Historical Newspapers in the General Regionally Annotated Corpus of Ukrainian (GRAC): Current State and PressMint Integration Prospects

Maria Shvedova[1,2], **Arsenii Lukashevskyi**[1]
[1] National Technical University "Kharkiv Polytechnic Institute"
Kyrpychova str. 2, 61002, Kharkiv, Ukraine
[2] Friedrich Schiller University Jena
Fürstengraben 1, 07743 Jena, Germany
mariia.shvedova@khpi.edu.ua
arsenii.lukashevskyi@sgt.khpi.edu.ua

Abstract

This paper presents the historical newspaper collection of the General Regionally Annotated Corpus of Ukrainian (GRAC) and outlines its prospective integration into the PressMint infrastructure. The collection comprises 117 newspaper titles published before 1950, totaling 23.6 million tokens, and reflects the political fragmentation, regional variation, and orthographic diversity of Ukrainian-language press from the late nineteenth to mid-twentieth century. We describe the corpus composition, temporal and geographic distribution, and metadata architecture. Special attention is given to morphosyntactic annotation challenges arising from the historical Western Ukrainian orthography (Zhelekhivka), as well as issues related to annotating historical texts using the rule-based TagText parser and neural UDPipe2 models. The paper compares GRAC's vertical format and metadata system with the TEI-based PressMint standard, identifying technical and conceptual harmonization challenges. Integrating GRAC newspapers into PressMint will facilitate comparative research on language policy, regional standardization, and media discourse within a broader European context.

Keywords: newspaper corpus, Ukrainian language, historical newspapers, corpus annotation, metadata standards, GRAC, PressMint

1. Introduction

The General Regionally Annotated Corpus of Ukrainian (GRAC) (Maria Shvedova (2017–)) is a large representative corpus of standard Ukrainian covering the 19th-21st centuries and effectively serving as the national corpus of the Ukrainian language. The newspaper collection is an important part of GRAC, enabling researchers to track and accurately date language change and study the influence of language policy and ideology. At the same time, newspaper texts are challenging to process, not only because of access and digitization issues, but also (in the case of Ukrainian) because of different orthographies and language norms at different times and in different territories. Tools designed for the modern Ukrainian language are of limited use for parsing historical texts.

Integration into PressMint: Interoperable Corpora of Historical Newspapers (CLARIN ERIC, 2025) will enable the study of Ukrainian newspapers in comparison with press materials from neighboring linguistic and political contexts, opening up new opportunities for research into language contact and shared European historical discourse.

This paper is structured as follows: Section 2 provides the historical and linguistic context of Ukrainian press from the late nineteenth to the twentieth century; Section 3 describes the GRAC newspaper collection's composition and sources; Section 4 presents the metadata architecture that captures regional and orthographic variation; Section 5 addresses morphosyntactic annotation challenges; Section 6 demonstrates research applications of the collection; Section 7 analyzes the alignment between GRAC and PressMint metadata standards; and Section 8 offers concluding remarks.

2. Historical and Linguistic Context

Before World War I, Ukrainian-speaking territories were politically divided between the Russian and Austro-Hungarian empires, and thereafter between the Ukrainian Soviet Socialist Republic (Ukrainian SSR), Poland, Czechoslovakia, and Romania[1]. By 1945, following the post-war territorial settlements, almost all these territories had been incorporated into the Ukrainian SSR. Policies regarding the Ukrainian language varied from country to country. Ukrainians in Austria had a fairly developed press since the end of the 19th century (in 1900 there were 25 periodicals published in Galicia and six in Bukovina (Shevelov, 2008)), while in the Russian Empire, Ukrainian-language publishing was severely restricted and a Ukrainian-language press did not exist until 1905.

[1] Historical map of Ukraine from (Magocsi, 1987).

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 21–26

Until the end of World War II, the Ukrainian language was shaped by political partition. The language of editions published in the Russian-controlled and later Soviet part of Ukraine (with its main cultural center in Kyiv, and from 1919 to 1934 in Kharkiv) differed substantially from the language of Western Ukraine, which was mostly culturally oriented toward Lviv. Researchers describe distinct regional variants of literary Ukrainian for this period, which developed under the influence of local dialects and different dominant languages (Hrytsenko, 1993; Franko, 1995; Matvijas, 1998), with differences in lexical and grammatical norms and different orthographic standards until the 1920s.

