General Regionally Annotated Corpus of Ukrainian: Recent Developments and Future Plans

Maria Shvedova
National Technical University
"Kharkiv Polytechnic Institute"
Kyrpychova str. 2, 61002, Kharkiv, Ukraine
Friedrich Schiller University Jena
Fürstengraben 1 07743, Jena, Germany
mariia.shvedova@khpi.edu.ua

Abstract

The General Regionally Annotated Corpus of Ukrainian (GRAC) effectively serves as a national corpus. GRAC v.19 (2025) contains 2 billion tokens from over 800,000 texts (1816–2025). The corpus has multi-level annotations: rich metadata including regional tags, morphological annotation based on the VESUM dictionary, and partial semantic annotation. GRAC is the source of several derivative projects, including UD_Ukrainian_ParlaMint, ParaRook parallel corpora, Rada_Trees, and others.

Keywords: Ukrainian language; national corpus; regional annotation; morphological annotation; semantic annotation; corpus development; digital linguistic infrastructure

1. Introduction

The General Regionally Annotated Corpus of Ukrainian (GRAC) (Maria Shvedova (2017–)) is currently the largest manually curated and annotated corpus of Ukrainian.

When GRAC was initiated in 2016, Ukrainian corpus linguistics needed a larger, more comprehensively annotated resource to support varied research tasks and serve as national infrastructure. The latest version (GRAC.v.19, 2025) contains 2 billion tokens from over 800,000 texts by approximately 35,000 authors, covering modern Ukrainian from 1816 to 2025, including texts from both Ukraine and the diaspora. GRAC includes a wide range of text types: fiction (FIC), non-fiction (NOF), academic writing (ACA), journalism (JOU), official documents (OFF), ego-documents such as memoirs, diaries, and correspondence (EGO), transcribed public (SPU) and private speech (SPR), poetry (POE), and internet communication (ICM), drawn from printed, recorded, handwritten, and web sources (see Appendix A for proportions).

GRAC operates as an open volunteer project hosted at Jena University (Germany), with partial funding through university research grants (2019-2024), and contributed by students and specialists from multiple Ukrainian universities, including Kharkiv Polytechnic Institute, Ukrainian Catholic University, Kyiv-Mohyla Academy, and others. Since 2018, over 1,600 Ukrainian students from more than 15 universities have contributed to development through university courses and volunteer initiatives. Since 2024, GRAC has been receiving selected web texts from PAWUK, courtesy of IPI PAN (Kieraś et al., 2025).

Beyond serving as a reference corpus, GRAC functions as foundational infrastructure for specialized resources. Recent derivative projects include: a Universal Dependencies treebank *UD_Ukrainian_ParlaMint* (Shvedova and Lukashevskyi (2024-2025)) (Shvedova et al., 2025); *ParaRook* parallel corpora (Shvedova and Lukashevskyi (2023–)); *PluG* (copyright-free Ukrainian texts from GRAC, for download) (Shvedova and Lukashevskyi (2024)); *Rada_Trees* (parliamentary transcripts, 1990-2024, annotated with UDPipe2 and TagText) (Arsenii Lukashevskyi (2025)); *ParlMix-UA-RU* (parliamentary code-mixing dataset) (Olha Kanishcheva (2025)) (Kanishcheva et al., 2026); *ParaFarm* (English-Ukrainian multiple-translation corpus of George Orwell's *Animal Farm*) (Maslij (Kalashnyk) and Shvedova (2025)); and *PressMint-UA* (Ukrainian component within comparable corpora of historical newspapers, work in progress) (CLARIN ERIC, 2025).

This paper presents the annotation architecture underlying GRAC, discusses the methodological challenges encountered in its development, and outlines planned enhancements.

2. Pipeline and Technical Infrastructure

The general pipeline for updating GRAC comprises the following stages: text collection and metatextual annotation, text preprocessing, lemmatization, morphological tagging and semantic annotation, and corpus compilation.

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 66–70. Palma, Mallorca. May 11, 2026.

2.1. Text collection and metatextual annotation

Texts are collected through multiple resources: digitized printed sources (OCR with manual correction), transcribed audio recordings, and downloads from online sources including news portals (such as hromadske.ua, zaxid.net, procherk.info) and digital libraries (ukrlib.com.ua for fiction, libraria.ua for historical press, scc.knu.ua for doctoral theses, and others).

