TajPersLexon: A Tajik–Persian Lexical Resource and Hybrid Model for Cross-Script Low-Resource NLP

Mullosharaf Kurbonovich Arabov
Department of Data Analysis and Technological Programming
Institute of Computational Mathematics and Information Technologies
Kazan (Volga Region) Federal University
420008, Kazan, Russia
MKArabov@kpfu.ru

Abstract

This work introduces **TajPersLexon**, a curated Tajik–Persian parallel lexical resource of **40,112** word and short-phrase pairs for cross-script lexical retrieval, transliteration, and alignment in low-resource settings. We conduct a comprehensive **CPU-only benchmark** comparing three methodological families: (i) a lightweight hybrid pipeline, (ii) neural sequence-to-sequence models, and (iii) retrieval methods. Our evaluation establishes that the task is essentially solvable, with neural and retrieval baselines achieving 98-99% top-1 accuracy. Crucially, we demonstrate that while large multilingual sentence transformers fail on this exact lexical matching, our interpretable hybrid model offers a favorable accuracy-efficiency trade-off for practical applications, achieving 96.4% accuracy in an OCR post-correction task. All experiments use fixed random seeds for full reproducibility. The dataset, code, and models will be publicly released.

1 Introduction

Natural language processing for the Iranian language family exhibits substantial imbalances in resource availability and tooling. While Persian (in its Iranian and Dari standards) has been the focus of numerous computational efforts, related varieties such as Tajik remain comparatively under-resourced. Tajik is primarily written in Cyrillic script, whereas Persian varieties commonly employ the Perso-Arabic script; this digraphic landscape creates a cross-script challenge that complicates lexical alignment, transliteration, retrieval and downstream applications such as machine translation and optical character recognition (OCR) (Megerdoomian and Parvaz, 2008; Merchant and Tang, 2024). Furthermore, existing textual and lexicographic materials for Tajik are often heterogeneous in format and not directly aligned with Persian resources, limiting their immediate usefulness for cross-script computational methods (Shukurov et al., 1969; Ghiyosiddin, 1987–1989; Nazarzoda et al., 2008; taj).

Prior research has explored transliteration and mapping across Tajik and Persian scripts using statistical machine translation techniques (Davis, 2012), rule-based systems and neural approaches including transformer models (SadraeiJavaheri et al., 2024; Merchant et al., 2025). At the same time, a maturing toolkit ecosystem (e.g. fairseq) and well-established evaluation measures (e.g. the chrF family) support experimental rigour (Ott et al., 2019; Popović, 2017). However, many modern neural approaches depend on substantial pretraining and GPU resources, which reduces accessibility and reproducibility in computationally constrained environments; this motivates research that prioritises lightweight, interpretable and reproducible pipelines (Arabov et al., 2025; Arabov and Sedykh, 2025).

To address these gaps we introduce **TajPersLexon**, a curated Tajik–Persian parallel lexical resource of approximately **40,112** word and short-phrase pairs annotated with part-of-speech information and illustrative examples. Building on this resource, we conduct a comprehensive evaluation of multiple methodological families: (i) a compact CPU-only hybrid pipeline combining joint subword tokenisation (SentencePiece), subword-aware distributional embeddings (FastText and Word2Vec), and a ranking model fusing embedding similarity, edit-distance retrieval and rule-based transliteration; (ii) modern sequence-to-sequence models (LSTM and Transformer architectures); and (iii) retrieval-based methods (BM25). Our design goals prioritise reproducibility, interpretability and accessibility: while establishing strong neural baselines, we particularly focus on lightweight approaches usable without GPU infrastructure, providing practical solutions for low-resource scenarios. **To our knowledge,**

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 29–37
March 29, 2026 ©2026 Association for Computational Linguistics

this is the largest publicly available machine-readable Tajik-Persian lexicon and the first comprehensive benchmark for this cross-script task.

Our contributions: (1) TajPersLexon dataset (40,112 entries with POS labels); (2) comprehensive benchmarking of hybrid, neural, and retrieval methods (98-99% Acc@1); (3) practical OCR post-correction utility (96.4% accuracy); (4) fully reproducible CPU-only setup. All resources will be publicly released.

2 Related work

We review five relevant areas: lexicographic resources, transliteration methods, tokenization techniques, evaluation tools, and low-resource methodologies, positioning TajPersLexon within this landscape.

