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

Creating an Annotated Dataset for Coreference Resolution in Uzbek Texts Based on the CoNLL-2012 Format

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

Jumayeva Dilnoza Baxshulloyevna
Department of primary education, Navoi State University
Navoi, Uzbekistan
professional-5555@mail.ru

Abstract — This article presents a methodology for creating a specially annotated linguistic dataset aimed at implementing automatic coreference resolution for texts in the Uzbek language. The dataset covers over 1,000 texts from literary, scientific, journalistic, and official styles, and is annotated in the CoNLL-2012 format, which is a standard in international NLP research. The annotation process was carried out both automatically and through manual verification, and the stages of this process are described in detail.

Keywords: Coreference, NLP, annotation, CoNLL-2012, linguistic corpus, annotated dataset, annotators.

I. INTRODUCTION

In a text, the phenomenon of coreference refers to the process where multiple linguistic units (nouns or pronouns) refer to the same referent (a person, object, event, or concept). For example, in the sentence *"Alisher ertalab maktabga ketdi. U kech qaytdi" (*"Alisher went to school in the morning. He came back late") both "Alisher" and "u" ("he") refer to the same person – Alisher – and are thus considered coreferent expressions. Identifying coreference is a crucial step in semantic analysis and understanding of natural language texts. It plays a significant role in many NLP tasks, including information retrieval, question answering systems, and text summarization [1, 2].

When applying machine learning methods to coreference resolution, a large annotated dataset is required. To date, many coreference corpora have been created in languages such as English [3], Chinese, Arabic, German, Russian, and Turkish [4]. In particular, the OntoNotes 5.0 [5] corpus, developed for English, Chinese, and Arabic, served as the primary data source for coreference resolution in the CoNLL shared tasks held in 2011–2012 [6]. A key feature of the OntoNotes corpus is that it is not limited to a specific type of noun or object – it includes all anaphoric relations in the text (including events and various noun phrases). At present, there is no comprehensive annotated coreference corpus for the Uzbek language. This article addresses the issue of developing an annotated dataset for coreference resolution in Uzbek texts. The article sequentially discusses the annotation format and methodology, tools and protocols used in the process, inter-annotator agreement metrics.

A. CoNLL Format and the OntoNotes Style

The CoNLL traditional format was used as the annotation standard for the corpus. This format, developed for the coreference layer of the OntoNotes corpus as part of the CoNLL-2012 shared task, is widely considered the de facto global standard in the field.

In the CoNLL format, each document is represented as a separate structure:

– The document begins with a line such as *#begin document* (DOC_ID); which specifies the document identifier (DOC_ID).

– The text is written token-by-token, line by line. Each token (word or punctuation mark) appears on a separate line, along with several columns of annotations.

– These columns typically include: document ID, paragraph and sentence numbers, the word itself, part-of-speech (POS) tag, syntactic role in the parse tree, and the coreference label.

– Sentences are separated by an empty line. The document ends with a line: #end document, indicating the end of the file.

In our case, the coreference column is of primary interest. If a token is part of a coreference chain, it is marked with the chain identifier in parentheses. In the OntoNotes style, the first mention in a chain is marked with an opening bracket (**n**, the last mention with a closing bracket **n**), and if the mention appears only once, it is labeled (**n**). Here, **n** is a unique number assigned to each coreference chain within the document. For example, consider the sentence: *"Ali ashaddiy kitobxon. U har kuni yangi kitob o'qiydi"* (*"Ali is a passionate reader. He reads a new book every day").* In this case, "*Ali*" in the first sentence and *"U"("He")* in the second sentence refer to the same entity and belong to the same chain (*Ali = U* (*He)*). If this chain is assigned the number **5**, then the token "*Ali*" would be annotated in the coreference column with (**5**, and "U" ("He") with **5**).

