A coreference corpus of Turkish situated dialogs

Faruk Büyüktekin and **Umut Özge**
Informatics Institute, Department of Cognitive Science,
Middle East Technical University
faruk.buyuktekin@metu.edu.tr, umozge@metu.edu.tr

Abstract

The paper introduces a publicly available corpus of Turkish situated dialogs annotated for coreference. We developed an annotation scheme for coreference annotation in Turkish, a language with pro-drop and rich agglutinating morphology. The annotation scheme is tailored for these aspects of the language, making it potentially applicable to similar languages. The corpus comprises 60 dialogs containing in total 3900 sentences, 18360 words, and 6120 mentions.

1 Introduction

Coreference annotation and corpus research have attracted significant attention among NLP researchers, cognitive scientists, and linguists, as understanding referring expressions and the relations between them is fundamental to natural language understanding. Numerous NLP tasks, including information retrieval, question answering, and summarization, require coreference resolution for effective performance. This need has resulted in an increase in the number of corpora annotated for coreference relations in recent decades, particularly with the success of data-driven techniques, especially for widely-studied languages like English (Weischedel et al., 2011; Zeldes, 2017; Uryupina et al., 2020) and German (Lapshinova-Koltunski and Ferreira, 2022; Bourgonje and Stede, 2020).

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

In this connection, adapting existing annotation schemes to Turkish poses numerous challenges and it is particularly challenging to offer a universal scheme for all languages when the complexity of the anaphoric phenomena is taken into consideration as stated by Poesio (2004). For instance, the treatment of morphological information, such as suffixes that carry referential information, is often overlooked. Similarly, the handling of phonologically null elements, which are pervasive in Turkish, is not sufficiently addressed. This inadequacy can lead to a loss of critical information necessary for accurate coreference resolution. As a result, there is a need for developing a specialized annotation scheme that can accommodate the unique features of Turkish and similar languages, ensuring more robust and reliable coreference annotation.

This study is driven by the necessity to develop a coreference dataset in Turkish, a language with relatively limited resources. It proposes a novel annotation scheme for coreference annotation, addressing the challenges encountered when adapting existing schemes designed for languages such as English. The paper is organized as follows: Sections 2 outlines the basic terminology related to coreference. Section 3 reviews the related work in coreference corpora. Section 4 describes the initial steps in corpus development. Section 5 introduces the proposed annotation scheme. Section 6 provides descriptive statistics of the resulting corpus. Section 7 ends with a summary and outlines future research directions.

2 Basic terminology

Coreference can be better understood within the larger picture of cohesion and concepts related to it. Cohesion itself is based on the idea that the spoken or written communication is usually a united whole rather than unrelated utterances or sentences. For cohesion to occur, the interpretation of some

Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024), pages 42–52

linguistic element in the discourse sometimes depends on previously mentioned items in the text (Halliday and Hasan, 1976). A closely related notion to cohesion is reference. It is the relationship between a linguistic expression and an entity in the world. There are two main types of reference. Exophoric reference refers to an entity which is outside the text. On the contrary, endophoric reference refers to another expression in the preceding discourse segment. Endophora is further divided into two types. Anaphora can be described as an item which relates back to a previous item in some way. The element which is referring back is called anaphor and the previously mentioned entity which then anaphor refers to or is related to is its antecedent. The process of linking the anaphor with its antecedent is called anaphora resolution. Cataphora, on the other hand, points to an item in the following discourse segment.

There are a variety of anaphora which are observed in written or oral language based on the form of the anaphor (Mitkov, 2014). Lexical noun phrase anaphora could appear as proper names and definite descriptions. Pronominal anaphora is one of the most studied and therefore understood type of anaphora in the literature. Anaphors in this type can be in the form of personal pronouns, possessive pronouns, reflexive pronouns, and demonstrative pronouns. Another type of anaphora is zero anaphora. It is considered to be one of the most challenging types of anaphora to resolve since they are not physically realized at the surface level. Although they are invisible, they do not damage the cohesion of the discourse but strengthen it. They are decoded by the reader or hearer without any loss during the comprehension of the discourse. If the anaphor and antecedent refer to the same entity, they are thought to be coreferential. This relation is also called identity anaphora, as in (1).

(1) A man came. **He** brought a book.