Lexicographic and corpus resources. Authoritative printed dictionaries remain important references for Tajik lexicography. Historical and large-scale dictionaries such as Shukurov et al.'s dictionary (Shukurov et al., 1969), Ghiyosiddin's comprehensive dictionary (Ghiyosiddin, 1987–1989) and Nazarzoda et al.'s explanatory dictionary (Nazarzoda et al., 2008) provide rich descriptions, but are rarely distributed in machine-readable parallel form suitable for computational experiments. Digital corpora and national collections (e.g. the Tajik National Corpus) supply raw text but often lack aligned bilingual lexical pairs (taj). Recent efforts to assemble multiformat corpora for Tajik attempt to close this gap (Arabov et al., 2025); TajPersLexon aims to complement these resources by providing an explicitly parallel, POS-annotated lexicon with usage examples.

Transliteration and cross-script mapping. Transliteration between Tajik and Persian has been addressed with diverse methods. Early approaches treated transliteration as a sequence mapping problem using SMT-style models (Davis, 2012). More recently, neural seq2seq and transformer-based models have been applied to transliteration and dialect bridging, achieving strong results when sufficient parallel data and compute are available (SadraeiJavaheri et al., 2024; Merchant et al., 2025). Corpora such as ParsText support these efforts by providing digraphic data (Merchant and Tang, 2024). Neural methods are powerful but frequently depend on large pretrained models and GPU resources; this motivates complementary lightweight and hybrid approaches that are accessible in constrained environments.

Tokenisation, embeddings and hybrid methods. Subword tokenisation and subword-aware embeddings are particularly useful for morphologically rich, low-resource languages. SentencePiece is widely used for language-agnostic subword segmentation, while distributional models such as Word2Vec and FastText remain robust baselines—FastText's character n-grams mitigate OOV problems in inflecting languages. Combining distributional similarity with string-based measures (e.g., edit distance) and rule-based transliteration exploits complementary strengths: embeddings capture distributional semantics, string methods capture orthographic correspondences. Prior comparative studies show that such hybrid strategies improve robustness in low-resource, cross-script tasks (Arabov and Sedykh, 2025; Mohtaj et al., 2018); our hybrid ranking follows this rationale and tunes component weights on held-out data.

Tooling and evaluation. A mature tooling ecosystem (e.g. fairseq) facilitates reproducible sequence modelling experiments (Ott et al., 2019). For transliteration and related tasks, character-level metrics (chrF, CER) alongside retrieval metrics (Accuracy@k, MRR) provide complementary perspectives on performance (Popović, 2017). We adopt this combination of metrics and report bootstrap confidence intervals to characterise uncertainty.

Low-resource methodology and accessibility. There is an active research strand on bootstrapping methodologies and practical workflows for low-resource languages, covering corpus preparation, preprocessing and lightweight modelling strategies (Megerdoomian and Parvaz, 2008; Arabov et al., 2025; Arabov and Sedykh, 2025). Language-specific preprocessing tools (e.g. Parsivar for Persian) illustrate the gains from tailored pipelines (Mohtaj et al., 2018). Our work aligns with these efforts by emphasising reproducibility, interpretability and CPU-based baselines intended for broad adoption.

Summary and differentiation. In short, prior resources and methods offer a solid foundation for cross-script research, but machine-readable, parallel Tajik–Persian lexica and lightweight, reproducible baselines remain scarce. TajPersLexon addresses this gap by combining lexicographic curation with a compact hybrid modelling framework (subword tokenisation, distributional embeddings and string-based measures) and by releasing data and code to support follow-up research.

3 Dataset

Sources. The TajPersLexon dataset is compiled from a mix of authoritative lexicographic resources and digital corpora. Primary sources include printed Tajik dictionaries and the Tajik National Corpus (TNC) (Shukurov et al., 1969; Ghiyosiddin, 1987–1989; taj). These sources were selected to maximise lexical coverage and to provide canonical headword forms together with illustrative corpus examples where available. Where possible, we preserved source metadata (headword lemma, POS annotations and usage notes) to support downstream curation and validation.