According to the OntoNotes style, only true coreference relations should be annotated. This means only entities mentioned at least twice in the text form a chain. Singular (non-repeated) mentions – names or pronouns mentioned only once – are not annotated and are excluded from the coreference chains. For example, if a sentence says *"Yusuf kelmoqda" ("Yusuf is coming")* and Yusuf is not mentioned again later, then *"Yusuf"* is not given a chain ID. Alternatively, single mentions are not treated as separate chains. This approach ensures the corpus annotation focuses specifically on anaphoric relationships, and during model evaluation, only real coreferent pairs are taken into account – just like in the OntoNotes standard.

The CoNLL-style annotation described above serves as the standard output format of our corpus. One major advantage of this format is its compatibility with widely used CoNLL scorer tools, and its ease of integration with various external libraries such as AllenNLP or coreference readers in the Hugging Face ecosystem. As a result, the UzCoref model can be trained or evaluated directly using data in this CoNLL format.

The following Table I presents annotation examples from the Uzbek coreference corpus in the extended CoNLL-2012 format.

TABLE I. UZBEK COREFERENCE CORPUS IN THE EXTENDED CONLL-2012 FORMAT

Doc_ID	Part №	Sentence№	Token №	Token	POS	Parse	Lemma	NE	Coreference
texts_001	0	1	0	Navoiy	N	NP	Navoiy	PER	(1
texts_001	0	1	1	15-asrda	N	ADVP	15-asrda	DATE	-
texts_001	0	1	2	yashagan	VB	VP	yashamoq	-	-
texts_001	0	1	3	buyuk	JJ	ADJP	buyuk	-	-
texts_001	0	1	4	o‘zbek	N	NP	o‘zbek	NORP	-
texts_001	0	1	5	shoiri	N	NP	shoir	TITLE	1)
texts_001	0	1	6	.	.	O	.	-	-
texts_001	0	2	7	U	P	NP	u	-	1
texts_001	0	2	8	Hirotda	N	PP	Hirot	LOC	(2
texts_001	0	2	9	tug‘ilgan	VB	VP	tug‘ilmoq	-	-
texts_001	0	2	10	.	.	O	.	-	2)
science_015	0	1	0	Suv	NN	NP	suv	-	(3
science_015	0	1	1	0°C	CD	ADVP	0°C	-	-
science_015	0	1	2	da	IN	PP	da	-	-
science_015	0	1	3	muzga	NN	NP	muz	-	-
science_015	0	1	4	aylanadi	VBZ	VP	aylanmoq	-	3)
science_015	0	1	5	.	.	O	.	-	-
science_015	0	2	6	Bu	DT	NP	bu	-	(4
science_015	0	2	7	jarayon	NN	NP	jarayon	-	4)
science_015	0	2	8	juda	RB	ADVP	juda	-	-
science_015	0	2	9	muhimdir	JJ	ADJP	muhim	-	-
science_015	0	2	10	.	.	O	.	-	-

In this table:

- POS: Part of Speech (e.g., N – noun, P – pronoun, VB – verb, etc.)
- Parse: Simplified syntactic structure indicator
- Lemma: Base form of the word (lemmas)
- NE: Named Entity type (LOC – location, PER – person, ORG – organization, etc.)
- Coreference: Coreference chain identifiers. An opening bracket "(" indicates the beginning of a chain, and a closing bracket ")" marks its end. A singleton mention is annotated as (n).

The texts in this table are annotated in the standard CoNLL-2012 format, which is the same structure used in practice by coreference resolution models.

II. ANNOTATION METHODOLOGY: AUTOMATIC AND MANUAL TAGGING

A two-step approach was used to create the coreference annotations for the corpus:

First, automatic tagging (pre-annotation) was applied, and then the tagged data was manually reviewed and corrected by expert linguists. This approach helps save human labor, ensures consistent tagging, and speeds up the process. The steps are detailed below:

A. Automatic Pre-annotation

Initially, all collected texts were automatically tagged with coreference markers using specialized software. For this, we used existing models and multilingual tools trained on various languages. In particular:

– *Multilingual Model:* based on modern approaches, we used an open-source end-to-end coreference model originally trained on English and later adapted for multilingual use. The model is based on XLM-RoBERTa, a multilingual transformer language model, and implements the architecture proposed by Lee et al. [7]. Since this model may not produce satisfactory results directly on Uzbek texts, we adapted some of its parameters and used relatively low confidence *thresholds* during pre-annotation. The model thus only tagged highly confident coreferent pairs, while uncertain cases were left untagged.

