УДК 004.932

Arabov M.K.
Candidate of Physical and Mathematical Sciences, Associate Professor,
Kazan (Volga Region) Federal University,
Institute of Computational Mathematics and Information Technologies,
Department of Data Analysis and Programming Technologies
420008, Kazan, Kremlevskaya St., 35.
E-mail: MKArabov@kpfu.ru

DEVELOPING THE TAJIK LANGUAGE IN THE ERA OF LARGE LANGUAGE MODELS: CORPUS INFRASTRUCTURE, LINGUISTIC CHALLENGES, AND SAFETY ALIGNMENT

Abstract.

The rapid progress of large language models (LLMs) has reshaped natural language processing, yet this progress has reinforced existing inequalities between high-resource and low-resource languages. Tajik, despite its long-standing literary tradition and official status, remains largely absent from contemporary LLM ecosystems. At the present stage, the language lacks publicly accessible, standardised and computationally usable corpora and datasets suitable for training, adaptation or evaluation of modern language models. Although a National Corpus of the Tajik Language is often cited, its internal structure, annotation formats and access conditions do not allow its effective use in reproducible NLP research. This paper adopts a theoretical and infrastructural perspective and analyses the structural reasons for this situation. The study identifies three interrelated domains that constrain the development of Tajik LLM technologies: data availability and quality, linguistic representation, and research infrastructure. Particular attention is paid to the discrepancy between classical linguistic proximity and functional technological compatibility, especially with respect to cross-lingual transfer from Persian. The paper does not present new datasets or empirical experiments; instead, it formulates a conceptual framework and preparatory research agenda intended to guide future corpus construction, linguistic preprocessing and safety-aware model adaptation for the Tajik language.

Ключевые слова: таджикский язык; большие языковые модели; языки с низким уровнем ресурсов; корпусная инфраструктура; морфологическая богатство; токенизация; переключение кодов; языковая безопасность; детоксикация; цифровое неравенство.

Keywords: tajik language; large language models; low-resource languages; corpus infrastructure; morphological richness; tokenisation; code-switching; language safety; detoxification; digital inequality.

1. Relevance

The relevance of this study is determined by the increasing role of large language models as a foundational technology for education, research, public administration and access to information, alongside the persistent exclusion of many low-resource languages from these systems. At present, the Tajik language is effectively absent from the global LLM landscape. It lacks publicly available, standardised and computationally usable corpora and datasets that could support training, adaptation or evaluation of modern language models. Although the National Corpus of the Tajik Language is frequently cited as a major linguistic resource, it does not provide open, machine-accessible data or documented interfaces suitable for reproducible NLP research. Consequently, Tajik remains without a functional corpus infrastructure that meets contemporary scientific and technological requirements.

An additional source of relevance arises from a widespread but methodologically oversimplified assumption that the close linguistic relationship between Tajik and Persian allows direct reuse of Persian language models and resources. While these languages share a common grammatical foundation and historical continuity, their modern functional registers have diverged significantly. Contemporary Tajik scientific, technical and administrative discourse has been shaped predominantly by Russian-mediated terminology, whereas modern Persian relies largely on Arabic, English and French borrowings. This divergence limits the effectiveness of direct cross-lingual transfer and exposes a critical gap between classical linguistic proximity and practical technological compatibility. As a result, uncritical reliance on donor-language models risks producing systems that perform poorly in key applied domains and generate inaccurate or inappropriate outputs.

Under these conditions, a theoretically grounded analysis that systematises existing limitations and formulates a coherent research agenda is a necessary preliminary step. Without such a framework, empirical work risks being fragmented, non-reproducible and disconnected from linguistic and socio-cultural realities. The present study addresses this need by articulating an infrastructure-oriented perspective that clarifies priorities, dependencies and design constraints for the future integration of the Tajik language into large language model ecosystems, thereby contributing to its long-term technological visibility and sustainability.

