Ukrainian NLP Contributions within the UniDive Project

Olha Kanishcheva[1,2], **Maria Shvedova**[3,4], **Hanna Sytar**[5], **Olha Yatsyshyna**[6], **Arsenii Lukashevskyi**[3]
[1]Heidelberg University, Germany, [2]SET University, Ukraine
[3]National Technical University "Kharkiv Polytechnic Institute", Ukraine
[4]Friedrich Schiller University Jena, Germany
[5]Institute of Slavonic Studies of the Czech Academy of Sciences, Czech
[6]Ternopil Volodymyr Hnatiuk National Pedagogical University, Ukraine

Relevant UniDive working groups: WG1, WG2, WG3

1 Introduction

This abstract presents the outcomes of four years of collaborative work within the UniDive COST Action CA21167, focusing on the development of linguistic and computational resources and tools for the Ukrainian language. The activities carried out within UniDive have contributed to advancing Ukrainian NLP across several interconnected areas, including syntactic annotation, multiword expression corpora, paraphrasing of multiword expressions, UA-XMPIE Multimodal dataset, ELEXIS-WSD, and bilingual code-switching datasets. The described resources were developed in close collaboration with international partners within the UniDive network and are publicly available, ensuring their accessibility for future research and integration into multilingual NLP pipelines.

2 Ukrainian UD Treebank

We developed a new Universal Dependencies (UD) treebank based on Ukrainian parliamentary transcripts (Shvedova et al., 2025), complementing the existing UD resources for Ukrainian.[1] The corpus includes manually annotated texts from key historical sessions of the Verkhovna Rada, capturing not only official rhetoric but also features of colloquial spoken language. The annotation pipeline combines the UDPipe2 (Milan Straka, 2016–) and a dictionary-based TagText[2] parsers with subsequent manual correction to ensure syntactic and morphological accuracy.

To demonstrate the applicability of the resource, we examined vocative and nominative case variation in direct address using a large-scale UD-annotated corpus of 88 million tokens of parliamentary transcripts from 1990 to 2024.[3] The corpus-based study was further extended to investigate the influence of Ukrainian-Russian bilingualism on case choice in direct addresses (Shvedova and Kanishcheva, 2025), as well as temporal trends in vocative usage across the full span of parliamentary transcripts (Shvedova and Lukashevskyi, 2025).

3 Corpus of Multiword Expressions

As part of the PARSEME 2.0 shared task,[4] we developed the first comprehensive annotated corpus of Ukrainian multiword expressions (MWEs)[5] (Sytar et al., 2026). The corpus comprises 198,555 tokens and 5,993 annotated MWEs across four major categories: verbal, nominal, adjectival/adverbial, and functional expressions.

The data consist of interview texts from the *Ukrainian Week* newspaper (2013-2016), selected from the General Regionally Annotated Corpus of Ukrainian (GRAC).[6] The corpus was created through a combination of automatic annotation using the UDPipe2 model and subsequent manual validation and annotation by linguists on the FLAT platform, following the PARSEME 2.0 guidelines. The resource is released in the standard `.cupt` format, ensuring compatibility with Universal Dependencies resources.[7]

The analysis shows a notable imbalance in the distribution of MWE types, with inherently adpositional verbs (IAV) emerging as the most frequent category, exceeding 1,200 instances. In addition to conventional idiomatic expressions, the corpus captures contemporary media neologisms (e.g., *"little*

[1] https://universaldependencies.org/treebanks/uk_parlamint/index.html
[2] https://github.com/brown-uk/nlp_uk
[3] https://huggingface.co/datasets/uacorpus/Rada_Trees
[4] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/
[5] https://lindat.mff.cuni.cz/repository/items/a89aaca0-9151-4f06-ba00-eedb4282ffe8
[6] https://uacorpus.org/
[7] https://gitlab.com/parseme/corpora

green men", *"deep concern"*) as well as colloquial variants that are often absent from idiomatic dictionaries.

4 Paraphrasing of Ukrainian Multiword Expressions

In addition to MWE corpus development (Sec.3), we explored the task of paraphrasing multiword expressions. A subset of approximately 150 sentences containing MWEs was selected and manually paraphrased. The resulting paraphrases were subsequently annotated and carefully validated by linguists to ensure semantic equivalence and linguistic correctness. This effort resulted in a curated dataset that reflects the variability of MWE usage and can support further research in paraphrase generation and evaluation.

5 Creation of the UA-XMPIE Multimodal Dataset

As part of the AdMIRe 2.0 project[8] (Arslan et al., 2026), we developed a Ukrainian multimodal dataset aimed at supporting the identification and interpretation of figurative language, with a particular focus on idiomatic expressions. These expressions present a well-known challenge for NLP due to their non-compositional nature, where the overall meaning cannot be directly inferred from individual components. To address the lack of multimodal resources for Ukrainian, we created the UA-XMPIE dataset, aligning idiomatic expressions with visual representations.

The dataset was constructed following a four-stage pipeline: (i) selection of English idioms and their alignment with Ukrainian equivalents, (ii) manual linguistic annotation, (iii) generation of image–text pairs using the Midjourney model, and (iv) collection of contextual usage examples. The final dataset comprises 63 validated English-Ukrainian idiomatic pairs and 315 images representing literal, figurative, and distractor interpretations. In addition, 274 contextual sentences were included, primarily extracted from the GRAC and complemented with automatically generated examples.

All data were manually reviewed and validated by expert linguists to ensure consistency and cultural adequacy.

