Mining Native Ukrainian Paraphrases: A Multi-Source Comparison

Vladyslav Fesenko, Hanna Dydyk-Meush, Volodymyr Mudryi
Ukrainian Catholic University, Lviv, Ukraine
{fesenko.pn, hanna_dydykmeush, mudryi.pn}@ucu.edu.ua

Abstract

We introduce a Ukrainian paraphrase dataset mined from event-aligned news headlines and compare it with translated and LLM-generated data sources. Candidate pairs are retrieved from native Ukrainian news titles and filtered using semantic and lexical constraints to form a training corpus in a semi-automatic pipeline. Human evaluation indicates that the sources differ in useful ways: LLM-generated paraphrases are generally stronger in meaning preservation, whereas news-mined pairs offer greater lexical variation while remaining fluent and meaning-preserving. We tune mT5-large and mT0-large and evaluate them on several held-out test sets, including a human-validated subset. Relative to Spivavtor-large, the models achieve comparable semantic preservation with lower copying on the combined and human-validated sets. Overall, the findings highlight the value of naturally mined Ukrainian paraphrases as supervision for low-resource paraphrase generation.

1 Introduction

Paraphrase resources serve two distinct roles in NLP: they supervise generation models and they test whether systems can preserve meaning under non-trivial reformulation. The strongest datasets for these purposes remain concentrated in English, including news-derived corpora such as MRPC (Dolan and Brockett, 2005), question-duplicate collections such as QQP and WikiAnswers (DataCanary et al., 2017; Fader et al., 2013), and challenge sets such as PAWS (Zhang et al., 2019). Ukrainian does not yet have a comparable resource designed specifically for paraphrase generation and evaluation.

The missing piece is not only scale but also a systematic comparison of native and synthetic paraphrase sources for Ukrainian. In low-resource settings, paraphrase supervision is often created by translating English corpora or prompting Large Language Models (LLMs) to rewrite text. Both strategies are useful, but they introduce different biases. Translation-based supervision scales cheaply, yet translated text often shows translationese effects and lower linguistic richness than native text (Zhang and Toral, 2019; Vanmassenhove et al., 2021). LLM-based rewriting avoids direct translation transfer, but unconstrained decoding in paraphrase generation can produce trivial copies or near copies, so explicit diversity controls are needed to obtain stronger lexical and structural variation (Thompson and Post, 2020; Holtzman et al., 2019). The result is a familiar trade-off between semantic reliability, naturalness, and lexical change.

We address this problem by mining paraphrases from independent Ukrainian news reports about the same event, which provides semantic anchoring. This idea follows the intuition behind naturally occurring paraphrase resources while adapting it to a low-resource language setting in which native supervision is scarce.

Following recent work on paraphrastic robustness (Srikanth et al., 2024), we treat semantic equivalence and variation in wording and structure as separate properties rather than collapsing them into a single similarity score. This distinction is especially important for Ukrainian, where prior robustness studies show that even small lexical substitutions can substantially affect model behavior and semantic fidelity (Mudryi and Ignatenko, 2025; Mudryi and Laba, 2025). This perspective motivates both our dataset construction choices and our evaluation setup.

Our contributions are:

- We construct a Ukrainian paraphrase corpus by mining news titles about the same event and filtering candidates with semantic and lexical constraints.
- We compare news-mined pairs against translated and LLM-generated data using human

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 199–208

judgments of meaning preservation, fluency, and lexical divergence.

- We evaluate adapter-tuned models trained on a balanced multi-source corpus and show competitive performance against Spivavtor baselines, with lower copying on the combined and human-validated sets.

2 Related Work

Paraphrase datasets differ mainly in how they obtain semantically aligned texts. Some rely on naturally occurring comparable descriptions, such as parallel news reports in MRPC (Dolan and Brockett, 2005), tweets linked to the same URL in Twitter URL (Lan et al., 2017), multiple captions for the same image in MS COCO, or duplicate questions in WikiAnswers and QQP (Lin et al., 2014; Fader et al., 2013; DataCanary et al., 2017). Across these settings, paraphrases are most natural when they arise from independent descriptions of the same underlying event, scene, or intent.