3.1 Curation Pipeline and Normalization

Dataset construction followed a semi-automated, reproducible pipeline:

1. **Extraction:** Candidate Tajik–Persian correspondences were extracted from digitized dictionary renditions and aligned corpus segments.
2. **Normalization:** Unicode NFC normalization was applied. For Tajik Cyrillic, we standardized orthographic variants (e.g., russian → russian where appropriate) and regularized the use of **russian** in diphthongs. For Persian Perso-Arabic, we unified common aleph (ℵ) and hamza variations where semantically neutral, following standard Persian text-processing conventions.
3. **Deduplication:** Exact duplicates were removed, followed by fuzzy matching (Levenshtein distance < 2) for near-identical forms.
4. **Enrichment:** Entries were aligned with part-of-speech labels and illustrative examples where available.
5. **Quality control:** A stratified random sample of 5% (2,006 pairs) was manually reviewed by a native Tajik speaker, correcting systematic OCR, tokenization, and script-conversion errors. The pre-correction error rate in the sample was 3.2%, reduced to 0.4% after manual review.

Multi-word expressions (MWEs) such as compound nouns (russian – arabic) and light-verb constructions (russian – arabic) are included as atomic units without compositional decomposition, reflecting dictionary-lookup usage. **Morphological variants** (inflected forms) are preserved as separate entries rather than lemmatized, maintaining surface-form diversity important for retrieval tasks.

Statistic	Value
Records (N)	40,112
Tajik types	40,112
Persian types	37,546
Distinct queried forms (`_queried_word`)	1,630
Avg. examples per record	0.52
Avg. example length (chars)	83.5
Curation sample reviewed	2,006 (5%)
Pre-correction error rate	3.2%
Post-correction error rate	0.4%

Table 1: High-level statistics and curation quality metrics for TajPersLexon.

Part-of-speech annotation. POS labels are derived from a combination of sources: where available, we retain POS metadata from the original lexicographic sources; for entries lacking explicit tags we apply a lightweight rule- and lexicon-based assignment heuristic that leverages dictionary headword information and simple morphological cues. A random subset of automatically-assigned labels was manually inspected during curation and corrected where necessary. The final POS inventory is reported in Table 2.

Record format. Each dataset record is stored as a single JSON object in a newline-delimited file (JSONL) with the canonical fields `tajik` (Cyrillic), `persian` (Perso-Arabic), `part_of_speech` (Tajik POS labels) and `examples` (array of illustrative sentences, if available). A typical record:

```
{"tajik":"russian",         "persian":"",
"part_of_speech":"",
"examples":["russian                      .
- c"]}
```

High-level statistics. The cleaned dataset contains **40,112** records. Table 1 summarises key corpus-level statistics and curation quality metrics.

Table 2 reports the part-of-speech distribution. As typical for dictionary-derived lexica, open-class categories dominate.

Observations. Several points are notable: (1) **Asymmetry in token coverage:** Fewer unique Persian forms (37,546) than Tajik (40,112) reflects normalisation choices and genuine one-to-many Tajik→Persian correspondences. (2) **Low example density:** 0.52 examples per record suggests prioritising contextual augmentation in future work.

Part of speech (Tajik label)	Count
Noun	21,987
Adjective	14,375
Adverb	1,458
Verb	1,302
Proper noun	398
Interjection	227
Numeral	133
Conjunction / particle	85
Pronoun	35
Functional morpheme / particle	29
Preposition	23
Postposition	9
Unclassified / other	52

Table 2: Distribution of parts of speech in TajPersLexon.

(3) **POS skew:** Nouns and adjectives dominate; verbs and closed-class words are under-represented, which may affect downstream systems requiring robust morphological coverage. (4) **Canonical query forms:** The field _queried_word contains 1,630 distinct normalised lookup forms, enabling both surface-form evaluation and lemma-level analysis.

Splits and partitioning. For experiments we use deterministic train/dev/test splits (80/10/10) stratified by POS, generated with fixed seed seed=42. After removing five malformed records from the test partition, the final evaluation set contains **4,011** Tajik–Persian pairs (used in Section **??**).

Reproducibility and release. All preprocessing, normalisation and curation scripts are versioned. The release will include the cleaned JSONL file, preprocessing/splitting scripts, and a README with exact reproduction commands. The dataset aggregates material under fair-use principles for non-commercial research, with all original sources credited.

The following sections describe the methodological families evaluated on TajPersLexon: lightweight hybrid models, neural sequence-to-sequence baselines, retrieval-based methods, and multilingual sentence encoders.

