Towards Metadata-enriched Literary Corpora in Line with FAIR Principles: 19/20MetaPNC

Rosiński, Cezary
cezary.rosinski@ibl.waw.pl
The Institute of Literary Research of the Polish Academy of Sciences, Poland

Karlińska, Agnieszka
agnieszka.karlinska@ibl.waw.pl
NASK National Research Institute

Kubis, Marek
marek.kubis@amu.edu.pl
Adam Mickiewicz University in Poznan, Poland

Hubar, Patryk
patryk.hubar@ibl.waw.pl
The Institute of Literary Research of the Polish Academy of Sciences, Poland

Wieczorek, Jan
jan.wieczorek@pwr.edu.pl
Wroclaw University of Science and Technology, Poland

We aim to introduce a comprehensive workflow for the enrichment and linking the metadata of a literary corpus, including an implementation of FAIR principles, which have been developed in the field of scientific data management. The metadata compiled following the proposed workflow enables the evaluation of the corpus representativeness and its balancing according to different features. The proposed formula is very adaptable and open to new applications by easy extension of the metadata schema and, as a consequence, the corpus is highly customisable to the possibly wide spectrum of research questions and theoretical frameworks. We will present the application of the workflow in practice using the example of 19/20MetaPNC, a corpus of Polish novels first published in book form between 1864 and 1939.

The practice of building corpora for literary research is relatively new, and model procedures for designing collections of literary texts or describing them with metadata have not yet been developed. Procedures founded in experience derived from linguistic corpora cannot be applied uncritically in this case, as the functions and expectations of literary and linguistic corpora are distinct.

In Poland, there is no representative and balanced corpus of novels. Most Polish literary corpora were designed for particular research projects, have not been robustly described with metadata and do not meet the criteria of interoperability and reusability. Evaluation of existing corpora shows a common lack of metadata verification. This is caused, on the one hand, by an over-reliance on institutional sources of metadata and the assumption that they should be accurate, and, on the other hand, by the lack of possibility to compare and validate metadata with other sources (the lack or insufficient use of LOD leads to isolation of sources).

For the purpose of building the 19/20MetaPNC corpus, we collected 5326 texts from four sources: the ELTeC corpus, the Wolne Lektury website, the Polish edition of the Wikisource project and the Polona digital library. The texts were imported into the CoNLL-U Plus format. We corrected OCR errors and carried out diachronic normalisation. Our goal was to create a representative corpus of Polish prose for the purpose of conducting spatial-diachronic literary research (Karlińska et al 2022). Therefore, we described this dataset with a rich set of metadata, which we used to balance version 1.0 of the 19/20MetaPNC corpus (Fig. 1).

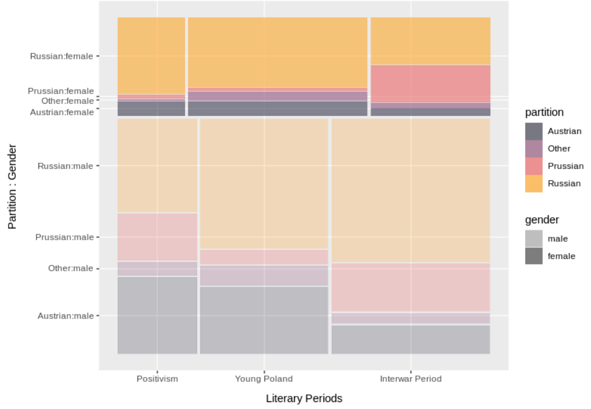

Figure 1: Mosaic diagram of the relationships between the characteristics of the novels in 19/20MetaPNC.

The original metadata that come with the source texts were supplemented both in automatic and manual manner. We linked texts to metadata from library catalogues, using the services of the National Library of Poland, and then enriched the entities with permanent identifiers (PIDs) of widely used databases: VIAF, Wikidata, and Geonames. Simultaneously, the collection of texts was manually annotated, which covered the time of the novel's action and also verified for original language and genre.

The metadata with which we have used to describe the corpus eludes the traditional model of archiving and sharing collections, which are well-known from domain bibliographies or library catalogues. None of the commonly used bibliographic data formats is capacious enough to capture information crucial to our research, such as the geographical coordinates of the places described in the novel.

However, neither the arbitrary format of metadata notation that we adopted, nor other available formats allow for the scientific reuse of a literary corpus. Our goal is to devise a procedure that can be easily adapted to other kinds of literary research. In order to fulfil this, it is necessary to provide both a universal, open and flexible way of adding further metadata categories, corresponding to current research needs, and to propose methods of structuring them in such a way that they are intuitive for subsequent researchers unaware of the previous shape of the metadata.

For this purpose, we propose a workflow for the meta-description of the corpus, which is based on state-of-the-art solutions in line with FAIR principles for making data available on the Internet. We propose the use of the Resource Description Framework (RDF) metadata format and the open-source Wikibase software, enabling the storing and sharing of data compatible with Linked Open Data. We use the Wikidata service as a graph database, thus giving a wide spectrum of possibilities for developing links with interdisciplinary datasets, e.g. VIAF, Worldcat, Library of Congress, ISNI. We provide a Semantic Web-ready solution for creating interchangeable and machine-readable metadata for literary corpora description, particularly tailored to the needs of a diverse scientific community.

Bibliography

Herrmann, Berenika / Lauer, Gerhard (2018): “Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne”, in: *Osnabrücker Beiträge zur Sprachtheorie* 92: 127–56.

Karlińska, Agnieszka / Rosiński, Cezary / Wieczorek, Jan / Hubar, Patryk / Kocoń, Jan / Kubis, Marek / Woźniak, Stanisław / Margraf, Arkadiusz / Walentynowicz, Wiktor (2022): “Towards a Contextualised Spatial-Diachronic History of Literature: Mapping Emotional Representations of the City and the Country in Polish Fiction from 1864 to 1939”, in: *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, International Conference on Computational Linguistics*, 115–25.

Lindemann, David (2021): “LOD-ification of Bibliographical Data Using Free Software: CLB-LOD Wikibase” https://doi.org/10.5281/zenodo.7250730 [28.04.2023].

Schöch, Christof / Patraş, Roxana / Santos, Diana / Erjavec, Tomaž (2021): “Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives”, in: *Modern Languages Open* https://doi.org/10.3828/mlo.v0i0.364 [28.04.2023].