Other resources are constructed synthetically. PAWS, for example, creates high-overlap paraphrase and non-paraphrase pairs through word swapping and back-translation, followed by human validation (Zhang et al., 2019). In low-resource settings, similar synthetic strategies often take the form of translated English corpora or LLM-generated rewrites. These approaches scale well, but translated text may exhibit translationese and reduced linguistic richness (Zhang and Toral, 2019; Vanmassenhove et al., 2021), while unconstrained neural generation often produces conservative rewrites unless diversity is explicitly encouraged (Holtzman et al., 2019; Thompson and Post, 2020).

Low-resource paraphrase generation has largely focused on transfer rather than data construction. LAPA, for instance, adapts pretrained sequence-to-sequence models with adapters and meta-learning under limited supervision (Li et al., 2022). We instead focus on which kind of target-language supervision is most useful when the training budget is fixed.

For Ukrainian, the closest prior resource is Spivavtor (Saini et al., 2024), a text-editing dataset and instruction-tuned model suite that includes paraphrasing alongside other rewriting tasks. It is therefore an important baseline, but not a dedicated paraphrase corpus built from naturally occurring Ukrainian text. We also draw on the evaluation perspective of ParaNlu (Srikanth et al., 2024), which shows that paraphrase quality should be assessed jointly in terms of semantic equivalence and surface variation rather than semantic similarity alone.

3 Dataset Methodology

3.1 Data Source and Candidate Generation

We build our corpus from the *Ukrainian News* dataset (zeusfsx, 2024), a collection of 22.5 million news titles collected from Ukrainian online media. News headlines naturally form paraphrases because independent outlets reporting on the same event write alternative descriptions of the same content.

To identify candidates, we encoded the corpus using a multilingual sentence-transformer (Reimers and Gurevych, 2019) and retrieved nearest neighbors for each title using an approximate nearest-neighbor index (Johnson et al., 2021). Although all retrieved neighbors were considered initially, we restricted the candidate pool to pairs with similarity of at least 0.5. The thresholds were selected conservatively through manual inspection of candidate pairs, prioritizing removal of obvious semantic drift, near-duplicates, large information mismatches, and factual inconsistencies while preserving lexically diverse reformulations. Applying the cutoff therefore reduced the candidate set substantially and prevented unnecessary computational overhead in downstream filtering.

3.2 Filtering Pipeline

Raw candidate pairs contain substantial noise. We therefore applied several filtering steps to remove unsuitable pairs. These include removing near-duplicates with high word-overlap Jaccard scores (> 0.8), pairs with low semantic similarity (cosine score < 0.6), large length mismatches ($\text{len\_ratio} < 0.67$), and pairs containing inconsistent numeric values. Additional heuristics removed mirror duplicates, mixed-language content, non-Cyrillic markup, and temporally distant pairs (publication dates differing by more than three days). Examples of retained pairs and discarded pairs for each rule are provided in Appendix A.1.4 and Appendix A.1.5.

4 Dataset Comparison and Human Evaluation

4.1 Dataset Comparison

We compare our news-mined paraphrase pairs with three alternative Ukrainian sources that differ in construction. The first consists of our *news pairs*, independently written headlines about the same event mined from the *Ukrainian News* corpus. The second includes *LLM-generated pairs*, created through instruction-based rewriting by a large language model. The third contains *translated pairs*, obtained by translating English sentence pairs sampled from ParaNlu into Ukrainian (Srikanth et al., 2024). Finally, we include *Spivavtor-derived pairs*, sampled from the paraphrasing subset of the Ukrainian text-editing dataset (Saini et al., 2024). These sources represent different strategies for constructing paraphrase supervision: native reformulations, LLM rewrites, translated pairs, and instruction-based editing data.

Source	n	Len.	Overlap
News (ours)	213	77.8	0.267
LLM	141	84.9	0.174
Translated	144	48.3	0.180
Spivavtor	152	70.7	0.277
Overall	**650**	**71.2**	**0.230**

Table 1: Validated pairs used in human evaluation, with average paraphrase length (chars) and token overlap.