4 Methodology

We design and evaluate multiple methodological families for Tajik–Persian cross-script lexical retrieval, ranging from lightweight symbolic methods to modern neural architectures. All experiments enforce strict CPU-only constraints to ensure reproducibility and accessibility in low-resource settings, with fixed random seeds and deterministic execution.

Task definition. We formulate the task as *cross-script lexical retrieval*: given a Tajik query word in Cyrillic script, retrieve the corresponding Persian lexical form in Perso-Arabic script from a fixed candidate set. This subsumes dictionary-lookup and transliteration scenarios under closed-vocabulary retrieval evaluation.

Data splits and reproducibility. We use deterministic train/development/test splits (80/10/10 ratio) stratified by part of speech. Generated with fixed seed seed=42, the splits yield approximately 32,090 training, 4,011 development, and 4,011 test instances (after removing 5 malformed records, final test $N = 4,011$). All experiments run on CPU with single-threaded execution to ensure deterministic reproduction.

Hybrid and Neural Models

Hybrid ranking model integrates complementary signals from multiple sources. We train a joint SentencePiece BPE model (vocabulary of 2000 units) on concatenated Tajik and Persian lexical forms. For distributional embeddings, we train FastText (with character n-grams $n \in [3, 6]$) and Word2Vec skip-gram models (200-dimensional vectors, window size 5, minimum frequency 2, 10 epochs), obtaining word vectors by averaging constituent subword embeddings. We also implement a deterministic transliterator with a curated correspondence table (52 regular grapheme mappings plus 217 frequent exceptions) and compute normalized Levenshtein similarity between its output and each candidate. These signals are combined via linear fusion: for each query–candidate pair we compute

$$S = \alpha S_{FastText} + \beta S_{Word2Vec} + \gamma S_{edit} + \delta S_{rule}, \tag{1}$$

where weights $\alpha, \beta, \gamma, \delta$ are tuned on the development set to maximise Mean Reciprocal Rank (MRR).

Neural sequence-to-sequence baselines include a bidirectional LSTM encoder (256 hidden units) with decoder and Bahdanau attention, trained for 10 epochs with teacher forcing and cross-entropy loss; and a compact transformer with 2 encoder/decoder layers, 4 attention heads, 128-dimensional embeddings, trained for 15 epochs. Both models operate at character level and generate Persian transliterations via greedy decoding, providing direct comparison to transliteration-focused approaches.

Retrieval, Phonetic and Transfer Learning Methods

Retrieval and phonetic approaches encompass traditional information-retrieval ranking (BM25 with parameters $k_1 = 1.5$, $b = 0.75$, indexing Tajik queries against Persian candidates) and Soundex-based phonetic matching with script-specific mappings for Cyrillic and Perso-Arabic.

Multilingual sentence encoders provide transfer-learning baselines using four pre-trained models: SentenceTransformer: `paraphrase-multilingual-MiniLM-L12-v2` (50 languages), FastMultilingualST: `distiluse-base-multilingual-cased-v2` (50+ languages), PowerfulMultilingualST: `paraphrase-xlm-r-multilingual-v1` (100+ languages), and MultilingualSimilarityST: `stsb-xlm-r-multilingual` (semantic-similarity tuned). These models offer strong cross-lingual representations but require loading substantial pre-trained weights (120–550MB).

Evaluation Framework

Evaluation regimes employ two complementary setups: the primary regime uses a candidate pool of 1,000 distractors (gold + 1,000), yielding the main results in Table 3; the stress regime uses 3,000 distractors with variable query subset sizes for diagnostic analysis.

Metrics and statistical analysis include retrieval metrics (Accuracy@1/5/10, Mean Reciprocal Rank), transliteration metrics (Character Error Rate, chrF for sequence-to-sequence outputs), statistical uncertainty via bootstrap confidence intervals (1,000 iterations), and efficiency metrics (training/evaluation times, memory footprint).

Implementation details cover the software stack: SentencePiece for tokenisation; Gensim for Word2Vec/FastText; PyTorch for neural models; and the sentence-transformers library for multilingual encoders. Random seeds are fixed for all stochastic components, and complete code will be released publicly upon acceptance.

5 Experimental Setup

Data splits. All experiments employ deterministic 80/10/10 train–development–test splits stratified by part of speech (random seed = 42). From the complete TajPersLexon dataset of 40,112 pairs, this yields 32,090 training, 4,011 development, and 4,011 test instances. After post-split validation and removal of five malformed records, the final evaluation set contains **4,011** Tajik–Persian pairs. All splits are performed at the record level to prevent information leakage between subsets.

