Big Data, Small Literatures

International conference

Masovni podatki, male literature

Mednarodna konferenca

Program konference in zbornik povzetkov

Conference Programme and Book of Abstracts

3.–4. november 2023 3–4 November 2023

Prešernova dvorana SAZU, Novi trg 4, Ljubljana

Masovni podatki, male literature / *Big Data, Small Literatures*

Ljubljana, 3.–4. november 2023 / *3–4 November 2023*

Program konference in povzetki referatov / *Conference Programme and Book of Abstracts*

Uredili / *Edited by*: Jernej Habjan, Lucija Mandić, Ivana Zajc

Organizacija posveta / *Conference organisation*: Jernej Habjan, Lucija Mandić, Ivana Zajc

Leto izida: 2023

Oblikovanje in prelom / *Design and layout*: Ivana Zajc, Lucija Mandić, Darko Ilin

Založnik / *Publisher*: Raziskovalni center za humanistiko Univerze v Novi Gorici / *Research Centre for the Humanities at the University of Nova Gorica*

Tisk / *Print*: Arttech

Naklada / *Print run*: 30

Brezplačna publikacija / *Publication free of charge*

MASOVNI PODATKI, MALE LITERATURE / *BIG DATA, SMALL LITERATURES*

Mednarodna konferenca / *International conference*

Program konference in zbornik povzetkov / *Conference Programme and Book of Abstracts*

3.–4. november 2023 / *3–4 November 2023*

Prešernova dvorana SAZU, Novi trg 4, Ljubljana

Založnik / *Publisher*: Raziskovalni center za humanistiko Univerze v Novi Gorici / *Research Centre for the Humanities at the University of Nova Gorica*

2023

1 Uvod

2 Program

Petek, 3. november 2023 / Friday, 3 November 2023

10.00–10.30 Pozdravni nagovori / ***Opening speeches***

- **Blaž Zabel**
- **Lucija Mandić**
- **Ivana Zajc**

10.30–11.30 1. sekcija / ***Session 1***

- **Oleg Sobchuk, Artjoms Šeļa:** Computational Thematics: Comparing Algorithms for Clustering the Genres of Literary Fiction
- **Silvie Cinková, Petr Plecháč:** Rhymes and Syntax. Morpho-syntactic Analysis in a Czech Nineteenth-Century Poetry Corpus
- **Ranka Stanković, Cvetana Krstev, Duško Vitas:** SrpELTeC: Serbian Literary Corpus for Distant Reading

11.30–12.00 Diskusija / ***Discussion***

12.00–13.30 Kosilo / ***Lunch***

13.30–14.30 2. sekcija / ***Section 2***

- **Jan Rybicki:** Between Textual Data and Metadata. Distant Reading 10,000 Books in Polish
- **Marko Pranjić, Andrejka Žejn, Senja Pollak:** Comparing Josip Jurčič and Ivan Cankar using Computational Semantic Change Detection Methods

- **Ivana Zajc:** Stylometric Analysis of 90 Slovenian Novels and the Oeuvre of Ivan Cankar

14.30–15.00 ***Diskusija* / *Discussion***

15.00–15.30 ***Odmor za kavo* / *Coffee break***

15.30–16.30 ***3. sekcija* / *Section 3***

- **Dominika Werońska:** Basking in Basque Prose: A Stylometric Glance at Novels in Euskara
- **Andrei Terian:** The Subgenres of the Romanian Novel (1845–1947): A Distant Reading Based on GDRN 2.0
- **Vlad Pojoga:** The Last Bastion of European Romanticism: A Quantitative Analysis of Poetry in Romania's *Convorbiri literare* (1867–1944)

16.30–17.00 Diskusija / ***Discussion***

19.00 Večerja / ***Dinner***

Sobota, 4. november 2023 / ***Saturday, 4 November 2023***

9.30–10.30 4. sekcija / ***Session 4***

- **Mojca Šorli:** Named Entities in Slovenian Modernist Literature: A Functional View
- **Darko Ilin**, **Katja Mihurko**, **Ivana Zajc**, **Mila Marinković:** Shaping Electronic Collection LETTERS: Discovering Epistolary Exchange and Navigating Correspondence Metadata
- **Lucija Mandić:** Exploring Social Institution Relationships in Nineteenth-Century Slovenian Narrative Fiction through Word Embeddings