4.2 Human Evaluation Protocol and Results

For human evaluation, we sampled pairs from all four sources, merged them, shuffled, and removed source labels before annotation. Each pair was evaluated along three dimensions: *meaning preservation*, *lexical divergence*, and *fluency*. Meaning was scored on a three-point scale (0 = different meaning, 1 = similar meaning, 2 = identical meaning), lexical divergence on a three-point scale reflecting word overlap (0 = mostly the same words, 1 = partial overlap, 2 = different wording), and fluency on a binary scale (0 = not natural, 1 = natural). Annotation was performed by a professional linguist. To assess reliability, a subset of the data was independently annotated by a second annotator. Under this scheme, agreement exceeded 80% for meaning and fluency, and was around 75% for lexical divergence. In cases of disagreement, we adopted the labels of the primary annotator due to their linguistic expertise in Ukrainian. After validation and removal of incomplete entries, 650 pairs remained for analysis.

Source	Meaning	Lex. div.	Fluency	Total	n
LLM-gen.	1.94	0.12	0.91	2.98	141
News pairs	1.48	0.36	0.94	2.78	213
Translated	1.39	0.49	0.70	2.58	144
Spivavtor	1.44	0.44	0.65	2.53	152

Table 2: Human evaluation average scores by source family. Total is the sum of Meaning, Lexical divergence, and Fluency.

Table 2 shows a clear ordering: LLM-generated pairs have the highest aggregate score, while Spivavtor-derived pairs have the lowest. This also reveals an important trade-off: LLM outputs are strong on meaning preservation, but lexical divergence is the lowest (0.12), which indicates conservative, safe rewrites.

To check whether this pattern persists among acceptable paraphrases, we apply a stricter subset filter (meaning score ≥ 1 and fluency $= 1$) and recompute lexical divergence.

Source	Lexical div.	n
News pairs	0.259	185
Translated	0.209	91
Spivavtor	0.153	98
LLM-generated	0.078	128

Table 3: Lexical divergence on the strict subset with meaning score ≥ 1 and fluency $= 1$.

Table 3 shows that, under this strict filter, news pairs have the highest lexical divergence and LLM-generated pairs remain the lowest. This confirms that LLM rewrites are often safe edits even when meaning and fluency are acceptable.

To test whether score distributions differ across source families, we ran Kruskal–Wallis tests for meaning, lexical divergence, fluency, and total score. All tests were significant (all $p < 10^{-6}$), which indicates that the differences between data sources are systematic rather than random variation.

Overall, LLM-generated pairs obtain the highest aggregate score, but lexical changes are minimal. In this sense, they are not the strongest paraphrases when the goal is non-trivial reformulation rather than safe editing. At the same time, news-mined data offers the strongest balance between acceptable meaning, fluent language, and non-trivial lexical variation.

5 Fine-Tuning

5.1 Setup and Metrics

For downstream experiments, we train all models on one combined dataset of about 16,000 pairs. We sample the data evenly from four sources (news-mined, LLM-generated, translated, and Spivavtor-derived) to cover a broader paraphrase distribution. This training size is also comparable to the Spivavtor paraphrasing subset scale.

The two adapter-tuned models are mT5-large (Xue et al., 2021) and mT0-large (Muennighoff et al., 2023). We compare them with Spivavtor-large and Spivavtor-XXL. Spivavtor-large follows the same mT5/T5-style encoder–decoder family, while Spivavtor-XXL belongs to a much larger parameter class (13B class), so we treat it as a high-capacity reference.

Evaluation is performed on three held-out sets: the combined test split ($n = 1600$), the Spivavtor test split ($n = 1563$), and a human-validated subset ($n = 502$) containing pairs with meaning score ≥ 1 and fluency $= 1$.

BLEU against a single reference is not sufficient for paraphrase evaluation because many valid rewrites can differ from the gold surface form. We therefore report two reference-based overlap metrics, $BLEU_{ref}$ and $chrF^{++}_{ref}$ (Popović, 2017), together with input-based semantic similarity measured by *BERTScore* (Zhang et al., 2020) using *XLM-RoBERTa-large* (Conneau et al., 2020) and copying measured by $BLEU_{in}$. We do not include METEOR (Banerjee and Lavie, 2005) in the main table because its standard formulation depends on language-specific stemming and synonym resources that are not well-defined for Ukrainian in our setup; a surface-form fallback adds little beyond the other reference-based metrics. We also considered ParaScore (Shen et al., 2022), but do not include it because it is not directly adapted to Ukrainian out of the box; validating paraphrase-specific learned metrics for Ukrainian remains future work.

