

Branko Tošović (ur./Hg.)

ANDRIĆEVO PISMO

ANDRIĆ'S BRIEF

Andrić-Initiative

18

**Institut für Slawistik der Karl-Franzens-Universität Graz
Narodna i univerzitetska biblioteka Republike Srpske u Banjaluci
Komisija za stilistiku Međunarodnog slavističkog komiteta
Svet knjige Beograd
Aletea Beograd
2025**

Gedruckt mit Unterstützung der Universität Graz
Štampanje je finansirao Univerzitet u Gracu.



Herausgeber • Urednik

Verlag • Izdavač

Institut für Slawistik der Karl-Franzens-Universität Graz

Merangasse 70, 8010 Graz, Österreich/Austria, +43 316 380 25 24, branko.tosovic@uni-graz.at; arno.wonisch@uni-graz.at, <http://www.gewi.uni-graz.at/gralis>

Narodna i univerzitetska biblioteka Republike Srpske

Jevrejska 30, 78 000 Banjaluka, BiH, Republika Srpska, direkcija@nub.rs, tel. +387 51 215 894, www.nub.rs

Svet knjige d.o.o.

Bulevar despota Stefana 15, 11000 Beograd, Srbija, tel. +381 11 324 73 07, svetknjige@svetknjige.net,
www.svetknjige.net

Aletea

Generala Štefanika 2, 11000 Beograd, Srbija, +381 62 311 300

Für die Herausgeber • Za izdavače

Branko Tošović,

Arno Wonisch, Ljilja Petrović-Zečić, Stevo Ćosović,

Milan Tasić

Satz, Umschlag, Lektorat und Korrektur

Prelom, korice, lektorisanje i korektura

Branko Tošović

Lektorat und Übersetzung der deutschen Inhalte

Arno Wonisch

Umschlagillustration • Ilustracija na koricama

<https://www.midjourney.com>

Tiraž: 200

Druck • Štampa

Excellent print, Balkanska 35–39, Beograd

Tošović, Branko (ur./Hg.). *Andrićevo pismo / Andrić's Brief*. Graz/Grac – Beograd – Banjaluka:
Institut für Slawistik der Karl-Franzens-Universität Graz – Narodna i univerzitetska biblioteka
Republike Srpske – Komisija za stilistiku Međunarodnog slavističkog komiteta – Svet knjige –
Aletea, 2025. 650 s./S. [Andrić-Initiative 18]

© Branko Tošović, Graz 2025

Alle Rechte vorbehalten. Sva prava zadržana.

ISBN 978-3-9519973-4-6 (Institut für Slawistik der Karl-Franzens-Universität Graz)

ISBN (Svet knjige Beograd)

ISBN (Narodna i univerzitetska biblioteka Republike Srpske Banjaluka)

ISBN (Aletea)

Marija Đokić Petrović (Graz)
Dragana Bečejski Vujaklija (Belgrade)
Jasmina Đorđević (Niš)
Jelena Mitrović (Passau)

Unveiling Literary Legacies: Integrating Named Entity Recognition and SPARQL to Analyse Andrić's Letters

Natural Language Processing (NLP) has completely revolutionized the analysis of textual data, offering new avenues for understanding different textual content. This paper uses the NLP technique for analysing letters written by Ivo Andrić, a Yugoslavian writer, diplomat, and recipient of the Nobel Prize for Literature. Employing Named Entity Recognition, an important NLP technique, we systematically extract significant entities such as persons, dates, and locations from Andrić's letters. These entities serve as inputs for predefined SPARQL queries to retrieve contextual information from DBpedia. Our queries focus on historical events in specific years, profiles of pertinent individuals, and literary works from particular timeframes. This approach allows us to correlate the query results with broader historical, cultural, and personal contexts, which had potential influences on Andrić's literary output. By integrating NLP with the Semantic Web, we aim to reveal more profound influences on Andrić's writing beyond traditional analysis. We believe this is a pioneering approach for analysing Andrić's letters.

