TASHKENT UNIVERSITY OF
INFORMATION TECHNOLOGIES
NAMED AFTER MUHAMMAD AL-KHWARIZMI

MUHAMMAD AL-XORAZMIY NOMIDAGI
TOSHKENT AXBOROT TEXNOLOGIYALARI
UNIVERSITETI

BULLETIN OF TUIT:

MANAGEMENT AND COMMUNICATION TECHNOLOGIES



ISSN:2181-1083 Date:2025-06-10 DOI:
Maksud S. Sharipov, Ogabek O. Sobirov
№ 1(8)2025

Bulletin of TUIT: Management and Communication Technologies

MODELING THE DEVELOPMENT OF LEMMATIZATION ALGORITHMS FOR THE UZBEK LANGUAGE

Maksud S. Sharipov
Department of Computer Sciences, Urgench State University named after Abu Rayhan Biruni, Urgench, Uzbekistan
maqsbek72@gmail.com
0000-0002-2363-6533

Ogabek O. Sobirov
Department of Computer Sciences, Urgench State University named after Abu Rayhan Biruni, Urgench, Uzbekistan
sobirovogabek0409@gmail.com
0000-0003-1770-8550

Abstract **-** Lemmatization is a vital task in Natural Language Processing (NLP), essential for various applications. However, the Uzbek language, categorized as a low-resource language in NLP, lacks dedicated lemmatization systems. This deficiency hinders the development of advanced NLP tools for Uzbek texts. To address this, we present a structured approach using IDEF0 and IDEF1X models to design an Uzbek lemmatization system. The IDEF0 model details the system's functionality, outlining the lemmatization process, while the IDEF1X model illustrates the relationships between data components. Furthermore, we developed a rule-based system that leverages Uzbek morphology to accurately determine word lemmas. This research contributes a well-defined architectural framework and a practical rule-based solution for lemmatization, aiding in the advancement of NLP for Uzbek and similar low-resource, morphologically rich languages.

Keywords *– lemma, lemmatization, IDEF0, IDEF1X, Morphological Analysis, NLP, Rule-based Systems.*

I. INTRODUCTION

Natural Language Processing (NLP) plays a critical role in enabling machines to understand, process, and generate human languages. While NLP technologies have achieved considerable success for well-resourced languages like English, their application to morphologically rich and low-resource languages, such as those in the Turkic family, remains a significant challenge. Uzbek, a Turkic language spoken by over 30 million people [1], is characterized by agglutinative morphology, where words are formed by attaching multiple suffixes to a root. This structure increases the vocabulary size exponentially and complicates fundamental NLP tasks such as part-of-speech tagging, syntactic parsing, and lemmatization.

Lemmatization, the task of reducing a word to its base or dictionary form (lemma), is essential in many NLP applications, including search engines, machine translation, and question-answering systems [2], [3], [4]. In agglutinative languages like Uzbek, a single lemma can yield hundreds of surface word forms due to the stacking of grammatical affixes. This makes rule-based lemmatization not only relevant but often

necessary, especially in the absence of large annotated corpora. The design of accurate lemmatization systems for Uzbek requires not only a deep understanding of its morphological structure but also a clear and functional representation of the processes involved.

To model the functional behavior of lemmatization algorithms and analyze system components systematically, structured modeling techniques are essential. IDEF0 (Integration Definition for Function Modeling) is a graphical modeling method used to describe the functions of a system and their interrelationships. It supports hierarchical decomposition of processes and emphasizes inputs, controls, outputs, and mechanisms (ICOM). IDEF1X, on the other hand, focuses on information modeling and is used for data structure design in relational databases. These methodologies are widely recognized in systems engineering and process design:

"IDEF0 is a method designed to model the decisions, actions, and activities of an organization or system, while IDEF1X is used to model data relationships" [5].

Recent studies have explored various aspects of lemmatization and morphological analysis in the context of low-resource and agglutinative languages. Authors [6]proposed BERT-based models for Russian and other morphologically complex languages, while other scientists [7] emphasized the importance of linguistic preprocessing, including lemmatization, in improving neural NLP models. For Turkic languages, projects like TRmorph [8] have aimed at rule-based morphological analyzers, which serve as a basis for lemmatization tasks.

