ISSN:2828-6235
Impact factor:13.77

International Journal of Literature and Language Studies(IJLLS)

SEMANTIC ANALYSIS OF UZBEK LANGUAGE TEXTS USING THE INCEPTION PLATFORM

Matyoqubova Shaydo Maqsud qizi
Master's Student Urgench State University
Email: shaydomatyoqubova522@gmail.com
Doi: https://doi.org/10.5281/zenodo.23387602

Abstract: This article examines the application of the INCEpTION platform to the semantic analysis of Uzbek-language texts, emphasizing linguistic annotation, knowledge representation, and the development of computational language resources. INCEpTION is a web-based semantic annotation platform that supports collaborative annotation, knowledge-assisted labeling, and machine-learning-assisted annotation workflows. The study investigates how its functionalities can be adapted to the linguistic characteristics of Uzbek, an agglutinative Turkic language with extensive morphological variation and relatively flexible word order. Particular attention is given to named entity recognition, semantic relation annotation, word sense disambiguation, and the identification of context-dependent meanings. The research adopts a descriptive and comparative methodological approach based on established studies in natural language processing, semantic annotation, and computational linguistics. The article proposes an annotation framework for Uzbek texts and identifies methodological challenges associated with morphological complexity, contextual ambiguity, annotation consistency, and limited annotated resources. The findings indicate that INCEpTION provides a technically suitable environment for organizing human-supervised semantic annotation projects, although its effectiveness for Uzbek depends on the availability of appropriate linguistic guidelines, annotated datasets, and compatible language-processing models. The study contributes to the methodological foundations of Uzbek natural language processing and outlines practical directions for developing semantically annotated Uzbek corpora.

Keywords: INCEpTION platform, Uzbek language, semantic analysis, natural language processing, semantic annotation, corpus linguistics, named entity recognition, word sense disambiguation, machine learning, linguistic resources.

Introduction

The rapid development of natural language processing (NLP) has significantly expanded opportunities for the computational analysis of human language. Contemporary NLP research includes automatic text classification, information extraction, machine translation, semantic search, question answering, and contextual language understanding. These technologies rely on linguistic resources that represent not only the formal structure of texts but also their meanings and semantic relationships [1].

Semantic analysis is particularly important because linguistic meaning cannot always be determined through isolated lexical units. Words acquire different interpretations depending on syntactic structure, discourse context, and communicative purpose. Consequently, computational semantic analysis requires systematic approaches to representing entities, concepts, relations, and contextual meanings [2].

The Uzbek language presents distinctive challenges for computational linguistic analysis. As a Turkic language, Uzbek possesses an agglutinative morphological system in which grammatical meanings are frequently expressed through sequences of suffixes. Its linguistic structure differs considerably from that of English, which has historically received substantially greater attention in NLP research.

For example, the Uzbek word kitoblarimizdan can be segmented into kitob (book), -lar (plural), -imiz (first-person plural possessive), and -dan (ablative case). Such morphological structures demonstrate why computational analysis must distinguish lexical meaning from grammatical information.

Another important issue is the availability of annotated linguistic resources. Supervised NLP methods generally require training examples with reliable labels. For languages with comparatively limited digital resources, the creation of annotated corpora remains an essential research priority [3].

The INCEpTION platform offers a practical environment for addressing annotation-related challenges. Developed as a web-based semantic annotation system, INCEpTION combines manual annotation, collaborative project management, knowledge-oriented annotation, and machine-learning-assisted workflows [4].

Unlike systems restricted to basic text labeling, INCEpTION supports multiple annotation layers and structured relationships between annotated elements. These capabilities are relevant to Uzbek because semantic interpretation frequently depends on interactions among morphology, syntax, and context.

The purpose of this article is to examine the methodological possibilities of using INCEpTION for Uzbek semantic annotation, evaluate its relevance to linguistic resource development, and propose an academically grounded framework for future annotation projects.

The research focuses on three principal questions: how semantic annotation can be organized for Uzbek texts, which INCEpTION functionalities are applicable to this task, and what linguistic and technical limitations must be considered.

