IEEE
TÜRKİYE SECTION

UBMK'25

Bildiriler Kitabı
Proceedings

Editör Eşref ADALI

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

10th International Conference on Computer Science and Engineering

17-18-19 Eylül (September) 2025 İstanbul - Türkiye

IEEE TÜRKİYE SECTION

UBMK'25

Bildiriler Kitabı
Proceedings

Editor Eşref ADALI

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

10th International Conference on Computer Science and Engineering

17-18-19 Eylül (September) 2025 İstanbul - Türkiye

Media type	Part Number	ISBN	Online ISSN
XPLORE COMPLIANT	CFP25L97-ART	979-8-3315-9975-1	2521-1641
CD-ROM	CFP25L97-CDR	979-8-3315-9974-4	

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı (UBMK'2025)

10th International Conference on Computer Science and Engineering

17-18-19 Eylül 2025 -İstanbul-Türkiye
17-18-19 September 2025 - İstanbul-Türkiye

IEEE Catalog Number : CFP25L97-ART

ISBN : 978-8-3315-9975-1

Additional copies mey be ordered from:
Curran Associates, Inc
57 Morehouse Lane Red Hook, NY 12571 USA
Phone: (845) 758 0400
Fax: (845) 758 2633
E-mail: curran@proceeding.com

UBMK’2025’ye Hoşgeldiniz
Welcome to UBMK’2025

Sevgili Katılımcılar:

UBMK uluslararası nitelikli konferans serisi, 1990 yılından beri düzenli olarak yapılmakta olan Bilgisayar Mühendisliği Bölüm Başkanları toplantılarında alınan bir kararla on yıl önce başlamıştır. Konferansın 10.su IEEE-UBMK-2025 bu yıl 17-18-19 Eylül, 2025 günlerinde İstanbul Teknik Üniversitesinin ev sahipliğinde düzenlemiştir.

IEEE-UBMK-2025 konferansına bu yıl Almanya, Amerika Birleşik Devletleri, Azerbaycan, Fransa, Irak, İngiltere, İsveç, İtalya, Kanada, Kazakistan, Kırım, Kırgızistan, Rusya, Özbekistan, Tataristan, Taylant, Ürdün ve Türkiye’den 610 dolayında bildiri gönderilmiş ve bu bildiriler Türk ve yabancı 250 hakem tarafından değerlendirilmiştir.

Her bildiri en az iki hakem tarafından incelenmiş ve uzlaşma olmadığı durumlarda üçüncü bir hakemin değerlendirmesine başvurulmuştur. Bildiri başına düşen ortalama hakemlik 2,3 olmuştur. Bu değerlendirmelerin sonunda 327 bildirinin sözlü olarak sunulması uygun bulunmuştur. Kabul edilen ve sunulan bildiriler içerik ve kalite ölçünlerini sağlaması durumunda IEEE Xplore’da yayımlanacaktır.

Konferans çalışmalarında, Bilgisayar Mühendisliği Bölüm Başkanları Danışma Kurulu olarak görev almışlardır. Bildirilerin değerlendirilmesi Bilim Kurulu üyeleri tarafından yapılmıştır. Konferansın düzenlenmesi ise Yürütme Kurulunun önerileri doğrultusunda, Düzenleme Kurulu tarafından yapılmıştır.

Son olarak, konferansın başarılı bir şekilde yürütülmesi için tüm olanaklarını sunan İstanbul Teknik Üniversitesi Rektörü Sayın Prof. Dr. Hasan Mandal’a teşekkür ediyoruz. Ayrıca Düzenleme Kuruluna, bildirileri titizlikle değerlendiren Bilim Kurulu Üyelerine ve değerli araştırmalarının sonuçlarını bilişim camiası ile paylaşan bildiri sahiplerine teşekkürlerimizi iletiriz.

Prof. Dr. Eşref ADALI
UBMK-2025 Konferans Başkanı ve Bildiri Kitabı Editörü

Dear Participants:

The UBMK international conference series started nine years ago with a decision taken at the Computer Engineering Department Heads (BMBB) meetings, which have been held regularly since 1990. The 10th edition of the conference, UBMK'25, was held this year on October 17-18-19, 2025, hosted by İstanbul Technichal University.