The Ukrainian language coexisted in each territory with other languages that enjoyed greater social prestige: in the large cities of Central and Eastern Ukraine it was Russian, in Galicia it was Polish, in Bukovina it was German and/or Romanian, and in Transcarpathia it was Hungarian (Shevelov, 2008). Therefore, in early Ukrainian newspapers, we observe a significant influence of dominant languages both in the Ukrainian language itself, saturated with borrowings at the level of vocabulary and syntax (Shvedova and von Waldenfels, 2021), and in the form of code-switching (since the newspaper audience was predominantly bilingual). Some early 20th-century Ukrainian newspapers published entire texts or columns in other languages. For example, the Ukrainian newspaper *Rada* (Kyiv, 1906-1919) contained advertisements in Russian. In the Ukrainian SSR before World War II, some newspapers published articles in different languages within a single issue (Ukrainian-Polish *Radianska Volyn* 'Soviet Volyn' (1924, Zhytomyr), Ukrainian-Yiddish-Russian *Chervona Shvachka* 'Red seamstress' (1932, Kyiv)).

This linguistic heterogeneity directly shapes the metadata and annotation challenges addressed in the following sections.

3. The GRAC Newspaper Collection: Composition and Sources

The GRAC newspaper collection contains only texts in Ukrainian (although there may be some cases of code-switching that are not currently specifically tagged). The collection consists of digitized newspaper texts: OCR output subsequently verified by human annotators. This paper focuses on the historical component: 117 newspaper titles published before 1950, totaling 23.6 million tokens.

The distribution of old newspaper texts across macroregions reveals significant temporal and geographic imbalances in corpus coverage (Figure 1). Western Ukrainian newspapers (macroregion W) constitute the earliest materials in the collection, with substantial coverage beginning in the 1880s and continuing through the interwar period. In contrast, newspapers from the Kyiv region (macroregion KYV) enter the corpus only after 1905, due to censorship in Russian Empire. The bulk of KYV materials dates from 1910 onward, when orthographic practices became more standardized and closer to modern conventions, facilitating corpus processing and linguistic annotation. The 1920s exhibit more diverse geographic representation, coinciding with the relatively liberal Ukrainization policy in Soviet Ukraine, which encouraged Ukrainian-language publishing, with materials from Central (C), Eastern (E), Northern (N), and Southern (S) regions appearing alongside continued coverage of Western and Kyiv publications. The World War II period (1941–1945) is particularly well represented, with substantial materials from both German occupation newspapers and underground press, as well as Soviet publications. The immediate post-war years (1945–1946) maintain strong coverage, particularly in Western Ukraine.

Quantitatively, Western Ukraine dominates our pre-1910 materials, with peaks exceeding 700,000 tokens in the early 1890s and consistent coverage through the interwar decades. Our collection for 1924-1925 shows exceptional volume across multiple regions, with Northern Ukraine contributing over 1.3 million tokens and Eastern Ukraine nearly 450,000 tokens. The WWII years represent our best-documented period, with approximately 4.4 million tokens assembled across all regions.

Newspaper texts in GRAC have been collected from multiple sources. The core Western Ukrainian collection derives from the historical press archive curated by Orest Drul on the Zbruch portal (zbruc.eu), featuring Galician newspapers from the late 19th–mid-20th centuries (Drul, 2014–). Most texts from the 1910s–1940s were prepared by university students during research practicum projects, working from scans provided by LIBRARIA, a digital archive of Ukrainian periodicals (Arkhivni Informatsijni Systemy, 2017–), or downloaded from the Archive of Old Newspapers (Old, 2010–2016). The WWII collection was systematically compiled by Anna Bordovska, encompassing newspapers from German occupation authorities, Soviet publications, and underground OUN-UPA press (Bordovska, 2024). The shape of the collection reflects not only historical publishing activity but also, to a large extent, the priorities of the digitisation projects from which these materials were sourced, and the specific research interests of the corpus compilers.

These diverse sources and the resulting uneven temporal-geographic distribution must be considered when designing comparative studies or assessing the representativeness of linguistic patterns across regions and periods.

