

UDC 004.8:81'322

**AN INTELLIGENT INFORMATION SYSTEM FOR UZBEK TEXT PROCESSING
AND GRAMMATICAL QUESTION-ANSWERING**

Sharipov Maksud Siddiqovich – *Candidate of Technical Sciences, independent researcher*
maqsbek72@gmail.com

Tashkent University of Information Technologies named after Muhammad al-Khwarizmi

**O‘ZBEK TILI MATNLARIGA INTELLEKTUAL ISHLOV BERISH
VA GRAMMATIK SAVOL-JAVOB AXBOROT TIZIMI**

Sharipov Maqsud Siddiqovich – *texnika fanlari nomzodi, mustaqil izlanuvchi*
Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti

**ИНФОРМАЦИОННАЯ СИСТЕМА ДЛЯ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ
ТЕКСТОВ НА УЗБЕКСКОМ ЯЗЫКЕ И ГРАММАТИЧЕСКОГО
ВОПРОСНО-ОТВЕТНОГО ВЗАИМОДЕЙСТВИЯ**

Шарипов Максуд Сиддиқович – *кандидат технических наук, самостоятельный соискатель*
Ташкентский университет информационных технологий имени Мухаммада аль-Хорезми

Tayanch so‘zlar: o‘zbek tili NLP, ByT5, Viterbi, mDeBERTa-v3, CRF, gibrid RAG, BM25.

Rezyume. “UZAI Tahlil” dasturiy majmuasi o‘zbek matnlarini morfologik, sintaktik va leksik-statistik tahlil qilishni manbaga tayangan grammatik savol-javob bilan birlashtiradi. Tizim UzbekTaggerBERT, ByT5, Viterbi morfotaktik tekshiruvi, mDeBERTa-v3 + CRF hamda BM25/MPNet gibrid RAG komponentlaridan foydalanadi. Tajribada morfotaktik tahlil F1 = 96,8%, sintaktik klassifikatsiya Macro F1 = 91,4% va grammatik qidiruv Hit@4 = 97,1% natijani ko‘rsatdi.

Ключевые слова: узбекский NLP, ByT5, Витерби, mDeBERTa-v3, CRF, гибридный RAG, BM25.

Резюме. Программный комплекс «UZAI Tahlil» объединяет морфологический, синтаксический и лексико-статистический анализ узбекских текстов с грамматическим вопросно-ответным поиском по проверенным источникам. Система использует UzbekTaggerBERT, ByT5, морфотактическую проверку Витерби, mDeBERTa-v3 + CRF и гибридный RAG BM25/MPNet. Эксперименты показали F1 = 96,8% для морфотактики, Macro F1 = 91,4% для синтаксической классификации и Hit@4 = 97,1% для грамматического поиска.

Key words: Uzbek NLP, ByT5, Viterbi, mDeBERTa-v3, CRF, hybrid RAG, BM25.

Summary. “UZAI Tahlil” integrates morphological, syntactic and lexical-statistical processing of Uzbek texts with source-grounded grammatical question answering. The system combines UzbekTaggerBERT, ByT5, Viterbi morphotactic validation, mDeBERTa-v3 + CRF, and hybrid BM25/MPNet retrieval-augmented generation. Experiments achieved F1 = 96.8% for morphotactics, macro F1 = 91.4% for syntactic classification, and Hit@4 = 97.1% for grammar retrieval.

Introduction. Uzbek’s agglutinative morphology, long affix chains, morphophonemic alternations and homonymous suffixes complicate automatic analysis. Existing Uzbek NLP work spans integrated resource surveys [1:2], linguistic modules for automatic analysis [2:15], explicit morphotactics [3:84], truncation-based morphology [4:190], UzMorphAnalyser [5:2] and lemmatization/stemming methods [6:10]. Together, these results motivate a neuro-symbolic design that combines learned representations with explicit linguistic constraints.

Contextual and structured models provide the second basis of the system: UzbekPOS supplies verified multi-domain POS data [7:3], UzUDT supports dependency analysis [8:11643], ByT5 enables byte-level sequence generation [9:292], and DeBERTaV3 improves contextual representations [10:3]. Viterbi decoding [11:269] and CRF sequence modeling [12:283] are used to enforce globally consistent morphological and syntactic decisions.

For grammar question answering, RAG grounds generation in external evidence [13:9460], BM25 provides lexical relevance [14:345], and Sentence-BERT supports dense semantic retrieval [15:3983]. Lexical richness is measured at token, stem and lemma levels with awareness of TTR’s length sensitivity [16:95]. This study integrates these components in the “UZAI Tahlil” suite and evaluates the resulting end-to-end pipeline.