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

3 Related work

The earliest attempts to develop annotation schemes for coreference annotation could be traced back to the Message Understanding Conference (MUC) information extraction tasks (Hirschman et al., 1997). The task was created to group all the mentions of an entity together and the scheme specified the basic task criteria, the markables to be annotated and the relations to be established in English. The task evolved with Automatic Content Extraction (ACE) Program (Doddington et al., 2004) enriching the coverage with entity, relation, and event annotation in English, Chinese and Arabic. The MATE/GNOME proposals (Poesio, 2004) were geared towards being more linguistically oriented than previous schemes, making a discourse model assumption. It also included bridging anaphora in addition to identity relations.

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

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

The development of annotation schemes have paved the way for the construction of many corpora in different languages. The initial products were naturally produced in the languages of the schemes mentioned above. One of the well-known and largest coreference corpora is the OntoNotes project (Weischedel et al., 2013). It consists of

various genres such as news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows. It was annotated for syntax and predicate argument structure and word sense and coreference in English, Chinese, and Arabic. ARRAU (Uryupina et al., 2020) is another multi-genre corpus which contains around 350K tokens. Unlike many corpora, it accepts nonreferential NPs and singletons as markable. It was annotated for different kinds of anaphoric relations including coreference, bridging anaphora and discourse deixis.

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

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

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

4 Corpus creation

4.1 Genre selection

We selected situated dialogs as the genre for our corpus. Most coreference corpora started with texts like news, and continued with articles, and stories (Uryupina et al., 2020). However, we chose to annotate situated dialogs with spontaneous speech. The language in this genre exhibit certain features. The utterances/sentences are relatively short compared to the genres like news and articles and therefore grammatically less complicated. Speakers might often produce ungrammatical forms and add disfluencies, which is associated with cognitive load and planning.

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

4.2 The source of our texts

Our dialogs were taken from an experimental setting where pairs were expected to solve tangram

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

Table 1: Anchors for the entities

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

4.3 Text preparation

We firstly transcribed the speech between the participants in the form of dialogs, indicating the roles of the pairs (ie. presenter and operator). Then, we manually split them into sentences and added punctuation where necessary. We added anchors for the entities available in the physical context at the beginning of each dialog. These are discourse participants and seven geometric including shapes two small triangles, one middle triangle, and two big triangles, small square, small parallelogram (see Table 1). We then encoded the dialogs in `JSON` format.

4.4 Tool choice

As a result of our evaluation of various annotation tools, we decided to use Labelbox (2024). It offers inherent templates for conversational texts, relatively easy annotation, and most importantly allows morpheme and character selection to capture morphological and null elements. It was also the most suitable tool to work with dyadic dialog data (see Figure 1 for a sample annotation).

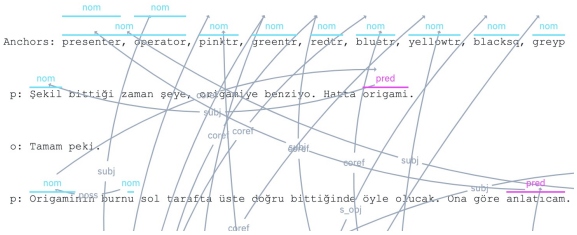

Figure 1: Annotation sample from the annotation tool

4.5 Training the annotators

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

5 Annotation scheme development

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

5.1 Scope

Our annotation scheme mainly focuses on how to annotate co-referring expressions. Our definition of coreference follows Deemter and Kibble (2000): "α_1 and α_2 corefer if and only if Referent(α_1) = Referent(α_2)"

However, certain anaphoric relations fall beyond the scope of this work. Discourse deixis (Webber, 1988) can be associated with coreference but discourse deictic expressions may refer to preceding or succeeding discourse segments, such as clauses or sentences. Given that the antecedents in these cases are non-nominal expressions (Çallı, 2012), we do not include them in our dataset. Additionally, there exists another type of relationship between an anaphor and its antecedent beyond identity relations. Bridging anaphora (Clark, 1975), also known as associative anaphora (Hawkins, 2015), requires the hearer or reader to establish an indirect connection between the anaphor and its antecedent,

drawing on their world knowledge. In sentence (2), the cover functions as the anaphor and *a book* as its antecedent. The reader infers the relationship because it is commonly understood that covers are parts of books.

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

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

- referentiality
- strict coreference.

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

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

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

5.2 Markables

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

(3) **Bir kitap** okuyorum. (Indefinite NP)
a book read.PROG.1sg
'I am reading **a book**.'