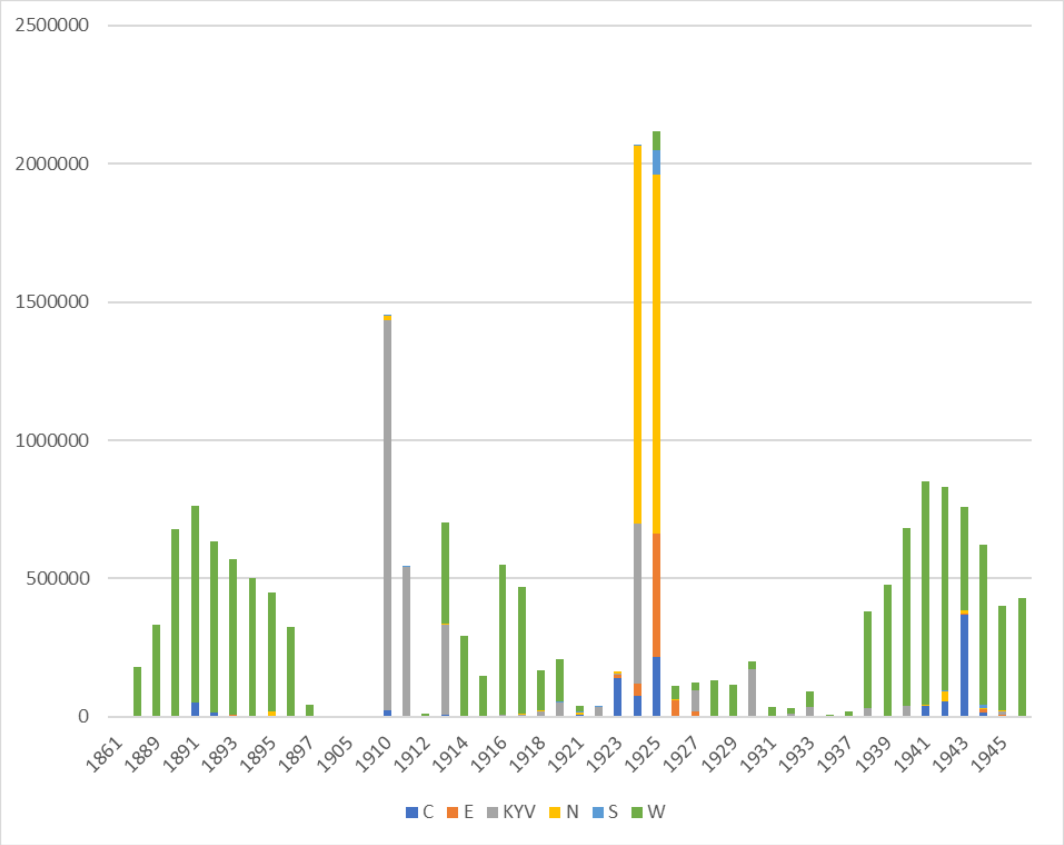

Figure 1: Tokens of old newspaper texts by year and macroregion: West, East, Center, South, North, Kyiv.

4. Metadata Schema in GRAC

The GRAC metadata schema for periodicals, described in detail in (Shvedova, 2020) and on the GRAC site, captures the complex political and administrative landscape of Ukrainian-language press across multiple states and regimes, reflecting the sociolinguistic heterogeneity described in Section 2. Periodicals are classified by type through the `doc.mediaType` attribute (*newspapers* and *magazines* for pre-1950 materials, with additional categories for later periods), with newspapers representing the most diverse category.

All periodicals in GRAC are annotated for the political entity that controlled the publication. This classification is encoded in the `doc.mediaAdmin` attribute and reflects the political fragmentation of Ukrainian territories throughout the corpus timespan, see Table 1.

Regional attribution follows different logic depending on whether individual authorship is available at the article level: when authors are known and annotated, regional tags reflect the author's origin; when author information is unavailable (as is common in newspaper materials) regional attribution is assigned based on the place of publication (Shvedova and von Waldenfels, 2021).

Text segmentation and the associated metadata vary by time period: pre-WWII newspapers were added article-by-article, resulting in text-level metadata (author, title, and regional attribution based on author's origin, where known). Newspapers from the WWII period onward were predominantly digitised as complete issues (with the exception of Western Ukrainian publications from 1945–1946, which retain article-level metadata), and therefore lack text-specific metadata fields.