2. Related Work

Recent scholarship on language technologies has emphasised two complementary themes that are particularly salient for low-resource languages: first, the limits of scale as a remedy for representational and social harms; and second, the conditional utility of cross-lingual transfer as a practical strategy for capacity building. Foundational critiques argue that increasing dataset size and model capacity does not by itself eliminate ethical, factual or representational problems, and may indeed amplify existing biases if provenance, documentation and filtering are neglected [1]. Empirical and community-oriented work on dataset infrastructure has reinforced this argument by demonstrating the necessity of transparent curation, versioning and licensing practices for reproducible model development and for meaningful cross-study comparisons [2,3]. Together, these lines of work provide a methodological rationale for prioritising corpus quality, provenance metadata and repeatable pipelines in any programme that seeks to deploy large language models for under-resourced languages.

Multilingual pretraining and cross-lingual representation learning have emerged as central technical tools for addressing data scarcity. Studies of large-scale unsupervised cross-lingual models show that shared representational structures can enable zero- or few-shot transfer across typologically related languages, yielding practical gains on a variety of downstream tasks [4,6]. Nonetheless, a substantial empirical literature documents important caveats: zero-shot transfer often degrades on tasks that require fine-grained cultural or pragmatic knowledge, and performance collapses more readily when source and target languages diverge in script, morphology or domain-specific terminology [7]. Work on Persian — including the development of language-specific pretrained models and treebanks — illustrates both the opportunities and the limits of donor-language strategies: where lexical, morphological and script mappings are reliable, transfer can be effective; where terminological or orthographic mismatches occur, naive transfer introduces systematic errors that require targeted correction [12,17,18,21].

Concurrently, methodological advances have focused on lowering the technical barrier to language adaptation. Parameter-efficient fine-tuning approaches permit substantial model specialisation with a modest number of additional parameters, enabling feasible adaptation in resource-constrained settings [5]. Instruction tuning and human-feedback paradigms have improved controllability and alignment for interactive systems [9], and recent proposals for direct preference optimisation offer an alternative route to incorporate comparative judgements without a separate reward model [16]. These adaptation techniques are promising for low-resource contexts, but their practical success depends on the availability of language-specific evaluation data and culturally informed preference annotations; absent such resources, alignment mechanisms risk inheriting the blind spots of donor or multilingual models.

Research addressing model safety has likewise matured along multiple axes. Data-level interventions (filtering, targeted augmentation), adversarial evaluation (red-teaming) and stress-testing have been proposed and empirically studied as complementary measures to detect and mitigate harmful behaviours in model outputs [8]. Investigations into cross-lingual generalisation for safety-sensitive tasks (for example, hate-speech and toxicity detection) consistently find that zero-shot classifiers perform poorly when cultural nuance and local pragmatic conventions are decisive, thereby underscoring the need for language-specific annotated datasets for safety evaluation and mitigation [7]. These findings imply that safety cannot be treated as a downstream add-on but must be integrated into corpus design and adaptation workflows from the outset.

Work specific to Tajik and closely related languages is comparatively sparse and fragmented, but offers relevant linguistic and technical foundations. National and regional studies have documented the systemic difficulties of digitising Tajik and compiling representative text collections, emphasising orthographic heterogeneity and the uneven availability of machine-readable sources [10,13]. Linguistic analyses and dissertations provide formal descriptions of Tajik morphology and affix inventories, which are essential inputs for any morphologically aware preprocessing or tokenisation strategy [14]. Applied projects — such as spelling and orthography correction systems — demonstrate that rule-based and hybrid approaches can address concrete normalisation problems, though they also reveal the gap between task-specific tools and the demands of large-scale pretraining [15]. Research on automatic transliteration and script conversion between Persian and Tajik scripts addresses a further practical obstacle to corpus aggregation, particularly for historical and cross-regional texts, and highlights the non-trivial engineering required to reconcile heterogeneous orthographies [20]. Studies on data augmentation and synthetic resource generation for low-resource languages propose techniques that may expand effective training material, while simultaneously warning about the risk of introducing artefacts that degrade downstream generalisation [19].