10.30–11.00 Diskusija / *Discussion*

11.00–11.30 Odmor za kavo / *Coffee break*

11.30–12.30 Okrogla miza / *Roundtable: Defining Small Literatures by Numbers*

- **Marko Juvan**
- **Jeanne Glesener**
- **Benedikts Kalnačs**

12.30–13.00 Diskusija in zaključek / *Discussion and concluding remarks*

13.00 Kosilo / *Lunch*

4 Povzetki / *Summaries*

Oleg Sobchuk (presenter) and Artjoms Šeļa

Computational Thematics: Comparing Algorithms for Clustering the Genres of Literary Fiction

What are the best methods of capturing thematic similarity between literary texts? Knowing the answer to this question would be useful for automatic clustering of book genres, or any other thematic grouping. This paper compares a variety of algorithms for unsupervised learning of thematic similarities between texts, which we call "computational thematics." These algorithms belong to three steps of analysis: text preprocessing, extraction of text features, and measuring distances between the lists of features. Each of these steps includes a variety of options. We test all the possible combinations of these options: every combination of algorithms is given a task to cluster a corpus of books belonging to four pre-tagged genres of fiction. This clustering is then validated against the "ground truth" genre labels. Such a comparison of algorithms allows us to learn the best and the worst combinations for computational thematic analysis.

Silvie Cinková (presenter) and Petr Plecháč

Rhymes and Syntax: Morpho-syntactic Analysis in a Czech Nineteenth-Century Poetry Corpus

A linguistically informed automatic analysis and modeling of text presupposes a decent performance of NLP tools. We describe our evaluation of the UDPipe parser on a manually annotated nineteenth-century sample from the Corpus of the Czech verse in the following steps: (1) creation of a documented training data set for this domain (poetry, nineteenth century, Czech); (2) error analysis. Our baseline was the best currently available Czech language model, by which we preprocessed a random sample of 29 poems or congruent poem parts totalling 5,000 tokens and had one annotator edit the automatic annotation node by node to come as close as possible to a manual annotation from scratch and record phenomena difficult to match with the annotation standard for contemporary Czech non-fiction prose, on which the model had been trained. Then we compared the automatic annotation with the manual one.

Ranka Stanković (presenter), Cvetana Krstev and Duško Vitas

SrpELTeC: Serbian Literary Corpus for Distant Reading

This paper will present the corpus SrpELTeC, developed within COST Action CA16204 Distant Reading for European Literary History. All novels in SrpELTeC were selected, prepared and annotated using the common principles established for all language collections in European Literary Text Collection - ELTeC. The challenges and solutions in preparing SrpELTeC collection from scratch will be presented. All novels are manually encoded in TEI with rich metadata and structural annotation. The automatic annotation included POS-tagging, lemmatisation and named entity recognition, relying on NLP resources developed and maintained by JeRTeh Society for Language Resources and Technologies. SrpELTeC integration with Wikidata is supported with a set of SPARQL queries for retrieval of metadata with different visualisation options. Recent activity is related to a linked data version of SrpELTeC using NIF (NLP Interchange Format), within COST Action NexusLinguarum – "European network for Webcentered linguistic data science" (CA18209). All versions of SrpELTeC are freely available, under CC-BY licence.

Jan Rybicki

Between Textual Data and Metadata. Distant Reading 10,000 Books in Polish

This paper offers an introduction to a collection of 10,000 literary texts in Polish: prose, poetry, and drama spanning the entire history of writing in Polish; half are original texts by Polish authors, half are translations from more than twenty languages. While such a quantity of literary material is produced in Poland in terms of new titles every four years, this is the most representative collection of stylometrically analysed literary material in Polish to date. Network analysis based on hierarchically clustered comparisons of most-frequent-word frequencies in the texts is combined with such metatextual features as author and translator data, source language, chronology, sentiment, keywords etc., to produce new insights into Polish literary history.