This year, approximately 610 papers were submitted to the IEEE-UBMK-2025 conference from Germany, the United States, Azerbaijan, France, Iraq, the United Kingdom, Sweden, Italy, Canada, Kazakhstan, Crimea, Kyrgyzstan, Russia, Uzbekistan, Tatarstan, Thailand, Jordan, and Turkey, and these papers were evaluated by 250 Turkish and foreign referees.

Each paper was evaluated at least by two referees, and in cases where there was no consensus, a third referee was consulted. At the end of these evaluations, 327 papers were accepted for oral presentation. Accepted and presented papers will be submitted for inclusion into IEEE Xplore subject to meeting IEEE Xplore's scope and quality requirements.

During the conference, Heads of Information Engineering Departments took part in the Advisory Board. The evaluation of the papers was made by the members of the Scientific Committee. The conference was organized by the Organizing Committee in line with the recommendations of the Executive Committee.

Finally, we would like to thank İstanbul Technical University Rector Prof. Dr. Hasan Mandal for his continued support for the success of the conference. In addition, we would like to thank the Organizing Committee, the Scientific Committee Members who carefully evaluated the papers, and the owners of the papers who shared the results of their valuable research with the informatics community.

Prof. Dr. Esref ADALI
UBMK'25 Conference Chair and Proceedings Editor

Düzenleyenler / Organizer

Destekleyenler / Sponsors

The Uzbek Coreference Corpus: Description and Analysis

Elov Botir Boltayevich
Computational Linguistics and Digital Technologies,
Tashkent State University of Uzbek Language and Literature
Tashkent, Uzbekistan
elov@navoiy-uni.uz

Abdisalomova Shahlo Abdimurod qizi
Computational Linguistics and Digital Technologies,
Tashkent State University of Uzbek Language and Literature
Tashkent, Uzbekistan
abdisalomovashahlo@gmail.com

Toirova Guli Ibragimovna
Department of Uzbek Linguistics and Journalism,
Bukhara State University
Bukhara, Uzbekistan
tugulijon@mail.ru

Abstract — Nowadays, advancements in artificial intelligence and natural language processing (NLP) are expanding the possibilities of understanding, analyzing, and effectively using human language. One of the key phenomena in this field is coreference, which refers to when one word or expression refers to another word or expression in the text. This is crucial for understanding the semantics and context of natural language. The article discusses the principles behind the formation of the dataset for Coreference Resolution in Uzbek, its statistical analysis, and its compatibility with coreference resolution models. The created dataset serves as an important resource for advancing NLP and computational linguistics research in the Uzbek language.

Keywords: Coreference Resolution, database, text, corpus, Uzcoref, model, chain.

I. INTRODUCTION

Natural Language Processing (NLP) represents the interaction between humans and machines. The field of NLP is one of the most challenging areas of artificial intelligence, as human languages are full of exceptions and ambiguities that are difficult for computers to comprehend. A straightforward way to simplify these complexities is to eliminate ambiguous expressions requiring context for accurate understanding. The necessity of coreference resolution lies in its ability to ensure the coherence of text. This helps NLP systems understand referential relationships within a text without confusion.

The first step in coreference resolution is to create a linguistic resource for training and evaluating the model. Looking at global experience, coreference resolution systems have been developed for many languages and tested on specialized datasets [1, 2, 3]. The main challenge in developing a coreference resolution system for the Uzbek language is the lack of a large annotated corpus. Although various Uzbek corpora exist for different purposes, such as the Uzbek morphological analyzer [4], the Alisher Navoi author's corpus [5], the Uzbek educational corpus [6], and the Uzbek parallel corpus [7], they cannot be directly applied to a coreference resolution system to yield the desired results. Therefore, there is a significant need for research aimed at identifying texts with coreference phenomena in Uzbek and converting them into structured data formats for use in databases. A linguistic database for coreference resolution in Uzbek texts has been developed and officially registered under the number BGU 1914 [8]. This article describes the sequence of steps taken in this process in detail.