The aggregate picture emerging from this literature is clear. Empirical and theoretical work cautions that scale without rigorous curation is insufficient and may be counterproductive [1,2,3]. Cross-lingual transfer holds practical promise but is highly contingent on lexical, morphological and scriptual compatibilities as well as on terminological alignment between donor and target languages [4,6,7,12,17,18,21]. Parameter-efficient adaptation and preference-based alignment lower computational barriers [5,9,16], yet their effectiveness is constrained by the scarcity of language-specific evaluation and safety datasets. Local studies provide necessary linguistic description and a handful of applied tools [10,13–15,20], but do not yet constitute an integrated, publicly accessible infrastructure that would support reproducible LLM research for Tajik. These persistent gaps—particularly the absence of standardised, richly annotated corpora and of benchmarks for culturally specific safety tasks—frame the central empirical and infrastructural challenges that follow in subsequent sections.

3. Methodology and Strategic Programme for Overcoming Key Limitations

The comparative analysis of existing scholarship and available resources allows the systemic constraints affecting the development of Tajik language technologies to be organised into three interdependent groups. The combined effect of these constraints produces a “digital barrier” that impedes the integration of Tajik within contemporary language-technology ecosystems. This barrier is paradoxical in light of a strategic opportunity afforded by the genetic and structural proximity of Tajik to Persian (Farsi). The established NLP infrastructure for Persian, however, has not automatically catalysed Tajik development owing to a constellation of infrastructural, linguistic and socio-technical divergences.

3.1. Critical data deficit: volume, heterogeneity and annotation

Existing digital corpora for Tajik fall markedly short of the conditions typically required for pretraining or robust adaptation of modern large language models, both in quantitative scale and in qualitative properties [1,3,4,10,13]. Available resources are often orders of magnitude smaller than the corpora commonly employed for large-scale pretraining and lack standardised, machine-readable formats and comprehensive provenance metadata [3,4]. Qualitative deficiencies are equally consequential. Historical orthographic variation and the persistence of multiple writing conventions demand careful normalisation; pervasive intra-textual code-switching (notably with Russian) fragments language signals and complicates language-identification and segmentation; and the near absence of systematic genre, domain and register labelling prevents controlled sampling and domain-aware training. In this context, naïve aggregation of heterogeneous materials risks producing noisy pretraining signals and amplifying artifacts rather than improving downstream capabilities [1,3].

3.2. Structural mismatch: morphology and subword tokenisation

Tajik exhibits rich inflectional morphology and productive derivation that increase surface-form variability. Dominant subword tokenisation schemes (BPE, WordPiece and related algorithms) were largely developed with languages exhibiting less synthetic morphology in mind and therefore can be suboptimal for Tajik [4,6,17]. The interaction between aggressive subword splitting and high morpheme productivity yields several operational problems: an inflated effective vocabulary, elevated rates of rare or out-of-vocabulary tokens in specialised domains, and inefficient utilisation of model context windows where grammatical morphemes occupy capacity that would be better allocated to semantically salient lexical items. These phenomena argue for the systematic evaluation of morphologically informed or hybrid tokenisation strategies prior to large-scale adaptation, and for empirical comparison of such strategies on Tajik data where feasible [4,17,21].

3.3. Infrastructure vacuum for safety and cultural alignment

Perhaps the most acute infrastructural gap concerns resources and processes for safety, cultural alignment and governance. Beyond the lack of basic annotated corpora for syntactic and semantic tasks, there is an absence of curated datasets for detoxification, bias measurement and culturally contextualised evaluation that reflect local socio-historical realities [7,8,19,20]. The lack of such resources prevents the establishment of a responsible development cycle: without language-specific safety benchmarks and expert-mediated annotations, direct deployment or uncritical adaptation of donor models risks producing outputs that are culturally inappropriate, misleading or harmful. This deficit thus constitutes both a technical and an ethical impediment to meaningful LLM work for Tajik.

3.4. Strategic opportunity and the conditions for cross-lingual transfer

The typological proximity of Tajik and Persian creates a significant strategic opportunity: Persian resources, models and toolchains can potentially serve as donor assets for Tajik adaptation [12,17,18,21]. The Persian ecosystem—comprising pretrained models, treebanks and annotated corpora—demonstrates the practical gains available when donor resources are well developed. Nevertheless, realising this potential depends on addressing two principal obstacles. First, a cross-script barrier exists (Cyrillic Tajik versus Perso-Arabic script), which necessitates robust, reversible script conversion and transliteration methods to enable corpus aggregation and aligned modelling [20]. Second, terminological divergence in modern technical and administrative registers reduces lexical equivalence in domain-specific contexts and requires explicit lexicon mapping and terminological alignment to avoid systematic semantic mismatches [10,20].