Following Zou et al. (2025), we compute *iBScore* as a combined measure of semantic preservation and anti-copying behavior:

$$iBScore = BERTScore_{in} - \frac{BLEU_{in}}{100}.$$

Higher *iBScore* indicates a better balance between semantic preservation and non-trivial reformulation.

5.2 Results

Table 4 shows a consistent pattern. Spivavtor-XXL obtains the highest *iBScore* on all three evaluation sets, so we treat it as a high-capacity reference rather than a directly comparable baseline. Among the non-XXL models, mT5-large is best on the combined and human-validated sets, while Spivavtor-large is best on the Spivavtor test split. At the same time, all models keep $BERTScore_{in}$ close to 0.98, which indicates similar semantic preservation.

The reference-based metrics add little extra separation. $BLEU_{ref}$ and $chrF^{++}_{ref}$ (Popović, 2017) show the same broad behavior: Spivavtor-XXL remains strongest on all splits, while the remaining systems stay close on the combined and human-validated sets and differ only modestly on the Spivavtor split. In other words, reference-based overlap metrics do not expose meaningful differences beyond the main copying-aware picture captured by *iBScore*.

The main differences come from input copying. On the combined test split, the adapter-tuned mT5/mT0 models show lower $BLEU_{in}$ than Spivavtor-large (roughly 68–69 vs. 69.46). On the human-validated subset, the gap is larger (roughly 69.7–70.2 vs. 71.99). On the Spivavtor test split, Spivavtor-large has lower copying than the two adapter-tuned models, consistent with an in-domain advantage.

To complement the automatic metrics, we manually evaluated 100 generated outputs using the same annotation dimensions as in the dataset evaluation: meaning preservation, lexical divergence, and fluency. The outputs show moderate meaning preservation: 27% were judged to have different meaning, 39% similar meaning, and 34% identical meaning, giving a mean meaning score of 1.07. Lexical divergence was relatively strong: 19% of outputs used mostly the same words, 45% showed partial wording change, and 36% used different wording, giving a mean lexical divergence score of 1.17. Fluency was the weakest dimension, with 63% of outputs judged natural and 37% judged not natural. Overall, 47% of outputs satisfied the criterion meaning ≥ 1 and fluency $= 1$, while 31% reached the stricter tier of identical meaning and fluent wording. During annotation, we also observed that many outputs with meaning changes or poor fluency came from the translated subset. This suggests that fluency issues in translated source texts can propagate to generated paraphrases. By

Eval set	Model	n	$iBScore \uparrow$	$BERTScore_{in} \uparrow$	$BLEU_{in} \downarrow$	$BLEU_{ref} \uparrow$	$chrF^{++}_{ref} \uparrow$
Combined test	Spivavtor-XXL	1600	0.3797	0.9779	59.82	13.37	36.87
Combined test	mT5-large (tuned)	1600	**0.2984**	0.9807	68.23	12.48	35.49
Combined test	mT0-large (tuned)	1600	0.2947	0.9818	68.71	12.48	35.52
Combined test	Spivavtor-large	1600	0.2867	0.9814	69.46	12.48	35.82
Spivavtor test	Spivavtor-XXL	1563	0.3777	0.9770	59.93	20.90	46.15
Spivavtor test	mT5-large (tuned)	1563	0.2854	0.9812	69.58	18.14	43.31
Spivavtor test	mT0-large (tuned)	1563	0.2701	0.9827	71.25	17.94	43.52
Spivavtor test	Spivavtor-large	1563	**0.3038**	0.9807	67.70	20.19	45.38
Human-validated subset	Spivavtor-XXL	502	0.3525	0.9795	62.70	15.74	41.15
Human-validated subset	mT5-large (tuned)	502	**0.2838**	0.9810	69.73	15.44	40.33
Human-validated subset	mT0-large (tuned)	502	0.2811	0.9827	70.16	15.41	40.42
Human-validated subset	Spivavtor-large	502	0.2625	0.9824	71.99	14.91	40.57

Table 4: Fine-tuning and baseline results on three evaluation sets. Lower $BLEU_{in}$ indicates less copying from the input. Bold marks the best result among non-XXL models; Spivavtor-XXL is included as a high-capacity reference.

contrast, the manually inspected outputs from the news-mined subset were generally stronger, which is consistent with our motivation for using native Ukrainian news text as supervision. These results suggest that the main remaining bottleneck is fluency rather than lexical variation.