Methodology

The study employs descriptive, comparative, and analytical research methods to investigate the application of INCEpTION to Uzbek-language semantic annotation. The methodological foundation combines established NLP principles with the documented architecture of INCEpTION and published research on Uzbek computational linguistics.

The research does not claim to have conducted a new experimental annotation campaign. Instead, it synthesizes existing empirical evidence and develops a reproducible methodological framework.

The proposed annotation procedure consists of corpus preparation, linguistic preprocessing, annotation schema development, semantic labeling, quality assessment, and data export.

Corpus preparation. Uzbek texts should be collected from legally accessible literary, educational, journalistic, and scientific materials. Corpus diversity is necessary to represent differences in vocabulary, sentence structure, terminology, and discourse conventions. Documents should retain metadata describing their sources, genres, and publication characteristics.

Linguistic preprocessing. Text normalization should address encoding consistency, punctuation, tokenization, and orthographic variation. Uzbek materials may contain Latin- and Cyrillic-script representations, making script normalization an important consideration. However, original textual forms should be preserved when orthographic variation is relevant to the research.

Annotation schema development. INCEpTION supports configurable annotation layers that can represent text spans, linguistic categories, and relationships between annotated elements [4]. For Uzbek semantic analysis, an annotation schema may include named entities, semantic roles, conceptual categories, and contextual lexical meanings.

Semantic annotation. Human annotators identify meaningful linguistic units and assign labels according to predefined guidelines. Named entities may be classified into categories such as persons, organizations, locations, and temporal expressions. Semantic relations may represent interactions between entities or participants in events.

Quality assessment. Annotation reliability can be evaluated through inter-annotator agreement. Cohen's kappa provides a chance-corrected agreement measure for two annotators working with categorical labels [5]. Krippendorff's alpha offers a more general framework applicable to different annotation settings [6].

Data validation. Completed annotations should undergo consistency checking and expert adjudication before being used as reference data for computational models.

The methodology emphasizes reproducibility, transparent annotation criteria, and separation between experimentally demonstrated findings and proposed applications.

Results

The documentary analysis identifies several established findings relevant to the application of INCEpTION in Uzbek computational linguistics.

First, INCEpTION provides a modular architecture for interactive annotation, concept linking, fact linking, and knowledge-oriented semantic annotation. Its original technical description was published by Klie and colleagues in 2018 in the proceedings of COLING, on pages 5–9 [4].

Second, the applicability of INCEpTION to Uzbek has been demonstrated in a published annotation project. Matlatipov and colleagues reported an Uzbek Universal Dependencies dataset containing 681 sentences and 7,542 tokens from literary and educational texts. The annotation process involved five contributors, including three linguists and two NLP engineers, working in the INCEpTION environment [7].

This constitutes direct evidence that INCEpTION can support structured annotation of Uzbek linguistic data. However, the reported dataset primarily concerns morphological and syntactic annotation rather than a comprehensive semantic annotation benchmark.

Third, research on Uzbek named entity recognition has produced an annotated dataset containing 1,160 sentences and approximately 19,000 word forms. Mengliev and colleagues presented this resource in 2024, demonstrating the feasibility of developing structured annotation resources for Uzbek entity recognition [8].

Fourth, the availability of Uzbek morphological resources has expanded. Sharipov's 2026 dataset includes 46,307 unique stems and 64,052 corresponding lemmas organized into 12 grammatical categories [9]. Such resources can support preprocessing stages preceding semantic annotation.

These published statistics describe existing linguistic datasets; they are not performance measurements obtained from a new experiment in this study.

Table 1. Documented linguistic resources relevant to Uzbek annotation

Resource	Published data	Relevance
Uzbek Universal Dependencies dataset, 2026	681 sentences; 7,542 tokens	Morphological and syntactic annotation using INCEpTION
Uzbek NER dataset, 2024	1,160 sentences; approximately 19,000 word forms	Named entity identification
Uzbek stems and lemmas dataset, 2026	46,307 stems; 64,052 lemmas	Morphological preprocessing

The results establish that INCEpTION is technically applicable to Uzbek annotation and that published linguistic datasets provide foundations for further semantic research.

Nevertheless, the reviewed evidence does not establish a universal accuracy rate, processing speed, or semantic annotation performance level for Uzbek texts processed through INCEpTION.