Marko Pranjić, Andrejka Žejn and Senja Pollak (presenter)

Comparing Josip Jurčič and Ivan Cankar using Computational Semantic Change Detection Methods

In our paper, the semantic change detection method using contextual embeddings is applied to the analysis of literary works. The contextual word representations are leveraged in comparison and identification of words that differ the most between two selected authors. We showcase the method's potential on the comparative analysis of two eminent Slovenian authors, Josip Jurčič and Ivan Cankar. We show that the approach based on contextual embeddings can be used for this purpose with satisfactory results. Second, we offer new insights in the discourse of two eminent Slovenian authors, as we show that the transition from realism to modernism is expressed by words semantically connected with movement and social and psychological actions and processes.

Ivana Zajc

Stylometric Analysis of 90 Slovenian Novels and the Works of Ivan Cankar

The proposed contribution shows the results of a computational stylometric analysis of 90 Slovenian novels from the periods of Slovenian realism and "moderna" made with the package "Stylometry with R". The used corpus is based on the works in the Slovenian ELTeC (European Literary Text Collection) and the works on Wikivir. I analyse the clustering of the novels searching for the most obvious signals. What follows is the research of the oeuvre of the Slovenian canonised prose writer Ivan Cankar (1876–1918): first I check the clustering of his works in the larger analysis of different novels, then I synthesise the literary-historical periodisation of his literary work to date, and finally I treat his oeuvre using the method of computational stylometry in the R program with the package "Stylometry with R." Using stylometry as a method of remote reading, I verify whether the findings of literary history about the periodisation of his works are confirmed by computer reading.

Dominika Werońska

Basking in Basque Prose: A Stylometric Glance at Novels in Euskara

While Euskara (Basque) is proclaimed possibly the oldest language on the European continent, it appears in written form only in the sixteenth century. The first Basque novel appears over 300 years later and, to this day, the genre lacks exhaustive research. The following paper sets as its aim a stylometric analysis of a selection of twentieth- to twenty-first-century Basque novels sourced from the online platform Booktegi. These are analysed based on the frequency of the MFW's (most frequent words) measured using cluster analysis and set against a backdrop of foreign novels translated into Euskara. The results show that novels originally in Euskara remain distinct from translated works pointing to the unique linguistic character of the Basque novel. Some linguistic patterns potentially responsible for this distinction are presented. The results are visualised on a map revealing the chronological evolution and the contribution of the Basque novel to the broader literary landscape.

Andrei Terian

The Subgenres of the Romanian Novel (1845–1947): A Distant Reading Based on GDRN 2.0

The present paper proposes a quantitative analysis of the Romanian novel (1845–1947) during the first century of its evolution from the standpoint of the subgenres that governed its dynamics. In this sense, my study makes use of metadata collected from the second edition of GDRN (General Dictionary of the Romanian Novel), which indexes and classifies all Romanian novels published until the year 2000 (of which approximately 1550 before 1948). The century of Romanian novelistic production under investigation is divided into five segments (1845–1877; 1878–1900; 1901–1917; 1918–1932; 1933–1947); for each of these periods, three parameters will be analysed: subgenres (the number and proportion occupied by each subgenre in the overall novelistic production), reissues (the share of reissued titles and the degree of patrimonialisation of certain subgenres and/or titles), and critical reception (the number of critical comments distributed according to titles and subgenres). Based on these investigations, my paper addresses broader aspects of literary history, such as: contextualising the national inferiority complex, expressed by Nicolae Iorga in 1890 through the question "Why don't we

have a novel of our own?", by considering the dominant subgenres prior to the establishment of communism; retracing the genesis of the modern Romanian literary field by institutionalising the opposition drawn between aesthetic and popular fiction; the possibility of explaining the canonisation process of certain subgenres and/or authors with the help of scientometric indicators.

