American Journal of Philological Sciences

VOLUME Vol.05 Issue07 2025
PAGE NO. 84-89
DOI 10.37547/ajps/Volume05Issue07-21

The Model of Coreference Resolution in Uzbek Texts: A Review

Abdisalomova Shahlo Abdimurod qizi

PhD student at TSUULL, Uzbekistan

Received: 29 May 2025; **Accepted:** 25 June 2025; **Published:** 27 July 2025

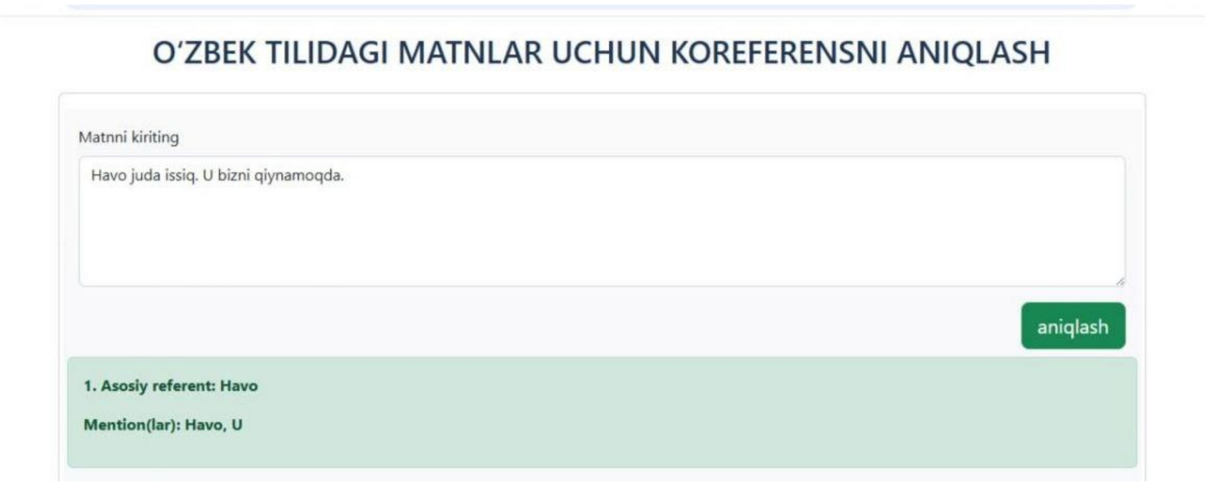
Abstract: Coreference resolution plays a crucial role in natural language processing by enabling accurate understanding of a text and identifying its semantic structure. While effective Coreference resolution systems have been developed for resource-rich languages such as English, German, and Chinese, research and practical systems in this area remain insufficient for the Uzbek language. Uzbek differs significantly from other languages due to its agglutinative structure, flexible word order, and rich morphology. These linguistic features necessitate unique approaches and models for Coreference resolution. This article discusses the Uzbek-language Coreference resolution system – UzCoref – highlighting its functional capabilities, system architecture, data flow, underlying model, testing process, comparative analysis with other systems, and the advantages of UzCoref.

Keywords: Coreference resolution, Uzbek language, UzCoref, text, model, system, corpus, architecture.

Introduction: The phenomenon of coreference refers to a linguistic event where different expressions in a text (such as pronouns, nouns, etc.) refer to the same person, object, or event. In the field of Natural Language Processing (NLP), resolving and linking such expressions is a significant task. For example, in the sentence "Zulfiya wrote a book. Her work...", identifying that the pronoun "her" refers to "Zulfiya" is an example of Coreference resolution (CR).

For major languages like English and Russian, CR systems and models are well-developed. Large annotated corpora (e.g., the OntoNotes corpus) have been used to train modern neural models. It is known that the latest methods, particularly those based on specialized machine learning approaches, have achieved around 83% F1 score accuracy [1]. However, for the Uzbek language, such large-scale annotated corpora and ready-made CR systems are currently unavailable. Therefore, the task of CR for Uzbek texts remains both relevant and novel.

