European International Journal of Philological Sciences (ISSN: 2751-1715)

Received: 14 April 2026 | Accepted: 02 May 2026 | Published: 31 May 2026

DOI: 10.55640/eijps-06-05-20

RESEARCH ARTICLE

Comparative Analysis of Named Entity Recognition Models for Russian And Uzbek

Djuraeva Zulkhumor Radjabovna

DSc, Professor, Department of Russian language and literature, Bukhara State University, Uzbekistan

VOLUME: Vol.06 Issue05 2026

PAGE: 95-100

Abstract

This study compares named entity recognition systems for Russian and Uzbek. The Russian line of work rests on 6 established datasets and on the transformer models Slovnet BERT NER and DeepPavlov RuBERT-CRF, whose F1 reaches roughly 0.92, whereas Uzbek resources only appeared from 2023 onward and remain an order of magnitude smaller. We examine the UZNER and BERTbek corpora and the Mengliev datasets, F1 figures on the WikiANN and XTREME benchmarks and typological obstacles such as agglutination and dual script. Data quality outweighs sheer size for Uzbek and a single-number comparison of the 2 languages is misleading because their annotation schemes differ.

KEYWORDS

Named entity recognition, Russian, Uzbek, transformers, BERT, RuBERT, BERTbek, XLM-R, corpus, cross-lingual transfer, low-resource languages, NEREL.

INTRODUCTION

Named entity recognition stands among the oldest applied tasks in computational linguistics, yet the resources that feed it are distributed across languages with striking inequality. For Russian the situation is comfortable. 6 public benchmarks have accumulated since 2013, from the early collection of N.Gareev through FactRuEval-2016 and Collection3 to the nested-entity set NEREL released by N.Loukachevitch, which alone contains 56000 annotated entities of 29 types. Transformer encoders trained on Russian news, notably Slovnet BERT NER (A.Kukushkin) and the DeepPavlov RuBERT-CRF pipeline, report token-level F1 of about 0.92 averaged across the standard test sets. Uzbek tells a different story. Its first manually annotated NER corpora appeared only in 2023 and 2024, in the work of A.Yusufu and D.Mengliev and the first monolingual encoder evaluated on the task, BERTbek by E.Kuriyozov.

Comparing the 2 languages is therefore neither symmetric nor mechanical. Russian is a fusional language written in a single script; Uzbek is agglutinative, Turkic and split between Latin and Cyrillic orthographies, which fractures any lexicon a model might rely on. The entity inventories also diverge, since the richest Russian set distinguishes 29 categories while most Uzbek corpora cover only the classical triad of person, location and organization. Reported numbers further mix silver-standard and human-annotated regimes.

METHODS AND REVIEW OF THE LITERATURE

The study used comparative bibliographic analysis, systematic extraction of reported F1, precision and recall values from each model's original publication, harmonisation of those figures by test set and annotation scheme, descriptive statistics on corpus size and entity-type inventories, qualitative typological analysis of Russian and Uzbek morphology and orthography and verbatim quotation of authors' own

formulations checked against the publishers' texts.

The Russian benchmark landscape is anchored by NEREL, whose authors describe it as significantly larger than its predecessors and unusual in annotating nested entities and relations at the discourse level. Monolingual adaptation of multilingual encoders for Russian was demonstrated by M.Arkhipov on the Balto-Slavic shared task, building on the RuBERT model of Y.Kuratov and M.Arkhipov, while the practical Slovnet and Natasha toolchain of A.Kukushkin made high-quality Russian tagging widely deployable. For Uzbek the first BERT encoder was introduced by B.Mansurov and A.Mansurov, followed by the BERTbek models and the UzNER set of E.Kuriyozov, the UZNER benchmark of A.Yusufu and the annotated dataset of D.Mengliev. Cross-lingual coverage of both languages traces back to the silver-standard resource of X.Pan, from which the WikiANN splits and the later XTREME evaluation derive.