(4) **Kitap** okuyorum. (Bare noun)
book read.PROG.1sg
'I am reading.'

(5) **Kitabı** okuyorum. (Definite NP)
book.Acc read.PROG.1sg
'I am reading **the book**.'

(6) Adamın okuduğu kitap (modified NPs)
man.Gen read.Rel.Agr book
'The book that the man read'

(7) Adamın okuduğu (Headless relative)
man.Gen read.Rel.Agr
'The one that the man read'

(8) Masa-da-ki (kitap) (Pron. locative)
table-Loc-Rel book
'The book/one on the table.'

(9) Proper names, pronouns and demonstratives and demonstrative NPs.

Null nominals:

(10) Kitap geldi. **ø** Eskiydi. (Subject drop)
book.Nom came **it** old.Past.3sg
'The book came. **It** was old.'

(11) Elma vardı. Ali **ø** yedi. (Object drop)
apple.Nom existed Ali **it** eat.Past.3sg
'There was an apple. Ali ate **it**.'

(12) ø ev-**i** güzel. (Possessor drop)
his/her house-Poss beautiful.Cop
'**His/her** house is beautiful.'

(13) Ben ø **okurken** uyudum. (Converbs)
I I/she read.Conv slept
'I fell asleep, while I/she was reading.'

5.3 Non-markables

We did not annotate the following categories:

Singletons: We left out mentions that occur only once throughout the text and therefore do not take part in a coreferential chain. Zhu et al. (2023) showed that incorporating singleton information along with entity type and information status could help coreference models generalize better. We are planning to enrich our dataset with singletons in the future.

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

(14) Ali **öğretmen** oldu.
Ali **teacher** become.Past.3sg

'Ali became **teacher**.'

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

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

5.4 Various issues in coreference annotation

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

5.4.1 Embedded mentions

One issue that complicated the annotation process was the annotation of embedded mentions. As a principle stated above, we annotate the whole noun phrase but sometimes the phrase can consist of other mentions. For instance in a form like in (15), the markable *the book* is embedded in *the man who brough the book*. We annotated the embedded mention along with the larger one, in cases where there is a reference back to the embedded markable in the text.

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

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

5.4.2 Appositives

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

5.4.3 Genitive-possessive constructions

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

5.4.4 Grammatical coreference

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

5.4.5 Split anaphora

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

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

5.4.6 Null elements

Turkish is a pro drop language, where zero pronouns are abundant in both spoken and written text, and get involved in coreference chains (see Section 6. In cases where a null markable has an overt morphosyntactic agreement like a verbal inflection or a possessive suffix, we annotated the corresponding suffix in lieu of the markable itself. However, when it comes to null objects there is no overt agreement correlate. It is still a matter of discussion how to treat such cases in annotation. For instance, Pradhan et al. (2012) inserted a small *pro* into the place the null element is omitted, but the detection of the correct place is also problematic on its own. We employed a convention of

marking the space character just before the head predicate to represent a dropped object. The information concerning both types of null anaphora is recovered during post-processing, abstracting away from the conventions we employed in the annotation process.

5.5 Annotation procedure

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

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

6 Analysis

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

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

6.1 Inter-Annotator Agreement

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

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

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (1)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (2)$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Our annotators achieved high precision, recall and F1 scores (0.96) for mentions and (0.90) for relations on average, which is quite satisfactory for corefererence annotation task (See Table 2 for interannotator agreement scores).

	Precision	**Recall**	**F1**
Mentions	0.96	0.96	096
Relations	0.90	090	0.90

Table 2: Interannotator agreement scores

6.2 Corpus statistics

We annotated 60 situated dialogues of participants solving a puzzle. Our dialogues have an average of 306 tokens. The dialogue with minimum number of words has 127 words and the one with maximum number of words has 961 (See Table 3 for average number of words in our dialogues).

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

Table 3: Counts of the corpus. Statistics are per dialog.

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

We also looked at the grammatical functions of the mentions. We found out that 50.5% of the mentions occupy a subject position in a sentence. 49.5% of the mentions occupy an object position or part of a genitive possessive construction. (See Table 4 for the percentage of grammatical roles of mentions)

%subject	%non-subject
50.5	49.5

Table 4: Percentage of the grammatical roles of mentions

We also analyzed the form of our referring expressions. We observed a relatively close distribution of null and overt form in our mentions. The percentage of mentions which have overt forms is 57.3% while the percentage of null forms is 42.7% (See Table 5 for the percentage of forms of the mentions).