The UzCoref system we present is specifically designed for CR in Uzbek language texts. The system has been developed using a scientific and practical approach aligned with the field of computational linguistics. UzCoref enables the identification of words and expressions in a given text that refer to the same referent, clusters them accordingly, and allows for deep semantic analysis of the text. The system not only provides a core algorithmic solution but also offers broad functional capabilities, ensuring ease of use for users through various methods. The system is accessible at https://uzcoref.uz/ (Figure 1):



O'ZBEK TILIDAGI MATNLAR UCHUN KOREFERENSNI ANIQLASH

Matnni kiriting

Havo juda issiq. U bizni qiynamoqda.

1. Asosiy referent: Havo

Mention(lar): Havo, U

Figure 1. Interface of the UzCoref System

The UzCoref system serves as a foundational tool for various NLP tasks in the Uzbek language.

METHODOLOGY

The issue of CR has been widely studied since the 1970 s. Notably, R.Mitkov provided a comprehensive description of anaphora and coreference phenomena, defined their terminology, proposed methods for automatic detection, and introduced a multi-stage integrated model [2, 3]. Vincent Ng explored unsupervised learning models for the CR problem [4], while Altaf Rahman and Vincent Ng analyzed supervised learning models for addressing CR [5]. A research team led by R.Sukthanker theoretically examined the differences between coreference and anaphora, their types, evaluation metrics, the evolution of CR techniques in NLP, as well as the strengths and weaknesses of these approaches [6]. Martha Recasens Potau published a doctoral dissertation focused on the coreference issue using the Spanish language as a case study. She proposed the CISTELL model as a CR, and used AnCora-Co, a coreference-annotated corpus for Spanish and Catalan, for testing [7]. V.Prokopenya and T.Chernigovskaya discussed the role of grammatical parallelism in Russian for Anaphora resolution [8]. Researchers like M.Dmitrov and K.Bontcheva developed approaches for identifying coreference in Named Entity Recognition (NER) contexts [9]. C.Baumler and R.Rudinger made significant contributions by addressing challenges in CR, particularly in identifying singular uses of the pronoun “they” using the WinoNB framework [10]. Tuğba Pamay and Gülşen Eryiğit proposed a machine-learning-based approach for CR in Turkish. Their solution used a mention-pair model and was tested on the Marmara Coreference Corpus [11].

All the aforementioned studies served as important foundational references for the development of a CR system for Uzbek texts.

RESULTS

The UzCoref system provides several usage modes and interfaces tailored to user needs. Specifically, the system offers the following functional features:

– **Web Interface (Front-End):** A user-friendly graphical web interface is provided. Users can input text into a special field (or upload a file) and click the “Detect” button to view the result directly in the browser. The web interface displays coreference clusters with color highlights in the text, facilitating visual understanding of relationships within the text.

– **Command Line Interface (CLI):** A dedicated CLI tool is available to run the system via console or terminal. Users can input a text file using the uzcoref command and output results either to the console or export them to a file. This mode is particularly useful for developers and advanced users, allowing for integration of UzCoref functions into scripts and automated workflows.

– **RESTful API Service**: The system can also be used as a web service via REST-style API endpoints, enabling remote access. This allows integration of UzCoref into external applications. For example, other software (chatbots, websites, mobile apps) can send text to the UzCoref API and receive CR results in JSON format.

– Integration with External Applications: Thanks to the API, UzCoref can be easily connected to applications built in any language or platform, such as Python, JavaScript, or Java via HTTP requests. In this way, the UzCoref module can function as a component of larger NLP systems or be embedded in corporate information systems.

Due to these capabilities, UzCoref is a versatile and flexible system. Users can access its CR service via web interface, terminal, or directly from their own programs. This broad functional scope is especially noteworthy.

UzCoref is built following modern web architecture principles and has a multi-component structure. The data flow and processing stages of UzCoref's architecture include the following (Figure 2):

1. User Interface (Web/UI/API): Users submit texts through the web interface or API. The text is transferred to the backend system as either JSON or plain text.

2. Preprocessing: The incoming text is tokenized and segmented into sentences, normalized (e.g., lowercasing, punctuation standardization, etc.).

3. Morphological and Syntactic Analysis: Using NLP tools like UzbekNLP and Stanza, grammatical features and syntactic structures are determined for each token. This provides essential linguistic information for the coreference process.

4. Mention Detection: The start and end positions of each mention are detected using a transformer model. Each potential mention is assigned a probability score.