– *Rule-based Methods:* additionally, simple heuristic rules were employed to assist with automatic tagging. For example, identically written proper nouns (e.g., multiple mentions of "Khorezm") were automatically linked into a single chain. Similarly, pronouns with clear morphological markers were matched to their antecedents using rules: pronouns like "u" (he/she), which do not reflect gender, were linked to the nearest compatible noun phrase based on sentence distance and boundaries. Demonstrative pronouns like "o‘sha" (that) and "bu" (this) were linked to nearby matching nouns. These rules were written based on the intuition and expertise of linguists and helped fill gaps where no trained automatic model existed.

– *Morphological Analysis Integration:* given the agglutinative nature of the Uzbek language, certain coreference relations were identified using morphological analysis [8]. For instance, references expressed via possessive suffixes (e.g., “kitobi” – “his/her book,” where the suffix -i implies third-person possession) were annotated as coreference mentions. Programmatically, this required splitting words into morphemes and implementing modules that could recognize referential components. For example, the automatic annotator interpreted “kitobi” as [“kitob” (book) + “i” (his/her)], treating the suffix as a hidden pronoun (“his”) and tagging it accordingly. This technique has been successfully applied in Turkish, such as in the Marmara Turkish Coreference Corpus, where dropped pronouns (zero pronouns) were reconstructed using affixes embedded in verbs or nouns [9]. In Uzbek, possessive and verb agreement suffixes also carry such referential information. During automatic tagging, these suffixes were treated as separate mentions and connected to the appropriate coreference chain. Our text corpus was annotated with coreference links using the UzCoref system, developed by B.Elov, Sh.Abdisalomova, R.Alayev [10]. At the end of the automatic phase described above, an initial set of coreference chains was generated for each document. Naturally, this phase may contain errors: some links might be incorrectly established, or some necessary links might have been omitted. Therefore, in the next phase, a team of qualified linguistic annotators thoroughly reviewed and edited the automatic annotations.

B. Manual Annotation and Review

In the second phase, the annotation team – consisting of several linguists experienced in Uzbek syntax and semantics – individually reviewed the automatically tagged documents. Each document was examined by at least two annotators: first, one annotator edited and completed the automatic tags, then a second annotator re-reviewed the document to correct any mistakes or oversights made by the first annotator. Afterward, final decisions in complex or ambiguous cases were made through collaborative discussion within the team. During the manual review process, the annotators followed specific guidelines and rules. These annotation instructions were based on standards used in OntoNotes and other international corpora, while also incorporating the unique features of the Uzbek language.

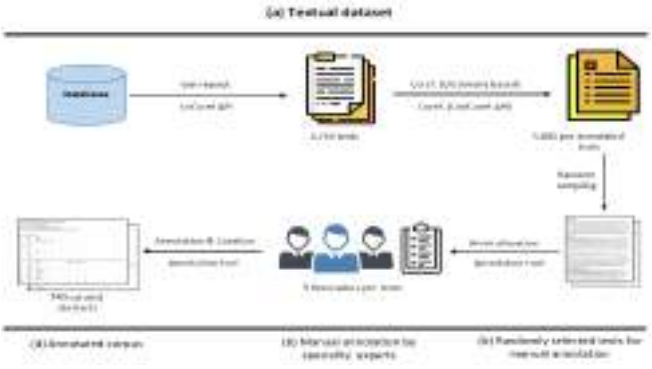

Fig. 1. Annotation Methodology: Automatic and Manual Tagging.