Vlad Pojoga

The Last Bastion of European Romanticism: A Quantitative Analysis of Poetry in Romania's *Convorbiri literare* (1867–1944)

This study puts forward a quantitative analysis of poetry published in the best known and most visible literary periodical in modern Romania, *Convorbiri literare* (*CL* – Literary Conversations), from its founding, in 1867, to its hiatus in 1944. Often seen as the last major magazine of European Romanticism, *CL* appeared at a time when poetry was a privileged genre in Romanian literature, and it hosted in its pages both the national poet, Mihai Eminescu, and the most prominent literary critic of the nineteenth century, Titu Maiorescu. My research will explore how the publication and theorisation of poetry was an integral part of one of the last nation-building processes in Europe by indexing and quantitatively analysing all forms of poetry and what can be designated as "poetic influences" in *CL*. By using ARCANUM's digital archive of more than 66.000 pages of *Convorbiri literare*, I will extract and examine three strata of poetry-related metadata: 1. the local production of poetry – who published in *CL*, when, and how much they did; 2. poetry imports – whose poetry was translated into Romanian in *CL*, when, and what were the authors' origins; 3. influences network – which were the foreign names most mentioned in regards to local production of poetry in essays, critical contributions or commentaries. Thus, my paper aims to chart the national regime of relevance that applies to Romanian modern poetry by uncovering the international network of authors that were at the centre of literary debates in *CL*.

Mojca Šorli

Named Entities in Slovenian Modernist Literature: A Functional View

This paper explores a new approach to the functions of proper names in modern literary discourse, highlighting their semantic, pragmatic and syntactic aspects. A corpus analysis of the representational meanings of personal and place names is followed by an interpretation of their use in terms of a "proprial" continuum that distinguishes prototypical uses from less conventionalised uses, e.g. those in which names are used as common nouns. We first briefly describe the corpus and introduce the annotation scheme, then focus on the results of the proper name analyses and conclude with further challenges and open questions in relation to established NER systems. Different functions of proper names are presented and some examples of preliminary analyses are given, focusing on the annotation of literary character names and place names. This paper is a continuation of research published in the JTDH 2022 Conference Proceedings on manual semantic annotation of named entities based on a proposed set of annotations for a corpus of modernist literary texts.

Darko Ilin (presenter), Katja Mihurko, Ivana Zajc and Mila Marinković

Shaping the Electronic Collection *LETTERS*: Discovering Epistolary Exchange and Navigating Correspondence Metadata

The paper aims to showcase the ongoing development of the Electronic Collection PISMA (LETTERS) at the Research Centre for Humanities, University of Nova Gorica. We will begin by offering a comprehensive overview of the process involved in creating the collection, including its inception and the methodology behind data entry supported by the principles of citizen science. The second part of the presentation will delve into the research implications that have unfolded due to the existence of this collection. We will illustrate how we have harnessed its metadata and classifications for a range of research projects, exploring diverse aspects of epistolary correspondence, including themes like love, intimacy, friendship, food, and more. In the final segment, we will highlight the prospects for student research, exemplified by a Computer Science student's work on interactive visualizations of the collection's metadata. This presentation demonstrates the evolving landscape of digital humanities in Slovenia and

how the Electronic Collection PISMA has become a valuable resource for research projects in Research Centre in Humanities and in the pedagogical process at the University of Nova Gorica.

Lucija Mandić

Exploring Social Institution Relationships in Nineteenth-Century Slovenian Narrative Fiction through Word Embeddings

In my paper, I employ computational techniques to analyse the relationships between discursive domains in nineteenth-century Slovenian narrative fiction. Using word embeddings and the Gensim package in Python, I conduct the analysis on the KDSP corpus of Slovenian narrative prose. My approach involves constructing discursive fields for five social institutions: economy, politics, culture, family, and religion. This is achieved by identifying the top 100 words with the highest cosine similarity to each institution's vector representation. This computational methodology provides a quantitative basis for exploring the relationships between these social institutions, as reflected in the literary works of the era. Notably, my findings reveal a significant overlap between the semantic fields of politics and culture, which serves as a quantitative foundation for an investigation of what traditional literary history has called the "Slovenian cultural syndrome" or the "Prešernian structure."

