. *The Semantics-Pragmatics Boundary in Philosophy*, 261.
- Charlow, S. & Sharvit, Y. (2014). Bound'de re'pronouns and the lfs of attitude reports. *Semantics and Pragmatics*, 7, 3–1.
- Chen, D. & Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 190–200).
- Chen, Q., Zhu, X., Ling, Z.-H., Inkpen, D., & Wei, S. (2017). Neural natural language inference models enhanced with external knowledge. *arXiv preprint arXiv:1711.04289*.
- Condoravdi, C., Crouch, D., De Paiva, V., Stolle, R., & Bobrow, D. (2003). Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning* (pp. 38–45).
- Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.
- Cooper, R., Crouch, D., van Eijck, J., Fox, C., van Genabith, J., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., et al. (1996a). A strategy for building a framework for computational semantics (the way forward).
- Cooper, R., Crouch, D., Van Eijck, J., Fox, C., Van Genabith, J., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., et al. (1996b). *Using the framework*. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.
- Cooper, R., Crouch, R., van Eijck, J., Fox, C., van Genabith, J., Jaspars, J., Kamp, H., Pinkal, M., Milward, D., Poesio, M., et al. (1996c). Building the framework. *FraCaS Deliverable D*, 15, 62–051.
- Cooper, R., Crouch, R., van Eijck, J., Fox, C., van Genabith, J., Jaspars, J., Kamp, H., Pinkal, M., Poesio, M., & Pulman, S. (1995). Evaluating the state of the art. *FraCaS Deliverable D*, 10.
- Cooper, R., Crouch, R., van Eijck, J., Fox, C., van Genabith, J., Jaspars, J., Kamp, H., Pinkal, M., Poesio, M., Pulman, S., et al. (1994). The state of the art in computational semantics: Evaluating the descriptive capabilities of semantic theories. *Deliverable D9, Centre of Cognitive Science, Edinburgh*.
- Cresswell, M. J. (2012). *Entities and indices*, volume 41. Springer Science & Business Media.

- Dagan, I., Glickman, O., & Magnini, B. (2005). The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop* (pp. 177–190).: Springer.
- Dagan, I., Glickman, O., & Magnini, B. (2006). The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers* (pp. 177–190).: Springer.
- De Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, 47(2), 255–308.
- Dede, M. (1986). Definiteness and referentiality in turkish verbal sentences. *Studies in Turkish Linguistics. Amsterdam: Benjamins*, (pp. 147–164).
- Demirok, Ö. (2020). A pied-piping theory of exceptional de re: Scoping after all.
- Devitt, M. & Sterelny, K. (1987). *Language and reality: An introduction to the philosophy of language*. Blackwell.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dinneen, F. P. (1990). *Peter of Spain: Language in Dispute. An English translation of Peter of Spain's' Tractatus' called afterwards SUMMULAE LOGICALES, based on the critical edition by LM de Rijk*, volume 39. John Benjamins Publishing.
- Drozdovskaia, E. (2021). De dicto / de re in russian propositional attitude reports: an empirical study.
- Elbourne, P. (2013). *Definite descriptions*, volume 1. Oxford University Press.
- Enç, M. (1991). The semantics of specificity. *Linguistic Inquiry*, (pp. 1–25).
- Fodor, J. D. (1970). *The linguistic description of opaque contexts*. (Unpublished PhD thesis). Massachusetts Institute of Technology.
- Frege, G. (1892). Über begriff und gegenstand.
- Ghasemi, Z. & Keyvanrad, M. A. (2021). Farsick: A persian semantic textual similarity and natural language inference dataset. In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)* (pp. 194–199).: IEEE.
- Giampiccolo, D., Dang, H. T., Magnini, B., Dagan, I., Cabrio, E., & Dolan, B. (2008). The fourth pascal recognizing textual entailment challenge. In *TAC*.
- Giampiccolo, D., Magnini, B., Dagan, I., & Dolan, W. B. (2007). The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (pp. 1–9).