Materials and methods. The “UZAI Tahlil” platform uses a multilayer Django MVT architecture (Fig. 1): query/session handling feeds a linguistic path (normalization, POS, morphology, syntax) and a retrieval path (BM25, ChromaDB, metadata), after which the generator returns a source-grounded response. Relational entities are listed in Table 1.

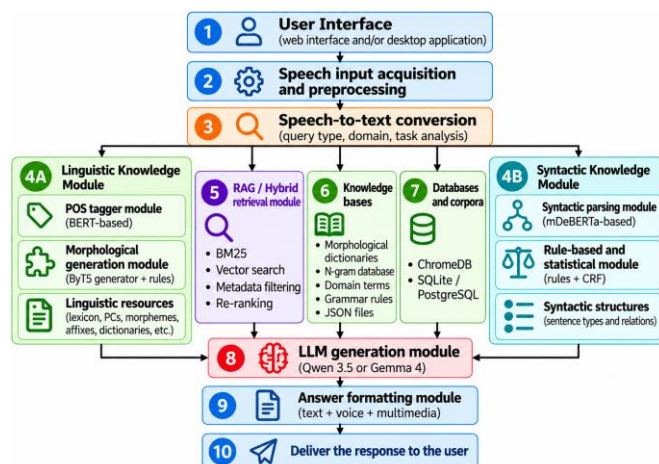


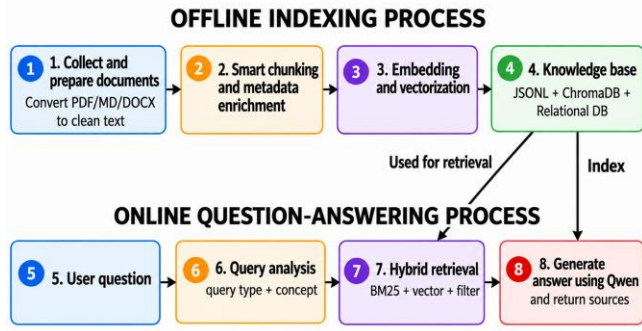
Fig. 1. Functional module architecture of the AI information system for Uzbek grammar.

The two paths share session, storage and response layers.

Table 1. Main relational database models of the software suite

Model	Appli-cation	Purpose and main fields
User & Profile	accounts	User account, role (student, teacher, researcher), organization, and demographic information
Convers ation	analyzer	Morphological and syntactic analysis session title, user ForeignKey, and creation time
Message	analyzer	Analysis messages (user/assistant), input text, and structured result stored as JSONField
QnaSess ion	qna_chat	AI grammar question-answering session, title, and user relation
QnaMes sage	qna_chat	Grammar question/answer text, query type, and sources stored as JSONField

Offline, grammar sources are cleaned, structurally chunked, metadata-enriched and indexed in JSONL/BM25 and ChromaDB; online, the query is classified and Top-4 evidence is retrieved for generation (Fig. 2).


Fig. 5.2. Data flow in the offline indexing and online question-answering modes.
Fig. 2. Data flow in the offline indexing and online question-answering modes.

Metadata enables concept/domain filtering and reranking.

The morphological pipeline performs normalization, tokenization, contextual POS tagging, ByT5 stem-affix segmentation, interpretation with 360 allomorph mappings, Viterbi path selection and a rule-based fallback. ByT5 estimates:

The ByT5 sequence-to-sequence model estimates the segmented output autoregressively:

$$P(Y|X) = \prod_{j=1}^N P(y_j | y_{<j}, X), \quad \hat{Y} = \operatorname{argmax}_Y P(Y|X)$$

For the resulting segment sequence, candidate grammatical tags are generated for each segment:

$$S = [s_0, s_1, s_2, \dots, s_k], \quad C(s_j) = \{g_{j,1}, g_{j,2}, \dots, g_{j,m}\}$$

For a tag sequence $G = [g_0, g_1, \dots, g_k]$, the Viterbi objective is:

$$\operatorname{Score}(G) = \sum_{j=1}^k \operatorname{Transition}(g_{j-1}, g_j) + \sum_{j=0}^k \operatorname{Intrinsic}(g_j)$$

The compatibility of adjacent morphemes is scored from their category slots and rule constraints:

$$\operatorname{Transition}(g_{j-1}, g_j) = \begin{cases} +18, & \text{if } \operatorname{Slot}(g_{j-1}) \leq \operatorname{Slot}(g_j) \text{ [correct morphotactic order]} \\ -45, & \text{if } \operatorname{Slot}(g_{j-1}) > \operatorname{Slot}(g_j) \text{ [strict penalty: order violation]} \\ +4, & \text{for weak morphotactic transitions} \\ -15, & \text{for unknown transitions} \end{cases}$$