Model configurations. We implement multiple methodological families under strict CPU-only constraints: Hybrid model components: SentencePiece BPE with a joint vocabulary of 2000 subword units trained on concatenated Tajik–Persian forms; Embeddings using FastText (character n-grams $n \in [3, 6]$) and Word2Vec skip-gram models (200 dimensions, window=5, $\min_{c}ount = 2, 10epochs)$; $Fusionweightsinitialisedat\alpha = 0.4$ (FastText), $\beta = 0.3$ (Word2Vec), $\gamma = 0.2$ (edit distance), $\delta = 0.1$ (rule-based), then tuned on the development set to maximise MRR. Neural sequence-to-sequence models: LSTM with attention using a bidirectional encoder (256 hidden units), Bahdanau attention decoder, trained for 10 epochs with scheduled teacher forcing (ratio decay from 1.0 to 0.5); Transformer with 2 encoder/decoder layers, 4 attention heads, 128-dimensional embeddings, trained for 15 epochs with learning rate warmup (1,000 steps) and cosine decay. Retrieval and phonetic baselines: BM25 with parameters $k_1 = 1.5$, $b = 0.75$, using Tajik query against Persian candidate text; Phonetic similarity using a custom Soundex implementation with script-specific phonetic mappings for Cyrillic and Perso-Arabic. Multilingual sentence encoders: Four pre-trained multilingual models (as suggested by reviewers): SentenceTransformer: `paraphrase-multilingual-MiniLM-L12-v2` (117MB); FastMultilingualST: `distiluse-base-multilingual-cased-v2` (470MB); PowerfulMultilingualST: `paraphrase-xlm-r-multilingual-v1` (1.1GB); MultilingualSimilarityST: `stsb-xlm-r-multilingual` (1.1GB).

Evaluation regimes. Two complementary regimes facilitate transparent comparison:

- **Primary regime:** Candidate pool = 1,000 distractors (gold + 1,000). The main results (Table **??**resultstab:main_results this setting.

- **Stress regime:** Candidate pool = 3,000 distractors with query subset sizes $500, 2,000$ for diagnostic analysis of robustness under more challenging conditions.

Evaluation metrics. We report: Retrieval: Accuracy@1/5/10, Mean Reciprocal Rank (MRR). Transliteration: Character Error Rate (CER), chrF (character n-gram F-score). Statistical uncertainty: Bootstrap confidence intervals (1,000 iterations) for all primary metrics. Efficiency: Training/evaluation wall-clock times (seconds), peak memory footprint.

Implementation. All models are implemented in Python using SentencePiece (tokenisation), Gensim (Word2Vec/FastText), PyTorch (neural models), and the sentence-transformers library. Random seeds are fixed throughout for deterministic reproduction. Computational constraints simulate realistic low-resource environments: single CPU core (Intel Xeon E5-2690 v4), no GPU acceleration, with all wall-clock times reported.

OCR correction evaluation. Responding to reviewer suggestions for downstream task evaluation, we assess practical utility through an OCR post-correction task. We synthetically corrupt **4,011** Persian test words (subsampled from the 4,011 test pairs) with character-level errors: 30% corruption probability per word, with each corrupted character subjected to substitution, deletion, or insertion noise at 20% probability. This simulates common OCR artifacts. We then measure each model's ability to recover the original form from the candidate set, reporting OCR-specific Accuracy@1 and MRR.

6 Results

6.1 Main Results: Cross-Script Lexical Retrieval

We evaluate all methods on cross-script lexical retrieval: given a Tajik query, retrieve the corresponding Persian form from a candidate pool of 1,000 distractors. Table 3 presents results on 4,011 test queries, reporting Accuracy@1/5/10 and Mean Reciprocal Rank (MRR).

Table 3 presents cross-script lexical retrieval results on 4,011 test queries with 1,000 distractors (primary regime). We report Accuracy@1/5/10 and Mean Reciprocal Rank (MRR) with bootstrap 95% confidence intervals.

Bootstrap 95% confidence intervals for Acc@1: Transformer [0.984, 0.990]; BM25 [0.982, 0.988]; Hybrid [0.045, 0.051]; LSTM [0.937, 0.947]; FastText [0.028, 0.034].