3.5. Lexico-terminological divergence as a semantic barrier

In spite of shared grammatical foundations, Tajik and Iranian Persian have undergone divergent contact histories that have produced distinct terminological strata in scientific, technical and administrative domains. Tajik technical lexicon has been substantially influenced by Russian-mediated forms and calques, whereas Persian contemporary registers incorporate Arabic and Western loanwords and calques. This divergence creates not merely orthographic or surface-level differences but substantive semantic mismatches that manifest in vector-space representations and in downstream task performance; consequently, transfer methods must incorporate lexicon-level alignment and domain-aware mapping procedures to be effective in applied settings [10,12,20].

3.6. Priority I: building a representative mega-corpus as a public good

The primary strategic priority is the construction of a large, representative, normalised and openly documented mega-corpus for Tajik (a Tajik LLM MegaCorpus, target scale ≥ 1e9 tokens as a planning aspiration). The corpus should be conceived and delivered as a public digital good with a reproducible ingestion and processing pipeline. Essential components of this pipeline include standardised orthographic normalisation under expert supervision; explicit handling and annotation of code-switching; multi-dimensional metadata (source, genre, date, register and licence); and robust deduplication and quality filtering (for example, locality-aware hashing and semantic deduplication). All stages must be accompanied by provenance tracking and versioned releases to support reproducibility and accountable reuse [3].

3.7. Priority II: establishing core benchmark datasets for evaluation and safety

Parallel to corpus construction, a curated suite of high-quality annotated resources is required to serve as benchmarking and evaluation material. This Tajik Benchmark Suite should include gold-standard annotations for morphological analysis, POS tagging, lemmatisation, dependency syntax and named-entity recognition, together with a dedicated Tajik Safety & Alignment Dataset that models culturally specific harms, biased narratives and adversarial prompts. The safety dataset must be developed in collaboration with native-speaking experts and social scientists and must support paired preference annotations to enable supervised alignment techniques [7,8,16,19].

3.8. Priority III: adapted modelling methodology and transfer protocol

The modelling programme must reconcile the constraints enumerated above through a staged transfer and adaptation protocol. Recommended components are: comparative evaluation of tokenisation strategies (morphologically informed segmentation, hybrid subword/morpheme schemes, and byte-level fallbacks) to identify the most effective preprocessing pipeline for Tajik; a three-stage adaptation protocol comprising (i) donor initialisation using suitable Persian or multilingual checkpoints, (ii) consolidation via further unsupervised pretraining on the Tajik mega-corpus, and (iii) targeted specialisation using parameter-efficient fine-tuning methods (e.g. LoRA or adapter modules) to reduce computational cost while permitting rapid iteration [5,12,17]. Safety and cultural alignment must be integrated throughout this lifecycle: from data-level controls and targeted augmentation to preference-based alignment and post-hoc evaluation. Where comparative preference data are available, methods such as Direct Preference Optimisation offer a principled mechanism to incorporate annotated preferences without an intermediate reward model [16]; adversarial testing and red-teaming should be institutionalised as part of the validation regime [8].

4. Scope and Limitations

This study primarily addresses the theoretical and infrastructural foundations necessary for the development of Tajik large language models (LLMs). Its scope is deliberately bounded, reflecting the current absence of publicly accessible, richly annotated corpora, benchmark datasets, and end-to-end pretraining pipelines for Tajik [10,13,14]. As such, the research does not present empirical experiments, model training, or quantitative evaluation of adapted LLMs. Rather, the focus lies on mapping systemic constraints, synthesising relevant linguistic and technical literature, and proposing a structured roadmap to overcome key limitations in a reproducible and culturally sensitive manner [1,3,4,17,20].

The limitations of this work stem from three interrelated factors. First, the scarcity and heterogeneity of Tajik digital resources preclude immediate empirical validation. Although initiatives such as the National Corpus of Tajik Language provide a foundation, access restrictions and incomplete metadata reduce their utility for large-scale NLP experimentation [10]. Second, the proposed methodological framework relies conceptually on cross-lingual transfer from Persian and multilingual models. While this strategy leverages typological proximity, the effectiveness of such transfer remains contingent on orthographic, terminological, and morphological alignment that cannot yet be fully quantified in the absence of experimental data [12,17,18,21]. Third, socio-cultural and ethical alignment considerations are largely prospective. The paper identifies safety and bias risks inherent in deploying donor or adapted models, but without annotated local datasets, mitigation strategies remain theoretical [7,8,16,20].