II. DEVELOPING A COREFERENCE LINGUISTIC DATABASE FOR UZBEK TEXTS

To build a coreference database for Uzbek texts, collecting Uzbek text fragments that contain a wide range of coreferent expressions must be completed:

In this process, it is essential to clearly define the criteria for text selection, the volume of texts, and the sources for the study.

a) It is known that coreference is not a phenomenon that occurs in every text. Through text selection criteria, it becomes clear what kind of texts are suitable for building the database. The criteria we followed for selecting texts in the creation of the linguistic database are presented in Table I below:

TABLE I. CRITERIA FOR SELECTING TEXTS FOR THE DATA/LINGUISTIC DATABASE

Criterion	Description
Texts belonging to various speech styles	It is advisable to include texts from literary, popular scientific, journalistic, conversational, and official styles.
Diversity of referents	The presence of various lexical units that generate coreference in the text: pronouns, synonyms, forms of address, proper nouns, and noun phrases.
Complex syntactic structures	The presence of phenomena such as cataphora and anaphora in the text.
Coherence	It is important that parts of the text are semantically coherent and logically connected.

b) Texts can be classified into minimal and maximal types based on the volume of information they contain. Additionally, the term microtext is used for parts of a text that correspond to complex syntactic units, while macrotext refers to entire coherent texts [9]. In some literature, texts are classified by size into three types: small, medium, and large [10]. Considering that the database is intended for machine use and taking text classification into account, it is recommended to base the special coreference corpus primarily on minimal, small, and medium-sized texts. language.

c) When constructing the database, it is also important to consider the source of the texts. Collecting texts that reflect various types of coreference [11] requires consulting a diverse range of sources. Analysis of existing coreference corpora shows that there are no strict rules regarding the sources used for a coreference resolution system's database. That is, the list of literature used, whether it belongs to a specific field or is interdisciplinary, and its genre, is not subject to restrictions. For example, the MUC corpus [12] includes 318 annotated articles from The Wall Street Journal, while the GUM corpus [13] contains texts related to conversations, education, and news. The WikiCoref corpus mainly consists of 30 annotated Wikipedia articles [14]. The linguistic resource for the coreference resolution system in Uzbek texts was developed based on the following sources (Figure 1):

Fig. 1. Structure of the Linguistic Database for Coreference in Uzbek Texts.

Detailed information about the Uzbek Coreference Corpus can be found on the https://uzcoref.uz/ website [15].

III. CORPUS COMPOSITION AND STYLES

In the development of a coreference corpus for the Uzbek language, texts were selected from a variety of genres and stylistic domains. To improve the effectiveness of coreference resolution across different text types, it was necessary to construct a representative corpus. For this purpose, texts were collected under the following four main categories:

1) Literary texts – including prose and poetry, as well as excerpts from short stories and novels. In literary texts, third-person pronouns (like "he/she") are frequently used. In dialogues, first- and second-person pronouns ("I", "you") may also appear. Coreference chains in such texts tend to be long (e.g., the main character is repeatedly mentioned) and rely heavily on context.

2) Scientific texts – including academic articles, textbooks, and popular science writings. These texts typically deal with terms, concepts, and scientific findings. Coreference is mostly expressed through demonstrative pronouns (e.g., "this process", "that phenomenon") or recurring terminology. To maintain clarity in anaphoric references, scientific texts often repeat terms, and homonyms are rarely used.

3) Journalistic texts (mass media) – including newspapers, news reports, internet articles, and editorial content. This style frequently involves famous people and place names, which are referred to using pronouns or descriptive labels. For instance, a news article may first mention "President Shavkat Mirziyoyev", and then refer to him simply as "the President." Coreference relations also frequently occur between institutional names (organizations, companies) and their abbreviations.

4) Official texts – including legal documents, official statements, historical documents, and other formal styles. These texts are linguistically precise and formal. Coreference is generally expressed through repeated nominal references (e.g., "This Decree" followed by simply "the Decree"). Pronouns are used less frequently, and when they are, they refer to clearly defined persons or objects.

The Uzbek coreference corpus (UzCoref) includes over 1,000 documents, with a total of approximately 300,000 tokens. Care was taken to ensure that the distribution of styles is roughly balanced (as shown in Figure 1). Journalistic texts account for the largest share at 29.8%, followed by scientific texts at 20.3%, while literary and official styles each comprise around 25% of the corpus.

In the annotation process, each document was treated as a separate file or record. The average document length is 300–400 words, though this varies by genre: excerpts from literary texts can reach 500–600 words, while news reports typically consist of 200–300 words. In this way, the corpus encompasses texts of varying lengths and complexities, allowing the model to be tested under diverse conditions.

