Generalization in Slavic Multi-Word Expression Lemmatization: A Comparative Study of Polish, Russian and Ukrainian

Anonymous ACL submission

Abstract

Multi-word expression (MWE) lemmatization requires modeling syntactic headship, agreement, and orthographic conventions across tokens, making it harder than single-token lemmatization. Yet it remains underexplored for morphologically rich Slavic languages and, for Ukrainian, essentially unstudied. We present a comparative study of MWE lemmatization for Polish, Russian, and Ukrainian, introducing RULEM for Russian and UALEM for Ukrainian, the first dedicated Ukrainian MWE lemmatization benchmark. We evaluate monolingual and multilingual encoder–decoder models, as well as instruction-tuned LLMs, under supervised, silver-transfer, and cross-lingual settings. Results show that left-context conditioning and silver transfer, previously effective for Polish, do not carry over uniformly to Russian and Ukrainian. Joint multilingual training provides a competitive single-model alternative, although its gains are not uniform across languages and evaluation settings. Instruction-tuned LLMs offer no consistent advantage over specialized fine-tuned models across languages and configurations, although the strongest LLM approaches fine-tuned performance in selected settings.

1 Introduction

Lemmatization is a core normalization step for morphologically rich languages, where surface forms vary across case, number, gender, and derivation. The task is harder for multi-word expressions (MWEs), including names, institutional names, titles, and semi-fixed nominal constructions, because their lemmas often cannot be recovered by normalizing tokens independently. Correct MWE lemmatization requires modeling syntactic headship, agreement, lexicalized structure, and orthographic conventions.

These difficulties are especially acute in Slavic languages, where rich inflection, relatively free word order, and frequent multi-word nominal phrases complicate canonicalization. Yet phrase-level MWE lemmatization remains much less studied than single-token contextual lemmatization, despite its relevance to named entity normalization, linking, and retrieval. Polish has recently received focused attention, but comparable benchmarks and results for Russian and Ukrainian remain limited.

This paper studies MWE lemmatization for Polish, Ukrainian, and Russian in a unified framework. We conduct new supervised and silver-transfer experiments for Russian and Ukrainian, compare against previously published Polish results (Król et al., 2025), and evaluate cross-lingual transfer and instruction-tuned LLMs. Our goal is to determine when fine-tuned encoder–decoder models remain stronger than larger generative models.

1.1 Lemmatization in Morphologically Rich Languages

Modern lemmatization research has moved from rule-based and edit-script methods toward contextual neural models. The SIGMORPHON 2019 shared task established contextual lemmatization and cross-lingual transfer as central benchmark problems for morphologically rich languages (McCarthy et al., 2019), with strong systems based on contextualized embeddings, morphological regularization, and multilingual fine-tuning (Straka et al., 2019; Kondratyuk, 2019).

Recent studies further show that lemmatization quality depends not only on model size, but also on explicit morphological information and task representation (Toporkov and Agerri, 2024b,a). Although Toporkov et al. (2025) report strong LLM performance for single-token

contextual lemmatization, phrase-level MWE lemmatization poses different challenges: the model must resolve syntactic headship, agreement, and multi-token canonicalization simultaneously. This distinction motivates our empirical comparison.

1.2 Multi-Word Expression Lemmatization

MWE lemmatization has received much less attention than single-token lemmatization. Early work was largely rule-based and language-specific, including Polish person names (Piskorski et al., 2007) and Polish multi-word common noun phrases and named entities (Marcińczuk, 2017). Schmitt and Constant (2019) introduced a language-independent encoder–decoder approach and showed that neural MWE lemmatization is feasible, while also identifying Polish as a difficult setting.

