Contribution to the ELTeC-ukr subcorpus

Keywords: ELTeC, Ukrainian, Ukrainian literature, distant reading

The European Literary Text Collection (ELTeC) consists of a multilingual corpus of novels produced by the COST Action *Distant Reading for European Literary History* (CA16204). It is designed as a collection of literary texts that are comparable in nature, scope, and quality across several European languages. The corpus covers novels published between 1840 and 1920 by both male and female authors, a period selected because of the substantial increase in literary production across Europe. ELTeC-ukr constitutes the Ukrainian subcorpus within ELTeC and is devoted to novels originally written in Ukrainian. Given the long history of suppression and restriction of the Ukrainian language, especially during the Russian Empire, the creation of this subcorpus responds to the need to preserve both the scholarly and cultural heritage of Ukrainian literature.

The present corpus study focuses on the ELTeC-ukr subcorpus. At the time of the study, ELTeC-ukr comprised 50 novels, encoded at levels 0 and 1, and authored predominantly by men. The main aims of the current work were to (i) expand ELTeC-ukr by identifying additional novels, with extensive focus on female authors; (ii) improve the balance of the subcorpus in terms of author gender and text length; and (iii) document the distribution of Ukrainian novels by gender and across different time periods. The principal selection criterion was a minimum length of 10,000 words. In addition, the research covered four time slots (T1: 1840–59, T2: 1860–79, T3: 1880–99, T4: 1900–1920), included only texts originally written in Ukrainian, focused on medium and long novels—especially by female authors—and favoured works first published in books rather than periodicals. Both canonical and non-canonical novels were included.

Data collection was conducted between July and September 2025 and involved an extensive search of online databases for texts freely available under reusable public licences. In total, 103 texts were identified: 35 male-authored novels by 17 male authors and 68 female-authored novels by 19 female authors. The results indicate that no suitable texts were found in the earliest period (T1: 1840–59), while female authors began to publish extensively from the 1880s onwards, with the highest concentration of works appearing after 1900. This pattern is consistent with broader changes in the political situation and the rise of feminist movements. Across the corpus, short texts clearly predominate, particularly among female authors, whereas medium and long novels are comparatively rare for both male and female authors.

Overall, male authors still prevail numerically over female authors in the final corpus, and some texts by both male and female writers are still missing. The next steps will involve validating the collected texts for their suitability for inclusion in the Ukrainian corpus and linguistically annotating them at TEI level 2 for integration into the final ELTeC-ukr subcorpus. In addition, further in-depth research, including in-person work in relevant archives, will be required to identify missing texts, especially by female authors. The present work contributes to the enlargement of the ELTeC-ukr subcorpus by enriching it with additional material and documenting text-encoding practices, while helping to preserve the cultural heritage of Ukrainian literary production and shedding light on the historical imbalance between male and female authors during this period.