The resulting metadata schema provides the foundation for regional and diachronic comparative

Code	Administrative Entity
AHI	Austro-Hungarian Empire
RUI	Russian Empire
POL	Interwar Poland
CZE	Interwar Czechoslovakia
ROM	Interwar Romania
ZUNR	West Ukrainian People's Republic
MAX	Makhno Movement
SOV	Ukrainian SSR
RUK	Reichskommissariat Ukraine
OUN	Organization of Ukrainian Nationalists
UKR	Contemporary Ukraine
DIA	Diaspora

Table 1: Administrative and political classification codes for periodicals in GRAC.

analyses, though its internal heterogeneity must be accounted for when formulating research queries or comparing findings across regions and time periods.

5. Morphosyntactic Annotation Challenges

Morphological annotation in GRAC is performed automatically using the TagText program, a dictionary-based tagger for modern Ukrainian based on the VESUM open-access dictionary (Starko and Rysin, 2022). The primary challenge for automatic morphosyntactic analysis arises from orthographic variation in Western Ukrainian materials. Newspapers from Austrian Galicia and interwar Poland (approximately half of the historical newspaper subcorpus) represent the Western Ukrainian regional variant of the literary standard, characterized by distinctive lexicon, some specific grammatical forms, and orthographic system that differ from both the modern standard and Eastern Ukrainian practices, leading to reduced morphosyntactic annotation accuracy when processed with tools developed for contemporary Ukrainian. Western Ukrainian historical newspapers in GRAC are represented in the Western Ukrainian orthographic standard of the late 19th–early 20th century (Zhelekhivka)—partly in its original form and partly in reconstructed form (the oldest Zbruč collection materials (Drul, 2014–), originally published in the etymological orthography, or Maksymovychivka, have been converted to Zhelekhivka). To handle Zhelekhivka texts, GRAC employs an additional rule-based module that adapts the standard morphological analyzer. However, this approach does not yield optimal results, partly due to inherent variability within Zhelekhivka itself (Chemerys et al., 2023).

UDPipe2 (Milan Straka and Hajič (2016–)), a neural network-based multilingual morphosyntactic parser, is planned for use in PressMint corpus annotation. Both Ukrainian models within UDPipe2 were trained on modern Ukrainian orthography data. As a neural network-based system, UDPipe2 handles out-of-vocabulary items and orthographic variation more robustly than dictionary-based approaches, mostly successfully processing the distinctive lexicon and spelling conventions of historical texts. However, this approach differs from TagText in lemmatization strategy: TagText's rule-based normalization module maps historical orthographic variants to standardized lemma forms, while UDPipe2 generates lemmas that preserve the input orthography. This results in distinct lemma representations for orthographic variants of the same lexeme—for instance, modern *svit* 'world' and historical Western Ukrainian *svït* 'world' would be lemmatized as separate entries. While this preserves orthographic information, it fragments lexeme frequencies across spelling variants and complicates cross-period lexical queries without post-processing normalization.

A more significant challenge for UDPipe2 arises from differences in word segmentation conventions between historical and modern orthographies. Zhelekhivka wrote clitics separately from their host words, as in *byty mut' sja* (modern *bytymut'sja* 'will fight'), where the reflexive marker *sja* and future tense marker *mut'* appear as independent tokens. Word boundaries for adverbs and prepositions also differ systematically: Zhelekhivka wrote *do domu* 'homeward' and *v oseny* 'in the fall' as two words (modern *dodomu*, *voseny*), while *vkinci* 'at the end' was written as one word (modern *v kinci*). The preposition *popry* 'despite' exhibits additional variability, appearing in Zhelekhivka as *popry*, *po-pry*, or *po pry*. Particles *že*, *ž*, *by*, *b* were attached directly to the preceding word in Zhelekhivka, preventing accurate recognition of both the host word and the particle.

Empirical evaluation confirms these challenges: manual verification of UDPipe2 part-of-speech tags in a 572-token sample from *Bil'shovyk Poltavshchyny* (Poltava, 1924) versus a 504-token sample from *Dilo* (Lviv, 1889, Zhelekhivka) revealed 96.7% accuracy for Soviet Ukrainian text but only 92.7% for the Lviv text, with errors concentrated in non-standard spelling and segmentation.