Recent work framed Polish MWE lemmatization as a text-to-text generation task with transformer models (Król et al., 2025). We extend this line of research to Russian, Ukrainian, and cross-lingual settings. Beyond Polish, phrase-level MWE lemmatization remains largely unexplored. Russian appears mainly in single-token lemmatization and Slavic named-entity normalization work, where proper-name MWEs are normalized but broader common-noun phrases are not the main target. To our knowledge, no prior study reports supervised text-to-text results on a Russian gold benchmark for phrase-level MWE lemmatization. For Ukrainian, the PARSEME Ukrainian MWE corpus (Sytar et al., 2026) provides the source material for our gold data, but Ukrainian MWE lemmatization has not previously been studied as a stand-alone task.

1.3 Slavic Normalization, LLMs, and Transfer

Slavic named-entity normalization is relevant because it covers one important real-world subclass of MWE lemmatization. The Slavic shared tasks established multilingual evaluation settings for inflectionally rich languages (Piskorski et al., 2017, 2019, 2021; Yangarber et al., 2023), and recent work introduced broader cross-lingual Slavic named-entity resources (Piskorski et al., 2024). Other studies explored BERT-, T5-, and LLM-based approaches for Slavic named entity recognition, linking, and lemmatization (Pałka and Nowakowski, 2023; Vīksna et al., 2023). However, these settings do not answer whether larger instruction-tuned models generalize better than fine-tuned encoder–decoder models.

Recent LLM work suggests that generative models can capture multi-word structure, for example in idiomatic MWE identification (Hashiloni et al., 2025), while related seq2seq work shows that lemmatization can benefit from transfer between close language varieties (Saeed and Habash, 2025). Together with recent LLM results for contextual lemmatization (Toporkov et al., 2025), this motivates our central question: whether larger instruction-tuned models generalize better than fine-tuned encoder–decoder models for phrase-level lemmatization in related Slavic languages..

2 Research Questions and Contributions

We organize the study around the following five research questions:

- RQ1 (In-language performance): How well do fine-tuned encoder–decoder models perform on in-language MWE lemmatization for Russian and Ukrainian, compared with existing Polish reference results?
- RQ2 (Effect of silver supervision): Does silver transfer learning improve MWE lemmatization quality for Russian and Ukrainian, as previously found for Polish, and is its effect stronger in noisier or lower-resource settings?
- RQ3 (Cross-lingual transfer): Can knowledge learned from one Slavic language help another through multilingual fine-tuning?
- RQ4 (LLMs vs. fine-tuned models): Are instruction-tuned LLMs better at extrapolating to unseen MWE patterns than fine-tuned encoder–decoder models, especially in low-resource and out-of-domain settings?
- RQ5 (Benchmark validity): What kinds of apparent model failures reflect genuine generalization problems, and what kinds instead point to reference ambiguity or benchmark artifacts?

To address these questions, the paper makes four contributions. (i) We present a comparative study of MWE lemmatization across Polish, Russian, and Ukrainian, with new controlled experiments for Russian and Ukrainian and previously published Polish results as a reference point. (ii) We introduce two new resource families, RULEM (Russian) and UALEM (Ukrainian), each pairing manually annotated gold data with Wikipedia-derived silver data. (iii) We test whether the left-context and silver transfer gains reported for Polish (Król et al., 2025) carry over to Russian and Ukrainian, and measure cross-lingual transfer among the three languages. (iv) We compare instruction-tuned LLMs with fine-tuned encoder–decoder models, finding that LLMs offer no consistent advantage on phrase-level MWE lemmatization.

3 Datasets and Data Preparation

We use one gold benchmark and one silver corpus per language. For Polish, we compare against previously published resources and results (Król et al., 2025). For Russian and Ukrainian, we introduce two new resource families: RULEM and UALEM. In all three languages, gold data is used for supervised training and evaluation, while silver data is used for an intermediate silver-transfer stage before gold fine-tuning.

The silver corpora are derived from Wikipedia hyperlinks. A hyperlink connects an inflected mention in running text to an article title; the anchor text is treated as the inflected MWE and the article title as its target lemma. This yields large numbers of candidate (inflected form, lemma) pairs, optionally with preceding left context. Since hyperlink anchors are noisy, we retain only pairs whose corresponding tokens have compatible morphological normal forms, with language-specific filtering described below.