Jeanne E. Glesener, Marko Juvan, and Benedikts Kalnačs

Defining Small Literatures by Numbers: Summary of the Conference Roundtable

The study of small and minority literatures is gradually gaining momentum both in national literature studies and comparative research. However, the historical development of small and minority literature in Europe in a methodologically coherent comparative perspective is still poorly researched. A number of issues need to be addressed, including (a) the community/nation-building function attributed to literature; (b) linguistic and aesthetic developments such as the major steps in the creation, recognition, and legitimization of respective vernaculars as literary languages in the context of sociocultural multilingualism; (c) the genre of literary historiography which was essential to nation-building and its methodology.

In this roundtable discussion, we will focus in particular on the question of what data are necessary and can be used in order to help define the phenomenon of small literatures and their specificities.

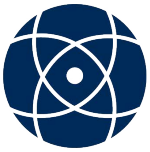
5 Seznam sodelujočih / *List of contributors*

- **Silvie Cinková**, Univerzita Karlova v Praze (cinkova@ufal.mff.cuni.cz)
- **Jeanne E. Glesener**, Université du Luxembourg (jeanne.glesener@uni.lu)
- **Jernej Habjan**, Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti (jernej.habjan@zrc-sazu.si)
- **Darko Ilin**, Univerza v Novi Gorici (darko.ilin@ung.si)
- **Marko Juvan**, Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti (marko.juvan@zrc-sazu.si)
- **Benedikts Kalnačs**, Latvijas Universitāte (benedikts.kalnacs@lulfmi.lv)
- **Cvetana Krstev**, Univerzitet u Beogradu (cvetana@matf.bg.ac.rs)
- **Lucija Mandić**, Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti (lucija.mandic@zrc-sazu.si)
- **Mila Marinković**, Univerza v Ljubljani (mm9136@student.uni-lj.si)
- **Katja Mihurko**, Univerza v Novi Gorici (katja.mihurko.poniz@ung.si)
- **Petr Plecháč**, Akademie věd České republiky (plechac@ucl.cas.cz)
- **Vlad Pojoga**, Universitatea "Lucian Blaga" din Sibiu (vlad.pojoga@ulbsibiu.ro)
- **Senja Pollak**, Institut "Jožef Stefan" (senja.pollak@ijs.si)
- **Marko Pranjić**, Institut "Jožef Stefan" (marko.pranjic@ijs.si)
- **Jan Rybicki**, Uniwersytet Jagielloński (jan.rybicki@uj.edu.pl)
- **Oleg Sobchuk**, Max-Planck-Institut für evolutionäre Anthropologie (oleg_sobchuk@eva.mpg.de)
- **Ranka Stanković**, Univerzitet u Beogradu (ranka@rgf.rs)
- **Artjoms Šeļa**, Polska Akademia Nauk (artjoms.sela@ijp.pan.pl)
- **Mojca Šorli**, Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti (mojca.sorli@zrc-sazu.si)
- **Andrei Terian**, Universitatea "Lucian Blaga" din Sibiu (andrei.terian@ulbsibiu.ro)
- **Duško Vitas**, Univerzitet u Beogradu (vitas@matf.bg.ac.rs)

- **Dominika Werońska**, Uniwersytet Jagielloński (dominika.weronska@doctoral.uj.edu.pl)
- **Ivana Zajc**, Univerza v Novi Gorici (ivana.zajc@ung.si)
- **Andrejka Žejn**, Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti (andrejka.zejn@zrc-sazu.si)

Konferenco prireja Slovensko društvo za primerjalno književnost, sofinancirajo pa jo Javna agencija za knjigo Republike Slovenije, Inštitut za slovensko literaturo in literarne vede ZRC SAZU ter Raziskovalni center za humanistiko Univerze v Novi Gorici.

Hosted by the Slovenian Comparative Literature Association, this conference is co-funded by the Slovenian Book Agency, the ZRC SAZU Institute of Slovenian Literature and Literary Studies, and the Research Center for the Humanities at the University of Nova Gorica.



ZRC SAZU
Inštitut za slovensko
literaturo in literarne vede