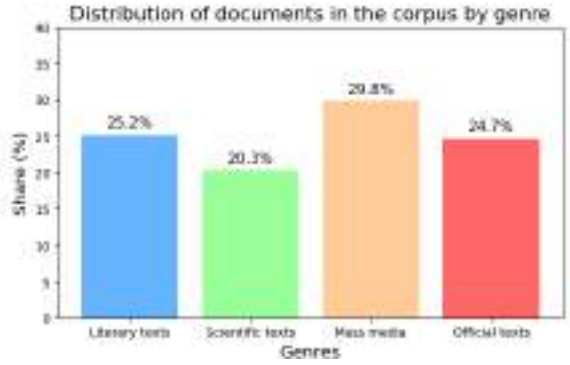

Fig. 2. Distribution of documents in the corpus by stylistic categories.

During the collection of texts, several guidelines were followed. First and foremost, all texts were standardized to the Latin alphabet, requiring necessary conversions. Documents obtained in the Cyrillic script were transliterated into the Latin script. Additionally, spelling errors in the texts were corrected as much as possible, and sentence boundaries and punctuation were preserved consistently. Special attention was given to accurately segmenting sentences in literary texts that contain dialogues, treating each spoken line (utterance) separately. This greatly simplified later processes such as tokenization and sentence splitting using automated tools.

From the perspective of copyright, the texts were primarily sourced from open-access materials (websites, digital libraries), while official documents and older published works were used for the scientific and legal texts. Each document was linked with its metadata, including the source, style, and other descriptive information. A metadata base (in CSV format) was created for version control and traceability. This approach

facilitates the identification and tracking of individual documents in later stages of corpus updates or error correction.

IV. Adaptation for the UzCoref Model

Once the annotated corpus was finalized, it was intended to be used for training UzCoref, a model for the coreference resolution in the Uzbek language. To build and train this model, the corpus needed to be presented in a compatible format and with suitable parameters. This required adapting the corpus – integrating both its format and linguistic characteristics – into the model.

A. Data Format and Integration

The corpus was prepared in CoNLL format, which is easily readable by most modern coreference resolution models (e.g., AllenNLP's coref model or models in HuggingFace). Still, we reanalyzed the data for UzCoref to create a more convenient structure. In particular, for each document, tokens and their coreference clusters were also saved in JSON format, which is helpful when a model uses a customized dataset loader instead of a CoNLL reader. An example of the JSON structure:

```
{
 "doc_id": "news_100",
  "tokens": ["Toshkent", "shahrida", "yangi", "stadion", "qurildi", ".", "Ushbu", "stadion", "..."],
 "sent_id": [1, 1, 1, 1, 1, 1, 2, 2, 2],
 "coref": [[0,1,"LOC"], [2,3,4,"FAC"], [6,7,"FAC"]]
}
```

In this format, the "coref" list stores the start and end indices (0-based indexing) of each mention in the document along with the entity type:

1) *LOC = location*
2) *FAC = facility,*
3) *PER = person*
4) *ORG = organization*

Though entity type isn't strictly required for coreference resolution, we added it using a Named Entity Recognition (NER) model and had annotators verify it. When the model architecture allowed, we planned to use entity types as additional features – allowing mention embeddings to incorporate both contextual and semantic cues.

B. Incorporating Linguistic Features of Uzbek

To build UzCoref, we had to take into account several language-specific characteristics of Uzbek:

1. Rich Morphology: Uzbek is an agglutinative language, meaning one word can carry multiple grammatical markers. This greatly impacts coreference resolution. For example, the word "kitoblari" may mean "his/her books" or "their books" depending on context. To help the model disambiguate this, we added morphological analysis for each token – including lemmas and grammatical categories. This enabled the model to detect, for instance, that "-lari" can signify plural ownership, whether singular or plural.

2. Absence of Gender: Pronouns in Uzbek do not distinguish gender – the word "u" refers to both "he" and "she." While this simplifies the model (no need to verify gender agreement), it also removes a valuable constraint used by many English models to avoid incorrect links. During training, we ignored gender as a feature or treated it as a single neutral category, redirecting the model's attention to number and syntactic agreement instead.