Gold annotation and splits. The Russian and Ukrainian gold sets were manually annotated and split into train and test partitions by grouping examples according to the target MWE lemma, so that all examples with the same lemma occur in the same partition. The Russian gold set contains 3,661 training and 1,219 test examples; the Ukrainian gold set contains 3,767 training and 1,255 test examples. Because the gold sets are small, no separate gold development split is used; model selection relies on manually corrected silver-derived validation material. The target lemma is treated as the canonical dictionary-like form of the full MWE, with nominal heads normalized to nominative singular where applicable and capitalization preserved for named entities.

Table 1: Dataset sizes used in the experiments.

Lang.	Gold train	Gold test	Silver train
Russian	3,661	1,219	797K / 10K
Ukrainian	3,767	1,255	9.5K
Polish	–	–	published reference

3.1 Polish Data

For Polish, we use the benchmark and published experimental results of Król et al. (2025), derived from the PolEval 2019 task on lemmatization of multi-word expressions and proper names, together with a Polish Wikipedia-derived silver corpus. Polish results are therefore used as a previously established reference point rather than re-trained in this study. Because the Polish results are taken from a previously published reference setting rather than re-trained under the exact same experimental pipeline, cross-language comparisons involving Polish should be interpreted as reference comparisons rather than fully controlled head-to-head comparisons.

3.2 Ukrainian Data

The UALEM family consists of a manually annotated gold set and a Wikipedia-derived silver set. The gold set was annotated from Ukrainian PARSEME[1] interview transcripts (Sytar et al., 2026), covering politics, religion, history, and culture, and mostly contains proper and common nominal phrases with left context. The source corpus is publicly available through the PARSEME shared-task repository. The Ukrainian silver corpus is much smaller than the Russian one (9.5K vs. 797K), because Ukrainian Wikipedia yields fewer candidate hyperlink pairs and fewer inflected MWE anchors; larger noisy silver variants caused regression in

[1] https://gitlab.com/parseme/sharedtask-data/-/tree/master/2.0/subtask1/UK

preliminary experiments, so we retained only a high-confidence filtered subset.

3.3 Russian Data

The RULEM family pairs a manually curated gold set with a Wikipedia-derived silver set. For Russian, candidate silver pairs were extracted from a Russian Wikipedia dump and required heavier cleaning than the Ukrainian silver data: one-word mentions and markup artifacts were discarded before morphological filtering.

4 Method

We formulate MWE lemmatization as text-to-text generation: given an inflected MWE and optionally its left context, the model generates the canonical lemma of the full expression.

4.1 Task Variants

We distinguish three input settings: no-format, where the model receives only the target MWE; formatted no-context, where the task template is preserved but the context field is empty; and left-context, where the same template includes a fixed-size preceding context. Models are trained in the formatted left-context setting unless stated otherwise. At evaluation time, the formatted no-context variant serves as a diagnostic condition separating the effect of task formatting from actual contextual evidence.

4.2 Model Families

We compare three model families. Monolingual encoder–decoder models are fine-tuned separately for Russian and Ukrainian, using T5-style base and large variants; Polish is represented by previously published monolingual reference results. Multilingual encoder–decoder models, such as mT5-large, test whether shared pretraining improves cross-lingual transfer. Instruction-tuned LLMs are evaluated in zero-shot and few-shot settings to test whether general-purpose generative models extrapolate better under low supervision, domain shift, or reduced lexical overlap (Toporkov et al., 2025; Hashiloni et al., 2025).

4.3 Language Conditioning

In multilingual joint training, we compare a combined variant trained on pooled Polish, Russian, and Ukrainian data without language tags, and a conditioned variant that prepends a language tag, e.g. <pl>, <ru>, or <ua>. This tests whether explicit language identity helps separate shared Slavic regularities from language-specific canonicalization conventions.