Taken together, the automatic and human evaluations show that the adapter-tuned models can reduce copying while preserving meaning in many cases, but generation quality remains limited by fluency and requires further improvement.

6 Conclusion

We construct a Ukrainian paraphrase corpus by mining event-aligned news headlines and filtering candidates with semantic and lexical constraints. We compare this data with translated and LLM-generated pairs using human evaluation of meaning, fluency, and lexical divergence. We then fine-tune two adapter-based generation models and evaluate them against strong baselines on three held-out sets with $BLEU_{ref}$, $chrF^{++}_{ref}$, $BERTScore_{in}$, $BLEU_{in}$, and $iBScore$. Across analyses, news-mined pairs show the strongest lexical divergence under quality constraints, and balanced multi-source training supports competitive meaning-preserving generation with reduced copying on the combined and human-validated sets. The two reference-based metrics show the same broad behavior and do not materially separate the non-XXL models.

Data and Code Availability

For reproducibility, we provide public project repositories for the code, dataset, and model checkpoints. The implementation for data construction, fine-tuning, and evaluation is released at GitHub. The dataset and trained checkpoints are released through Hugging Face at dataset repository and model repository.

Limitations

This study has four main limitations. First, we fine-tuned only two large-model backbones (mT5-large and mT0-large, approximately 1.2B class), so broader architectural coverage remains open.

Second, the dataset is built entirely from news headlines. Headlines are short, compressed, and strongly event-centered, so they do not cover longer sentence structures, paragraph-level context, or other text types such as conversational, instructional, or literary prose. As a result, models trained on this data may not transfer reliably to longer inputs or broader domains. Our mining setup also assumes a one-to-one relation between two headlines. Natural paraphrasing can be more flexible: one sentence may correspond to several sentences, several sentences may be compressed into one, and meaning can be redistributed. Extending the approach beyond headline-level one-to-one pairs remains future work.

Third, due to time and compute limits, we trained on the combined training corpus only. We did not fine-tune separate models for each data source family, so we cannot directly test whether source-specific training reproduces the same ordering observed in human evaluation.

Fourth, we could not fine-tune models in the same scale class as Spivavtor-XXL (10B+), including Aya-101-scale systems. This limits conclusions about the best achievable quality under unconstrained compute.

Ethical Considerations

Our data source consists of publicly available Ukrainian news headlines. Although headlines are short, they can include offensive language, insensitive framing, and descriptions of violence or traumatic events. We therefore treat the corpus as potentially harmful content and assume that downstream models may reproduce such language.

Because headlines are collected from publisher websites, source-specific licensing and reuse terms may apply. We use the corpus for research and evaluation, and we recommend that any dataset release follows the legal and attribution requirements of the original sources.

The corpus can also reflect editorial bias, political framing, and factual inconsistency across outlets. Mining paraphrases from event-aligned headlines improves lexical diversity, but it does not remove these source-level biases. Our filtering pipeline reduces semantic drift and obvious mismatches, yet it cannot guarantee factual correctness of the underlying claims.

Human annotation introduces a second risk: annotators may be exposed to disturbing or offensive text. In this study, annotation was performed by a trained linguist, which improved judgment consistency on meaning and fluency. Future annotation rounds should keep clear skip rules and workload limits to reduce potential harm.

The resulting models can be used for beneficial tasks such as rewriting and educational support, but they can also be misused to rephrase misleading or manipulative content. For this reason, we frame the dataset and models as research resources, report their limitations explicitly, and encourage careful deployment with moderation policies.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

DataCanary, hilfialkaff, Lili Jiang, Meg Risdal, Nikhil Dandekar, and tomtung. 2017. Quora question pairs. `https://kaggle.com/competitions/quora-question-pairs`. Kaggle.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Zhigen Li, Yanmeng Wang, Rizhao Fan, Ye Wang, Jianfeng Li, and Shaojun Wang. 2022. Learning to adapt to low-resource paraphrase generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1014–1022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing.

Volodymyr Mudryi and Oleksii Ignatenko. 2025. Precision vs. perturbation: Robustness analysis of synonym attacks in Ukrainian NLP. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 131–146, Vienna, Austria (online). Association for Computational Linguistics.