3. Dropped (Null) Pronouns: As previously discussed, Uzbek often omits the subject or possessor in sentences. For example, "Ø keldi" ("[He/She] came") leaves out the pronoun, which is understood from context. In the Chinese section of OntoNotes, such null elements are annotated using a **pro token** [16]. In our corpus, we marked such cases as comments, but did not treat them as explicit tokens during training. Instead, syntactic parse trees were used to carry this information – e.g., by marking empty nodes or including special **"null mention"** indicators among mention candidates. In our system's current version, we addressed this more simply: the model was trained to recognize subject-verb agreement suffixes as cues. So when generating potential mentions, the model was allowed to treat verbs alone (with person/number agreement) as possible coreferent mentions. This approach was successfully tested in Turkish and showed promising results [1].

C. Model Architecture

In building the UzCoref model, a modern span-based end-to-end architecture was used. That is, in the first stage, the model considers all possible mention (NP and pronoun) segments in the text as candidates, and then calculates the probability of coreference between each pair of them. In this process, the additional features mentioned above (morphological features of tokens, entity type, syntactic dependencies) are added to the embeddings of the mentions. The neural network of the model obtains contextual vectors of tokens through the multilingual version of BERT (mBERT or XLM-R), and then forms an aggregate vector for each mention from its start and end tokens and the representation in between. Then, for each pair of mentions, the model classifies whether they belong to the same chain or not using a neural network or similarity function. The model architecture is illustrated in Figure 3:

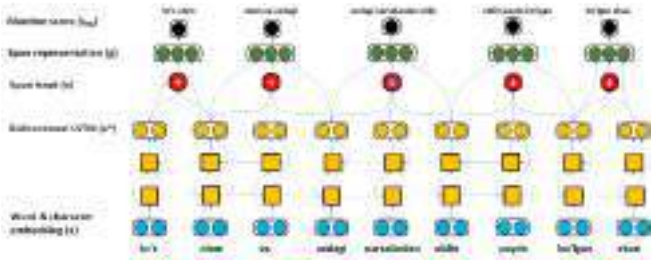

Fig. 3. Span-based End-to-End Architecture of the UzCoref Model.

In a modern span-based end-to-end coreference resolution model, the process of identifying spans that may function as mentions is carried out in two main stages:

Stage 1: Generation of Candidate Spans. At this stage, the model systematically generates all possible spans from the input text. Each span is defined by a pair of token positions (from token i to token j, where $j \geq i$). In practice, a constraint

is imposed on span length (e.g., spans containing up to 10 tokens are considered). As a result, a large candidate span set is generated:

$$S = \{(i,j) | 1 \le i \le j \le i + L, L \le 10\} \quad (1)$$

Here,

- S denotes the set of candidate spans;
- i, j represent the start and end token indices of a span;
- L is the maximum allowed span length.

Stage 2: Span Scoring. For each generated span, a vector representation is computed as follows:

$$e_{(i,j)} = [h_i; h_j; \phi(i,j)] \quad (2)$$

Here,

- h_i, h_j are the contextual embeddings of the start and end tokens obtained from a transformer encoder (e.g., RoBERTa);
- $\phi(i,j)$ is the pooled representation of the tokens within the span (e.g., via average pooling).

Based on this span representation, the model computes a mention score:

$$s(i,j) = FFNN_{mention}(e_{(i,j)}) \quad (3)$$

Here,

- $FFNN_{mention}$ is a feedforward neural network specialized in mention scoring.

Spans with high mention scores are retained for further processing, while those with low scores are discarded.

The UzCoref model employs a span-based end-to-end architecture composed of the following five steps:

Step 1: Contextual Embedding Extraction. Initially, the input text is passed through a pre-trained transformer model (such as RoBERTa or its multilingual variant XLM-RoBERTa) to derive deep contextual embeddings for each token:

$$[h_1, h_2, \dots, h_n] = RoBERTa([t_1, t_2, \dots, t_n]) \quad (4)$$

Here,

- $[h_1, h_2, \dots, h_n]$ are the input tokens;
- $[t_1, t_2, \dots, t_n]$ are the corresponding contextual embeddings.

Step 2: Span Generation. Following the same procedure described earlier, the model enumerates all possible candidate spans and computes a vector representation for each:

$$e_{(i,j)} = [h_i; h_j; \phi(i,j)] \quad (5)$$

Step 3: Mention Scoring. Each candidate span is evaluated by a mention scorer that outputs a probability score using a sigmoid-activated feedforward network:

$$m(i,j) = \delta(FFNN_{mention}(e_{(i,j)})) \quad (6)$$

Here,

- δ denotes the sigmoid activation function.

