Digital Literary Memory in Central-East-Europe: Analysing Wikipedia with LLMs

Botond Szemes, Kata Dobás
HUN-REN BTK Institute for Literary Studies, Budapest, Hungary
szemes.botond@abtk.hu, dobas.kata@abtk.hu

Introduction

This research investigates contemporary literary knowledge within Central and Eastern Europe in order to explore the region's connections and cultural unity/diversity. In the research, the term "literary memory" is used to address precisely these questions, i.e. the institutionalised and contemporary knowledge of a community—in the narrow sense of Jan Assmann's concept of cultural memory (Assmann 2011), referring only to the field of literature. It is also closely tied to the question of digital memory, as contemporary memory culture cannot be separated from the function of digital archives, as well as in the digital space, narrative forms of memory are increasingly being replaced by databases and algorithms. These require not only narrative theory but also data science and computational tools to effectively investigate how literary memory is structured and transmitted today (Mandolessi 2023).

In light of this, we believe that analyzing multilingual online encyclopedias and databases offers the most effective approach to understanding the structure of literary memory in the region. The advantage of analyzing online databases lies in their ability to map the contemporary knowledge and interests of a community that extends beyond the professional actors of cultural mediation (such as translators, scholars etc.).

However, this approach is far from straightforward. In addition to collecting data from semantic databases, information must also be extracted directly from encyclopedia articles to ensure reliable results. Manually performing this task is only feasible to a limited extent. Therefore, we aim to automate the information retrieval process using large language models.

Research Question

What do different cultures know about each other's literary histories? To what extent is a literary tradition represented in the collective memory of the region? How thoroughly does one culture remember the literature of another? The broader research project explores these questions by analyzing multilingual online encyclopedias—primarily Wikipedia and Wikidata—to map the structure of literary memory across Central and Eastern Europe.

The present, more focused phase of the research investigates how large language models can support large-scale data collection and data cleaning. It aims to address the following sub-questions:

- Which language model is best suited for this task?
- How can model performance be reliably measured?
- What is the optimal balance between maximizing data coverage and ensuring result accuracy?
- Which prompts can be used to improve the performance of the models?
- How can the methodology created for a specific task be used for other research in the future?

Methodology

1. Data collection

Our primary inspiration comes from the 2023 special issue of the Journal of Cultural Analytics, Wikipedia, Wikidata, and World Literature, which "revolves around encyclopedic data and interlinked facts that can provide novel sources and tools for studying the reception of world literature.' (Fischer 2023) This issue builds upon earlier studies that also recognized Wikipedia as "a prominent example of a source for peeking into the wisdom of the masses, rather than the preferences of a few." (Rosendahl 2019) With these considerations in mind, we have analyzed both Wikipedia and the Wikidata database, particularly due to their multilingual nature and extensive data on contemporary knowledge structures. These platforms reflect not only the perspectives of researchers and professionals but also the interests of the wider public, as they are open for anyone to edit. More specifically, we assume that if an author from one literary tradition (e.g., Milan Kundera, who also published in Czech) has a Wikipedia page in another language (e.g., Polish), she/he can be considered part of that tradition's literary memory (Kundera of Polish literary memory)—at the very least, information about her/him is accessible in the most widely used encyclopedia within that linguistic and cultural context. By quantifying such instances across the Visegrad area we can measure the extent of literary memory in different traditions and map their relationships with one another.

To investigate this, we consulted the semantic database Wikidata, which catalogues personal name records according to various attributes and links them to the corresponding Wikipedia articles. For an efficient query, we had to clarify two fundamental questions: first, which attribute determines nationality, and second, which occupations qualify the category of 'author.' The answers to these questions were far from straightforward. Nationality can be defined linguistically, geographically, or in terms of cultural identity—each of which may yield different results. Similarly, the categories of 'literature' and 'authorship' have experienced significant historical changes, making classification increasingly complex. To ensure consistency, we developed general criteria to guide our decisions in individual cases. Nationality was defined on the basis of language (e.g. a work belongs to the Hungarian literary tradition if it was written in Hungarian, regardless of the place of writing and the nationality of the author). In addition, among the authors who lived and worked after 1800, only those were included in the research who had at least one work of fiction in the modern sense (novel, poem, short story, etc. and not memoir, religious text, essay etc.) All authors before 1800 were included without generic restriction, since literature was not yet separated from other written texts at that time.

2. Data cleaning

However, these aspects cannot be accurately filtered on Wikidata—only broad categories such as "writer" are available, which encompass a wide range of professions, from politicians to pop musicians. Including all individuals labeled as "writers" would distort our analysis and hinder us from answering our original question about the presence of fiction writers and, consequently, the structure of literary memory. Therefore, following the SPARQL queries, the data had to be further refined. This filtering was done manually in the first phase of the research, limiting the analysis to the Visegrad countries (Czech Republic, Hungary, Poland, and Slovakia). This restriction was due to practical reasons: there was simply too much data to review manually.

Both of us reviewed all authors retrieved through the SPARQL queries in each of the 12 final tables (four literary traditions, each represented in three foreign-language Wikipedias), consulting regional experts in debatable cases. This process resulted in a cleaned dataset that offered valuable insights into the literary and cultural relations within the Visegrad region. A paper based on these findings is forthcoming in the Q1 journal Revista Transilvania in 2025.