- Grant, D. A. (1948). The latin square principle in the design and analysis of psychological experiments. *Psychological Bulletin*, 45(5), 427.
- Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.
- Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1), 1–12.
- Heim, I. & Kratzer, A. (1998). *Semantics in generative grammar*, volume 1185. Blackwell Oxford.
- Hintikka, K. J. J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*.
- Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47, 853–899.
- Hoffman, B. A. (1995). *The computational analysis of the syntax and interpretation of "free" word order in Turkish*. University of Pennsylvania.
- Imambi, S., Prakash, K. B., & Kanagachidambaresan, G. (2021). Pytorch. *Programming with TensorFlow: Solution for Edge Computing Applications*, (pp. 87–104).
- İşsever, S. (2003). Information structure in Turkish: The word order–prosody interface. *Lingua*, 113(11), 1025–1053.
- Kamp, H. (2013). Formal properties of now. *Meaning and the Dynamics of Interpretation: Selected Papers of Hans Kamp*, 29, 11.
- Kara, N., Aslan, D. B., Marşan, B., Bakay, Ö., Ak, K., & Yıldız, O. T. (2020). TRopbank: Turkish Propbank v2. 0. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 2763–2772).
- Kawazoe, A., Tanaka, R., Mineshima, K., & Bekki, D. (2017). An inference problem set for evaluating semantic theories and semantic processing systems for japanese. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2015 Workshops, LENLS, JURISIN, AAA, HAT-MASH, TSDAA, ASD-HR, and SKL, Kanagawa, Japan, November 16-18, 2015, Revised Selected Papers* (pp. 58–65).: Springer.
- Kechriotis, Z. C. A. (2009). Referentiality in turkish: Np/dp. In *Essays on Turkish Linguistics: Proceedings of the 14th International Conference on Turkish Linguistics, August 6-8, 2008*, volume 79 (p.83).: Otto Harrassowitz Verlag.
- Kelepir, M. (2001). *Topics in Turkish syntax: Clausal structure and scope*. (Unpublished PhD thesis). Massachusetts Institute of Technology.
- Keshet, E. (2010). Split intensionality: A new scope theory of de re and de dicto. *Linguistics and Philosophy*, 33, 251–283.

- Keshet, E. & Schwarz, F. (2019). De re/de dicto. *The Oxford Handbook of Reference*, (pp. 167–202).
- Keshet, E. R. (2008). *Good intensions: Paving two roads to a theory of the de re/de dicto distinction*. (Unpublished PhD thesis). Massachusetts Institute of Technology.
- Khot, T., Sabharwal, A., & Clark, P. (2018). Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Kneale, W. (1962). Modality de dicto and de re,[w:] e. nagel, p. suppes, a. tarski (red.), logic, methodology and philosophy of science.
- Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.
- Kratzer, A. (2007). Situations in natural language semantics. *Stanford Encyclopedia of Philosophy*.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.
- Larson, B. (2017). Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 1–11).
- Leidner, J. L. & Plachouras, V. (2017). Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 30–40).
- Lev, I. (2006). Logic puzzles: A new test-suite for compositional semantics and reasoning.
- Lewis, C. I. (1932). *A survey of symbolic logic*. University of California press.
- Liu, P., Wang, X., Xiang, C., & Meng, W. (2020). A survey of text data augmentation. In *2020 International Conference on Computer Communication and Network Security (CCNS)* (pp. 191–195).: IEEE.
- Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3* (pp. 246–252).: Springer.
- Ljunglöf, P. & Siverbo, M. (2011). *A Bilingual Treebank for the FraCaS Test Suite CLT Project Report*. Technical report, Technical Report, University of Gothenburg.
- MacCartney, B. & Manning, C. D. (2007). Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (pp. 193–200).

- MacCartney, B. & Manning, C. D. (2008). Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)* (pp. 521–528).
- Manning, C. D. (2006). Local textual inference: Its hard to circumscribe, but you know it when you see it—and nlp needs it.
- Marcus, R. B. (1961). Modalities and intensional languages. *Synthese*, (pp. 303–322).
- Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., et al. (2014). A sick cure for the evaluation of compositional distributional semantic models. In *Language Resources and Evaluation Conference* (pp. 216–223).: Reykjavik.
- Marşan, B., Kara, N., Özçelik, M., Arıcan, B. N., Cesur, N., Kuzgun, A., Saniyar, E., Kuyrukçu, O., & Yıldız, O. T. (2021). Building the Turkish Framenet. In *Proceedings of the 11th Global Wordnet Conference* (pp. 118–125).
- Monz, C. & de Rijke, M. (2001). Light-weight entailment checking for computational semantics. In *Proceedings of the Third Workshop on Inference in Computational Semantics (ICoS-3)*.
- Morey, R. D. et al. (2008). Confidence intervals from normalized data: A correction to cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, 4(2), 61–64.
- Naik, A., Ravichander, A., Sadeh, N., Rose, C., & Neubig, G. (2018). Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*.
- Novaes, C. D. (2003). A medieval reformulation of the de dicto/de re distinction.
- Öztürel, A., Kayadelen, T., & Demirsahin, I. (2019). A syntactically expressive morphological analyzer for turkish. In *Proceedings of the 14th International Conference on Finite-state Methods and Natural Language Processing* (pp. 65–75).
- Ozturk, B. (2004). *Case, referentiality and phrase structure*. Harvard University.
- Pavlick, E. & Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7, 677–694.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.
- Percus, O. (2000). Constraints on some other variables in syntax. *Natural Language Semantics*, 8(3), 173–229.
- Pinkal, M. & Koller, A. (2019). Semantic research in computational linguistics. *Semantics-Typology, Diachrony and Processing*, (p. 366).

