***Shpak Anastasiia**, BA Student*
***Andrushenko Olena**, Dr. Habil., DSc (Philology), As. Prof.*
Zhytomyr Polytechnic State University

CORPUS OPPORTUNITIES FOR LANGUAGE LEARNERS AND RESEARCHERS: GRAC

As the field of digital linguistics advances quickly, work with annotated corpora is an essential part of the process. Annotated corpora are structured collections of texts with linguistic annotations that allow researchers, educators, and technologists to study and explore language in a data-based way as well as instinctively. In Ukrainian, this function is increasingly performed by GRAC — the General Regionally Annotated Corpus of Ukrainian [1]. GRAC is an open access digital resource that boasts a wealth of grammatically tagged data from real usage of Ukrainian language, offering new opportunities for the research of language, direct instruction of language, and language processing in digital modalities.

Introduced by a team of individuals including Maria Shvedova, Ruprecht von Waldenfels, and others, GRAC runs in response to a long-felt need for a modern, accessible, and linguistically substantial Ukrainian corpus [2, p. 2]. An important milestone in the history of Ukrainian digital resources, it represents a stride toward the representation of Ukrainian in the digital sphere, comparable to other significant world languages in terms of corpora and resources.

Corpus linguistics as a scientific method underwent resurgence in the late 20th century with the advent of linguistic research that utilized empirical and replicable data. Corpora like the British National Corpus and the Russian National Corpus have long been invaluable for academic and technological purposes. However, for many years, Ukrainian did not have a comparable large-scale and publically available corpus providing exhaustive grammatical tagging.

To address this gap, the GRAC project was established to gather a range of contemporary Ukrainian texts, whole texts, and apply a uniform morphological and syntactic annotation as described by international guidelines such as Universal Dependencies.

The GRAC is a balanced corpus of texts from multiple genres and registers. The genres and registers include journalistic writing, fiction, scientific writing, official writing, educational materials, and varieties of spoken or informal language. There are presently over 400,000,000 tokens in the corpus, covering almost all genres of written texts [2, p. 2].

This genre diversity guarantees that the corpus accurately represents both the standardized reference point of the Ukrainian language and the fluid use of the language in everyday contexts. This distinction is important for theoretical linguistics or even language technology applications.

The multiplicity guarantees that the corpus appropriately reflects the fixed, reference point of the Ukrainian language and the squirrely use of the language more broadly. This distinction is crucial for theoretical linguistics or even language technology.

Syntactically, the corpus is parsed using dependency grammar, which models sentence structure as a hierarchy of words connected by grammatical relationships. For example, the word “йшов” (was walking) is annotated as follows: йти VERB Aspect=Imp|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin. Such

annotation allows for advanced searches and linguistic analysis based on grammatical patterns, not just keywords.

In addition to morphological and syntactic tagging, GRAC has added a semantic annotation layer, using the Ukrainian Semantic Lexicon (USL) and the TagText tagger as a basis for that annotation. The most frequent lemmas, around 1,000, have been given semantic tags, making this the basic underlying lexicon in the corpus. This open the corpus up to more advanced searching and analysis features that are meaning based, which we hope will improve the corpus for linguistic and computational need [3, p. 4].

GRAC [1] is also available on an online site, which facilitates searching and analyzing the corpus in an easy manner. Users can search by lemma, word form, part of speech or grammatical features, while also filtering the results by genre or source, see examples of words being used in real life situations, and export their search results and analyze offline.

Developers and researchers in natural language processing (NLP) will appreciate GRAC's API capability that allows users to access its data in a computationally-based environment, such as a spell-checker, language translator, or text-classifier.

GRAC developers and researchers in natural language processing (NLP) will enjoy GRAC's API functionality, which enables users to access GRAC data within a computationally-based environment (e.g. a spell-checker, language translator, or text-classifier).

In education, GRAC is a source of authentic materials that are useful for the teaching of the Ukrainian language. For example, you can use GRAC to devise exercises, grammar explanations, vocabulary inventories, and readings that fit naturalistic contexts. This is particularly significant when we think about teaching Ukrainian as a foreign language, where you do have limited access to samples of natural, annotated language.

In the NLP context, GRAC provides raw data for training language models, developing syntactic parsers, and developing dialogue systems. It aids in the development of software using grammatical understanding, analysis, and generation of Ukrainian text - a requirement for modern AI apps.

Even though GRAC is already a strong and flexible resource, it represents a major opportunity to grow and improve. Some of the most exciting possibilities are:

- Increasing the representation of spoken Ukrainian in the form of transcribed conversations.
- Incorporating multimedia content (like audio, subtitled video)
- Developing bilingual and multilingual aligned corpora (like Ukrainian-english, Ukrainian-polish)
- Improving the user interface specifically for non-specialist users like school teachers and students
- Developing language learning applications based on the corpus like smart dictionaries, grammar trainers and AI tutors.

These areas of investment would position GRAC as a pillar of both academic linguistics and equitable and technologically driven education in languages.

The GRAC corpus represents a landmark achievement in the digital structuring of Ukrainian language. It freely searchable grammatically and semantically annotated information, well suited for linguists, educators, students and software developers alike. It combines linguistic rigor with technological accessibility, meaning that Ukrainian will have a future as a live and wholly participating language in the digital world. Its future

growth and expansion will be essential to enhancing the global presence of the Ukrainian language, and involve innovation in the areas of Slavic and computational linguistics.

REFERENCES

1. Shvedova, M., von Waldenfels, R., Yarygin, S., Rysin, A., Starko, V., Nikolenko, T., et al. (2017–2025). *General regionally annotated corpus of Ukrainian.* Kyiv, Lviv, Jena. Retrieved from http://uacorpus.org
2. Shvedova, M., von Waldenfels, R., Yarygin, S., Kruk, M., Rysin, A., Starko, V., & Woźniak, M. (2020). *The General Regionally Annotated Corpus of Ukrainian (GRAC): Architecture and Functionality*. CEUR Workshop Proceedings, 2604, 1–8. Retrieved from https://ceur-ws.org/Vol-2604/paper36.pdf
3. Starko, V. (2021). *Implementing Semantic Annotation in a Ukrainian Corpus*. CEUR Workshop Proceedings, 2870, 1–10. Retrieved from https://ceur-ws.org/Vol-2870/paper32.pdf