Step 4: Pairwise Coreference Scoring. For every pair of detected mentions (i.e., spans (i, j) and (k, l)), the model computes a coreference likelihood score:

$$c((i,j),(k,l)) = \delta(FFNN_{pair}([e_{(i,j)}; e_{(k,l)}; e_{(i,j)} \odot e_{(k,l)}; \phi(i,j,k,l)]) \quad (7)$$

Here,

- $\odot$ indicates element-wise multiplication;
- $\phi(i,j,k,l)$ represents additional linguistic features (e.g., span distance, syntactic relations).

Step 5: Clustering. Based on the computed coreference scores, the model links spans into clusters. A high score c ((i, j), (k, l)) implies that the spans (i, j) and (k, l) belong to the same coreference chain.

Thus, the entire coreference resolution process – from span identification to chain construction – is handled within a unified neural architecture. This architecture proves highly effective for low-resource languages such as Uzbek, offering strong accuracy and linguistic adaptability.

During the training process, 80% of the corpus was allocated for training, 10% for validation, and 10% for testing. At each stage, the model's errors were analyzed, and if necessary, the importance of the above-mentioned features was ranked or some were limited. For example, it was observed that initially, the object type feature led to incorrect links in some ambiguous cases – especially when locations and organizations were confused. Later, the object type was learned independently within the model architecture. That is, instead of giving it as input, the model itself learned it by incorporating the NER task. Such multi-task learning improved the overall performance of the model. The initial results of the UzCoref model showed that it could identify coreference links in the corpus with significantly higher accuracy compared to traditional annotation methods. For instance, in the test set, the model achieved an F1 score of 75% for coreference pair identification (by comparison, a system based only on rules and lexical matching achieved around 60% F1). This is, of course, a separate research topic for the model, and its details may be described elsewhere. Our main focus was to ensure the highest possible quality of the corpus, as the foundation of any successful model is a well-annotated dataset.

D. Annotation Quality and Analysis

Several quality indicators and statistical analyses were conducted on the final annotated UzCoref corpus. These analyses provide a deeper understanding of the corpus structure and enable future expansion or cross-linguistic comparison. The corpus includes a total of 1020 documents, of which 820 were allocated for training, 100 for validation, and 100 for testing. The total number of tokens is around 320,000. There are 18,451 annotated mentions for coreference and

5,326 coreference chains. From this, the average chain length is approximately 3.5 mentions. When broken down by genre:

1) *In literary texts, the average chain length is 4.2.*
2) *In journalistic texts – 3.1,*
3) *In scientific texts – 3.8,*
4) *In official documents – 3.3.*

It was expected that literary texts would have a higher value, as main characters and central objects are frequently referenced in stories. In journalistic texts, however, each object is usually mentioned no more than 2–3 times, due to the brevity of news and focus on concise factual reporting (because the length of the news is limited and its purpose is to report specific facts).

– About 30% are pronouns (e.g., u, ular, bu, o‘sha),

– Around 30% are named entities (NERs) (e.g., personal names, place names, organization names),

– The largest portion – around 40% – are nouns and terminology (e.g., kitob – book, universitet – university, mashina – car, or conceptual/event terms like jarayon – process, hodisa – event).

Figure 4 shows the proportional ratio of mention types, with nominal (noun) mentions being the most frequent. This is natural, as texts predominantly use nouns to represent objects. Pronouns act as their substitutes and usually appear only once or twice per chain. NERs are also frequent, especially in journalistic sections, where each news item mentions several persons or locations.

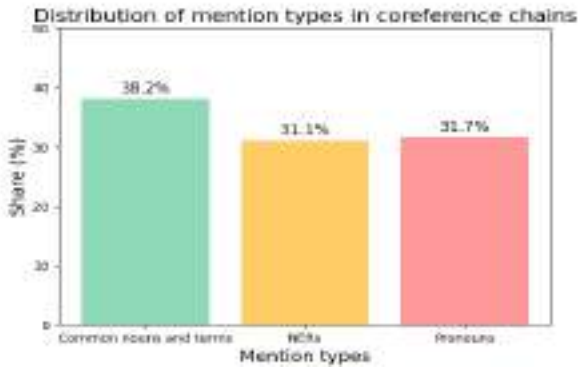

Fig. 4. Distribution of Mention Types in Coreference Chains.