These substantial linguistic differences suggest that historical Western Ukrainian newspapers should be treated as a distinct subcorpus within the PressMint framework. For optimal annotation quality, we plan to develop a dedicated UDPipe2 model trained specifically on Zhelekhivka-orthography texts, capable of capturing the linguistic and orthographic features of this historical regional standard. In the longer term, further harmonization in the treatment of orthographic and segmentation variation across varieties, including normalization of lemma forms, would be particularly beneficial for enabling consistent use of both Ukrainian corpora in linguistic research.

6. Use Cases and Research Applications

The GRAC newspaper collection has enabled diverse corpus-based studies investigating linguistic variation in Ukrainian press during critical periods of political transformation.

Shvedova (2021) compared three newspaper subcorpora (Soviet 1919–1933, Western Ukrainian 1937–1943, and Western Ukrainian Soviet 1939–1946) to trace the formation of a new journalistic lexical norm in the 1940s. Analysis of 117 synonymous sets demonstrated that the new norm had

an Eastern Ukrainian basis, with minimal Western Ukrainian influence. The integration into PressMint would allow extending this analysis beyond the Ukrainian component of the multilingual Galician discourse of the interwar period: systematic comparison with contemporary Polish-language press would enable distinguishing lexical items specific to Ukrainian journalistic norm from those shared with and reinforced by Polish.

Bordovska (2024) examined World War II newspapers of different political orientations (German occupation, Soviet, and underground OUN-UPA press), revealing systematic orthographic and lexical differences correlated with political ideology. German and underground newspapers gravitated toward 1928 Orthography norms, while Soviet publications consistently applied 1933 norms. At the lexical level, Soviet periodicals systematically preferred international vocabulary, while underground press favored Ukrainian equivalents.

Hleba (2025) investigated regional variation in spatial prepositions *pry*, *bilja* and *kolo* 'near' across Lviv and Kyiv newspapers and fiction texts, using profile-based and collocational analysis. Fiction texts showed 75% uniformity (indicating distinct regional differences), with Lviv preferring *pry* (Polish influence) and Kyiv preferring *bilja* and *kolo*. Newspapers, however, showed 98% uniformity, suggesting genre-specific standardization effects.

These studies relied on GRAC's detailed metadata (including `doc.mediaAdmin` and regional annotation) and full-text search, with linguistic patterns identified through manual analysis.

However, some documented patterns of regional variation reflect not only internal Ukrainian factors but also substantial influence from neighboring languages. Integration into PressMint would not only enable consistent morphosyntactic analysis across regional varieties through improved annotation tools, but also allow for systematic comparison of GRAC newspapers with comparable corpora from neighboring linguistic and political contexts from the same time period, revealing shared public terminology across multilingual discourses and patterns of cross-linguistic influence at multiple linguistic levels.

7. GRAC Metadata vs. PressMint Standards

Harmonization of GRAC with PressMint involves both technical and sociolinguistic challenges. Structurally, the two formats differ substantially. GRAC uses vertical files optimised for NoSketch Engine, where tokens carry positional attributes (word form, lemma, morphological tag, semantic annotation), and metadata is encoded as attributes of the <doc> element. PressMint, in contrast, follows the Text Encoding Initiative (TEI) XML standard, widely used for encoding historical documents in digital humanities, completely separating metadata from the text and supporting a richer set of structural elements, including article types, column divisions, and links to facsimiles. Most positional attributes and basic structural elements (`<doc>`, `<s>`) can be converted to TEI-XML without conceptual loss.

More significant difficulties arise from the historical and sociolinguistic complexity of Ukrainian press. The GRAC attribute `doc.mediaAdmin`, designed to reflect political affiliations of Ukrainian-language periodicals (Table 1), has no direct equivalent in PressMint and will require the development of a language-specific taxonomy. Regional metadata also requires normalization. All metadata must be translated into English for interoperability. Publication dates, currently often recorded at the year level, should be refined to the day of issue where possible.