RESULTS

The Russian side of the comparison rests on a layered history of annotation. Each new corpus widened the entity inventory or added structure that the previous one lacked, moving from 2 flat categories in the Gareev set to the deeply nested scheme of NEREL. The progression is summarised in Table 1, where the rightmost column marks the shift from flat to multi-level annotation that reshaped the field after 2021.

Table 1. Public Russian NER datasets ordered by entity-inventory growth

Dataset	Entity instances	Entity types	Annotation
Gareev (2013)	44000	2	flat
Collection3, Persons-1000	26400	3	flat
FactRuEval-2016	12000	3 + sub-types	2 levels
BSNLP-2019 (ru)	9000	5	flat
RuREBus	121000	5	flat
NEREL	56000	29	sixlevels

"In this paper, we present NEREL, a Russian dataset for named entity recognition and relation extraction. NEREL is significantly larger than existing Russian datasets: to date it contains 56K annotated named entities and 39K annotated relations. Its important difference from previous datasets is annotation of nested named entities, as well as relations within nested entities and at the discourse level."

Performance figures on these sets tell a consistent story once the architecture moves to transformers. The CRF baselines of the early 2010s already passed 0.75 F1 on news text, and the two-stage CRF of V.Mozharova and N.Loukachevitch reached 0.956 on Persons-1000, but the gains became reliable across domains only with contextual embeddings. Slovnet BERT NER, trained on a news-adapted RuBERT, holds the top average of 0.92 over the four standard test sets, with the DeepPavlov RuBERT-CRF pipeline a fraction behind.

A separate strand of recent Russian work has tested whether large language models can displace the fine-tuned encoder altogether. On the SPbLitGuide cultural-news collection, A.Levchenko found that GPT-4o reached 0.93 and GPT-4.1 reached 0.94 on the person class through structured JSON prompting, edging past the RoBERTa-Large fine-tune at 0.84 and the DeepPavlov Collection3 checkpoint at 0.78. The result is narrower than it first appears, since it covers a single entity type in one domain, yet it signals that prompting now competes with supervised tagging on at least part of the Russian task. No comparable LLM evaluation exists for Uzbek, and the absence is itself informative about where the two research communities stand.