The dynamic-programming layer computes the best partial path for candidate j at step i and reconstructs the optimum by backtracking:

$$\operatorname{DP}[i][j] = \max_p (\operatorname{DP}[i-1][p] + \operatorname{Transition}(g_{i-1,p}, g_{i,j})) + \operatorname{Intrinsic}(g_{i,j})$$

$$\hat{p}_k = \operatorname{argmax}_j \operatorname{DP}[k][j], \quad \hat{p}_{i-1} = \operatorname{Backpointer}[i][\hat{p}_i]$$

The family order is Stem $\rightarrow$ Voice $\rightarrow$ Negation $\rightarrow$ Participle/Converb $\rightarrow$ Tense $\rightarrow$ Person/Possessive $\rightarrow$ Case $\rightarrow$ Copula $\rightarrow$ Particle (slots 0–100).

UzbekSyntacticAnalyzer concatenates fine-tuned mDeBERTa-v3 representations with 12 linguistic features, then applies Dropout, a linear layer and CRF decoding over 33 BIO labels.

Seven post-processing rules correct modal/nominal predicates, copular predicates, accusative objects and converb + auxiliary constructions.

The qna_chat corpus contains grammar textbooks (grades 5–11), the Universal Handbook 2024 and monographs. Typed chunks carry concept/domain metadata; retrieval combines BM25, vector similarity and metadata boosting:

In the hybrid retrieval system, lexical BM25 and vector semantic similarity are combined using the following integrated ranking and metadata-boosting function:

$$\operatorname{Score}_{\text{final}}(d, q) = \alpha \cdot \operatorname{Score}_{\text{dense}}(d, q) + (1 - \alpha) \cdot \operatorname{Score}_{\text{BM25}}(d, q) + \operatorname{Boost}_{\text{metadata}}(d, q)$$

Here $\alpha = 0.65$ balances semantic-vector similarity and lexical BM25 relevance. BM25 inverse document frequency is calculated as:

$$\operatorname{IDF}(t) = \ln \left(1 + \frac{N - n(t) + 0.5}{n(t) + 0.5} \right)$$

With $\alpha = 0.65$, metadata adds +0.35 for concept, +0.20 for chunk type and +0.15 for domain; Top-4 passages are sent to Qwen/Qwen3.5-4B.

The corpus_statistics module computes word-form, stem and lemma TTR for texts up to 5,000 words:

The three Type-Token Ratio measures are calculated as:

$$\begin{aligned} \operatorname{TTR}_{\text{token}} &= \frac{V_{\text{token}}}{N} \times 100\%, & \operatorname{TTR}_{\text{stem}} &= \frac{V_{\text{stem}}}{N} \times 100\%, & \operatorname{TTR}_{\text{lemma}} &= \frac{V_{\text{lemma}}}{N} \times 100\% \end{aligned}$$

Agglutinative inflection and derivational activity are additionally summarized by the following differences:

$$\begin{aligned} D_{\text{inflect}} &= \operatorname{TTR}_{\text{token}} - \operatorname{TTR}_{\text{lemma}}, & D_{\text{derive}} &= \operatorname{TTR}_{\text{lemma}} - \operatorname{TTR}_{\text{stem}} \end{aligned}$$

Lemma TTR is grouped as A ($\geq 65\%$), B (45–65%), C (30–45%) and D ($< 30\%$).

Results. Evaluation used project test corpora, 02_bio_dataset_latin.conll and 350 grammar questions.

ByT5 + Viterbi was compared with rule-based and neural-only baselines (Table 2).

Table 2. Comparative performance of morphological-analysis and segmentation approaches

Model / Approach	Segmentation accuracy (%)	Homonymous-affix disambiguation (%)	Morphotactic validity (%)	F1-score (%)
Rule-based (Baseline)	81.2%	74.5%	86.0%	77.8%
Pure ByT5 segmenter	94.1%	79.8%	81.3%	86.4%
ByT5 + Viterbi morphotactics (Proposed)	97.6%	96.2%	98.9%	96.8%

UzbekSyntacticAnalyzer was evaluated on 02_bio_dataset_latin.conll (Table 3).

Table 3. Sentence-component classification results for UzbekSyntacticAnalyzer

Sentence component (syntactic role)	Tags	Precision (%)	Recall (%)	F1-score (%)
Subject	B-EG, I-EG	93.4 %	91.8 %	92.6 %
Verbal predicate	B-FK, I-FK	95.8 %	94.2 %	95.0 %
Nominal predicate	B-OK, I-OK	89.2 %	87.5 %	88.3 %
Direct object	B-VS, I-VS	92.1 %	90.6 %	91.3 %
Indirect object	B-VL, I-VL	88.4 %	86.9 %	87.6 %
Adverbial (all types)	OH, PH, MH, DH, SAH, VH	89.7 %	88.1 %	88.9 %
Modifier / Attribute	QA, SA, IA	94.5 %	92.7 %	93.6 %
Macro average	All tags	92.3%	90.5%	91.4%

For 350 grammar questions, lexical, dense and hybrid retrieval were compared (Table 4).