Analysis and Discussion

The application of INCEpTION to Uzbek semantic analysis must be considered in relation to linguistic structure, annotation methodology, machine-learning integration, and corpus quality.

Linguistic complexity and semantic interpretation

Uzbek agglutination creates challenges for systems that rely exclusively on surface word forms. A single lexical stem may appear with numerous combinations of inflectional and derivational suffixes. These combinations influence grammatical relationships and sometimes contribute to lexical interpretation.

For example, the sentence Talaba universitetda ilmiy tadqiqot olib bormoqda means “The student is conducting scientific research at the university.”

A semantic annotation scheme could identify talaba as the agent, ilmiy tadqiqot as the activity or research object, and universitetda as a location expression.

These labels illustrate a proposed analytical representation rather than the output of an experimentally tested model.

The distinction between syntactic and semantic relations is especially important. Dependency parsing identifies grammatical relationships, whereas semantic role labeling attempts to identify participants and their functions within events.

Research on semantic role labeling demonstrates that computational systems can represent predicate–argument relationships through structured linguistic annotations [10].

For Uzbek, morphological case marking can provide useful information for semantic interpretation. However, grammatical case does not always correspond directly to a single semantic role. Consequently, contextual interpretation remains necessary.

Named entity recognition and contextual ambiguity

Named entity recognition constitutes an important component of semantic analysis because it identifies references to people, institutions, geographical locations, and other entities.

In Uzbek, named entities may contain grammatical suffixes attached to their final components. For instance, Toshkentda refers to a location while also expressing the locative grammatical relationship.

An annotation system must therefore distinguish the entity itself from the grammatical information associated with its surface form.

INCEpTION allows researchers to configure annotation layers and define categories appropriate to particular projects. Its knowledge-oriented architecture also supports linking textual expressions to concepts represented in external knowledge resources [4].

However, successful entity linking requires appropriate knowledge bases, reliable normalization procedures, and sufficient coverage of Uzbek names and terminology.

The availability of annotated Uzbek NER data provides a basis for developing and evaluating these capabilities [8].

Machine-learning-assisted annotation

One significant characteristic of INCEpTION is its integration of annotation recommendation mechanisms.

Machine-learning-assisted annotation can reduce repetitive manual labeling by presenting suggested annotations for human review. Nevertheless, the usefulness of recommendations depends on the quality of the underlying models and training examples.

The original INCEpTION research describes a human-centered annotation environment in which machine-generated suggestions assist rather than replace annotators [4].

This distinction is particularly relevant to Uzbek. A recommendation model developed primarily for another language cannot be assumed to provide accurate Uzbek semantic annotations without suitable adaptation and evaluation.

Transfer learning offers a possible methodological direction. Multilingual transformer models, including multilingual BERT, demonstrate the broader feasibility of cross-lingual representation learning [11].

Similarly, XLM-R research established the effectiveness of large-scale multilingual pretraining for several cross-lingual NLP tasks [12].

Nevertheless, these findings should not be interpreted as direct evidence of specific INCEpTION performance on Uzbek semantic datasets.

Annotation reliability and research validity

The quality of semantic annotation depends on clear definitions, consistent interpretation, and systematic disagreement resolution.

Two annotators may interpret the same expression differently when semantic categories overlap or when context is insufficient.

For this reason, annotation guidelines should define category boundaries, provide representative examples, and establish procedures for resolving ambiguous cases.

Inter-annotator agreement measures can identify inconsistencies, although numerical agreement alone does not guarantee semantic correctness [5], [6].

A reliable workflow should combine quantitative agreement assessment with qualitative expert review.

The Uzbek Universal Dependencies project demonstrates the practical relevance of collaborative annotation, calibration, double annotation, and adjudication. Its published methodology documents these procedures within an INCEpTION-based workflow [7].

Prospects for Uzbek semantic corpus development

The integration of linguistic expertise, structured annotation, and computational tools offers several opportunities for Uzbek language technology.

Semantically annotated corpora could support information extraction from administrative documents, educational materials, scientific publications, and other text collections.

They could also contribute to the evaluation of semantic search systems, question-answering applications, and domain-specific language models.