The main rules for randomly selecting coreference spans were defined as follows:

1. Eligible coreference candidates should be restricted to noun (N), noun phrases (NPs) and pronouns only [11]. References formed with verbs or adjectives are considered special cases (such as event coreference) but the primary focus is placed on nominal references.

2. The minimal unit to be annotated as a mention was defined as a complete noun phrase (NP). For example, in the phrase “katta uy” (“big house”), if it is later referred to simply as “uy” (“house”), the initial mention “katta uy” is tagged as a single complete mention and linked to the later occurrence. That is, all modifiers of the noun (adjectives, numbers, determiners) were included in the span whenever possible.

3. Conjoined noun phrases (e.g., “Ali and Vali”) were treated with special attention. If a pronoun like “ular” (“they”) later refers to them jointly, the entire conjunction was considered a single mention and linked to the pronoun. However, if only Ali or Vali is mentioned again individually, each is annotated as a separate mention and belongs to a separate chain.

4. Nested mentions: Sometimes, a reference inside another reference may occur. In such cases, each mention was annotated separately. For example, in “Alining mashinasi” (“Ali’s car”), both “Ali” and “Ali’s car” refer to distinct entities. If later in the text there are separate mentions of “he” (Ali) and “car”, then these must be marked as separate chains. In this example, “Alining mashinasi” contains two mentions: Ali (inner mention) and Ali’s car (outer mention). This rule allowed annotators to layer mentions based on their syntactic nesting.

5. Demonstrative pronouns (“bu”, “shu”, “o‘sha”) can refer to prior events or situations. In such cases, if the anaphoric link is clear, the demonstrative was linked to the phrase representing the preceding event or situation. For example, in the earlier example “Bu jarayon” (“This process”), “bu” points to the entire event in the previous sentence. Thus, “Bu jarayon” was marked as coreferent with the phrase describing “water turning into ice”. These discourse-level anaphora were also included as much as possible, since OntoNotes also allows for event coreference, not just noun phrases.

6. Synonymous names: If an entity is referred to using different synonyms (e.g., “AQSH” and “Qo‘shma Shtatlar” [both “USA”], or “shahzoda” and “valiahd” [both “prince”]), they were included in the same chain if the context clearly indicated they referred to the same person or entity. This followed the LUI (Linguistic Unique Identifier) principle, where annotators relied on world knowledge and context.

7. Dummy pronouns: In semantically empty constructions like “bu yog‘ingarchilik bo‘lmoqda” (“it is raining”), even though Uzbek does not have a dummy subject like English “it”, the language does use formalized elements like “shunday” (“such”) in expressions like “shunday bo‘ldiki…” (“it happened that…”). These do not refer to real entities and were excluded from coreference annotations.

The annotator team manually edited each document following these rules. They ensured that each chain included only valid mentions, and that unnecessary or incorrect links

were removed. Weekly meetings were held to resolve difficult or ambiguous cases, and refinements were made to the annotation guidelines as necessary. For example, early on, there was debate about whether to annotate possessive suffixes as hidden pronouns. Some annotators argued that phrases like “uning kitobi” (“his book”) implicitly contained the pronoun “u” (“he”) and should be marked accordingly, while others believed no additional annotation was necessary. To resolve this, the team consulted practices from other languages (e.g., Turkish and Russian) and adopted a compromise: words with possessive suffixes were not given special coreference tags by default. However, if no ambiguity existed within the object phrase, the referent implied by the possessive suffix was documented in a comment to aid understanding. At the end of the manual annotation phase, the corpus was fully annotated and brought to **gold standard** quality. This means that all coreference chains for each document were verified by human experts, making the dataset a reliable resource for future research and model training.

III. Annotation Tools And Protocols

To organize the coreference annotation process effectively, dedicated annotation tools and structured protocols were employed.