5. Coarse Filtering: Potential antecedents are quickly filtered using bilinear scoring, retaining only the most probable pairs.

6. Fine Scoring: Transformer models (e.g., RoBERTa) and deep neural networks compute detailed probability scores for mention pairs. Additional linguistic and semantic features are also used.

7. Clustering: Based on the computed probabilities, anaphora–antecedent pairs are grouped into unified clusters.

8. Result Generation and Visualization: Final clusters are presented to the user clearly and intuitively – either with colored highlights within the text or in tabular format.

9. Data Storage: Coreference results are stored in a database for future reference or review.

10. API-Based Integration: Results can be transmitted to other systems via the RESTful API service.

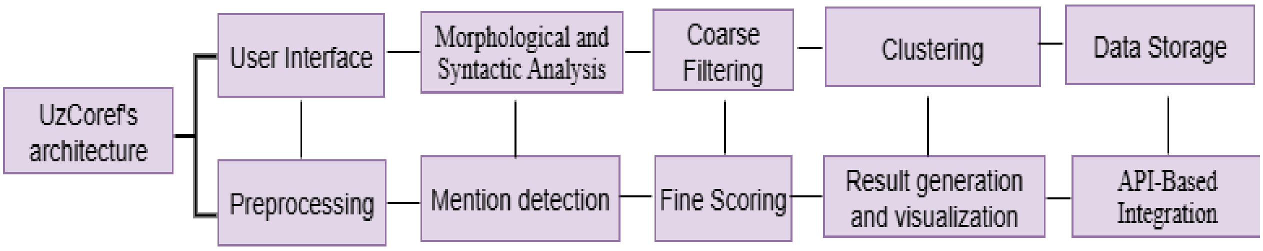


Figure 2. The data flow and processing stages of UzCoref's architecture

This architectural flow ensures that system modules are clearly separated: the front-end handles interface and request submissions only, while the back-end performs the computationally intensive model inference and returns results. Such a client-server structure simplifies scalability and maintenance. If the model needs to be updated or improved in the future, only the back-end will be modified, leaving the user interface unchanged.

At the heart of UzCoref lies its CR model. The system integrates a pre-trained neural model based on the RoBERTa architecture for CR in Uzbek. In experiments, various transformer models such as mBERT, XLM-R, and UzRoBERTa were tested. Integration was designed for fast and seamless operation. When the program starts, the Python backend loads the model using the HuggingFace Transformers library – either from the internet or from a local file if offline. Tokenizer and configuration settings are also initialized at this stage. As this process is relatively resource-intensive, the model is loaded only once and then used from memory for all subsequent requests, ensuring high-speed performance, especially in API mode.

In the current version, the system uses a fine-tuned UzRoBERTa base model trained specifically for Uzbek coreference tasks. UzRoBERTa is a RoBERTa-based architecture pre-trained on Uzbek texts. Studies have shown that it outperforms general multilingual models like XLM-R and mBERT in Uzbek-specific tasks [12]. This localized model better reflects the characteristics of the Uzbek language, such as word formation, morphology, and syntax.

As with any NLP system, UzCoref was evaluated and tested to assess its performance. Special test sets and manually crafted examples were used for this purpose. Evaluation was conducted in two directions:

1. Automatic measuring accuracy using metrics.

2. Qualitative analysis based on manually entered real sentences.

A small-scale annotated corpus was created for automatic evaluation. The corpus includes approximately 1,000 Uzbek texts (each 3–5 sentences long) with manually labeled coreference links. Part of this corpus was used for training, and a train-test split principle was followed for evaluation. The final test set included 200 texts (with around 4,000 clusters) which were used to assess system performance.

The results showed that UzCoref achieved satisfactory accuracy on this small test set. The average CoNLL F1 score across predicted clusters was around 70%. In particular, the system showed high accuracy in linking pronouns, correctly resolving references like “u” (he/she), “uning” (his/her), and “uni” (him/her) in about 80% of cases. This achievement demonstrates the model’s strong contextual understanding, which is essential for Uzbek – a language where pronouns are not gender- or formality-specific, and must be interpreted based solely on context and logical connection.