Metadata is entered manually into spreadsheets and subsequently transferred to a dedicated metadata database (developed by Sergey Yarygin), which enforces validation by checking that all values belong to predefined sets of allowed values and verifying the consistency between metadata records and text files. Text files are stored separately and are linked to their metadata records by filename.

All texts receive metatextual annotation including author information, dates of creation and publication, stylistic register, genre classification, information about the source medium and language of the original text.[1] A distinctive feature is GRAC's regional annotation, which tracks geographical variation through publication location and the author's region of origin. This is particularly important given Ukrainian's complex dialectal landscape and different historical varieties (Shvedova and von Waldenfels, 2021).

2.2. Text Preprocessing

Text preprocessing includes cleaning with the CleanText program[2] (Starko et al., 2021), which addresses a range of issues common in texts obtained via OCR or downloaded from the internet. These include erroneous apostrophe characters (replaced with the standard U+0027), Latin characters mixed into Cyrillic text (e.g., the Latin *i* substituting the Ukrainian *і*), digits used in place of visually similar letters, soft hyphens within words, and dangling or end-of-line hyphens that split words across lines.

Non-Ukrainian text (predominantly Russian) is excised and replaced with three hyphens (---). For parliamentary transcripts, this process has been automated using CleanText's language detection algorithm, which compares word counts matched against Ukrainian and Russian dictionaries. For other texts, the removal of non-Ukrainian fragments has in some cases been performed manually.

[1] https://uacorpus.org/en/rozmitka-tekstiv

[2] https://github.com/brown-uk/nlp_uk

2.3. Lemmatization, morphological tagging, and semantic annotation

Morphological annotation is performed automatically using the TagText program[2] based on the VESUM open-access dictionary (Starko and Rysin, 2022). Each token receives a lemma and a composite tag consisting of morphological features (part of speech, case, number, gender, tense, person, etc.) separated by colons. The format is *word/lemma/tag*, for example *korpusiv/korpus/noun:inanim:p:v_rod* (*korpusiv* 'corpus.GEN.PL', lemma *korpus*, noun, inanimate, plural, genitive). The system handles detailed morphological annotation using an extensive tagset and includes specialized tags for colloquialisms, archaisms, vulgarisms, and orthographic variants. Evaluation of the TagText tagger shows high precision: 99.3% for lemmas, 98.7% for pos, and 94.5% for full morphological tags. Specialized rule-based tools process non-standard orthography in historical texts, including the normalization of the Western Ukrainian Zhelekhivka orthographic system (Shvedova et al., 2021, 2022; Chemerys et al., 2023).

Morphological ambiguity presents an ongoing challenge. Most GRAC versions retain all possible analyses for ambiguous forms, though GRAC.v.17a and GRAC.v.19a implement automatic disambiguation. Following Rysin's description of the system (Shvedova et al., 2025), the disambiguation in TagText operates on three levels. First, rarely used word forms are discarded: for example, *rozpalenij* ('FIRE-PST.PASS.PTCP-ADJ-F.LOC.SG') could theoretically be parsed as an imperative verb form (*rozpalenij* 'INFLAME-IMP.2SG'), but is almost always an adjective and is tagged as such. Second, rule-based disambiguation handles both specific and general cases: for instance, only the locative case is retained in phrases like *v Ukrajini* ('in Ukraine.LOC'), and vocative forms are discarded after prepositions. Third, a statistical module draws on data from the manually disambiguated BrUK corpus (Starko and Rysin, 2023), using word form frequencies and morphological tags with contextual information (preceding and following token) to select the most probable analysis.[3] This hybrid approach shows promise but has not yet achieved complete reliability.

The corpus includes partial semantic annotation (Starko, 2021), which is stored in VESUM alongside morphological tags, and is thus assigned during the same tagging process. The annotation currently covers approximately 3,000 most frequent lemmas, as well as lemmas belonging to selected semantic groups, and continues to expand. A faceted approach is employed, allowing flexible tag com-

[3] https://github.com/brown-uk/nlp_uk/blob/master/doc/disambig.md

binations: lemmas receive one or more semantic features drawn from six major categories (concrete nouns, abstract nouns, proper nouns, adjectives, adverbs, and verbs), with separate tagsets developed for each category. For instance, the noun *ultras* receives the tags `conc:hum:group` (concrete noun, human, group). Some tags of a semantic nature — such as `prop` (proper noun) and its subtypes (e.g., `prop:fname` for first names, `prop:geo` for geographical names) — are incorporated into the morphological tag and assigned via the *tag* attribute, whereas the remaining semantic annotation is stored separately and searchable via the *semtag* attribute.