1. Introduction

In literary studies, the question of how historical context and a writer's personal experiences influence their creative output remains a subject of enduring fascination and scholarly debate (Literature in Context 2024-www, Biographical Criticism 2024-www). The intricate relationship between an author's life and their literary works offers a window into understanding the cultural and historical milieu in which these works were conceived. This paper explores this correlation in the context of Ivo Andrić (Tošović 2012), a Nobel laureate and Yugoslavian writer.

Andrić, known for his profound narratives interweaving personal and historical themes, has been the subject of extensive literary analysis (Tošović 2014, Tošović 2015, Mihailovich & Andrić 1966). Yet, modern Artificial Intelligence (AI) technologies have not been applied to a large extent for the analysis of his works, especially not for the analysis of his personal letters. This study uses NLP (Khurana et al. 2023) technique, specifically Named Entity Recognition (NER)

(Nasar et al. 2021) in conjunction with Semantic Web (SW) technology, specifically SPARQL (Polleres 2014), in order to analyse Andrić's personal letters. NER, an important NLP technique, detects significant entities like persons, dates, and locations in letters. These entities are then used for predefined SPARQL queries over the invaluable DBpedia database (Auer et al. 2007) to uncover significant information about the socio-political climate and Andrić's social milieu that could influence his creativity.

This paper not only aims to enhance our understanding of Andrić's creativity work but also seeks to demonstrate the effectiveness of integrating modern AI techniques in literary analysis, enriching the field known as Digital Humanities (DH) (Helmi 2021). By bridging traditional literary criticism with innovative analytical tools, this study aspires to provide a more nuanced view of Andrić's work, illuminating the intricate web of influences that shape a writer's narrative world.

Our contributions are summarized as follows:

- The paper represents a methodological approach by combining NER with predefined SPARQL queries for analysing Andrić's letters (Andrić 2000).
- By applying NER, key entities such as persons, dates, and locations are extracted from the letters.
- Extracted NER entities are used as inputs for predefined SPARQL queries over DBpedia. The query results are connected with historical events, cultural dynamics, and Andrić's social milieu, offering insights into their potential impact on his writing.

The remainder is organized as follows. We review the related work in Section 2, while important preliminaries are given in Section 3. We introduce our methodology in Section 4 and demonstrate the application of our approach in Section 5. In Section 6, we highlight the limitations of our work. Section 7 summarizes the conclusions and proposes recommendations for future work.

2. Related Work

The NLP community has shown a growing interest in DH in recent years (McGillivray et al. 2020). The first major step in DH is mostly digitization, the process of converting physical documents into digital (machine-readable) formats, allowing global access and sharing. For this purpose, technologies like OCR (Awel & Abidi 2019) or Handwritten Text Recognition (HTR) (Retsinas et al. 2022) can be used. However, the success of the application of these technologies depends on the physical state of documents or language variations in texts (Awel & Abidi 2019). This paves the way for the quality application of other techniques to analyse textual data.

The field of literature has undergone an important transformation with the advent of DH. One particularly promising NLP technique in this area is NER. For example, the role of NER in extracting meaningful information from texts has been explored in (Tambuscio and Andrews 2021), showing its efficacy in identifying key thematic elements and characters in literary works. Many approaches apply off-the-shelf NER tools to novels and then only retain the *Person* entities as stated in (Labatut & Bost 2019). Authors (van Dalen-Oskam et al. 2014) were focused on mapping names in modern Dutch literature through the lens of literary onomastics – the study of proper names in fictional texts. To achieve this, authors trained existing NER tools on literary texts and developed a new NE tagger that extends standard name categories to distinguish between first and family names. Authors (Dekker et al. 2019) assess tools for NLP on old and modern novels for automatically extracting social networks from novels to gain insights into the community structures and social interactions depicted in literary works.

In many approaches, it is taken a step further, and entities extracted with NER are linked with specific knowledge bases or datasets (Tedeschi et al. 2021) or used for further analysis of character, theme, graph creation, etc. For example, authors in (van Dalen-Oskam et al. 2014) link NER results, a proper names in literary texts, to Wikipedia entries. Authors in (Dekker et al. 2019) evaluate NER tools for the automatic extraction of social networks from novels as well as their network structure. Authors (Ikonić Nešić et al. 2022) process the European Literary Text Collection (ELTeC), a multilingual corpus of novels from 1840 to 1920, and detail a pipeline that includes retrieving metadata, performing NER, transforming and mapping data, populating Wikidata, and linking entities before exporting the data to the NLP Interchange Format (NIF). Focusing on seven ELTeC sub-collections (English, Portuguese, French, Slovenian, German, Hungarian, and Serbian), authors aim to automate data preparation and import into Wikidata using specific tools. The paper also provides examples of SPARQL queries for retrieving authors, novel titles, publication places, and other metadata, showcasing various visualization options and statistical summaries.