5 Experimental Design

Our experiments are organized around four complementary settings corresponding to the research questions introduced in Section 2. Throughout, we evaluate with exact match in a case-sensitive variant (EM-CS) and a case-insensitive variant (EM-CI); the LLM comparison additionally reports token-level F_1, computed over whitespace-tokenized predicted and reference lemmas.

In preliminary experiments, the formatted no-context condition consistently outperformed the unformatted input condition. We therefore use the formatted no-context condition as the main no-context baseline and retain the unformatted setting only as a diagnostic check. For subsequent experiments, we continue with the three strongest models per language to keep the comparison tractable.

5.1 Experiment 1: In-language gold-data-only Supervised Learning

The first experiment evaluates supervised in-language performance for Russian and Ukrainian, with previously published Polish results used as a reference point.

We fine-tune the selected Russian and Ukrainian models on the corresponding gold data using the formatted left-context setting.

Table 2: Experiment 1: in-language supervised models fine-tuned on gold data, evaluated on the with-context and no-context test sets.

	With context		No context	
Model	EM-CS	EM-CI	EM-CS	EM-CI
Russian				
enruT5-base	92.5	93.4	91.9	93.2
mT5-large	95.6	96.1	94.4	95.3
ruT5-large	96.3	96.6	95.1	95.7
Ukrainian				
T5-large-ua-news	93.2	93.6	92.4	92.8
mT5-large	94.2	94.4	92.7	93.2
spivavtor-large	93.1	93.5	91.6	92.1

Table 2 reports the gold-supervised results: ruT5-large is strongest for Russian and mT5-large for Ukrainian, with every model scoring

higher on the with-context test set than without context.

5.2 Experiment 2: Silver Transfer Learning

The second experiment measures the effect of silver supervision. For Russian and Ukrainian, each model is first fine-tuned on the newly prepared silver dataset and then fine-tuned on the gold data. We compare these silver-then-gold models against the gold-only models of Experiment 1 (Table 2), reporting the main gains and regressions in the discussion rather than adding separate delta columns.

Table 3: Experiment 2: models trained on silver data and then fine-tuned on gold, evaluated on the with-context and no-context test sets. Best scores in each language and condition are bolded.

	With context		No context	
Model	EM-CS	EM-CI	EM-CS	EM-CI
Russian				
enruT5-base	92.5	93.5	90.4	92.8
ruT5-large	96.4	97.2	95.9	96.9
mT5-large	96.1	96.6	94.4	95.7
Ukrainian				
T5-large-ua-news	93.4	93.8	93.1	93.6
mT5-large	94.9	95.2	92.7	93.2
spivavtor-large	93.6	93.9	93.1	93.6

Table 3 compares the silver-then-gold models with the gold-only models of Table 2. The effect of silver transfer is mostly positive but uneven. On Ukrainian, all three models improve or hold steady, with spivavtor-large gaining about 1.5 points without context. On Russian, the strongest model, ruT5-large, improves on all four metrics – by up to +1.6 EM-CI without context – and mT5-large improves on three; only enruT5-base regresses on the no-context test set, losing 1.5 EM-CS. Silver supervision therefore sharpens the better systems but does not rescue the weakest, and it leaves the best model per language unchanged from Experiment 1 (ruT5-large for Russian, mT5-large for Ukrainian).

5.3 Experiment 3: Cross-Lingual Transfer Between Polish, Russian and Ukrainian

We test cross-lingual transfer by jointly fine-tuning a single multilingual model on the combined Polish, Russian, and Ukrainian gold data, optionally with a language tag prepended to each input (language conditioning). Comparing both variants against the best monolingual reference model shows whether the three related Slavic languages share enough morphological and structural regularity for joint training to help.