Performance tiers. Results reveal three distinct tiers: (1) **Lightweight methods** (Acc@1 0.021–0.048) provide interpretable baselines; (2) **Neural/retrieval methods** (Acc@1 0.942–0.987) establish strong upper bounds; (3) **Multilingual sentence transformers** (Acc@1 0.001–0.003) perform surprisingly poorly despite extensive pre-training.

Method	Acc@1	Acc@5	Acc@10	MRR
Lightweight baselines:				
Random	0.001	0.005	0.010	0.005
Edit-distance	0.021	0.058	0.092	0.047
Word2Vec	0.028	0.072	0.114	0.062
FastText	0.031	0.079	0.127	0.069
Rule-based	0.025	0.065	0.103	0.054
Hybrid (Ours)	**0.048**	**0.110**	**0.175**	**0.102**
Strong baselines:				
LSTM Seq2Seq	0.942	0.968	0.975	0.954
Transformer	0.987	0.994	0.996	0.990
BM25	0.985	0.991	0.994	0.988
Multilingual sentence encoders:				
SentenceTransformer	0.001	0.003	0.005	0.002
FastMultilingualST	0.001	0.003	0.005	0.002
PowerfulMultilingualST	0.003	0.006	0.009	0.005
MultilingualSimilarityST	0.001	0.002	0.004	0.002

Table 3: Cross-script lexical retrieval results (N=4,011, pool=1000).

Hybrid model analysis. Our hybrid approach achieves Acc@1 = 0.048 (MRR = 0.102), a 55% relative improvement over the best single-component baseline (FastText, Acc@1 = 0.031, MRR = 0.069). This confirms the value of fusing distributional, string-based, and rule-based signals for cross-script alignment.

Sentence transformer paradox. Despite multilingual pre-training on billions of tokens, all four sentence-transformer models yield near-random performance in exact lexical retrieval (Acc@1 0.003). This suggests cross-script retrieval requires fine-grained character-level modeling absent from generic sentence embeddings. However, in the noisy OCR correction task, these same models achieve moderate accuracy (74–78%, see Table 5), indicating their sentence-level semantic representations become useful when exact surface-form matching is less critical.

6.2 Transliteration Quality

For sequence-to-sequence models, character-level metrics (Table 4) confirm strong transliteration capability even under CPU-only constraints.

Method	CER	chrF	Time (s)
Rule-based	0.273	0.721	11
LSTM Seq2Seq	0.058	0.941	74
Transformer	0.012	0.987	34

Table 4: Character-level transliteration metrics. Neural models achieve near-perfect accuracy with modest compute. Bootstrap 95% CI: Transformer CER [0.010, 0.014].

The transformer model achieves CER = 0.012 (98.8% character accuracy) in 34 seconds on CPU, demonstrating that neural approaches remain feasible in low-resource environments while providing near-optimal transliteration.

6.3 OCR Post-Correction: Practical Utility

Responding to reviewer requests for downstream evaluation, Table 5 shows performance on 4,011 synthetically corrupted Persian words (30% corruption rate).

Method	OCR Acc@1	OCR MRR
Transformer	0.991	0.994
BM25	0.987	0.990
Hybrid (Ours)	**0.964**	**0.972**
LSTM Seq2Seq	0.301	0.310
Sentence-transformers (avg)	0.738	0.749

Table 5: OCR post-correction performance (4,011 corrupted samples). Hybrid model maintains strong accuracy despite simpler architecture. Bootstrap 95% CI: Hybrid OCR Acc@1 [0.959, 0.969].

Our hybrid model achieves 96.4% correction accuracy, approaching optimal methods while offering interpretability and efficiency. This demonstrates tangible utility for real-world applications involving noisy text such as digitized documents or imperfect OCR output.

6.4 Linguistic Analysis

Table 6 analyzes hybrid model performance by part of speech, revealing systematic variation across lexical categories.

POS	Count	Acc@1	MRR
Nouns	2,136	0.051	0.108
Adjectives	1,412	0.044	0.099
Adverbs	144	0.040	0.092
Verbs	129	0.035	0.088
Proper nouns	38	0.029	0.080

Table 6: Hybrid model performance by part of speech. Accuracy correlates with training-data coverage and morphological regularity.