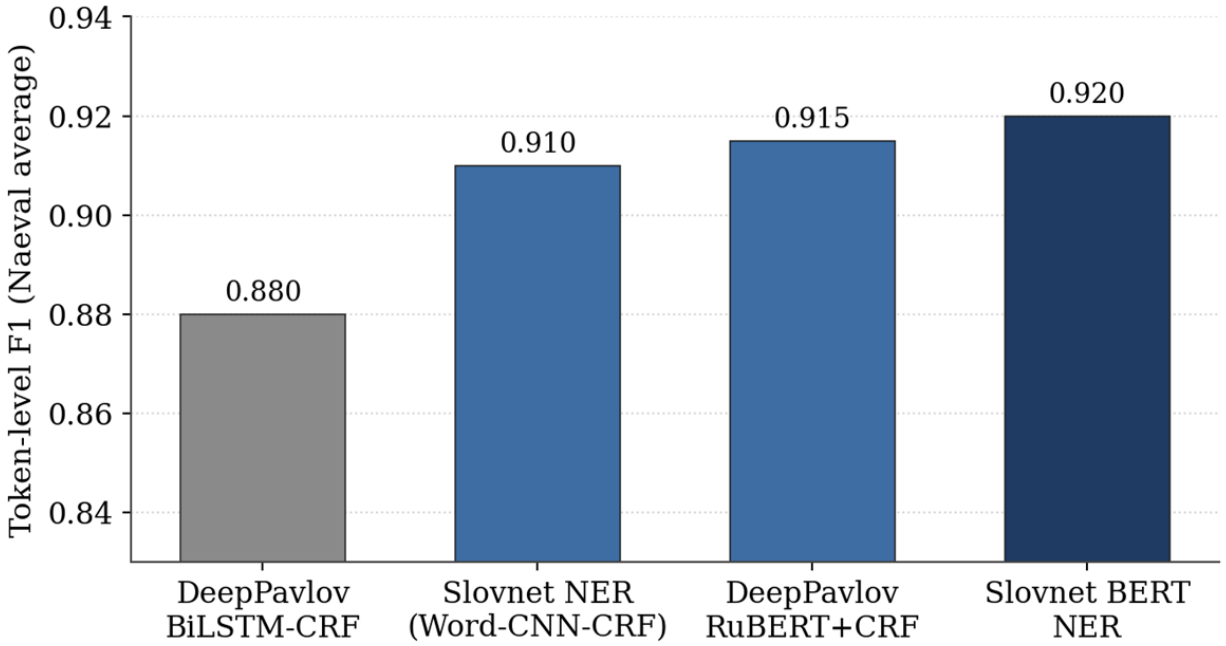

Figure 1. Token-level F1 of leading Russian NER systems (Naeval average over Collection5, Gareev, FactRuEval-2016, BSNLP-2019).

Why the monolingual encoders win is itself a finding rather than an assumption. When M.Arkhipov and colleagues continued pretraining a multilingual BERT on four Slavic languages and added a conditional random field head, they outperformed both plain multilingual BERT and the neural baselines that preceded it and they reported the gap in unambiguous language. Their result reframed the Russian task around transfer from a Slavic-restricted model rather than from the full hundred-language checkpoint. "Our paper addresses the problem of multilingual named entity recognition on the material of 4 languages: Russian, Bulgarian, Czech and Polish. We solve this task using the BERT model. We use a hundred languages multilingual model as base for transfer to the mentioned Slavic languages. Unsupervised pre-training of the BERT model on these 4 languages allows to significantly outperform baseline neural approaches and multilingual BERT."

The Uzbek record begins where the Russian one stood a decade earlier, yet it has compressed that decade into roughly two years. Dictionary and gazetteer methods of D.Mengliev gave way almost immediately to neural taggers and to manually built corpora of meaningful size. Table 2 lists the five publicly described Uzbek NER resources together with their schemes. The numbers are small against any Russian set, but the annotation discipline is comparable, and several of these corpora adopt the BIOES tagging that the Russian community converged on only gradually. The team behind UZNER states its motivation in terms that any researcher in a low-resource setting will recognise.

Table 2. Publicly described Uzbek NER corpora

Corpus	Size	Entity types	Scheme
Mengliev	1160 sent., ~19000 forms	PER/LOC/ORG + POS	BIOES
Mengliev	2000 sent., 25 865 words	PER/LOC/ORG	BIOES
UZNER (Yusufu, 2023)	~11000 sent., 402 K tokens	5+ (incl. discont.)	grid
Qalampir (Yusufu, 2025)	500 articles, 222536 tokens	PER/LOC/ORG	BIO
UzNER / BERTbek	300 articles, ~7000	6 categories	BIO

(2024)	entities		

Reported accuracy on these corpora climbs steeply once a monolingual encoder enters the pipeline. The CNN-plus-LSTM tagger of D.Mengliev reached F1 of 90.8 per cent on their second corpus and the grid-tagging BERT of A.Yusufu reported 91.7 per cent on UZNER, while on the harder UzNER split released with BERTbek the best model, BERTbek-News-Big, settled at 78.69 against 75.14 for multilingual BERT. The first dedicated Uzbek encoder had already shown the principle four years earlier, when B.Mansurov and A.Mansurov demonstrated that a single-language model could beat the multilingual default on intrinsic measures. "Pretrained language models based on the Transformer architecture have achieved state-of-the-art results in various natural language processing tasks such as part-of-speech tagging, named entity recognition, and question answering. However, no such monolingual model for the Uzbek language is publicly available. In this paper, we introduce UzBERT, a pretrained Uzbek language model based on the BERT architecture. Our model greatly outperforms multilingual BERT on masked language model accuracy."

The contrast between a multilingual encoder and a language-specific one is not peculiar to Uzbek. A.Conneau and colleagues observed the same dependence on data quality and quantity across the low-resource tail of their hundred-language model, and their finding bears directly on why a small but clean Uzbek news corpus can outperform a larger but noisier one. "We observed in Table 5 that pretraining on Wikipedia for Swahili and Urdu performed similarly to a randomly initialized model; most likely due to the small size of the data for these languages. On the other hand, pretraining on CC improved performance by up to 10 points. This confirms our assumption that mBERT and XLM-100 rely heavily on cross-lingual transfer but do not model the low-resource languages as well as XLM-R."