GRAC uses vertical files optimised for NoSketch Engine, where tokens carry positional attributes and structural metadata is encoded as attributes of the `<doc>` element. The compilation process consists of two stages: first, the TagText XML format is enriched with metadata and converted into a vertical file using XSLT; then, the resulting file is compiled into NoSketch Engine's indexed format using its CLI toolkit.

Users search via NoSketch Engine (Rychlý, 2007; Kilgarriff et al., 2014) using word forms, lemmas, tags, semtags, and complex CQL queries, with the ability to create random samples and obtain statistical information. Rich metadata enables fine-grained, multidimensional search queries.

The corpus is updated at least once a year. Before each update, metadata are validated and texts undergo preprocessing. Version-specific changes are documented on the project website.[4]

3. Challenges and Current Limitations

The corpus faces structural imbalances. Addressing representational gaps remains a long-term priority. Rather than aiming for a strictly balanced corpus in terms of historical versus contemporary coverage (which is unachievable given the uneven availability of sources) our goal is to ensure adequate representation across text types and regions. Contemporary online texts will likely continue to grow faster than historical collections, but targeted efforts to expand underrepresented text types will continue.

While morphological analysis handles standard contemporary Ukrainian effectively, several issues persist. Disambiguation accuracy requires continued improvement. Processing texts with non-standard orthography, particularly pre-standardization materials and regional variants, remains difficult despite specialized tools.

Mass-generated text poses a challenge for future corpus expansion. A related issue has already arisen with machine-translated content: many online media before 2014 published parallel Ukrainian and Russian versions, where the Ukrainian text may have been automatically translated, with or without post-editing. Since the origin of such texts could not be reliably determined, online media publishing two language versions were not included in the corpus. We are aware of the analogous risk posed by AI-generated content and are considering ways to address it, though no fully reliable detection method is currently available.

4. Future Plans

For GRAC v.20, we plan to implement syntactic dependency annotation using UDPipe2 (Milan Straka (2016–)). This will add a crucial layer to our annotation architecture, enabling more precise searches for grammatical phenomena and syntactic constructions. UD annotation should also improve disambiguation accuracy.

We continue working on corpus expansion with both historical and contemporary texts. Addressing the structural imbalances described above remains a long-term priority.

Despite ongoing challenges, GRAC has established itself as essential infrastructure for Ukrainian linguistics. With multi-level annotation, and growing derivative projects, it continues to serve research, education, and language technology.

5. Acknowledgements

We thank the reviewers for their valuable comments, which helped improve this paper. We are grateful to the GRAC team — Sergey Yarygin, Ruprecht von Waldenfels, Andriy Rysin, Vasyl Starko, Arsenii Lukashevskyi, and all others who have contributed to its development. We also acknowledge the support of the University of Jena and IPI PAN.

6. Bibliographical References

Yurii Chemerys, Olesia Nakhlik, Andriy Rysin, and Maria Shvedova. 2023. Normalization of a historic Western Ukrainian orthographic system Zhelekhivka in the Ukrainian language reference corpus (GRAC). In *Proceedings of the IEEE 18th International Conference on Computer Sciences and Information Technologies (CSIT)*, Lviv, Ukraine.

[4] https://uacorpus.org/en/informaciya-pro-grak/versiyi-korpusu

CLARIN ERIC. 2025. PressMint: Interoperable corpora of historical newspapers. Accessed: 2026-02-28.

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).

Witold Kieraś, Łukasz Kobyliński, Dorota Komosińska, Michał Rudolf, Maria Shvedova, and Anna Zwierzchowska. 2025. PAWUK: Extensive annotated web corpus of Ukrainian. In *Computational Science – ICCS 2025*, volume 15904 of *Lecture Notes in Computer Science*, Cham. Springer.

Adam Kilgarriff et al. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Pavel Rychlý. 2007. Manatee/Bonito-a modular corpus manager. In *RASLAN*, pages 65–70.

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. ACL.

Maria Shvedova, Nataliia Prydvorova, and Ilona Skibina. 2022. Normalization of early modern Ukrainian in GRAC: the case of Lesia Ukrainka's works. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022)*, Gliwice, Poland.

Maria Shvedova, Andriy Rysin, and Vasyl Starko. 2021. Handling of nonstandard spelling in GRAC. In *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, pages 105–108, Lviv, Ukraine.