Table 4: Experiment 3: multilingual mT5-large models compared with the strongest single-language reference models. Russian and Ukrainian scores are reported on test sets; Polish is reported on a previously published validation/reference setting because no directly comparable Polish test set is available.

Model	Split	EM-CS	EM-CI
Russian			
Multilingual	test	96.6	97.0
Conditioned	test	95.9	96.2
Best single-lang.	test	96.4	97.2
Ukrainian			
Multilingual	test	94.3	94.7
Conditioned	test	93.9	94.4
Best single-lang.	test	94.9	95.2
Polish			
Multilingual	valid/ref.	87.4	88.2
Conditioned	valid/ref.	88.8	89.5
Best single-lang.	valid/ref.	95.33	97.3

The Polish monolingual score is taken from a previously published Polish reference setting (Król et al., 2025).

Table 4 shows that joint multilingual fine-tuning is competitive with the strongest single-language systems for Russian and Ukrainian, but does not consistently improve over them. For Russian, the multilingual model reaches the best with-context EM-CS and nearly matches the best EM-CI score, while for Ukrainian it remains close to the best silver-transfer model. Language conditioning does not improve performance over the plain multilingual setting. Polish is included only as a previously published validation/reference comparison, since no directly comparable Polish test set is available. Overall, multilingual training offers a practical single-model alternative, but its advantage is not uniform across languages or evaluation settings.

5.4 Experiment 4: Fine-Tuned Models vs. LLMs

We compare six open-weight instruction-tuned LLMs (Table 5) against the fine-tuned seq2seq

models of Experiments 1–3. Each is prompted in English and in the dataset language, with and without left context, at 0, 3, 5, and 8 shots; Table 6 reports each model's best configuration. We therefore interpret the LLM results as a best-case prompting comparison rather than as evidence of configuration stability. Full prompt grids are used for model selection, but the main table reports only the peak EM-CI configuration to keep the comparison compact.

Table 5: Models evaluated in the experiments.

Name	HuggingFace ID
Bielik-7B	speakleash/Bielik-Minitron-7B-v3.0-Instruct
GigaChat-10B	ai-sage/GigaChat3.1-10B-A1.8B-bf16
Lapa-0.1.2	lapa-llm/lapa-v0.1.2-instruct
Mamay-12B	INSAIT-Institute/MamayLM-Gemma-3-12B-IT-v1.0
Gemma-4-31B	google/gemma-4-31B-it
Gemma-4-E4B	google/gemma-4-E4B-it

Table 6: Best LLM results on MWE lemmatisation for each language. The Split (C/NC), Prompt, and Shots columns report the configuration that achieved the peak EM-CI for each model.

Model	EM-CI	F_1	Split	Prompt	Shots
Russian					
Gemma-4-31B	93.2	95.9	NC	ru	5
Mamay-12B	76.2	84.6	NC	en	8
GigaChat-10B	69.7	79.9	NC	ru	5
Gemma-4-E4B	67.5	81.8	C	en	5
Lapa-0.1.2	55.4	69.7	NC	ru	8
Bielik-7B	54.7	73.9	NC	ru	8
Ukrainian					
Gemma-4-31B	90.7	94.7	NC	ua	8
Mamay-12B	63.1	74.1	C	ua	8
Gemma-4-E4B	61.1	78.8	NC	ua	5
GigaChat-10B	60.6	76.0	NC	ua	0
Bielik-7B	57.3	73.6	C	ua	5
Lapa-0.1.2	53.3	70.5	C	ua	8
Polish					
Gemma-4-31B	92.0	95.4	C	en	3
Lapa-0.1.2	85.5	92.1	NC	pl	8
Bielik-7B	83.0	91.6	C	pl	8
Gemma-4-E4B	79.3	88.8	C	en	3
Mamay-12B	75.8	84.0	C	pl	0
GigaChat-10B	74.5	85.1	C	en	8

Table 6 shows Gemma-4-31B far ahead of every other model – by 17–27 EM-CI points – in all three languages, while the other models vary widely by language. The best prompt configuration is also language-dependent: no context suits Russian and Ukrainian, Polish benefits from context, and the optimal shot count ranges from 3 to 8. The strongest LLM, Gemma-4-31B, approaches the specialized systems in some settings, especially for Russian, but this behavior is not consistent across models or languages.