Despite these limitations, the study establishes clear boundaries for subsequent research and provides a systematic foundation for the development of Tajik LLMs. By formalising the infrastructure requirements, annotation standards, tokenisation strategies, and safety considerations, this work delineates a reproducible path forward that is compatible with future empirical validation and scalable implementation.

5. Implications for Research and Language Policy

The theoretical and infrastructural framework proposed in this study carries several important implications for both research and language policy. From a scholarly perspective, the roadmap for Tajik LLM development provides a structured foundation for low-resource language research, offering methodological guidance on corpus construction, annotation standards, morphologically informed tokenisation, and culturally sensitive alignment. By synthesising insights from cross-lingual transfer, parameter-efficient adaptation, and safety-aware modelling, the work establishes a coherent agenda for future empirical studies, reducing redundancy and guiding resource allocation in Tajik computational linguistics [3,4,5,12,17,20].

From a policy standpoint, the findings underscore the necessity of national and institutional engagement in supporting digital language infrastructure. The creation of a publicly accessible Tajik MegaCorpus, together with benchmark datasets and standardised annotation pipelines, can serve as a strategic instrument for preserving linguistic sovereignty and promoting equitable participation in global NLP ecosystems. In particular, ensuring open access, reproducibility, and expert-mediated validation will strengthen both educational initiatives and research collaborations within Tajikistan and with international partners [10,13,14].

Moreover, the identification of lexico-terminological divergence and cultural alignment challenges highlights the need for interdisciplinary collaboration. Linguists, computational scientists, and social experts must jointly define standards for terminology, ethical guidelines, and culturally contextualised safety measures. The integration of such standards into national language policy would not only support responsible LLM development but also provide a model for other low-resource languages facing analogous barriers [7,8,16,20].

In summary, the study emphasises that advancing Tajik NLP is not solely a technical endeavour. It requires coordinated research, infrastructure investment, and policy initiatives to establish sustainable, culturally aligned, and publicly beneficial language technologies. The roadmap outlined here can inform strategic planning, prioritise resource development, and foster international engagement while ensuring that Tajik maintains visibility and competitiveness in the era of large language models.

6. Conclusion

This study has addressed the pressing need for systematic development of Tajik language technologies in the era of large language models. By mapping structural, data-related, and infrastructural constraints, it has elucidated the factors that contribute to the "digital barrier" preventing Tajik from achieving parity with high-resource languages. The work has also highlighted the strategic opportunities afforded by the typological proximity of Tajik to Persian, alongside the challenges posed by script divergence, terminological mismatch, and the absence of curated safety and alignment resources [10,12,17,20].

The principal contribution of this research lies in articulating a coherent, theoretically grounded roadmap for advancing Tajik NLP. This roadmap defines priorities for corpus construction, benchmark dataset development, morphologically informed tokenisation, cross-lingual transfer, and integrated safety protocols. Although empirical implementation remains a future endeavour, the framework provides a reproducible foundation that can guide both researchers and policymakers in establishing robust, culturally sensitive, and publicly accessible language technologies [3,4,5,13,16,20].

Ultimately, the study underscores that overcoming low-resource constraints requires more than technical solutions. It necessitates coordinated scholarly effort, institutional support, and policy engagement to ensure that Tajik language technologies are both effective and socially responsible. By formalising this agenda, the research positions Tajik to participate meaningfully in global NLP initiatives, fostering digital inclusion and linguistic sustainability in the age of large language models.

References

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21) (pp. 610-623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

2. Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M. The State and Fate of Linguistic Diversity and Inclusion in the NLP World // *arXiv preprint arXiv:2004.09095*. 2021. URL: https://arxiv.org/abs/2004.09095.