Maria Shvedova and Ruprecht von Waldenfels. 2021. Regional annotation within GRAC, a large reference corpus of Ukrainian: Issues and challenges. In *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021)*, pages 32–45, Kharkiv, Ukraine.

Vasyl Starko. 2021. Implementing semantic annotation in a Ukrainian corpus. In *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021)*, pages 435–447, Kharkiv, Ukraine.

Vasyl Starko and Andriy Rysin. 2022. VESUM: A large morphological dictionary of Ukrainian as a dynamic tool. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022)*, pages 71–80, Gliwice, Poland.

Vasyl Starko and Andriy Rysin. 2023. Creating a POS gold standard corpus of modern Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 91–95, Dubrovnik, Croatia. Association for Computational Linguistics. DOI: 10.18653/v1/2023.unlp-1.11.

Vasyl Starko, Andriy Rysin, and Maria Shvedova. 2021. Ukrainian text preprocessing in GRAC. In *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, volume 2, pages 101–104, Lviv, Ukraine.

7. Language Resource References

Arsenii Lukashevskyi, Kyrylo Zakharov, Maria Shvedova. 2025. *Rada_Trees: A Syntactically Annotated Corpus of Ukrainian Parliament Transcripts (1990–2024)*. Hugging Face Datasets.

Maria Shvedova, Ruprecht von Waldenfels, Sergey Yarygin, Andriy Rysin, Vasyl Starko, Tymofij Nikolajenko, Arsenii Lukashevskyi et al. 2017–. *GRAC: General Regionally Annotated Corpus of Ukrainian*.

Viktoriia Maslij (Kalashnyk) and Maria Shvedova. 2025. *ParaFarm: English-Ukrainian Multiple-Translation Corpus (1.1)*. Zenodo.

Milan Straka, Jana Straková, Jan Hajič. 2016–. *UDPipe Web Service (LINDAT/CLARIAH-CZ): Trainable Pipeline for Tokenization, Tagging, Lemmatization and Parsing*. LINDAT/CLARIAH-CZ, Institute of Formal and Applied Linguistics, Charles University.

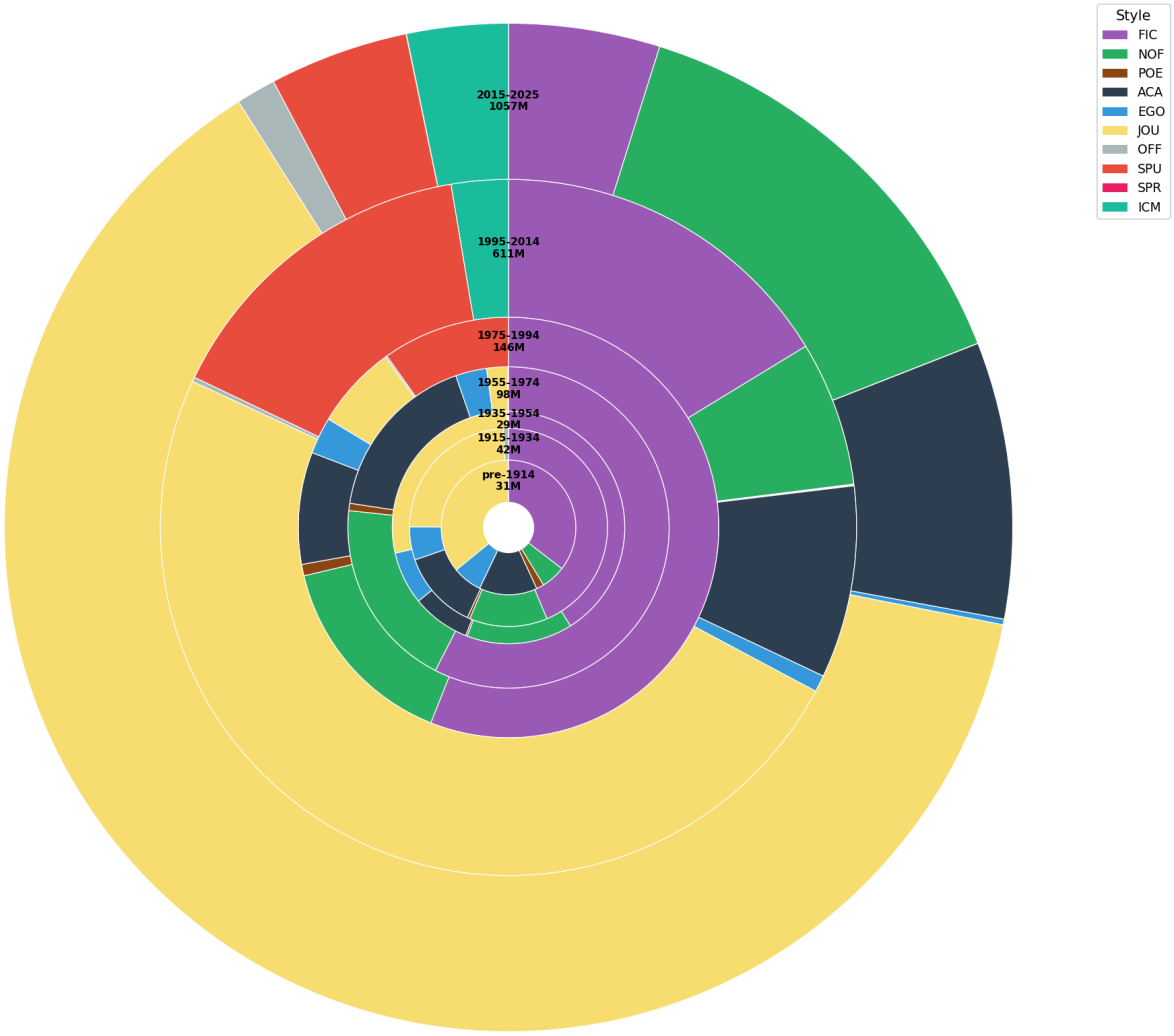
Olha Kanishcheva, Maria Shvedova, Liudmyla Dyka, Kristina Husenko. 2025. *ParlMix-UA-RU: Ukrainian Parliamentary Code-Mixing Dataset*. Zenodo.