Volodymyr Mudryi and Yurii Laba. 2025. From benchmark to better embeddings: Leveraging synonym substitution to enhance multimodal models in Ukrainian. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20458–20468, Suzhou, China. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Aman Saini, Artem Chernodub, Vipul Raheja, and Vivek Kulkarni. 2024. Spivavtor: An instruction tuned Ukrainian text editing model. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 95–108, Torino, Italia. ELRA and ICCL.

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 3178–3190.

Neha Srikanth, Marine Carpuat, and Rachel Rudinger. 2024. How often are errors in natural language reasoning due to paraphrastic variability? *Transactions of the Association for Computational Linguistics*, 12:1143–1162.

Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

zeusfsx. 2024. Ukrainian news dataset. https://huggingface.co/datasets/zeusfsx/ukrainian-news.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei Zou, Ziyuan Zhuang, Xiang Geng, Shujian Huang, Jia Liu, and Jiajun Chen. 2025. Improved paraphrase generation via controllable latent diffusion. *Frontiers of Computer Science*, 20(1).

A Appendix

A.1 Data Preparation

For candidate generation, we used the sentence encoder checkpoint `paraphrase-multilingual-mpnet-base-v2` (model card: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2).

Both LLM-based headline paraphrasing and English-to-Ukrainian translation were generated with Gemini 2.5 Flash.

A.1.1 LLM headline paraphrase prompt

For headline rewriting, we used the following system prompt.

System prompt:

> You are an expert news editor rewriting headlines.
> Your task is twofold:
> 1. **Clean**: Remove any artifacts from the input title (e.g., website names, "Read more", "Source: ...", truncated text "...").
> 2. **Paraphrase**: Write a new headline that conveys the EXACT SAME information as the cleaned title but uses different vocabulary and structure.
>
> Guidelines:
> - The paraphrase must be in Ukrainian.
> - Do not change names, numbers, or locations.
> - Avoid trivial changes (like just changing one word). Aim for structural variety.

Structured output schema:

- `cleaned_original`: The original title with artifacts removed.
- `paraphrase`: A semantic paraphrase of the cleaned title with different wording.

A.1.2 English-to-Ukrainian bucket translation prompt

The English source sentences were sampled from ParaNlu (Srikanth et al., 2024) and translated using the prompt below.

System prompt:

> You are an expert translator specializing in preserving semantic nuances and variation.
>
> Task: Translate the following list of English sentences into Ukrainian.
> Context: All these sentences are paraphrases of the SAME meaning.
>
> Constraints:
> 1. **Semantic Equivalence**: The meaning must be preserved.
> 2. **Diversity**: The translations must NOT be identical. You must vary the Ukrainian vocabulary and syntax to match the diversity of the English inputs.
> 3. **Quality & Filtering**: If an English sentence is nonsensical, weird, has artifacts, or would result in a broken/meaningless Ukrainian sentence, **SKIP IT**.
> - Do NOT translate nonsensical inputs.
> - It is strictly VALID (and encouraged) if the number of output Ukrainian sentences is smaller than the input list when some inputs are bad.

Structured output schema:

- `ukrainian_sentences`: list of valid Ukrainian translations; list length may be smaller than input due to filtering.

A.1.3 Ukrainian verbalizer prompts used in adapter training and generation.

During adapter fine-tuning and inference, we prepend one verbalizer sampled from the following fixed set:

- **Ukrainian text:** Перефразуй речення:
  English translation: Paraphrase the sentence:
- **Ukrainian text:** Перепиши речення іншими словами:
  English translation: Rewrite the sentence in other words:
- **Ukrainian text:** Перефразуй цей текст:
  English translation: Paraphrase this text:
- **Ukrainian text:** Перефразуй це речення:
  English translation: Paraphrase this sentence:
- **Ukrainian text:** Перефразуй:
  English translation: Paraphrase:
- **Ukrainian text:** Переформулюй це речення:
  English translation: Rephrase this sentence:
- **Ukrainian text:** Переформулюй цей текст:
  English translation: Rephrase this text:

A.1.4 Examples of retained paraphrase pairs

Table 5 shows examples of pairs retained after filtering. These examples illustrate the type of lexical and structural variation preserved in the final corpus.