3. Lhoest Q., Villanova del Moral A., Jernite Y., Thakur N., Michael J., Ma Y., Sajjad H., Scao T., Wolf T. Datasets: A Community Library for Natural Language Processing // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2021. P. 175–184. DOI: 10.18653/v1/2021.emnlp-demo.21.

4. Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V. Unsupervised Cross-lingual Representation Learning at Scale // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

5. Hu E. J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Chen W., Chen L. LoRA: Low-Rank Adaptation of Large Language Models // arXiv preprint arXiv:2106.09685. 2021.

6. Wu S., Conneau A., Li H., Chaudhary V., Artetxe M., Guzmán F., Ott M., Goyal N., Grave E., Hérissé O., Stoyanov V. Emerging Cross-lingual Structure in Pretrained Language Models // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 6022–6034. DOI: 10.18653/v1/2020.acl-main.536.
7. Nozza D. Exposing the Limits of Zero-Shot Cross-lingual Hate Speech Detection // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2021. P. 907–914. DOI: 10.18653/v1/2021.acl-short.114.
8. Ganguli D., Lovitt L., Kernion J., et al. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviours, and Lessons Learned // arXiv preprint arXiv:2209.07858. 2022.
9. Ouyang L., Wu J., Jiang X., et al. Training Language Models to Follow Instructions with Human Feedback // Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 27730–27744.
10. Arabov M.K., Sedykh V.V. Comparative Analysis of Methods for Modelling Semantic Word Representations under Low-Resource Language Conditions: The Case of Tajik // *Scientific and Technical Bulletin of the Volga Region*. 2025. No. 6. P. 196–198. EDN ZHBKFG.
11. Usmanov Z.D., Sharipov Sh.A., Davudov G.M. On the Variety of Word-Form Anagrams // Bulletin of the Technological University of Tajikistan. 2022. No. 1(48). P. 186–191. EDN XIIEFY.
12. Mohammadi M., Faili H. Cross-lingual Transfer Learning for Persian Dependency Parsing // Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). 2018. P. 2800–2807.
13. Arabov M.K., Makhmadaliev Kh.S., Khabibullozoda K.K. Creating a Multiformat Text Corpus for the Tajik Language to Train Modern Language Models // Science and Innovation. Series of Geological and Technical Sciences. 2025. No. 2. P. 131–136. EDN FJMXTF.
14. Davudov G.M. Computer Morphological Analysis of Tajik Word Forms: PhD Dissertation. Dushanbe, 2018. 186 p. EDN OKGAFH.
15. Soliev O.M., Khudoyberdiev Kh.A., Davudov G.M. Automatic Spell-Checking System for Tajik — TajSpell // Bulletin of the Technological University of Tajikistan. 2021. No. 3(46). P. 188–194. EDN WZYMGP.
16. Rafailov R., Sharma A., Mitchell E., et al. Direct Preference Optimisation: Your Language Model is Secretly a Reward Model // Advances in Neural Information Processing Systems (NeurIPS). 2023. Vol. 36. URL: https://arxiv.org/abs/2305.18290.
17. Farahani M., Gharachorloo M., Farahani M., Manthouri M. ParsBERT: Transformer-based Model for Persian Language Understanding // arXiv preprint arXiv:2005.12515. 2020. URL: https://arxiv.org/abs/2005.12515.
18. Abbasi M.A., Ghafouri A., Firouzmandi M., Naderi H., Minaei Bidgoli B. PersianLLaMA: Towards Building First Persian Large Language Model // *arXiv preprint arXiv:2312.15713*. 2023. URL: https://arxiv.org/abs/2312.15713.
19. Fedorov D.S., Kozhevnikov V.V. Text Data Augmentation for Low-Resource Languages Based on Neural Models // Artificial Intelligence and Decision Making. 2023. No. 1. P. 45–56. DOI: 10.14357/20718594230104.
20. Sarimsakova A.T., Yuldashev A.A. Methods of Automatic Transliteration for Languages with Different Writing Systems (Example: Persian and Tajik) // Proceedings of the International Conference “Corpus Linguistics-2023”. St. Petersburg, 2023. P. 112–125.
21. Seraji M., Jahani C., Megyesi B., Nivre J. A Persian Treebank with Stanford Typed Dependencies // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14). Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). P. 796–801. URL: https://aclanthology.org/L14-1326/