Maria Shvedova and Arsenii Lukashevskyi. 2023–. *ParaRook: Parallel Corpora Based on GRAC*.

Maria Shvedova and Arsenii Lukashevskyi. 2024. *PluG: Corpus of Old Ukrainian Texts Based on GRAC*.

Maria Shvedova and Arsenii Lukashevskyi. 2024-2025. *UD_Ukrainian_ParlaMint*.

A. Distribution of functional styles across periods in GRAC.v.19 (ring area proportional to token count)



Style	pre-1914	1915-1934	1935-1954	1955-1974	1975-1994	1995-2014	2015-2025	**Grand Total**
FIC	11.0M (35%)	18.3M (44%)	12.1M (41%)	56.2M (58%)	81.9M (56%)	99.6M (16%)	51.2M (5%)	**330.2M (16%)**
NOF	1.8M (6%)	5.3M (13%)	4.3M (15%)	18.7M (19%)	22.3M (15%)	41.1M (7%)	150.0M (14%)	**243.6M (12%)**
POE	559.8K (2%)	185.5K (0%)	71.6K (0%)	723.8K (1%)	1.3M (1%)	506.8K (0%)	102.0K (0%)	**3.4M (0%)**
ACA	4.3M (14%)	5.4M (13%)	2.4M (8%)	16.9M (17%)	12.6M (9%)	54.6M (9%)	93.6M (9%)	**189.7M (9%)**
EGO	2.2M (7%)	2.3M (5%)	2.2M (7%)	3.1M (3%)	4.1M (3%)	4.7M (1%)	1.9M (0%)	**20.5M (1%)**
JOU	11.1M (36%)	10.1M (24%)	8.4M (28%)	2.1M (2%)	9.4M (6%)	300.1M (49%)	664.7M (63%)	**1005.8M (50%)**
OFF	15.5K (0%)	322.6K (1%)	26.6K (0%)	39.8K (0%)	200.1K (0%)	1.3M (0%)	13.5M (1%)	**15.3M (1%)**
SPU	16.8K (0%)	13.5K (0%)	41.4K (0%)	44.7K (0%)	14.4M (10%)	93.3M (15%)	47.1M (4%)	**154.8M (8%)**
SPR	0.0K (0%)	0.0K (0%)	0.0K (0%)	0.0K (0%)	0.0K (0%)	0.0K (0%)	122.7K (0%)	**122.7K (0%)**
ICM	0.0K (0%)	0.0K (0%)	0.0K (0%)	0.0K (0%)	0.0K (0%)	16.2M (3%)	34.3M (3%)	**50.5M (3%)**
Total	**31M**	**42M**	**29M**	**98M**	**146M**	**611M**	**1057M**	**2014M (100%)**