Table 4. Retrieval and answer-generation performance of the hybrid RAG system

Retrieval method	Hit@1 (%)	Hit@4 (%)	MRR	Context relevance (%)	Hallucination (%)
BM25 only (Lexical)	68.2 %	81.4 %	0.742	76.5 %	18.2 %
Dense MPNet only (ChromaDB)	72.5 %	85.1 %	0.789	81.0 %	14.5 %
Hybrid (BM25 + MPNet + Boosting)	88.9 %	97.1 %	0.915	94.8 %	3.1 %

Discussion. Neural segmentation alone does not guarantee valid Uzbek morphotactics; Viterbi constraints raise the combined model to F1 = 96.8%.

mDeBERTa-v3 + linguistic features + CRF achieved macro F1 = 91.4%. Hybrid BM25 + MPNet + metadata reached Hit@4 = 97.1%, context relevance = 94.8% and hallucination = 3.1%.

These findings support neuro-symbolic integration of learned representations, explicit constraints and retrieved evidence.

Conclusion. “UZAI Tahlil” integrates Uzbek text processing and source-grounded grammar QA. Its developed algorithms, methods, AI models and linguistic rules support morphological and syntactic analysis of Uzbek texts and evaluation of an author’s writing style, while the RAG module answers questions on Uzbek grammar using retrieved sources. Results are F1 = 96.8% for ByT5 + Viterbi, macro F1 = 91.4% for mDeBERTa-v3 + features + CRF, and Hit@4 = 97.1% with 3.1% hallucination; Token/Stem/Lemma TTR is also computed in the Django MVT platform.

References

- Sharipov M., Avezmatov I., Kuriyozov E., Babayev S. Natural Language Processing for Uzbek and Turkmen: A Systematic Review of Resources, Methods, and Benchmarks. // 2025 IEEE XVII International Scientific and Technical Conference on Actual Problems of Electronic Instrument Engineering (APEIE). IEEE, 2025. – P. 1-6.
- Abjalova M. Tahrir va tahlil dasturlarining lingvistik modullari. – Tashkent: «Nodirabegim», 2020.
- Matlatipov G., Vetulani Z. Representation of Uzbek Morphology in Prolog. // Aspects of Natural Language Processing. LNCS 5070. Springer, 2009. – P. 83-110.
- Sharipov M., Salaev U., Matlatipov G. Morphological Analyzer of Uzbek Language Using Truncating Methods // Ilm Sarchashmalari. – Urgench, 2023. № 5/1. – P. 189-193.
- Salaev U. UzMorphAnalyser: A Morphological Analysis Model for the Uzbek Language Using Inflectional Endings. // AIP Conference Proceedings. 2024. Vol. 3244(1), 030058. – P. 1-7.
- Abjalova M. O‘zbek tilidagi matnlarni avtomatik morfologik tahlil qilishda lemmatizatsiya va stemming jarayoni. // O‘zbekiston: til va madaniyat. Lingvistika. 2024. № 3. – P. 6-21.
- Sharipov M., Kuriyozov E., Vičič J. UzbekPOS: A Multi-Domain Dataset for Uzbek Part-of-Speech Tagging. // Data in Brief. 2026. Vol. 66, Article 112640.
- Matlatipov S.G., Aripov M. UzUDT: Uzbek Universal Dependencies Treebank. // Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026). 2026. – P. 11642-11649.
- Xue L., Barua A., Constant N., et al. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. // Transactions of the Association for Computational Linguistics. 2022. Vol. 10. – P. 291-306.
- He P., Gao J., Chen W. DeBERTaV3: Improving DeBERTa Using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. // ICLR. 2023. – P. 1-17.

11. Forney G.D. The Viterbi Algorithm. // Proceedings of the IEEE. 1973. Vol. 61(3). – P. 268-278.
12. Lafferty J., McCallum A., Pereira F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data // Proceedings of ICML. 2001. – P. 282-289.
13. Lewis P., Perez E., Piktus A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. // Advances in Neural Information Processing Systems. 2020. Vol. 33. – P. 9459–9474.
14. Robertson S., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. // Foundations and Trends in Information Retrieval. 2009. Vol. 3(4). – P. 333-389.
15. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. // Proceedings of EMNLP-IJCNLP. 2019. – P. 3982-3992.
16. Covington M.A., McFall J.D. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). // Journal of Quantitative Linguistics. 2010. Vol. 17(2). – P. 94-100.