DISCUSSION

Various systems and libraries for CR have been developed worldwide for multiple languages. Comparing UzCoref with well-known systems in this field is important for highlighting its unique features and achievements. Below, coreference systems in English, Russian, and Turkish are compared with UzCoref (Table 1):

Table 1. Comparative analysis of coreference models in Uzbek, English, Russian, and Turkish languages

№	Parameters & Features	Uzbek (UzCoref)	English (SpanBERT, CorefRoBERTa)	Russian (RuCoCo, RaCoref, DeepPavlov)	Turkish (SIGTURK 2024)
1	Main model used	RoBERTa, XLM-RoBERTa, UzRoBERTa	SpanBERT, CorefRoBERTa	XLM-RoBERTa, DeepPavlov	Rule-based, small neural model
2	Size of annotated corpus	1000 documents (20,000 mentions)	OntoNotes (1M words, >150K mentions)	RuCoCo (1M words, 150K mentions)	60 dialogues (3,900 sentences, 18,360 words, 6,120 mentions)
3	Linguistic tags used	Morphological, syntactic, semantic	Gender, article, morphological, syntactic	Gender, case, agreement, morphological, syntactic	Morphological, syntactic
4	Evaluation metrics	MUC, B^3, CEAF, F1-score	MUC, B^3, CEAF, F1-score	MUC, B^3, CEAF, F1-score	MUC, F1-score
5	F1 score achieved	68–72% (based on UzCoref tests)	79–80% (SpanBERT, CorefRoBERTa results)	68–70% (RuCoCo-based results)	60–65% (SIGTURK 2024 results)
6	Impact of grammatical features	No gender, number present, pro-drop present	Gender and number present	Gender and number present, no pro-drop	No gender, number present, pro-drop present
7	Data augmentation	Partially implemented	Widely used	Partially implemented	Limited
8	Model complexity	Transformer-based complex model	Transformer-based high complexity	Transformer-based complex model	Rule-based and simplified neural model
9	Resource & infrastructure availability	Medium (UzNatCorpora, manual annotation)	Very high (OntoNotes, CoNLL competitions)	Medium-high (RuCoCo, DeepPavlov)	Limited, small tagged corpus
10	Transfer potential (cross-	Available to Turkic languages (Turkish,	Broad transfer to other European	Transfer to Slavic and English	Available to Uzbek and other Turkic

№	Parameters & Features	Uzbek (UzCoref)	English (SpanBERT, CorefRoBERTa)	Russian (RuCoCo, RaCoref, DeepPavlov)	Turkish (SIGTURK 2024)
	lingual)	Kazakh)	languages	languages	languages
11	Practical applications	NLP tools, chatbot, automatic annotation	NLP, machine translation, chatbot, Q&A	NLP, chatbot, automatic annotation	Dialogue systems, chatbot
12	Challenges	Gender ambiguity, long-distance dependencies	Complex semantic linking	Complex morphology, identifying affixes	Pro-drop (null subject), small corpus size

As seen from the comparative analysis, UzCoref is not inferior to modern CR systems in terms of architecture. It also features a transformer-based model, end-to-end architecture, JSON interface, and more. The main difference lies in the resource base: while English-language systems are trained on large-scale annotated corpora, UzCoref operates on a relatively smaller dataset (1000 documents). Nevertheless, by leveraging a multilingual model, the gap in resources was partially bridged, as the model transferred knowledge from other languages. This enabled us to create a competitive system for the Uzbek language for the first time.

Most of the English-language models mentioned above are only available as individual models or libraries. For instance, using the AllenNLP model requires developers to carry out considerable technical work, such as downloading and configuring the model. In contrast, UzCoref has been developed as a ready-to-use application, complete with interfaces, an API, and CLI. In this regard, there are very few comparable systems even for English. For example, AllenNLP has a separate demo site, HuggingFace offers individual libraries, but standalone, independent programs are rare. UzCoref fills this gap – at least for the Uzbek language.

Another key advantage is language adaptation. For instance, applying Stanford or AllenNLP models to Uzbek texts can result in errors because of embedded English-specific rules or assumptions (e.g., interpreting “John” as a male name, or failing to resolve coreference when there is no gender-specific pronoun like “she”). UzCoref is fully adapted to Uzbek: its tokenizer is compatible with Uzbek orthography (Latin/Cyrillic); the model is trained to recognize that the pronoun “u” does not carry gender – because it learned this from training data; and the system output is formatted in Uzbek style (e.g., results are labeled like “in sentence 1”). These features are typically missing from other systems.