In the second phase of our research—which is the focus of our current study—we aim to automate the data-cleaning process. This will not only accelerate our workflow but also allow us to extend the analysis to encompass the entire Central European region, or potentially an even larger area. This scaling up represents a significant step forward in the study of Europe's cultural history. Our method employs multilingual large language models, which have proven particularly effective in similar information-retrieval tasks (Tang et. al. 2024, Zhu et. al. 2023 and Wang et. al. 2024). In this approach, each author's Wikipedia entry in a given language is fed to the model in the form of a table, where each author appears in a separate row. The prompt includes our criteria for data filtering (see above). The model returns an expanded table with two additional columns: one indicating whether the author qualifies as a "writer" (True/False), and the other listing the relevant writing language(s) using ISO 639 codes. For example: Franz Kafka – writer = True, language = 'de'.

3. Evaluation

To validate the results, we use our previously compiled, manually cleaned list of over a thousand authors as a benchmark dataset. The models' performance is evaluated by language model, author nationality, and Wikipedia language, while also considering additional formal variables such as the length of the Wikipedia article. This phase of the research compares the performance of different language models. From the results, we can infer the language model's knowledge of different languages, and their accuracy in real-life information retrieval tasks.

Preliminary Findings

Model Performance

Preliminary results indicate that the most reliable free model is **LLaMA 3.3:70B**, accessed locally via the HUN-REN GenAI4Science API. Across 12 datasets (organized by author language and Wikipedia language), it achieves a **mean F1 score of 0.87**, with an average precision of 0.82 and recall of 0.93. Accuracy improves further when the evaluation is limited to authors for whom the language model assigns only a single language. In these cases—where the ambiguity introduced by multiple language tags is eliminated—the average F1 score rises to 0.91. However, this filtered subset represents only about 50% of the full dataset, which is a clear limitation of this approach.

Cultural History

This stage of the research demonstrates that large language models can be used not only for data cleaning and information retrieval but also for developing new methodological approaches in the digital humanities. By integrating quantitative techniques with cultural analysis, the project creates opportunities to examine cultural systems that were previously accessible only through qualitative or anecdotal case studies. The cleaned dataset on the Visegrad region has already yielded fresh insights into cultural relations and literary networks within the area (see example Fig. 1). Automating the data cleaning process will allow for the analysis of cultural connections on a broader scale, uncovering patterns and dynamics that have largely remained unexplored.

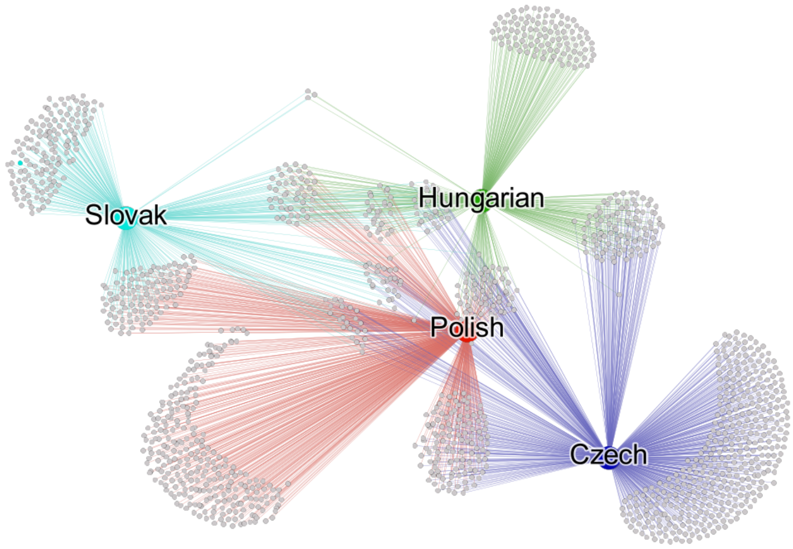

Figure 1: The network of literary memory in the Visegrad region. An author as nodes is linked to a tradition if he or she does not write in the language of that tradition but is listed in its Wikipedia.

Impact

The project raises critical ethical and methodological concerns: how we define national identity, authorship, and literary canons in computational research. Choices made in categorizing data (e.g., excluding memoirs or defining "writer") reflect not just technical needs but cultural values and historical biases—requiring ongoing critical reflection and professional responsibility.

In addition to the cultural-historical findings and the methodology developed, the cleaned database generated through this research will be made freely available for future studies via the Institute for Literary Studies' semantic database, ITIdata. The dataset for the Visegrad countries can be accessed here: https://n9.cl/ji5do.

References

[1] Applebaum, A. (2013). *Does Eastern Europe Still Exist?*. Prospect Magazine, 20.

[2] Assmann, J. (2011). *Cultural Memory and Early Civilization: Writing, Remembrance, and Political Imagination*. Cambridge University Press.

[3] Cobel-Tokarska, M. (2020). *Problems and contradictions in Polish postcolonial thought in relation to Central and Eastern Europe*. Postcolonial Studies, 24(1), 1–20.

[4] Fischer, F., et al. (2023). *Preface: World Literature in an Expanding Digital Space*. Journal of Cultural Analytics, 8(2).

[5] Thomsen, M. R. (2019). *Media and Method: The Digitized Library of Babel*. In M. R. Thomsen & S. Helgesson (Eds.), *Literature and the World* (pp. 109–130). Routledge.

[6] Mandolessi, S. (2023). *The Digital Turn in Memory Studies*. Memory Studies, 16(6).

[7] Tang, Q., et al. (2024). *Self-Retrieval: End-to-End Information Retrieval with One Large Language Model*. arXiv preprint arXiv:2403.17826.

[8] Zhu, Y., et al. (2023). *Large Language Models for Information Retrieval: A Survey*. arXiv preprint arXiv:2307.16892.

[9] Wang, Z., et al. (2024). *Redefining Information Retrieval of Structured Databases via Large Language Models*. arXiv preprint arXiv:2402.06033.