One result cuts against the usual reflex that more pretraining data is always better. The BERTbek experiments paired two encoders of identical token count, one trained on news and one on Wikipedia, and the news model won the NER task by three full points (76.88 against 73.85), with the larger news model reaching 78.69. E. Kuriyozov and colleagues traced the deficit to the composition of the Uzbek Wikipedia itself rather than to its volume. "Uzbek Wikipedia has many articles that were created by bots that used either automatically translated text or articles generated from predefined structures. Another downside of this source is the fact that the majority of the Uzbek Wikipedia articles were bulk imported from Uzbek Encyclopedia (Aminov et al., 2000-2006) directly, which were written in a terse style with an abundance of abbreviations to save printing space. All these factors mentioned above result a corpus with a lower data quality."

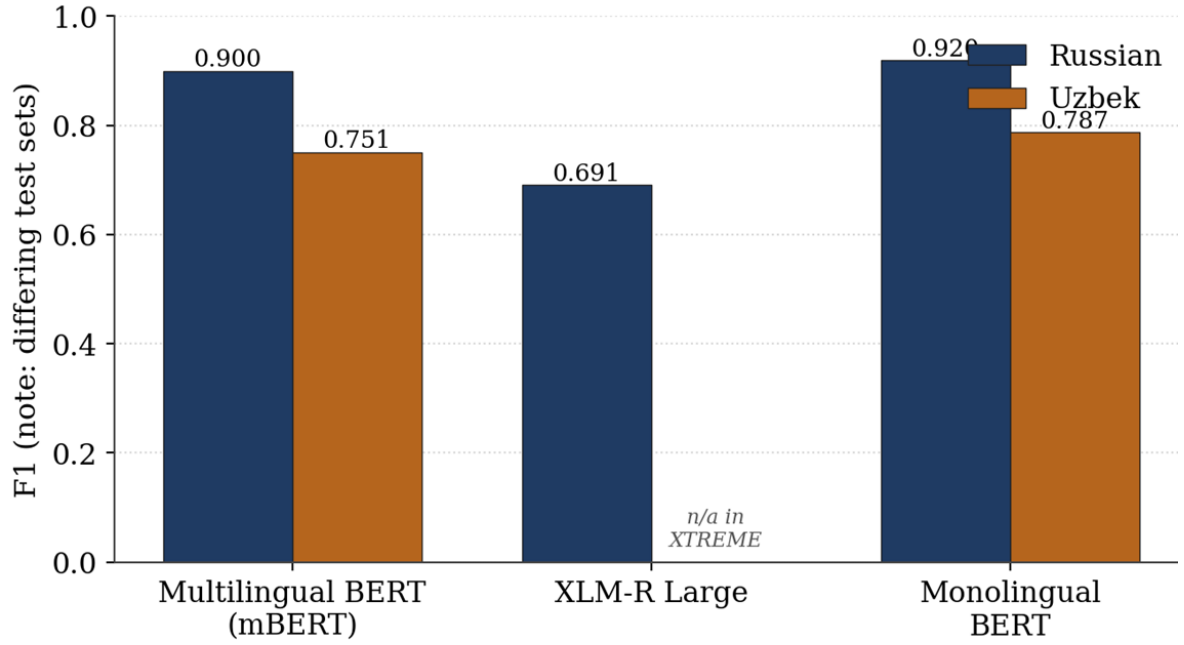

Figure 2. Russian and Uzbek F1 for three model families

Cross-lingual transfer is where the asymmetry hardens into a measurable boundary. The WikiANN resource of X. Pan et al. covers both languages, reporting silver-standard F1 of 90.1 for Russian over 1.4 million name mentions and 98.3 for Uzbek over a far smaller 92 000, figures that flatter Uzbek precisely because they are scored against Wikipedia-derived gold rather than human annotation. The method that produced them is described candidly by its authors. “We achieve this goal by performing a series of new KB mining methods: generating 'silver-standard' annotations by transferring annotations from English to other languages through cross-lingual links and KB properties, refining annotations through self-training and topic selection, deriving language-specific morphology features from anchor links, and mining word translation pairs from cross-lingual links.”