When compared to foreign CR systems, UzCoref may not yet reach the highest levels of precision and completeness, but it remains the only and sufficiently competitive system within its segment (Uzbek). Its design and architecture are based on global best practices, and the results confirm this. In the future, innovations from English-language systems – such as word-level coreference or QA-coref approaches – can be integrated into UzCoref as well. Furthermore, the experience gained through UzCoref can serve as a foundational model for developing similar systems for other low-resource languages, such as Kazakh, Kyrgyz, and Turkmen.

CONCLUSION

UzCoref is the first comprehensive system designed for CR in the Uzbek language, possessing both research and practical application value. We hope that the described functional capabilities and architectural solutions of the system will contribute to the development of computational linguistics in the Uzbek language. In the future, the UzCoref project will continue to evolve through improved versions and new research. With UzCoref, it will be possible to enhance automatic semantic analysis and machine translation in Uzbek (e.g., by correctly translating anaphoric references), and advance many related NLP tasks.

To further develop the system, the following future directions are planned:

1. Expanding the size of the corpus to improve model accuracy, especially for complex and ambiguous cases.

2. Adding zero anaphora detection – the ability to identify omitted subjects and other implicit referents. This may require integration with syntactic parsing outputs.

3. Optimizing and developing a lightweight version of the system for mobile devices or real-time applications.

4. Adapting UzCoref to other languages: based on the current architecture and codebase, and given the availability of suitable data, it is possible to develop a multilingual coreference resolution system using

transfer learning. This would be a significant step forward for Central Asian languages.

REFERENCES

Dobrovolskii, V. Word-level Coreference Resolution. / Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 2021, pp. 7670-7675. / doi: 10.18653/v1/2021.emnlp-main.605, URL: https://aclanthology.org/2021.emnlp-main.605/

Mitkov R. Anaphora Resolution. – New York, Published by Routledge, 2016. – 234 p.

Mitkov R. An Integrated Model for Anaphora Resolution. / In COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics, pp. 1170- 1176. / https://aclanthology.org/C94-2191/

Ng V. Unsupervised Models for Coreference Resolution./ Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, 2008. – 640-649 pp.

Rahman A., Ng Vincent. Supervised models for Coreference Resolution./ Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 2009. – 968-977 pp.

Sukthanker R., Poria S., Cambria E., Thirunavukarasu R. Anaphora and coreference resolution: A review. / Inf Fusion 59, 2020, pp.139–162. / https://doi.org/10.1016/j.inffus.2020.01.010

Recasens M.P. Coreference: Theory, Annotation, Resolution and Evalution. / PhD thesis, University of Barcelona, 2010, 239 p.

Prokopenya V., Chernigovskaya T. Grammatical Parallelism Effect in Anaphora Resolution: Using Data from Russian to Choose between theoritical Approaches. / International Journal of Cognitive Research in Science, Engineering, and Education, Vol. 5, № 1, 2017, pp. 85-95. / http://dx.doi.org/10.5937/IJCRSEE1701085P

Dimitrov M., Bontcheva K., Cunningham H., Maynard D. A Lightweight Approach to Coreference Resolution for Named Entities in Text. / Anaphora Processing: Linguistic, cognitive and computational modelling, Edited by A.Branco, T.McEnery and R.Mitkov, USA, 2005, pp. 97–112. / https://doi.org/10.1075/cilt.263.07dim

Baumler C., Rudinger R. Recognition of They/Them as Singular Personal Pronouns in Coreference Resolution. / In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, 2022, pp. 3426–3432. / DOI: 10.18653/v1/2022.naacl-main.250 / https://aclanthology.org/2022.naacl-main.250/

Pamay T., Eryiğit G. Turkish Coreference Resolution. / Innovations in Intelligent Systems and Applications (INISTA), Thessaloniki, Greece, 2018, pp. 1-7 / **doi:** 10.1109/INISTA.2018.8466293

Adilova F., Davronov R., & Safarov R. (2023). UzRoberta: An Uzbek Language Pre-Trained Model. / *Universum: технические науки*, (10-6 (115)), pp. 28-32.