A. Annotation Tools

Since the selected CoNLL format is a text-based format, it is technically possible to annotate using simple text editors. However, in practice, this becomes inconvenient – especially when dealing with dozens of chains and hundreds of mentions in a single document. Therefore, a specialized annotation interface with a graphical user experience was used. Initially, the GATE Annotation Editor was tested, but it presented compatibility issues with direct CoNLL support and complex coreference structures. As a result, web-based tools like BRAT (Brat Rapid Annotation Tool) and WebAnno (INCEpTION) were explored.

After testing, the team selected WebAnno/INCEpTION, which operates through a web browser and supports real-time remote collaboration by multiple annotators. Its standout features include: Color-coded display of coreference chains, the ability to link mentions visually by simply highlighting words or phrases with a mouse, importing of pre-annotated documents, so annotators only had to correct errors and add missing links. WebAnno also logged every annotator’s actions to a log file, and supported version control. Any modifications to documents were versioned automatically, allowing for comparison between different versions. In case of unexpected issues, previous document states could be restored to trace and resolve errors.

B. Protocols and Team Collaboration

The annotation team consisted of five linguists. Each document was assigned to two annotators (a primary and a reviewer). To avoid fatigue and ensure diversity of perspective, annotators rotated across genres – no one worked exclusively on literary or scientific texts. This ensured: exposure to a range of linguistic challenges, and that each genre benefited from at least two expert perspectives. The team held weekly meetings to review the annotated documents, exchange feedback, and clarify ambiguities in the guidelines. This iterative process improved overall annotation quality.

In the early stages, disagreements were common (especially concerning possessive pronouns and non-referential forms), but consensus was gradually reached. As a result, inter-annotator agreement (IAA) steadily improved (Figure 1).

To measure IAA formally, several documents were annotated independently by all annotators. The annotations were compared using metrics like Cohen's kappa and F1-distance. Results showed:

- 90% agreement on mention boundaries;
- 76% agreement on coreference links.

These values meet accepted standards for reliable annotation. Notably, literary texts showed lower agreement (~0.72 kappa) due to abstract or ambiguous references, while journalistic texts had the highest agreement (over 0.85) thanks to clearer referents and more structured writing. As per the team protocol, annotators marked uncertain or ambiguous cases using comments within WebAnno. The reviewer paid special attention to these flags and left suggestions. If agreement still couldn’t be reached, the team lead (project’s academic supervisor) made the final decision – minimizing unresolved issues in the final corpus.

C. Versioning the Annotation Process

Throughout the annotation process, multiple corpus versions were created step-by-step. The first 100 documents were saved and reviewed as Version 1. After annotating another 300 documents, Version 2 was created and compared with the previous version. This comparison followed two tracks:

1. Guideline Changes – the impact of revised rules was analyzed. For instance, in Version 1, the reflexive pronoun “o‘z” (“self”) was not annotated as a mention. Later, it was decided that if “o‘z” was anaphoric, it should be annotated. In Version 2, such mentions were added. The analysis showed only a 2% increase in mentions, which did not significantly impact overall statistics.

2. Error Detection – the comparison also helped identify systematic errors made by certain annotators. For example, in Version 1, some annotators incorrectly tagged coordinated noun phrases (groups in plural). These were corrected in Version 2. To ensure consistency, all documents from Version 1 were re-reviewed and updated as needed. By the time the corpus reached its final state, all documents adhered to a unified standard.

Once annotation was complete, the dataset was officially named UzCoref 1.0. All related metadata – version history, annotator list, and annotation guidelines – were packaged alongside the corpus. Version control was maintained using a system like Git, with each update recorded as a separate version. This approach ensured scientific traceability, allowing future researchers or thesis writers to track every change made in the corpus development.

