ÚRSCÉAL: Building and Analysing a Corpus of the early Irish-language Novel.

Tonra, Justin
justin.tonra@universityofgalway.ie
University of Galway, Ireland

ÚRSCEAL is a project to build an open-access TEI-encoded collection of the Irish-language novel for inclusion in the European Literary Text Collection (ELTeC), the state-of-the-art and standards-led multilingual corpus of the European novel (1840-1920) (Schöch et al. 2021). ELTeC provides a collection of novel corpora that are comparable in nature, scope, and quality across several European languages: an essential condition for the creation of multilingual tools and methods of computational literary analysis. To date, twenty European languages are represented in ELTeC, but Irish is not. This short presentation reports on the challenges and opportunities of developing an Irish-language corpus for ELTeC and considers questions for preliminary computational analysis of the collection.

ÚRSCÉAL's contribution to ELTeC solves a gap by completing and submitting a collection of the Irish-language novel collection to the corpus, providing the data and motivation for computational literary analysis of the Irish-language novel and its European comparators. Moreover, it seeks to remedy broader gaps in the availability of crucial texts from the early Irish-language novel tradition by publishing new and accurate digital texts in a range of accessible formats and examining the potential for developing an expanded corpus of the Irish-language novel beyond 1920.

Compared to other European languages, the tradition of the novel in Irish begins relatively late: in 1901. Renowned for a rich oral literary culture, the development of an Irish-language prose culture was delayed until the beginning of the twentieth century owing to a variety of complex and interlinked causes which lie in "the turbulent disjunctions of Irish colonial history" (O'Leary 1994). Warfare, famine, and penal legislation to suppress the use of Irish led to such an impoverished bibliographical culture that many of the founding authors of the late nineteenth-century Gaelic Revival movement had never read a book in the Irish language.

Irish, thus, is an outlier in the linguistic history of the European novel, but that uniqueness offers analytical opportunities of its own: how does a newly-developing novel culture compare to the more mature traditions of its European neighbours? How does it reflect its encounters with novels in English and other languages, as well as its own vernacular literature in established forms such as poetry? The presentation considers how computational analyses might be deployed to address these research questions.

Many of these computational analyses have been used by colleagues from South-Eastern Europe during and after COST Action CA16204 Distant Reading for European Literary History. Computational analysis of novel collections from Romania, Serbia, and Croatia prove instructive for making preliminary assessments of representative national novel corpora, while a study exploring titling practices in literary discourse which examined Romanian, Serbian, Slovenian, and Ukrainian collections alongside those of more culturally hegemonic Western European languages delivers a model for genuine comparative multilingual analysis of the European novel (Patras et al. 2019; Stanković et al. 2022; Krstev et al. 2019; Primorac 2019; Marinescu 2019; Patras et al. 2021).

This short presentation provides a report on the development of an important new resource that will act as a model, describing opportunities and challenges, for others who wish to a build a corpus in a minority or minoritised language. At the same time, it links to the broader collaborative development of ELTeC and, in considering questions for computational analysis of the collection, demonstrates fruitful exchanges and collaborations with South-Eastern European colleagues and their research, while offering advice for initial computational analysis of similar corpora. The opportunities presented by these collaborations will strengthen ELTeC as the basis for comparative analysis of the European novel and clear a path for dedicated computational analysis of the early tradition of the novel in Irish.

Bibliography

Krstev, Cvetana, et al. (2019): "Analysis of the First Serbian Literature Corpus of the Late 19th and Early 20th Century with the TXM Platform", in: Pálko, Gábor (ed.): *DH_Budapest_2019*, 2019, 36–37 http://elte-dh.hu/wp-content/uploads/2019/09/DH_BP_2019-Abstract-Booklet.pdf.

Marinescu, Luiza (2019): "From Close to Distant Reading of 100 Romanian Novels", in: *Studia Ubb Philologia* LXIV, 2: 239–50. DOI: 10.24193/subbphilo.2019.2.19.

O'Leary, Philip (1994): *The Prose Literature of the Gaelic Revival, 1881-1921: Ideology and Innovation*. University Park: Pennsylvania State University Press.

Patras, Roxana et al. (2019): "The Splendors and Mist(Eries) of Romanian Digital Literary Studies: A State-of-the-Art Just before Horizons 2020 Closes Off", in: *Hermeneia* 23: 207–22.

Patras, Roxana et al. (2021): "Thresholds to the 'Great Unread': Titling Practices in Eleven ELTeC Collections", in: *Interférences Littéraires/Literaire Interferenties* 25: 163–87.

Primorac, Antonija (2019): "Infrastructural Challenges for Large Scale Digital Text Corpora: A View From the European Margins", Humanities, Arts and Culture Data Summit and DARIAH Beyond Europe workshop, National Library of Australia, Canberra, 2019 https://www.humanities.org.au/special-event-1903/.

Schöch, Christof, et al. (2021): "Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives", in: *Modern Languages Open* 1: 1-19. DOI: 10.3828/mlo.v0i0.364.

Stanković, Ranka, et al. (2022): "Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection", in: European Language Resources Association (ed.): *Proceedings of the Language Resources and Evaluation Conference*, 2022, 3337–45 https://aclanthology.org/2022.lrec-1.356.