[8] https://www.codabench.org/competitions/10547/

6 ELEXIS-WSD Parallel Sense-Annotated Corpus

Within UniDive WG2 focused on developing and enhancing the *ELEXIS-WSD Parallel Sense-Annotated Corpus* (*ELEXIS-WSD*),[9] a small parallel corpus with translations of the same sentences in multiple languages. Each subcorpus includes manual tokenization, lemmatization, morphosyntactic tagging, and sense annotations linked to lexicographic resources. *ELEXIS-WSD* supports word-sense disambiguation and cross-lingual comparison. *ELEXIS-WSD* is currently being extended with new languages, including Ukrainian, and additional annotation layers.

7 Ukrainian Code-Switching Corpus

We developed and released a Ukrainian-Russian code-switching dataset based on parliamentary transcripts of the Verkhovna Rada (1990–2024), comprising over 200,000 manually annotated tokens. [10] Each token is labeled as Ukrainian, Russian, hybrid (surzhyk), number, other, or punctuation. Annotation guidelines addressed challenges specific to closely related languages sharing the Cyrillic script, including orthographically identical words at switch points, morphologically adapted hybrids, syntactic calques, and borrowings versus dialectal variants (Kanishcheva et al., 2026).

A token-level language identification model was developed for this dataset,[11] with a BERT-based system achieving the highest performance ($F_1 = 0.98$), surpassing CRF, BiLSTM, and existing language identification tools. Analysis of the corpus using quantitative code-switching metrics revealed that Ukrainian-Russian mixing has a higher switching density than many standard benchmarks, and that switch points follow systematic part-of-speech patterns consistent with morphosyntactic constraint theories (Kanishcheva and Shvedova, 2026).

8 Conclusions and Future Plans

This work summarizes a four-year initiative within the UniDive COST Action, resulting in a comprehensive suite of open-source resources for

[9] https://live.european-language-grid.eu/catalogue/corpus/22947

[10] https://zenodo.org/records/14724542

[11] https://huggingface.co/OKanishcheva/ParlLangID-UA-RU

Ukrainian NLP. By addressing syntactic annotation, MWEs, multimodality, and code-switching, we have established a solid framework for the language's integration into modern multilingual pipelines. Future research will focus on completing the creation of Ukrainian-language subcorpora within the ELEXIS-WSD Parallel Sense-Annotated Corpus and the Paraphrasing of Multiword Expressions, and conducting experiments with them. We also plan to improve tools for morphosyntactic annotation of Ukrainian corpora, particularly non-standard texts (historical and dialectal), through the further expansion and annotation of Ukrainian UD treebanks. Further directions include scaling these resources using large language models, enhancing multimodal integration, and ensuring the long-term sustainability of the datasets within global research infrastructures to further support the Ukrainian computational linguistics community.

9 Dissemination

The results presented in this abstract were disseminated at international conferences and incorporated into teaching courses at the National Technical University "Kharkiv Polytechnic Institute" (Kharkiv) and SET University (Kyiv). The resources were also used in educational projects in digital humanities for Ukrainian students: (i) *Ukraine digital: Ensuring academic success in times of crisis* (2022–2025),[12] Friedrich Schiller University Jena; (ii) *Digital Humanities: Ensuring Ukrainian University Teaching in Times of Crisis* (TFK, Ukraine programme 2025–2026), organised by the University of Helsinki.[13]

Acknowledgements

We would like to thank the reviewers for their time and effort in reviewing this abstract. We sincerely appreciate their valuable comments and suggestions, which greatly helped us improve the quality of the work. This research was partially funded by the Alexander von Humboldt Foundation, and this work received support from the COST Action CA21167 'UniDive'[14] (European Cooperation in Science and Technology). The authors are also grateful to Friedrich Schiller University Jena for providing the research facilities and support that made this work possible.

[12] https://www.ukr.uni-jena.de/

[13] https://blogs.helsinki.fi/digihum-ua/

[14] https://unidive.lisn.upsaclay.fr/

References

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoglu Selamet, Thomas Pickard, Aline Villavicencio, Adriana Silvina Pagano, and Gülşen Eryiğit. 2026. MWE-2026 shared task: AdMIRe 2 advancing multimodal idiomaticity representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 276–287, Rabat, Marocco. Association for Computational Linguistics.

Olha Kanishcheva and Maria Shvedova. 2026. Quantifying code-switching in a Ukrainian parliamentary dataset 1990–2021. In *Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026*, pages 2–12, Palma de Mallorca, Spain. European Language Resources Association.

Olha Kanishcheva, Maria Shvedova, Liudmyla Dyka, and Kristina Husenko. 2026. Study of language identification task on the token level for Ukrainian-Russian code-switching dataset. *Northern European Journal of Language Technology*, 12(1):55–68.

Jan Hajič Milan Straka, Jana Straková. 2016–. Udpipe web service (lindat/clariah-cz): Trainable pipeline for tokenization, tagging, lemmatization and parsing.

Maria Shvedova and Olha Kanishcheva. 2025. Case variation in Ukrainian direct addresses in a bilingual context. In *The 11th International Contrastive Linguistics Conference*, Prague, Czech Republic. Charles University.

Maria Shvedova and Arsenii Lukashevskyi. 2025. Case choice in Ukrainian vocative expressions: A study of parliamentary transcripts (1990–2024) annotated with Universal Dependencies. In *Grammar and Corpora. 10th International Conference. Book of Abstracts*, pages 106–108, Riga, Latvia. University of Latvia Press.

Maria Shvedova, Arsenii Lukashevskyi, and Andriy Rysin. 2025. Developing a Universal Dependencies treebank for Ukrainian parliamentary speech. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 55–63, Vienna, Austria (online). 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, Morocco. Association for Computational Linguistics.