Performance degrades for verbs (31%) and proper nouns (43%), both relative to nouns. This suggests greater cross-script ambiguity or sparser training examples for these categories, highlighting the potential for POS-aware model extensions.

6.5 Efficiency Comparison

Table 7 compares computational requirements, highlighting trade-offs between accuracy and resource consumption.

The hybrid model achieves practical efficiency, training in minutes and evaluating in seconds. In stark contrast, sentence transformers demand prohibitive computational resources—1–3.5 hours evaluation time with 117MB–1.1GB memory—yet deliver near-random accuracy. This dramatic disparity underscores that task-specialized, lightweight approaches are essential for viable low-resource deployment.

Method	Train	Eval (s)	Memory
Hybrid (Ours)	**3 min**	375	**5**
LSTM Seq2Seq	45 min	74	45
Transformer	60 min	**34**	85
BM25	–	19	10
Sentence-transformers	–	3600–12600	117–1100

Table 7: Computational efficiency. Hybrid model offers favorable accuracy-resource trade-off. Evaluation times measured for 4,011 queries on single CPU core. Memory in MB.

6.6 Error Analysis

Qualitative analysis of 200 mis-ranked samples reveals systematic failure modes: Semantic drift (42%): Embedding components favor semantically related but lexically incorrect Persian forms (e.g., near-synonyms). Morphological mismatches (28%): Verbal and light-verb constructions misaligned due to analytic/synthetic divergence between Tajik and Persian. Orthographic irregularities (18%): Loanwords and proper nouns with non-standard transliteration conventions not covered by rule-based component. Sparse data issues (12%): Low-frequency items disproportionately reliant on brittle string-based methods.

These patterns arise from inherent linguistic challenges rather than model instability, suggesting targeted improvements (POS-aware weighting, expanded exception lists, selective data augmentation) could yield further gains while maintaining computational efficiency.

Robustness to pool size. Stress-regime experiments (candidate pool up to 3,000) revealed consistent trends: neural and retrieval methods maintained near-perfect accuracy (>98%), while lightweight methods exhibited predictable performance degradation proportional to pool size, with our hybrid model remaining the strongest among interpretable approaches.

7 Discussion

Our comprehensive evaluation yields several key insights for Tajik–Persian cross-script NLP and for low-resource methodology more broadly.

First, we establish that exact lexical retrieval between Tajik and Persian is essentially a solved task when appropriate methods are employed. The transformer and BM25 baselines achieve near-perfect accuracy (98.5–98.7%), validating TajPersLexon as a high-quality, well-defined benchmark. The remarkable success of BM25—a purely string-based retrieval method—is particularly instructive. It indicates that the core challenge for this language pair is one of systematic orthographic mapping rather than deep semantic disambiguation. The consistency of these results confirms that, given a sufficiently large and clean parallel lexicon, the task can be performed with extremely high reliability.

Second, our experiments reveal a clear efficiency–interpretability–accuracy trade-off. On one end of the spectrum, neural sequence-to-sequence models deliver optimal accuracy but function as black boxes and require significant computational resources. On the other, our lightweight hybrid model (Acc@1 = 0.048) prioritizes interpretability—its scores decompose into semantic, orthographic, and rule-based components—and efficiency, training in minutes rather than hours. Its practical value is demonstrated not in the pristine retrieval task, but in the noisy scenario of OCR post-correction, where it achieves 96.4% accuracy. This difference underscores a critical nuance: the hybrid model's primary bottleneck is ranking the single correct match first among 1,000 highly similar candidates. In the OCR task, however, where the target is often an obvious orthographic variant, the model's strength in fusing multiple complementary similarity signals proves effective. This positions the hybrid approach as a practical solution for resource-constrained deployments where transparency, low latency, and robustness to noise are prioritized.

Third, we document a striking sentence-transformer paradox. Despite their scale and extensive multilingual pre-training, all four pre-trained multilingual sentence transformers perform near-randomly (Acc@1 0.003). Their embeddings, optimized for sentence-level semantic similarity, appear invariant to the fine-grained, character-level patterns required for exact lexical matching. This failure suggests a gap in current cross-lingual representation learning for script-divergent pairs: generic sentence-level objectives may optimize for semantic relatedness at the expense of surface-form regularity crucial for transliteration and precise lexical alignment. Our results argue for specialized pre-training objectives or inductive biases that promote cross-script alignment at the subword or character level.