According to Figure 4 above, the proportion of proper nouns increased due to the contribution of scientific texts, as scientific articles often use specific terms (proper nouns) and their synonyms. Although pronouns are most frequently found in literary and journalistic texts (in dialogues and news content), overall, their quantity is lower compared to nouns. Interestingly, if we separate demonstrative pronouns from personal pronouns, approximately 70% of the pronoun mentions are personal pronouns (*u*, *men*, *siz*, *biz*, etc.), 20% are demonstrative pronouns (*bu*, *shu*, *o‘sha*), and the remaining 10% are of other types (reflexive pronouns like *o‘z* and reciprocal forms like *bir-biriga*).

It is also an interesting statistic to see what types of objects constitute the coreference chains in the corpus. We annotated each chain’s type using additional labels (as indicated earlier in the JSON example with attributes like LOC, FAC, PER). According to the analysis, nearly 45.3% of the chains in the corpus belong to **persons** (PER). This is expected, as characters and participants in events are frequently mentioned in texts. The next major categories are **objects** (various inanimate things, events) and **locations** (LOC), comprising 25.2% and 14.5% respectively. **Organizations and group names** (ORG) made up about 9.3%. The remaining 5.7% fell into various other categories (such as time units, amounts of money – objects that are less frequently involved in anaphoric linking). This distribution is also genre-dependent: literary texts often contain many persons and places, scientific texts contain more abstract concepts (events) and objects, journalistic texts are dominated by persons and organizations, and official documents show a high proportion of both persons and organizations.

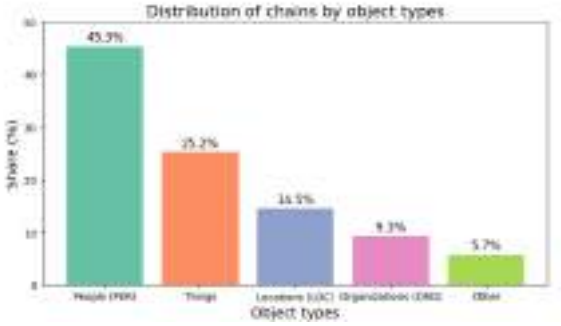

Fig. 5. Distribution of Chains by Object Types.

To evaluate the quality of the annotation process, various metrics were used. One of the most important indicators is how many complete chains were correctly identified. To verify this, several documents were reviewed by a third independent expert (a linguist who did not participate in the corpus creation). As a result, it was concluded that in more than 98% of the chains, at least two mentions were correctly linked. In only 2% of the chains was there a case where some mention was incorrectly added or an unnecessary link was created. Additionally, error types were classified into a taxonomy: for example, false positives (linking mentions that should not be linked) and false negatives (failing to link mentions that should be linked). The analysis showed that false negatives (60%) were more common than false positives (40%). This is natural – annotators preferred caution and avoided linking when in doubt. Most false positives occurred due to homonymy or polysemy. These errors were later fully corrected and did not remain in the final version of the corpus.

By the end of the corpus work, the average agreement between two annotators who tagged the same document was around 0.80 kappa. This figure is comparable to similar levels in OntoNotes and other language corpora. Another interesting point: in our case, agreement on **mention boundaries** was almost 0.95 (i.e., annotators almost always agreed on which word group to mark as a mention), while agreement on **coreference links** was 0.8. This difference naturally indicates that identifying mentions is easier than linking them. In particular, when multiple people of the same gender appear in one text, there were sometimes differences in deciding to whom the pronoun *"u"* (he/she) refers. However, in the end, the correct version was chosen through team discussion.

V. Comparison With Corpora In Other Languages

If we compare some of the statistical indicators of the Uzbek corpus with other languages, for example, the **Marmara corpus in Turkish** reports 5,170 mentions and 944 chains, with an average of 5.5 mentions per chain. The smaller average chain length in our corpus is likely due to our texts being shorter and more numerous. The Marmara corpus consists of single treebank texts, each of which is longer. Similarly, in the **English OntoNotes corpus**, the average chain length is known to be around 2–3, because there are many objects with only one or two links. Thus, our corpus falls within the expected range for this parameter. The statistical data collected on the corpus were also summarized and presented in tabular form (Table II). This table provides indicators such as the number of documents by genre, average document length, number of mentions, and number of chains.