Source	Sentence A	Sentence B
News	Рада ухвалила держбюджет-2022	Парламент затвердив державний бюджет на 2022 рік
EN	*The Verkhovna Rada adopted the 2022 state budget.*	*Parliament approved the state budget for 2022.*
News	Гелікоптер і літак, що належить родині Медведчука, передали для потреб ЗСУ.	Гелікоптер і літак Медведчука передали на потреби армії
EN	*A helicopter and an airplane belonging to Medvedchuk's family were transferred for the needs of the Armed Forces of Ukraine.*	*Medvedchuk's helicopter and airplane were transferred for the needs of the army.*
News	Буданов назвав дати оголошення нової мобілізації в РФ.	Буданов назвав дату, коли Росія розпочне нову мобілізацію
EN	*Budanov named the dates for the announcement of a new mobilization in Russia.*	*Budanov named the date when Russia will begin a new mobilization.*
News	Майже кожен другий українець незадоволений своєю роботою	Майже половина українців не задоволена своєю роботою
EN	*Almost every second Ukrainian is dissatisfied with their job.*	*Almost half of Ukrainians are dissatisfied with their job.*

Table 5: Examples of retained news-mined paraphrase pairs. English translations are provided for readability.

A.1.5 Filtering examples

Below we show examples of pairs discarded by the filtering pipeline. For each example, we provide both the original Ukrainian text and an English translation.

Near-duplicate filter (Jaccard > 0.8):

Sentence A (**Ukrainian text**): У Львові відреконструюють старе тролейбусне депо

Sentence B (**Ukrainian text**): У Львові реконструюють старе тролейбусне депо

Sentence A (**English translation**): The old trolleybus depot in Lviv will be reconstructed.

Sentence B (**English translation**): The old trolleybus depot in Lviv will be reconstructed.

Low semantic similarity (cosine < 0.6):

Sentence A (**Ukrainian text**): Кахім Перріс зіграв за збірну Ямайки проти Аргентини

Sentence B (**Ukrainian text**): Два ефектних голи Мессі дозволили Аргентині розгромити Ямайку

Sentence A (**English translation**): Kahim Perris played for the Jamaican national team against Argentina.

Sentence B (**English translation**): Two spectacular goals by Messi allowed Argentina to rout Jamaica.

Length disparity ($\text{len\_ratio} < 0.67$**):**

Sentence A (**Ukrainian text**): У Харкові відкрили новий реабілітаційний центр

Sentence B (**Ukrainian text**): У Харкові відкрили новий реабілітаційний центр для військових, який щодня прийматиме до 200 пацієнтів із різних регіонів України

Sentence A (**English translation**): A new rehabilitation center was opened in Kharkiv.

Sentence B (**English translation**): A new rehabilitation center for military personnel has opened in Kharkiv, which will treat up to 200 patients daily from various regions of Ukraine.

Number inconsistency:

Sentence A (**Ukrainian text**): Низка країн на чолі зі США підписали заяву щодо модернізації ППО України

Sentence B (**Ukrainian text**): США та 12 країн світу підписали заяву щодо нагальної модернізації ППО України

Sentence A (**English translation**): A number of countries, led by the United States, have signed a statement on the modernization of Ukraine's air defense system.

Sentence B (**English translation**): The United States and 12 other countries have signed a statement calling for the urgent modernization of Ukraine's air defense system.

A.2 Fine-Tuning Details and Hyperparameters

Both fine-tuned models (mT5-large and mT0-large) use adapters with bottleneck dimension 128 inserted in transformer attention blocks. We train adapters only, while keeping the base model parameters frozen.

The optimizer is AdamW with learning rate 5×10^{-6} and weight decay 0.01. Batch size is 4. We train up to 15 epochs because the training loss and validation metrics plateaued near this range. We evaluate every three epochs and use early stopping

with patience 3 evaluation rounds.

The learning-rate schedule is linear warmup/decay with warmup equal to 10% of total training steps. We apply gradient clipping with max norm 1.0. For generation during evaluation, we use beam search (num beams $= 5$), repetition penalty 1.5, no-repeat trigram constraint, and adaptive output length with approximately $1.3\times$ source-token budget.

Input lengths are truncated to 128 tokens. For metric computation we report $BLEU_{ref}$, $chrF^{++}_{ref}$, $BLEU_{in}$, $BERTScore_{in}$, and $iBScore$ as defined in Section 5.