In Uzbek NLP, the paper [9] presents a rule-based model for analyzing and predicting the placement of periods and commas in Uzbek texts, addressing the lack of such tools for this low-resource language and suggesting its potential benefit for related Turkic languages. The study [10] introduces the first publicly available POS-annotated dataset and a high-accuracy rule-based tagger for the agglutinative Uzbek language, using a suffix-stripping approach to support various NLP tasks and potentially aid related Turkic languages. The paper [11] proposes a rule-based method for detecting verbs in Uzbek texts using morphological suffix patterns, achieving a high F1-score of 0.97 and demonstrating its effectiveness for NLP tasks in this agglutinative language. This article [12] explores the extraction of basic vocabulary from the Uzbek elementary corpus, highlighting its benefits for language learners and linguists by focusing on core words prevalent in daily use and literature through high-frequency analysis. This work [13] introduced rule-based morphological analysis using suffix patterns, highlighting the significance of well-defined morpho-syntactic rules for accurate lemma extraction. However, there is a lack of formalized functional modeling to accompany such linguistic frameworks. This study addresses that gap by applying IDEF0 and IDEF1X models to systematically represent the process of lemmatization in Uzbek, using the open-source UzbekLemma project as a case study. The modeling approach aims to support better design,

implementation, and scalability of NLP systems for the Uzbek language and other Turkic languages with similar morphological characteristics.

II. METHODOLOGY

This study adopts a structured modeling approach to represent the lemmatization process for the Uzbek language, using IDEF0 for functional analysis and IDEF1X for data modeling. These modeling techniques enable clear visualization of processing steps, data flows, and component interactions within the rule-based system implemented in the UzbekLemma GitHub project.

2.1 Input Type Classification

Inputs to the system fall into multiple categories:

- **Complete sentences**: composed of one or more words.
- **Individual words**: either short (1–2 characters) or long (3+ characters).
- **Non-alphabetic inputs**: digits, symbols, or their combinations.
- **Compound expressions**: combinations of words with special roles.

The IDEF0 logic includes precise routing paths to determine which modules handle which input types. For example, **1–2 character words** are filtered early, while **3+ character words** are analyzed morphologically via A4 modules. Compound lemmas (e.g., proper nouns or phrasal verbs) are matched through dedicated submodules and lemma databases, such as the "*Ajratib yoziluvchi qo'shma so'zlar bazasi (Database of open compound words)*".

Table 1. The complete set of input types that can be provided by the user to the model.

Type of input		Example
Sentence		*Sayohatchi uzoq mamlakatlarga borib keldi.*
Word		*Type of word structure* [Table 2]
Number		*70*
Character(s)		*($$$)*
Combination	Word and number	*2025yil*
	Word and char	*parol123*
	Number and char	*93***4263*

The table above outlines the main types of input data that can be submitted to the lemmatization model. These include not only standard linguistic units such as sentences and individual words, but also numbers, characters, and their various combinations (e.g., word-number, word-character, number-character). Such input diversity reflects real-world usage scenarios and is essential for designing a robust and comprehensive lemmatization system.

Table 2. Types of word structure.

Structure type	Word

Soddа	*Tub*	*kitob (book)*	
	Yasama	*guldon (vase)*	
Qo‘shma		*mehnatsevar (hardworking)*	*o‘n bir (eleven)*
Juft		*aka-uka (brothers)*	
Takror		*uzun-uzun (long-long)*	
Atoqli ot		*Anvar (proper name)*	
Qisqartma		*BMT (abbreviation, United Nations - UN)*	
KFSQ		*aytib yuribdi (has been saying)*	
To‘liqsiz fe’l		*edi (was)*	
Sifat darajalari		*eng (most)*	

The table above summarizes the structural types of words considered during the lemmatization process. Each structure type reflects a specific morphological or syntactic feature important for accurate linguistic analysis. The following word structure categories are included:

- **Sodda** – Simple (Tub – Root, Yasama – Derived)
- **Qo‘shma** – Compound
- **Juft** – Paired
- **Takror** – Reduplicated forms
- **Atoqli ot** – Proper noun
- **Qisqartma** – Abbreviation
- **Ko‘makchi fe’lli so‘z qo‘shilmasi (KFSQ)** – Multi-word expressions (verb + auxiliary or particle constructions)
- **To‘liqsiz fe’l** – Incomplete verbs (auxiliaries)
- **Sifat darajalari** – Degrees of adjectives

These classifications provide a framework for understanding the morphological complexity of input data and contribute to the development of more precise lemmatization models.