IV. Data Preparation For Training The Coreference Resolution Model

We chose an end-to-end neural approach to detect coreference. This architecture examines the text from beginning to end, first identifying potential mention (anaphora) candidates, and then evaluating their probability of association and outputting them in the form of clusters. The working logic of the model is divided into the following steps:

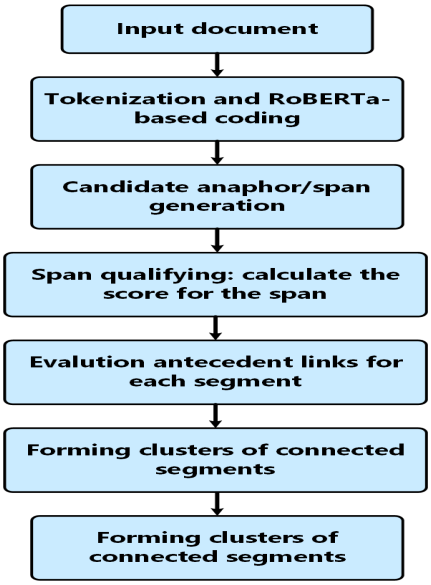

Fig. 2. Flowchart of the model algorithm for Coreference Resolution in Uzbek texts

The model training process was organized as follows: initially, a corpus of 1,000 texts tagged for coreference was split into train/validation/test sets in an 80/10/10 ratio. That is, 800 documents were used for training the model, 100 for validation, and the remaining 100 for final testing. The training data was stored in a custom JSON format, similar to the CoNLL format, where each document includes the full text, tokenized sentences, and cluster identifiers. The purpose of using this format is to ensure compatibility with existing SpanBERT code, as the BERT-Coref code released by Mandar Joshi et al. expects this format [12].

The analysis of the training corpus and data preparation process was organized as follows:

– *Manually Tagged Corpus:* The coreference links were annotated manually by experts to ensure accuracy. Each antecedent and all related anaphors were marked and stored in a special format.

– *Data Splitting:* To ensure objective model evaluation, the corpus was divided into training, validation, and test sets. This division allows for assessing how well the model generalizes to unseen data.

– *Data Augmentation:* Minor modifications were made to some data to improve the model's generalization ability. Care was taken not to distort the integrity and logic of the original annotations.

– *Cross-Lingual Transfer:* Data from the GAP corpus, originally developed for English coreference tasks, was translated and used in Uzbek. This was done to enhance the model's understanding of gender-related contextual relationships.

– *Linguistic Resources:* Additional dictionaries and lists of proper names were used to help the model better identify links through synonyms and name recognition. These resources significantly improved the model's accuracy.

A. Comparative Analysis of Coreference Models

We compare the presented approach with models in foreign languages:

1. Uzbek (UzCoref): Transformer-based models (RoBERTa, XLM-R, UzRoBERTa) have been used. The corpus and resources used are of medium size, with high-quality annotations, although limited in quantity. The Uzbek language lacks grammatical gender but includes pro-drop features and morphological complexity. The current results range between 68–72%.

*2. English (SpanBERT, CorefRoBERTa)***:** These models show the best results, supported by large and well-developed resources. English features include grammatical gender, articles, and other linguistic markers. The performance is high, ranging from 79–80% [12, 13].

3. Russian (RuCoCo, RaCoref, DeepPavlov): With access to large corpora, Russian coreference models address gender and grammatical agreement requirements. Based on the XLM-R transformer, the results are quite strong (68–70%). The models demonstrate a strong ability to handle complex morphology.

4. Turkish (SIGTURK 2024): Currently, only a small tagged corpus exists, and rule-based methods or small neural models are mostly used. Results are lower (60–65%), but Turkish shares close grammatical structure and linguistic complexity with Uzbek.

Turkish is one of the closest languages to Uzbek in terms of grammatical structure. In 2024, within the scope of the Turkic Languages Workshop (SIGTURK 2024), a Turkish coreference corpus was introduced. This corpus is relatively small: it includes only 60 dialogues, containing 3,900 sentences, 18,360 words, and 6,120 annotated mentions. Clearly, there is no large-scale corpus in Turkish yet, and only initial steps have been taken for the domain of dialogue. However, the rule-based and morphological characteristics of Turkish are very similar to Uzbek: both languages use a single third-person pronoun ("u" in Uzbek, "o" in Turkish) with no gender distinction; verbs include person and number suffixes, and subject omission (pro-drop) can occur. Therefore, conceptually, there is a degree of similarity between the Turkish models and the Uzbek model. The difference lies in the fact that we built a full end-to-end transformer-based model, whereas early Turkish efforts were likely limited to rule-based or small neural models. For instance, in SIGTURK 2024, a special annotation scheme was proposed, and rules adapted for agglutinative languages were introduced.