Error type	Typical pattern	Interpretation
Morphology	Wrong gender, number, case, or POS normalization	Genuine lemmatization error
Boundary	Modifier dropped, head-only output, or extra token generated	Genuine MWE-structure error
Lexical substitution	Related or stem-sharing word generated instead of the lemma	Semantic overgeneration, mostly LLM-like
Capitalization	Case differs from the reference, often in names or sentence-initial phrases	Mixed: model error or reference artifact
Degeneration	Repeated output, copied prompt fragments, or malformed string	Decoding/generation failure
Reference issue	Plausible prediction disagrees with noisy or inconsistent gold	Benchmark artifact

Table 7: Main error types observed in manual analysis of LLM and fine-tuned model outputs.

5.5 Manual Analysis

Automatic exact-match evaluation is insufficient for phrase-level lemmatization in noisy or inconsistently annotated settings. We therefore complemented the main evaluation with manual qualitative analysis, inspecting representative LLM and fine-tuned model errors across the main experimental conditions and grouping them by likely source. This analysis identifies recurring modeling failures and reference artifacts, but does not estimate their corpus-wide prevalence.

6 Limitations

Our study has several limitations. First, the three languages are not evaluated under fully identical conditions: Polish is included through

previously published reference results (Król et al., 2025), while Russian and Ukrainian are newly evaluated here. The benchmarks also differ in resource maturity, annotation policy, and construction, so cross-language differences should not be interpreted solely as linguistic difficulty. Second, the silver corpora are automatically derived from Wikipedia hyperlinks and remain noisy despite filtering and deduplication. Third, the LLM comparison is protocol-dependent, as results may vary with prompt language, context setting, shot count, decoding parameters, and example selection. Finally, exact-match metrics do not fully capture acceptable lemma variants, near-correct outputs, or reference inconsistencies. We therefore complement automatic scores with manual error analysis, while broader validation across additional Slavic languages and MWE types remains future work.

7 Conclusion

We presented a comparative study of MWE lemmatization for Polish, Russian, and Ukrainian, introducing RULEM for Russian and UALEM for Ukrainian. Together with previously published Polish reference results, these resources enable a cross-lingual comparison of phrase-level lemmatization in three morphologically rich Slavic languages. Across supervised, silver-transfer, cross-lingual, and prompting-based settings, our results show that MWE lemmatization does not simply follow trends observed for single-token contextual lemmatization.

Fine-tuned encoder–decoder models remain strongest overall, although Gemma-4-31B approaches them in selected configurations, especially for Russian. This suggests that large instruction-tuned models can capture part of the required multi-token structure, but they do not provide a consistent advantage across languages, model families, or prompting conditions. Silver transfer is useful but uneven: it improves the strongest Russian and Ukrainian systems, yet does not uniformly benefit all models. Joint multilingual training offers a competitive single-model alternative, but its effect varies across languages and evaluation settings.

Manual analysis further shows that exact-match errors mix genuine modeling failures with capitalization variation, plausible lemma variants, and reference inconsistencies. This is particularly important for MWE lemmatization, where correct canonicalization depends on phrase boundaries, syntactic headship, agreement, and named-entity conventions. Our findings therefore support evaluation protocols that combine automatic scores with linguistically informed error analysis, and point to the need for stronger Slavic MWE resources, especially for under-studied languages such as Ukrainian.

References

Kai Golan Hashiloni, Ofri Hefetz, and Kfir Bar. 2025. Easy as PIE? identifying multi-word expressions with LLMs. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23771–23790, Suzhou, China. Association for Computational Linguistics.

Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. In Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 12–18, Florence, Italy. Association for Computational Linguistics.

Magdalena Król, Aleksander Smywiński-Pohl, Zbigniew Kaleta, and Paweł Lewkowicz. 2025. Lemmatization of Polish multi-word expressions. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22148–22157, Suzhou, China. Association for Computational Linguistics.

Michał Marcińczuk. 2017. Lemmatization of multi-word common noun phrases and named entities in Polish. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 483–491, Varna, Bulgaria. INCOMA Ltd.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Gabriela Pałka and Artur Nowakowski. 2023. Exploring the use of foundation models for named entity recognition and lemmatization tasks in Slavic languages. In Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023

(SlavicNLP 2023), pages 165–171, Dubrovnik, Croatia. Association for Computational Linguistics.

Jakub Piskorski, Bogdan Babych, Zara Kancheva, Olga Kanishcheva, Maria Lebedeva, Michał Marcińczuk, Preslav Nakov, Petya Osenova, Lidia Pivovarova, Senja Pollak, Pavel Přibáň, Ivaylo Radev, Marko Robnik-Sikonja, Vasyl Starko, Josef Steinberger, and Roman Yangarber. 2021. Slav-NER: the 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, pages 122–133, Kiyv, Ukraine. Association for Computational Linguistics.

Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 2019. The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 63–74, Florence, Italy. Association for Computational Linguistics.

Jakub Piskorski, Michał Marcińczuk, and Roman Yangarber. 2024. Cross-lingual named entity corpus for slavic languages. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 4143–4157, Torino, Italia. ELRA and ICCL.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. The first cross-lingual challenge on recognition, normalization, and matching of named entities in Slavic languages. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 76–85, Valencia, Spain. Association for Computational Linguistics.

Jakub Piskorski, Marcin Sydow, and Anna Kupść. 2007. Lemmatization of Polish person names. In Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, pages 27–34, Prague, Czech Republic. Association for Computational Linguistics.

Mostafa Saeed and Nizar Habash. 2025. Lemmatizing dialectal arabic with sequence-to-sequence models. In Proceedings of the Third Arabic Natural Language Processing Conference, pages 117–129, Suzhou, China. Association for Computational Linguistics.

Marine Schmitt and Mathieu Constant. 2019. Neural lemmatization of multiword expressions. In Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 142–148, Florence, Italy. Association for Computational Linguistics.

Milan Straka, Jana Straková, and Jan Hajič. 2019. UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. In Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 95–103, Florence, Italy. Association for Computational Linguistics.

Hanna Sytar, Maria Shvedova, and Olha Kanishcheva. 2026. Ukrainian multiword expressions corpus: Creation, annotation, and linguistic analysis. In Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 38–47, Rabat, Marocco. Association for Computational Linguistics.

Olia Toporkov and Rodrigo Agerri. 2024a. Evaluating shortest edit script methods for contextual lemmatization. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 6451–6463, Torino, Italia. ELRA and ICCL.

Olia Toporkov and Rodrigo Agerri. 2024b. On the role of morphological information for contextual lemmatization. Computational Linguistics, 50(1):157–191.

Olia Toporkov, Alan Akbik, and Rodrigo Agerri. 2025. Lemma dilemma: On lemma generation without domain- or language-specific training data. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18219–18232, Suzhou, China. Association for Computational Linguistics.

Rinalds Vīksna, Inguna Skadiņa, Daiga Deksne, and Roberts Roziş. 2023. Large language models for multilingual slavic named entity linking. In Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023), pages 172–178, Dubrovnik, Croatia. Association for Computational Linguistics.

Roman Yangarber, Jakub Piskorski, Anna Dmitrieva, Michał Marcińczuk, Pavel Přibáň, Piotr Rybak, and Josef Steinberger. 2023. Slav-NER: the 4th cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023), pages 179–189, Dubrovnik, Croatia. Association for Computational Linguistics.