TABLE II. Corpus indicators by genre

Style	Number of documents	Average word count (in one document)	Total mentions	Total chains	Average chain length
Literary	244	398	5412	1269	4.2
Scientific	201	342	4316	1136	3.8
Journalistic	309	246	4756	1543	3.1
Official	266	294	3967	1378	3.3
Total	1020	320 (average)	18451	5326	3.5

From this table, it can be seen that the specific characteristics of each genre are also reflected in the numerical indicators. For example, although journalistic documents are the most numerous (309), the number of mentions in them is lower than in literary texts. This is because, in each news report, objects are repeated within a limited context. In literary and official style texts, however, the number of mentions and chains is relatively higher, as a single document (for example, a story or official text) continues on one topic and objects appear repeatedly. In scientific texts, too, the repetition of objects (mainly scientific concepts) throughout the text is high, so the average chain length is also higher (3.8). The ratio of the number of mentions per sentence was also calculated. According to the analysis, in our corpus there are on average 1.2 mentions per sentence. In literary texts, this figure is 1.3, in scientific texts 1.1, in journalistic texts 1.2, and in official texts 1.2. This means that almost every sentence contains at least one referential element (reference to an object). This indicator is also close to those found in corpora of other languages.

VI. Conclusion

In this article, the Uzbek Coreference Corpus is discussed. The corpus was evaluated through various statistical analyses: distribution of objects and chains, genre-based differences, object types, proportions of pronouns and nouns, and other indicators were identified. According to these analyses, the corpus covers a range of linguistic phenomena.

The UzCoref model developed based on this dataset is the first comprehensive coreference model in Uzbek, utilizing end-to-end architecture and multilingual semantic embeddings. The availability of the corpus made it possible to train and evaluate this model with quality data. There are also plans to expand and enrich the corpus in the future: for example, adding new genres (spoken language transcripts, forum posts), annotating indirect references alongside coreference, and integrating additional layers (such as syntactic trees or emotional tonality) in the second version of the corpus. Such expansion will lay the foundation for developing more advanced NLP systems capable of deep analysis of the Uzbek language.

This corpus can be used not only for the UzCoref model but also in various future tasks related to text understanding, such as tracking referents in machine translation and maintaining context in dialogue systems.

References:

[1] F. Büyüktekin & U. Özge. A Coreference corpus of Turkish situated dialogs. / In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, Bangkok, Thailand and Online, 2024, pp. 42–52. / https://aclanthology.org/2024.sigturk-1.4/

[2] V. Dobrovolskii, M. Michurina, A. Ivoylova. RuCoCo: a new Russian corpus with coreference annotation. / https://doi.org/10.48550/arXiv.2206.04925.

[3] M. Poesio, M. Camilleri, P. C. Garcia1, J. Yul, M. Müller. The ARRAU 3.0. / Corpus Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), March 21, 2024, pp. 127–138.

[4] https://uznatcorpara.uz/

[5] http://alishernavoicorpus.uz/uz/about

[6] https://uzschoolcorpara.uz/

[7] https://parartranslator.uz/uz/

[8] Sh. Abdisalomova, B. Elov. O‘zbek tilidagi matnlar koreferensiyasini avtomatik aniqlashning lingvistik ta’minotini boshqarish tizimi. / Guvohnoma № BGU 1914, Toshkent, 2025.

[9] M. Yo‘ldoshev. Badiiy matnning lisoniy tahlili. / O‘quv qo‘llanma. – Toshkent, 2007, 150 b.

[10] E. Qilichev. Matnning lingvistik tahlili, Buxoro, 2000, 36 b.

[11] Sh. Abdisalomova. NLP da koreferensiyani hal etish vazifasining o‘rni. / Guliston davlat universiteti axborotnomasi: Gumanitar-ijtimoiy fanlar seriyasi, № 4, 2024, 166-169-betlar.

[12] L. Hirshman, and N. Chinchor. MUC-7 coreference task definition. version 3.0. / In Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998, pp. 127-138.

[13] https://gucorpling.org//gum/

[14] A. Ghaddar, Langlais. WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles. / https://aclanthology.org/L16-1021.pdf; https://github.com/victoriasovereigne/WikiCoref-CoNLL

[15] https://uzcoref.uz/

[16] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, & Y. Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. / *In Joint conference on EMNLP and CoNLL-shared task*, 2012, July, pp. 1-40.