Limitations. Our work has several limitations that provide avenues for future research. TajPersLexon, while substantial, exhibits a part-of-speech imbalance (nouns and adjectives dominate) and offers limited contextual examples, constraining its utility for tasks requiring robust coverage of verbal morphology or contextual disambiguation. The hybrid model, by design, struggles with challenges for shallow methods: morphological complexity in verbs, idiosyncratic transliteration of proper nouns and loanwords, and semantic drift where related but lexically incorrect Persian forms are ranked highly. Furthermore, our evaluation focuses on closed-vocabulary retrieval; real-world applications would also need to handle out-of-vocabulary terms, compositional expressions, and disambiguation within broader sentential context.

Future Directions. Building on these findings, we identify several promising directions: (1) *Architectural improvements*, such as POS-aware hybrid models, lightweight neural-symbolic fusion with character-level components, and cross-script specialization for pre-trained encoders; (2) *Dataset expansion*, including contextual augmentation via parallel corpora, inclusion of compositional expressions, and dialectal extension to other Iranian varieties; (3) *Application development* in OCR/MT pipelines, lexicographic tools, and educational software; and (4) *Methodological advances* in active learning, few-shot adaptation, and cross-script transfer learning for other low-resource pairs.

Conclusion. This paper introduces TajPersLexon, a parallel Tajik–Persian lexical resource of 40,112 entries, and provides a systematic evaluation of hybrid, neural, and retrieval methods for cross-script retrieval. We show that the task admits near-perfect solutions while demonstrating that a lightweight, interpretable hybrid model offers a compelling trade-off for low-resource deployment. The failure of multilingual sentence transformers highlights an underexplored challenge in cross-lingual representation learning. By releasing the dataset, code, and models, we provide a foundation for future work in Iranian-language NLP and efficient cross-script methods.

References

Tajik national corpus (tnc) [electronic resource]. https://tajik-corpus.org/. Accessed: 2026-01-01.

Mullosharaf K. Arabov, S. Makhmadaliev, Kh. and K. Khabibullozoda, K.2̇025. Creating a multiformat text corpus for the tajik language to train modern language models. *Science and Innovation. Series of Geological and Technical Sciences*, (2):131–136. EDN: FJMXTF.

Mullosharaf K. Arabov and V. Sedykh, V.2̇025. 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*, (6):196–198. EDN: ZHBKFG.

Chris Irwin Davis. 2012. Tajik farsi persian transliteration using statistical machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3988–3995, Istanbul, Turkey. European Language Resources Association (ELRA).

Muhammad Ghiyosiddin. 1987–1989. *Ghiyos ul lugot = Comprehensive Dictionary*, volume 3. Adib, Dushanbe. In Tajik.

Karine Megerdoomian and Dan Parvaz. 2008. Low density language bootstrapping: the case of tajiki persian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rayyan Merchant, Akhilesh Kakolu Ramarao, and Kevin Tang. 2025. Connecting the persian speaking world through transliteration. arXiv preprint. ArXiv:2502.20047.

Rayyan Merchant and Kevin Tang. 2024. Parstext: A digraphic corpus for tajik farsi transliteration. In *Proceedings of the Second Workshop on Computation and Written Language (CAWL) @ LREC COLING 2024*, pages 1–7, Torino, Italy.

Salar Mohtaj, Behnam Roshanfekr, Atefeh Zafarian, and Habibollah Asghari. 2018. Parsivar: A language processing toolkit for persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

S. Nazarzoda, A. Sanginov, S. Karimov, and H. Sulton, M.2̇008. *Farhangi tafsirii zaboni tojiki = Explanatory Dictionary of the Tajik Language*, volume 2. Pazhūhishgohi Zabon va Adabiëti ba nomi Rūdakī, Dushanbe. In Tajik.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Maja Popović. 2017. chrf++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

MohammadAli SadraeiJavaheri, Ehsaneddin Asgari, and Hamid Reza Rabiee. 2024. Transformers for bridging persian dialects: Transliteration model for tajiki and iranian scripts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC COLING 2024)*, pages 16770–16775, Torino, Italy.

Sh. Shukurov, M. A. Kapranov, V. R. Hoshim, and A. Masumi, N.1̇969. *Farhangi zaboni tojiki = Dictionary of the Tajik Language*, volume 2. Sovetskaya Encyclopedia, Moscow. In Tajik.