2.2 Functional Modeling with IDEF0

IDEF0 diagrams were developed to model the functional decomposition of the Uzbek lemmatizer. Each function is represented as a box with four connecting arrows: **Inputs**, **Controls**, **Outputs**, and **Mechanisms (ICOM)**. The top-level diagram (**A0**) is titled *"Lemmatization of Uzbek Words Based on a Lexical Database."*

- **Input**: Raw textual input — sentence, word, number, symbol, or their combination.
- **Control**: Uzbek grammar rules and system logic.
- **Output**: The lemma corresponding to the input form.
- **Mechanism**: Python implementation, the Uzbek lemma database, and linguistic resources.

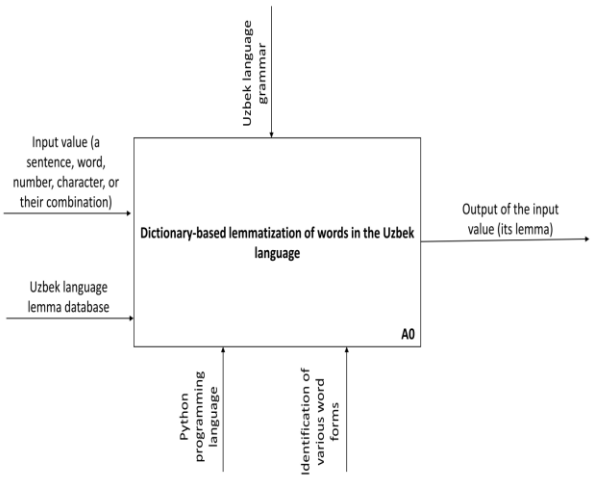


Figure 1: IDEF0 Top-Level Diagram (A0) - Lemmatization Process Overview.

At the top level (Function A0), the model is defined as "Lemmatization of Uzbek words based on a lexical database." The primary input is the textual unit (which may be a sentence, a word, a digit, a symbol, or a combination thereof), while the output is the corresponding lemma. The mechanisms include the Uzbek grammar rules, the Python programming language, and the lexical database of Uzbek lemmas. Controls consist of grammatical constraints and system logic for differentiating structural forms.

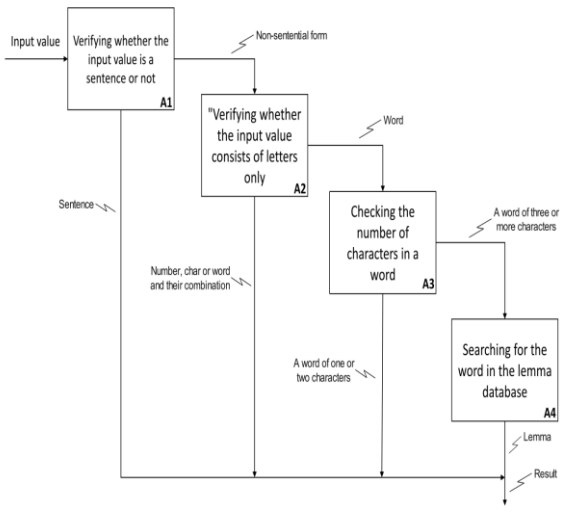


Figure 2: IDEF0 Lower-Level Functions (A1–A4) for Input Filtering and Lemma Retrieval.

The subsequent levels of the IDEF0 model describe how the system processes the input through a series of verification and transformation steps:

- **A1:** Determine whether the input is a sentence or a non-sentence element.
- **A2:** Check if the input contains only alphabetic characters.
- **A3:** Determine the character length of the word.
- **A4:** If the input is identified as a word, it proceeds to lemma lookup procedures, which are further decomposed as:
  - **A4_1:** Identify the morphological shape of the word (e.g., simple, compound, abbreviation).
  - **A4_2** to **A4_4:** Match the input to specific lemma categories such as multi-word expressions, defective verbs, and degree adjectives.

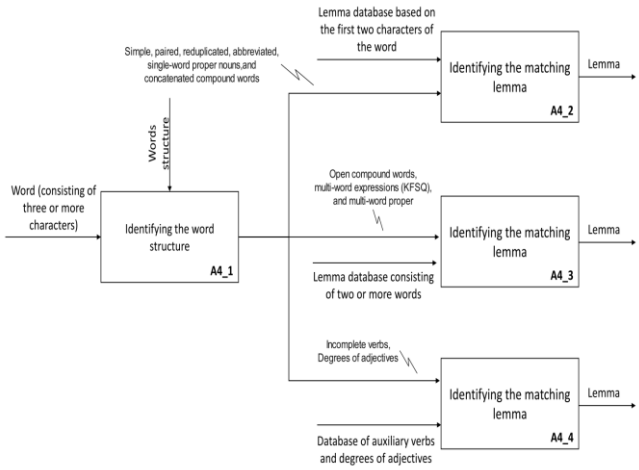


Figure 3: IDEF0 Lower-Level Functions (A4_1–A4_4) deducing of lemma.