However, these applications require careful attention to corpus representativeness, data licensing, personal information protection, and annotation validity.

Research on contextual language representations demonstrates that meaning can be modeled using information from surrounding linguistic context [13]. This provides a theoretical basis for integrating contextual language models with human-supervised semantic annotation.

The most appropriate approach for Uzbek is therefore not to rely exclusively on automatic processing, but to combine computational assistance with expert linguistic interpretation.

A further research priority is the establishment of publicly documented evaluation protocols. These should distinguish entity recognition accuracy, relation extraction performance, semantic role labeling quality, and annotation agreement rather than combining them into an undefined measure of semantic analysis accuracy.

Conclusion

The study establishes that the INCEpTION platform offers a suitable technological foundation for organizing semantic annotation projects involving Uzbek-language texts. Its configurable annotation layers, collaborative environment, knowledge-oriented functionalities, and machine-learning-assisted recommendations provide methodological advantages for structured linguistic resource development.

Published research demonstrates the practical use of INCEpTION in Uzbek morphological and syntactic annotation. Existing datasets for named entity recognition, stems, lemmas, and dependency relations further strengthen the resource base for Uzbek computational linguistics.

Nevertheless, the availability of these resources does not eliminate challenges associated with agglutinative morphology, contextual ambiguity, semantic category definition, and annotation consistency.

The principal contribution of this article is a methodological framework integrating corpus preparation, linguistic preprocessing, semantic annotation, expert validation, and computational evaluation.

Future research should implement this framework on a representative Uzbek corpus and report experimentally measured annotation agreement, model precision, recall, F1-score, and annotation efficiency.

Such work could support the development of more reliable Uzbek semantic processing technologies while strengthening the scientific infrastructure of computational linguistics in Uzbekistan.

The following bibliography contains published scholarly sources. Page ranges are supplied for paginated articles and conference papers; article identifiers are retained for publications that use electronic article numbering. Exact cited pages cannot be established for every source without a full-text page-by-page audit.

References

1. Jurafsky, D., Martin, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd ed. Pearson Prentice Hall, 2009. 1024 pp.
2. [2] Navigli, R. “Word Sense Disambiguation: A Survey.” ACM Computing Surveys, 2009, Vol. 41, No. 2, Article 10, pp. 1–69.
3. [3] Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M. “The State and Fate of Linguistic Diversity and Inclusion in the NLP World.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 6282–6293.
4. [4] Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., Gurevych, I. “The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation.” Proceedings of COLING 2018: System Demonstrations, 2018, pp. 5–9.
5. [5] Cohen, J. “A Coefficient of Agreement for Nominal Scales.” Educational and Psychological Measurement, 1960, Vol. 20, No. 1, pp. 37–46.
6. [6] Krippendorff, K. “Computing Krippendorff's Alpha-Reliability.” University of Pennsylvania, 2011, pp. 1–12.
7. [7] Matlatipov, S., Aripov, M., Bobokandov, M., Matlatipov, G. “Annotated Universal Dependencies Dataset for Literary and Educational Uzbek Texts.” Data in Brief, 2026, Vol. 66, Article 112857.
8. [8] Mengliev, D., Barakhnin, V., Abdurakhmonova, N., Eshkulov, M. “Developing Named Entity Recognition Algorithms for Uzbek: Dataset Insights and Implementation.” Data in Brief, 2024, Vol. 54, Article 110413.
9. [9] Sharipov, M. “A POS-Tagged Dataset of Uzbek Word Lemmas and Stems for Morphological and NLP Research.” Data in Brief, 2026, Vol. 68, Article 113072.
10. [10] Gildea, D., Jurafsky, D. “Automatic Labeling of Semantic Roles.” Computational Linguistics, 2002, Vol. 28, No. 3, pp. 245–288.
11. [11] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” Proceedings of NAACL-HLT 2019, 2019, pp. 4171–4186.
12. [12] Conneau, A., Khandelwal, K., Goyal, N., et al. “Unsupervised Cross-lingual Representation Learning at Scale.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

13. [13] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. "Deep Contextualized Word Representations." Proceedings of NAACL-HLT 2018, 2018, pp. 2227–2237.