%overt	%null
57.7	42.7

Table 5: Percentage of linguistic forms of mentions

We aligned the grammatical function of the referring expressions along with their forms to see if the grammatical function has a relation with the form. When we looked at the mentions which occupy the subject position, we observed that 61.91% of the expressions have null forms. However, when we looked at the non-subject positions including objects and all other positions, our analysis showed that only 23.34% of referring expressions have null forms (See Table 6 for null forms in subject and non-subject positions). Consequently, our data indicated that there can be a strong relationship between subjecthood and linguistic form of the mentions.

	subj	nonsubj
null	62.1	23.3
overt	37.9	76,7

Table 6: Distribution of linguistic forms according to function

7 Conclusion and future work

We introduced a new publicly available[1] corpus of situated dialogs manually annotated for mentions and coreference relations. Our work has made novel contributions in a number of ways. Our dataset comprises 60 conversational texts. To our knowledge, it has been the first dialog corpus, which has been annotated for mentions and coreference relations in Turkish. Another significant contribution is that it includes null elements, agreement markers, and possessive suffixes as realizations of entities in text in addition to overt noun phrases and pronouns.

We also proposed an annotation scheme about how to annotate coreferential phenomena including both overt and null mentions in a morphologically rich and pro drop Turkic language. The high inter-annotator agreement shows that our scheme can be reliably applied to languages similar to Turkish in the relevant respects.

We believe that our corpus and scheme can serve as a resource for researchers working in different fields such as linguists, computational linguists, and cognitive scientists. The scheme can be a model for researchers who want do develop an annotation scheme and create a coreference corpus in other Turkic languages and similar low resourced languages.

The corpus can be improved in various ways. The most critical is the accumulation of more annotations. Another direction for improvement would

[1]Please contact the corresponding author to obtain the corpus for research purposes.

be to enrich the corpus with further grammatical information.

Acknowledgements

This work was supported by the Middle East Technical University Scientific Research Projects Coordination Unit under the grant number GAP-704-2023-11066. The first author acknowledges the financial support by The Scientific and the Technological Research Council of Türkiye (TÜBİTAK) under the 2214-A International Research Fellowship Programme for a research stay at the University of Cologne. We would like to thank our annotators Derin Dinçer, Batuhan Karataş and Anıl Öğdül for their meticulous work during scheme development and corpus creation. We are grateful to Klaus von Heusinger and his team at the Collaborative Research Centre (CRC 1252) at the University of Cologne, İsmail Sengör Altıngövde, Murat Perit Çakır, and Asiye Tuba Özge for their valuable feedback during annotation scheme development. We are also thankful to the anonymous reviewers for their helpful comments and suggestions to improve our work.

References

Tuğba Pamay Arslan, Kutay Acar, and Gülşen Eryiğit. 2023. Neural end-to-end coreference resolution using morphological information. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 34–40.

Tuğba Pamay Arslan and Gülşen Eryiğit. 2023. Incorporating dropped pronouns into coreference resolution: the case for turkish. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 14–25.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Springer, Dordrecht, Holland.

Peter Bourgonje and Manfred Stede. 2020. The potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1061–1066.

Thorsten Brants. 2000. Inter-annotator agreement for a german newspaper corpus. In *LREC*. Citeseer.

Ayışığı B Sevdik Çallı. 2012. Demonstrative anaphora in turkish: A corpus based analysis. In *First workshop on language resources and technologies for turkic languages*, page 33. Citeseer.

Herbert H Clark. 1975. Bridging. In *Theoretical issues in natural language processing*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Kees van Deemter and Rodger Kibble. 2000. On coreferring: Coreference in muc and related annotation schemes. *Computational linguistics*, 26(4):629–637.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, Imre Solti, et al. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144. American Medical Informatics Association.

Şeniz Demir. 2023. Neural coreference resolution for turkish. *Journal of Intelligent Systems: Theory and Applications*, 6(1):85–95.

Robert MW Dixon. 2003. Demonstratives: A cross-linguistic typology. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 27(1):61–112.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Michael Götze, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, and Ruben Stoel. 2007. Information structure. *Interdisciplinary studies on information structure: ISIS*, (7):147–187.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*, 1st edition. Routledge.

John Hawkins. 2015. *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. Routledge.