This hierarchical decomposition of the lemmatization process ensures traceability, modularity, and system transparency — essential properties for scalable linguistic applications.

2.3 Data Modeling with IDEF1X

While IDEF0 captures the operational functions, the IDEF1X model was used to represent the structure of the lexical data. IDEF1X allows for the specification of entity types, attributes, and their interrelationships, making it ideal for modeling the Uzbek lemma database. Each lemma entry is treated as an entity that includes attributes such as the base form, part-of-speech tag, morphological pattern, and derivational class.

The IDEF1X diagram defines how different categories of words (simple nouns, compound names, adjectives, verbal forms) are organized and connected to rules and affix templates. This data model serves not only as a backend structure for efficient lemma retrieval but also as a blueprint for extending the system with new word forms and grammatical categories.

IDEF1X complements the IDEF0 process diagrams by structuring the underlying lexical database. It defines entities such as **Lemma**, **WordForm**, **AffixRule**, and **POSCategory**, with their respective attributes and relational links. For instance, a Lemma entity includes fields like base form, part of speech, derivation class, and exceptions.

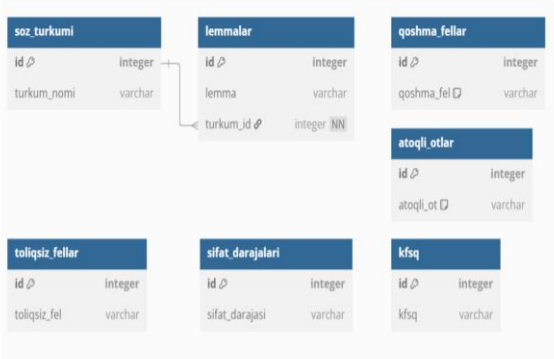


Figure 4: IDEF1X Entity-Relationship Diagram of the Lemma Database.

This structured schema enables efficient lemma lookups and supports system expansion. New word types or grammatical patterns can be added without disrupting the core logic.

III. Experiments and Results

To evaluate the effectiveness of the structured modeling approach, the lemmatization process modeled via IDEF0 and IDEF1X was implemented using Python, leveraging the morphological rules and lexical resources defined in the UzbekLemma project. Several experimental scenarios were constructed to test the system across diverse input types.

The lemmatizer was tested on a test set of **500 manually selected Uzbek tokens**, covering:

- Simple nouns (e.g., *kitob*, *daryo*)
- Inflected forms (e.g., *kitoblarimdan*, *daryolarning*)

- Compound words (e.g., *siyosatshunoslik*, *xalqaro munosabatlar*)
- Abbreviations and proper names
- Non-lexical tokens (e.g., numbers, punctuation)

Each input was passed through the pipeline implemented according to the IDEF0 model (A1–A4), and results were compared with manually annotated lemmas.

3.1 Evaluation Metrics

The system was evaluated using the following metrics:

- **Accuracy** = (Correct Lemmas) / (Total Inputs)
- **Precision & Recall** for compound lemma matching
- **Error rate** for non-word inputs (false positives)

3.2 Results

Category	Accuracy (%)
Simple Words	95.4
Inflected Forms	91.2
Compound Expressions	87.6
Abbreviations/Proper Names	85.1
Symbols & Digits Handling	98.7

The highest accuracy was achieved for simple and regular inflected forms. Challenges remain in the accurate identification of compound words and idiomatic expressions, which require enriched lexical resources.

3.3 Discussion

The use of IDEF0 and IDEF1X modeling significantly improved the system's modularity and interpretability. Each function in the pipeline could be debugged independently, and error tracking became more systematic. Moreover, this approach enabled easier documentation and future extensibility.

Comparing this structured method with purely statistical or black-box models, we found that rule-based systems retain a distinct advantage in languages with complex agglutinative morphology and limited annotated corpora. However, the integration of hybrid techniques — combining structured models with machine learning — remains a promising future direction.