Named entity information in GRAC is encoded at the token level as semantic features within morphological tags, with separate labels for personal name components (given names, patronymics, surnames), geographical names, and other proper nouns. PressMint requires multi-token named entities with explicit boundaries (e.g., *Taras Hryhorovych Shevchenko* as a single PERSON entity rather than separate tokens). Converting to PressMint's four-class system (PERSON, LOCATION, ORGANIZATION, MISC) will require post-processing to identify entity boundaries and merge multi-token entities.

At the morphosyntactic annotation level, orthographic heterogeneity must be explicitly addressed. We plan to treat Zhelekhivka materials as a dedicated subcorpus with a specialized UDPipe2 model trained on historical Western Ukrainian texts, which will improve annotation accuracy while preserving orthographic information for diachronic research.

8. Conclusion

The GRAC historical newspaper collection reflects the political fragmentation, regional diversity, and orthographic heterogeneity of Ukrainian press from the late nineteenth to the mid-twentieth century. While structural conversion to PressMint is technically feasible, successful integration depends primarily on careful metadata normalization, taxonomy development, and adaptation of annotation layers to account for historical variation.

Addressing these challenges will enable Ukrainian historical newspapers to function as a fully interoperable component of the PressMint infrastructure, supporting comparative research on media discourse, language policy, and regional standardization processes across Europe.

9. Acknowledgements

The authors are grateful to Orest Drul for his assistance in integrating his large Western Ukrainian newspaper collection into GRAC. The authors also thank the anonymous reviewers for their valuable comments, which helped to significantly improve the paper. The authors benefited from discussions within the Universal Dependencies community and participation in COST Action CA21167 "UniDive". The preparation of the interwar Soviet and WWII newspaper collections for GRAC was partially funded by Friedrich Schiller University Jena (2019–2023), with the support of Prof. Ruprecht von Waldenfels.

10. Bibliographical References

2010–2016. Archive of old newspapers. Digital archive of historical Ukrainian newspapers.

Arkhivni Informatsijni Systemy. 2017–. LIBRARIA: Archive of Ukrainian periodicals.

Anna Bordovska. 2024. Orthographic and lexical variation in the journalism of the Second World War period: Based on GRAC data. *Movni i kontseptual'ni kartyny svitu*, 1(75):36–59. In Ukrainian.

Yurij 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.

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

Orest Drul. 2014–. Western Ukrainian historical press collections. Zbruč Digital Archive. Three chronological collections: *125 Years Ago* (1897–1902), *100 Years Ago* (1922–1927), *75 Years Ago* (1947–1952).

Zynovija Franko. 1995. Variation or territorial distinction of the Ukrainian literary language. *Ukrans'ka istorična ta dialektna leksyka*, (2):169–173. In Ukrainian.

Anastasiia Hleba. 2025. Regional variation of the spatial Ukrainian prepositions in the 1920s–1930s: a corpus-based study. In *Synsémantické slovní druhy ve slovanských jazycích*, pages 147–166. Institute of Slavonic Studies of the Czech Academy of Sciences, Prague.

Pavlo Hrytsenko. 1993. Some remarks on the dialectal basis of the Ukrainian literary language. In *Philologia slavica: To the 70th Anniversary of Academician N.I. Tolstoy*, pages 284–294. Moscow. In Russian.

Paul Robert Magocsi. 1987. *Ukraine: A Historical Atlas*. University of Toronto Press, Toronto.

Ivan Matvijas. 1998. *Variants of the Ukrainian Literary Language*. Kyiv. In Ukrainian.

Yurij Shevelov. 2008. The Ukrainian language in the first half of the twentieth century (1900–1941): State and status. In *Selected Works: In 2 volumes. Vol. 1: Linguistics*, pages 26–279. Kyiv-Mohyla Academy, Kyiv. In Ukrainian.

Maria Shvedova. 2020. The General regionally annotated corpus of Ukrainian (GRAC, uacorpus.org): Architecture and functionality. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS)*, pages 489–506, Lviv, Ukraine.

Maria Shvedova. 2021. Lexical variation in the language of the Ukrainian press of the 1920s–1940s and the development of a new lexical norm: A corpus-based research. *Movoznavstvo*, (1):16–35. In Ukrainian.

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)*, pages 32–45, 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)*, pages 71–80, Gliwice, Poland.

11. Language Resource References

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*.

Milan Straka, Jana Straková and 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.