- Plantinga, A. (1976). Actualism and possible worlds. *Theoria*, 42(1-3), 139–160.
- Prabhumoye, S., Boldt, B., Salakhutdinov, R., & Black, A. W. (2020). Case study: Deontological ethics in nlp. *arXiv preprint arXiv:2010.04658*.
- Prior, A. N. (1952). Modality de dicto and modality de re. *Theoria*, 18(3), 174–180.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.
- Quine, W. V. (1956). Quantifiers and propositional attitudes. *The Journal of Philosophy*, 53(5), 177–187.
- Quine, W. V. O. (1971). *Reference and modality*. Oxford University Press, USA.
- Reppen, R., Ide, N., & Suderman, K. (2005). American national corpus (anc) second release. *Linguistic Data Consortium*.
- Richardson, L. (2007). Beautiful soup documentation.
- Rogers, A., Baldwin, T., & Leins, K. (2021). Just what do you think you’re doing, dave?’ a checklist for responsible data use in nlp. *arXiv preprint arXiv:2109.06598*.
- Russell, B. (1905). On denoting. *Mind*, 14(56), 479–493.
- Sak, H., Güngör, T., & Saraçlar, M. (2011). Resources for turkish morphological processing. *Language Resources and Evaluation*, 45, 249–261.
- Schwarz, F. (2012). Situation pronouns in determiner phrases. *Natural Language Semantics*, 20, 431–475.
- Schweter, S. (2020). Berturk-bert models for turkish, april 2020. URL <https://doi.org/10.5281/zenodo.3770924>.
- Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.
- Sharma, S., Santra, B., Jana, A., Santosh, T., Ganguly, N., & Goyal, P. (2019). Incorporating domain knowledge into medical nli using knowledge graphs. *arXiv preprint arXiv:1909.00160*.
- Sherwood, W. o., Kann, C., & Brands, H. (1995). *Einführung in die Logik. Introductiones in Logicam.: Lateinisch-Deutsch*. Felix Meiner Verlag.
- Shmueli, B., Fell, J., Ray, S., & Ku, L.-W. (2021). Beyond fair pay: Ethical implications of nlp crowdsourcing. *arXiv preprint arXiv:2104.10097*.
- Simard, P. Y., LeCun, Y. A., Denker, J. S., & Victorri, B. (1998). Transformation invariance in pattern recognition tangent distance and tangent propagation. In *Neural Networks: Tricks of the Trade* (pp. 239–274). Springer.
- Siverbo, M. (2012). Creating inference rules based on gf syntax trees for the fracas test suite.

- Soucy, P. & Mineau, G. W. (2001). A simple knn algorithm for text categorization. In *Proceedings 2001 IEEE International Conference on Data Mining* (pp. 647–648).: IEEE.
- Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1422–1432).
- Taylan, E. E. (1984). *The function of word order in Turkish grammar*, volume 106. University of California Press.
- Thalenberg, B., Guide, B., Silva, C., de Oliveira Lima, G., Câmara, I. C., Stanojevic, M., Souza, R., & de Paiva, V. (2018). Sick-br: A portuguese corpus for inference. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, volume 11122 (p. 303).: Springer.
- Tiskin, D. (2021). De re, de dicto and the problem of the default case.
- Tong, S. & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2(Nov), 45–66.
- van Benthem, J. F. (1977). Tense logic and standard logic. *Logique et Analyse*, 20(80), 395–437.
- Vanderwende, L., Menezes, A., & Snow, R. (2006). Microsoft research at rte-2: Syntactic contributions in the entailment task: an implementation. In *Proceedings of the Second PASCAL Challenges Workshop*.
- von Wright, G. H. (1951). An essay in modal logic.
- Wang, W. Y. & Yang, D. (2015). Thats so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2557–2563).
- Wang, X., Kapanipathi, P., Musa, R., Yu, M., Talamadupula, K., Abdelaziz, I., Chang, M., Fokoue, A., Makni, B., Mattei, N., et al. (2019). Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33 (pp. 7208–7215).
- Wang, Z., Li, L., & Zeng, D. (2020). Knowledge-enhanced natural language inference based on knowledge graphs. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6498–6508).
- Wei, J. & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

- (*EMNLP-IJCNLP*) (pp. 6383–6389). Hong Kong, China: Association for Computational Linguistics.
- Wickham, H. (2011). *ggplot2*. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2), 180–185.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.
- Wijnholds, G. & Moortgat, M. (2021). Sicknl: A dataset for dutch natural language inference. *arXiv preprint arXiv:2101.05716*.
- Williams, A., Nangia, N., & Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45).
- Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 67–78.
- Zadeh, A., Chan, M., Liang, P. P., Tong, E., & Morency, L.-P. (2019). Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8807–8817).
- Zaenen, A., Karttunen, L., & Crouch, R. (2005). Local textual inference: can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment* (pp. 31–36).
- Zhang, Y., Baldridge, J., & He, L. (2019). Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.
- Zhang, Y. & Davidson, K. (2021). De re interpretation in belief reports: An experimental investigation. *Experiments in Linguistic Meaning*, 1, 310–321.