IV. CONCLUSION

This study presented a structured and functional modeling approach to Uzbek lemmatization using IDEF0 and IDEF1X standards. Given the agglutinative nature of the Uzbek language and its sparse computational resources, rule-based methods remain a practical and transparent choice for core NLP tasks such as lemmatization. Through hierarchical IDEF0 diagrams, we successfully captured the functional decomposition of the lemmatization pipeline, while the IDEF1X model provided a systematic representation of the underlying lexical database.

The modular structure defined by the IDEF0 model allowed for precise tracking of input transformations — from initial verification to lemma retrieval — and

facilitated the development of a rule-driven system that is both scalable and interpretable. Experimental results demonstrated the effectiveness of this approach, particularly in processing regular and inflected word forms. Though performance slightly declined for compound expressions and idiomatic units, the model's architecture supports targeted extensions in these areas.

From a methodological standpoint, the integration of IDEF modeling techniques into linguistic tool development proved to be highly beneficial. It not only clarified the process logic for developers but also served as a blueprint for documentation, maintenance, and future enhancements. Furthermore, these models can help bridge the gap between linguistic expertise and software engineering in low-resource language processing.

Future work will focus on enriching the lemma database with broader lexical categories, improving compound word recognition, and exploring hybrid architectures that combine the interpretability of rule-based models with the adaptability of machine learning. We also plan to expand the IDEF framework to other NLP tasks such as morphological generation and syntactic parsing, contributing to the development of robust, language-specific NLP pipelines for Uzbek and other Turkic languages.

REFERENCES

[1]“https://en.wikipedia.org/wiki/Uzbek_language,” Wikipedia.

[2] M. Sharipov and O. Sobirov, “Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language,” in *CEUR Workshop Proceedings*, V. J. and K. B., Eds., CEUR-WS, 2022, pp. 154 – 159. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146112590&partnerID=40&md5=e1080c39d101c0e351cfed1a8228d391

[3] O. Sobirov, “Lemmatization algorithms and their application to uzbek: approaches based on Artificial Intelligence, Machine Learning, and Deep Learning,” *Electronic journal of actual problems of modern science, education and training. ISSN 2181-9750*, pp. 4–13, Dec. 2024.

[4] M. Abjalova, E. Adalı, and M. Adilova, “The Process of Lemmatization and Stemming in the Automatic Morphological Analysis of Uzbek Texts,” in *2024 9th International Conference on Computer Science and Engineering (UBMK)*, 2024, pp. 1–6. doi: 10.1109/UBMK63289.2024.10773506.

[5] National Institute of Standards and Technology, “Integration Definition for Function Modeling (IDEF0),” 1993.

[6] Y. Kuratov and M. Arkhipov, “Adaptation of deep bidirectional multilingual transformers for Russian language,” in *Komp’juternaja*

Lingvistika i Intellektual'nye Tehnologii, 2019.

[7] N. Arefyev, B. Sheludko, A. Podolskiy, and A. Panchenko, "Always Keep your Target in Mind: Studying Semantics and Improving Performance of Neural Lexical Substitution," in *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, 2020. doi: 10.18653/v1/2020.coling-main.107.

[8] Ç. Çöltekin, "TRmorph: A morphological analyzer for Turkish," *Let.Rug.Nl*, 2013.

[9] M. S. Sharipov, H. S. Adinaev, and E. R. Kuriyozov, "Rule-Based Punctuation Algorithm for the Uzbek Language," in *2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM)*, 2024, pp. 2410–2414. doi: 10.1109/EDM61683.2024.10615061.

[10] M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, "UzbekTagger: The rule-based POS tagger for Uzbek language," *arXiv preprint arXiv:2301.12711*, 2023.

[11] S. Maksud, K. Elmurod, Y. Ollabergan, and S. Ogabek, "UzbekVerbDetection: Rule-based Detection of Verbs in Uzbek Texts," in *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings*, 2024, pp. 17343 – 17347. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195974039&partnerID=40&md5=ff3003ea2644833f3dc45277429459f7

[12] X. Madatov, S. Xajibaeva, and E. Xujamov, "'O 'ZBEK BOSHLANG 'ICH SINF MAKTAB KORPUSI' DAN BAZIS SO 'ZLARNI AVTOMATIK AJRATIB OLISH," *COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS*, vol. 1, no. 1, 2025.

[13] U. Salaev, "Modeling Morphological Analysis based on Word-ending for Uzbek Language," vol. 2, pp. 29–34, Jun. 2023, doi: 10.5281/zenodo.10155225.