In our view, joint approaches to Coreference Resolution in Turkish and Uzbek (e.g., transfer learning) could be very beneficial – both languages face similar adaptation challenges. In fact, some studies have already tested cross-lingual transfer between Turkic languages and reported positive outcomes. Considering this, we also plan to fine-tune the Uzbek model on Turkish (or Kazakh) data in the future or, alternatively, attempt to build a multilingual Turkic CorefBERT model.

V. CONCLUSION

In this article, a coreference corpus for the Uzbek language was developed in the form of a dataset marking references across more than 1000 texts of various genres. The corpus was prepared in CoNLL (OntoNotes) format and named UzCoref. A combination of automatic and manual verification was used in the annotation process. As a result, high-quality, consistent annotation was achieved. The annotation rules were developed considering the morphological and syntactic features of the Uzbek language. The inter-annotator agreement was high (≈0.8 kappa), confirming the reliability of the annotation.

The experiences gained from this research show that building a coreference corpus requires in-depth study of linguistic features, close collaboration with annotators, and iterative improvement. In the case of Uzbek, we demonstrated that by adapting international practices and customizing guidelines to the language, a high-quality annotated corpus can be created.

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] B.Elov, Ş.Abdisalomova, X.Karimova. Özbek Dili Metinlerindeki Eşgönderge Çözümlemesi Algoritmasi. / *9. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı*, (IEEE – UBMK – 2024), Antalya, pp. 95-100.

[3] O. Uryupina, R. Artstein, A. Bristot, F. Cavicchio, F. Delogu, K. J. Rodriguez, & M. Poesio, Annotating abroad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. / *Natural Language Engineering*, *26*(1), 2020, pp. 95-128.

[4] A. Zeldes. The GUM corpus: Creating multilayer resources in the classroom. / *Language Resources and Evaluation*, *51*(3), 2017, pp. 581-612.

[5] R. Weischedel, S. Pradhan, L. Ramshaw, M. Palmer, N. Xue, M. Marcus, ... & A. Houston. Ontonotes release 4.0. LDC2011T03 / *Philadelphia, Penn: Linguistic Data Consortium*, *17*, 2011. / https://doi.org/10.35111/gfjf-7r50

[6] 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.

[7] K. Lee, L. He, M. Lewis, & L. Zettlemoyer. (2017). End-to-end Neural Coreference Resolution. / *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copengagen, Denmark, 2017, pp. 188-197.

[8] https://uznatcorpora.uz

[9] P. Schüller, K. Cıngıllı, F. Tunçer, B. G. Sürmeli, A. Pekel, A. H. Karatay, & H. E. Karakaş. (2017). Marmara Turkish coreference corpus and coreference resolution baseline. / arXiv:1706.01863v2 [cs.CL], 31 Jul 2018. / https://doi.org/10.48550/arXiv.1706.01863

[10] htpps://uzcoref.uz/

[11] U. Hamdamov, B. Elov, X. Ahmedova, Sh. Abdisalomova, etc. The Problem of Coreference in NLP. / Intelligent Sustainable Systems selected Papers of WorldS4 2024, Vol.3, Springer, pp. 151-160.

[12] M. Joshi, O. Levy, D. S. Weld, L. Zettlemoyer. BERT for Coreference Resolution: Baselines and Analysis. / arXiv: 1908.09091v4 [cs.CL], 22 Dec 2019.

[13] https://huggingface.co/nielsr/coref-roberta-large