DISCUSSION

Reading the two records together yields a clear methodological lesson about what a fair comparison can and cannot claim. The monolingual transformer is the right default for both languages, but the routes to it differ. Russian researchers inherit a settled choice between NEREL for nested evaluation and the Collection5 union for legacy comparability, a public Slovnet checkpoint that runs on modest hardware, and a decade of error analysis to lean on. An Uzbek group faces an open construction problem instead, where the training mass of UZNER must be combined with the domain robustness of the Mengliev legal set, where the encoder choice between BERTbek and UzBERT turns on whether the input arrives in Latin or Cyrillic, and where a morphological post-processor is not an optional refinement but a load-bearing component. The evidence from D. Mengliev et al. is concrete on this last point, since their hybrid post-processing layer raised mBERT to 95 per cent F1 on their three-thousand-sentence corpus from a baseline that merely matched existing tools.

The deeper caution concerns the comparison itself. A bare statement that Russian NER reaches 0.92 while Uzbek reaches 0.79 conceals more than it reveals, because the Russian figure comes from news text annotated with a mature scheme and the Uzbek figure from a six-category split that is harder in some respects and easier in others. Silver-standard WikiANN scores invert the apparent ranking, handing Uzbek a higher number that reflects evaluation against machine-built gold rather than genuine difficulty. A defensible comparison reports three axes at once, namely intrinsic monolingual F1 on each language's own benchmark, zero-shot WikiANN transfer where both are present, and entity-type-stratified F1 that exposes which categories drive the averages. Anything less invites a conclusion the data do not support. The threshold that would change the recommendation is also visible on the horizon, since a public Uzbek corpus exceeding fifty thousand manually annotated entities with ten or more fine-grained types would shift the whole Uzbek pipeline onto XLM-R Large and away from the present reliance on a small monolingual encoder.

CONCLUSION

Russian named entity recognition is a mature, transformer-saturated field with six benchmarks and a top monolingual F1 near 0.92, whereas Uzbek crossed the transformer threshold only in 2023 and 2024 and now reaches roughly 0.79 to 0.92 depending on the corpus and scheme. Monolingual encoders beat multilingual baselines in both languages, data quality outweighs raw token count for Uzbek, and the typological distance between a fusional Cyrillic language and an agglutinative dual-script one makes any single-number ranking unsafe.

REFERENCES

1. Arkhipov M. Tuning Multilingual Transformers for Language-Specific Named Entity Recognition / M. Arkhipov, M. Trofimova, Y. Kuratov, A. Sorokin // Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. – Florence: ACL, 2019. – P. 89-93.
2. Conneau A. Unsupervised Cross-lingual Representation Learning at Scale / A. Conneau, K. Khandelwal, N. Goyal [et al.] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – 2020. – P. 8440-8451.
3. Hu J. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation / J. Hu, S. Ruder, A. Siddhant [et al.] // Proceedings of the 37th International Conference on Machine Learning. – 2020. – P. 4411-4421.
4. Kuratov Y. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language / Y. Kuratov, M. Arkhipov // arXiv preprint. – 2019.
5. Kuriyozov E. BERTbek: A Pretrained Language Model for Uzbek / E. Kuriyozov, D. Vilares, C. Gómez-Rodríguez // Proceedings of the 3rd Workshop on NLP for Similar Languages, Varieties and Dialects (SIGUL) @ LREC-

COLING. – Torino, 2024. – P. 33-44.

6. Loukachevitch N. NEREL: A Russian Dataset with Nested Named Entities, Relations and Events / N. Loukachevitch, E. Artemova, T. Batura [et al.] // Proceedings of RANLP. – 2021. – P. 876-885.