Iris Hendrickx, Gosse Bouma, Frederik Coppens, Walter Daelemans, Veronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri Van Der Vloet, and Jean-Luc Verschelde. 2008. A coreference corpus and resolution system for dutch. In *LREC*.

Lynette Hirschman, Patricia Robinson, John Burger, and Marc Vilain. 1997. Automating coreference: The role of annotated training data. In *Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 118–121.

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

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Olga Krasavina and Christian Chiarcos. 2007. Pocos–potsdam coreference scheme.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30:61–70.

Labelbox. 2024. Labelbox. https://labelbox.com. Online; accessed 2024.

Ekaterina Lapshinova-Koltunski and Pedro Augusto Ferreira. 2022. *ParCorFull2. 0: A parallel corpus annotated with full coreference*. Saarländische Universitäts-und Landesbibliothek.

Özge Mançe-Çalışır. 2018. *Geniş Otizm Fenotipi Gösteren Erişkinlerde Sosyal Biliş: Bir Göz İzleme Çalışması [Social Cognition in Adults with Broad Autism Phenotype: An Eye-Tracking Study]*. Ph.D. thesis, Ankara University, Ankara.

Ruslan Mitkov. 2014. *Anaphora resolution*. Routledge.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovskỳ. 2016. Coreference in prague czech-english dependency treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 169–176.

Malvina Nissim, Shipra Dingare, Jean Carletta, Mark Steedman, et al. 2004. An annotation scheme for information status in dialogue. In *LREC*. Citeseer.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2016. Polish coreference corpus. In *Human Language Technology. Challenges for Computer Science and Linguistics: 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7-9, 2013. Revised Selected Papers 6*, pages 215–226. Springer.

Balkiz Ozturk. 2004. *Case, referentiality and phrase structure*. Harvard University.

Tuğba Pamay and Gülşen Eryiğit. 2018. Turkish coreference resolution. In *2018 Innovations in Intelligent Systems and Applications (INISTA)*, pages 1–7. IEEE.

Massimo Poesio. 2004. The mate/gnome proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 154–162.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Arndt Riester and Stefan Baumann. 2017. The reflex scheme-annotation guidelines.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Peter Schüller, Kübra Cıngıllı, Ferit Tunçer, Barış Gün Sürmeli, Ayşegül Pekel, Ayşe Hande Karatay, and Hacer Ezgi Karakaş. 2017. Marmara turkish coreference corpus and coreference resolution baseline. *arXiv preprint arXiv:1706.01863*.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*, volume 2008, pages 96–101.

Pınar Tüfekçi and Yılmaz Kılıçaslan. 2007. A computational model for resolving pronominal anaphora in turkish using hobbs-naïve algorithm. *International Journal of Computer and Information Engineering*, 1(5):1402–1405.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Natural Language Engineering*, 26(1):95–128.

Veronika Vincze, Klára Hegedűs, Alex Sliz-Nagy, and Richárd Farkas. 2018. Szegedkoref: A hungarian coreference corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Bonnie Webber. 1988. Discourse deixis: Reference to discourse segments. In *26th Annual Meeting of the Association for Computational Linguistics*, pages 113–122.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23:170.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, 17.

Savaş Yıldırım and Yılmaz Kılıçaslan. 2006. A machine learning approach to personal pronoun resolution in turkish. *Computational Linguistics*, 27(4):521–544.

Savaş Yıldırım, Yılmaz Kılıçaslan, and R Erman Aykaç. 2004. A computational model for anaphora resolution in turkish via centering theory: an initial approach. In *International Conference on Computational Intelligence*, pages 124–128.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2021. Stay together: A system for single and split-antecedent anaphora resolution. *arXiv preprint arXiv:2104.05320*.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Deniz Zeyrek, Işın Demirşahin, and Ayışığı B Sevdik Çallı. 2013. Turkish discourse bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue & Discourse*, 4(2):174–184.

Deniz Zeyrek, Isin Demirsahin, Ayisigi B Sevdik-Calli, Hale Ögel Balaban, İhsan Yalçinkaya, and Umit Deniz Turan. 2010. The annotation scheme of the turkish discourse bank and an evaluation of inconsistent annotations. In *Proceedings of the fourth linguistic annotation workshop*, pages 282–289.

Yilun Zhu, Siyao Peng, Sameer Pradhan, and Amir Zeldes. 2023. Incorporating singletons and mention-based features in coreference resolution via multi-task learning for better generalization. *arXiv preprint arXiv:2309.11582*.