3. Preliminaries

In this section, we describe some concepts that are used throughout the paper.

Optical Character Recognition (OCR) is a technology that transforms diverse documents, like images captured by a camera, PDF files, or scanned documents, into machine-readable text format (Awel & Abidi 2019). This technology is a field of study in AI, pattern recognition, and computer vision.

Named Entity Recognition (NER) is an NLP technique that identifies and categorizes specific entities in unstructured text, such as organizations, locations, person names, time expressions, monetary values, quantities, etc (Nasar et al. 2021).

SPARQL (Polleres 2014) is a query language for manipulating and retrieving data stored in Resource Description Framework (RDF) format (Pan 2009). Developed by the World Wide Web Consortium (W3C), SPARQL is beneficial for querying complex and large RDF datasets, and it is an important technology in many data-driven applications, including data analytics, search engines, knowledge management systems, and others.

Dbpedia (Auer et al. 2007) seeks to extract structured content from data produced as part of the Wikipedia effort. Basically, DBpedia transforms Wikipedia, which is mainly text-based, into a structured knowledge database. This content allows users and machines to access and query Wikipedia data more sophisticatedly and refinedly.

4. Methodology

The methodology employed in this research encompasses a systematic analysis of Ivo Andrić's letters, integrating OCR, NER, and SPARQL queries (Figure 1). Initially, letters available in scanned formats are digitized through OCR technology, converting them into machine-readable text. The OCR is performed using the *ScanDocFlow* platform (ScanDocFlow 2024-www), utilizing API calls for the Serbian language, implemented in Python. Subsequent to the digitization process, the text undergoes NER processing, which involves the identification and classification of key entities, such as personal names, locations, and dates within the letters. The NER was executed using *It-Sr-NER*, a CLARIN-compatible NER and geoparsing web service designed for processing texts written in Serbian (Krstev et al. 2014). The output of NER served as inputs for predefined SPARQL queries. These SPARQL queries are executed on DBpedia.

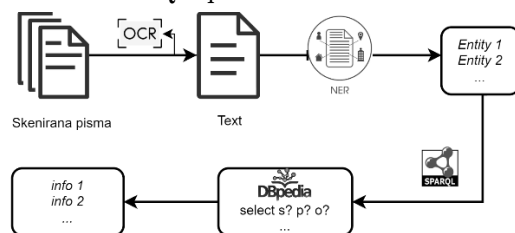


Figure 1. Methodology for analysis of Andrić's letters

Specifically, the SPARQL queries focus on three primary areas: identifying historical events occurring during the period in which the letters were written (Listing 1), extracting information about individuals mentioned in the letters

(Listing 2), and retrieving information on literary works published during the period in which the letters were written (Listing 3). All queries are focused on extracting information in English, due to the lack of information in the Serbian language. The results allowed us to draw correlations between actual historical events, literary trends, and Andrić's social milieu with his thematic focuses. In the following Section, we will see proposed methodology in action.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?eventLabel ?comment
WHERE {
  ?event rdf:type dbo:Event;
  dbo:date ?date;
          rdfs:comment ?comment;
          rdfs:label ?eventLabel.
  FILTER(YEAR(?date) >= %s and YEAR(?date) <= %s).
  FILTER(LANG(?eventLabel) = „en“ and LANG(?comment) = „en“).
  FILTER(regex(?comment, „%s“)).
}
```

Listing 1. The SPARQL query used to retrieve the labels and comments of events from DBpedia, filtering for events that occurred within a specified date range and contain a specific keyword in their comments, while ensuring that both the event labels and comments are in English.

The variables %s will be replaced with the given years and location.

```
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?abstract
WHERE {
  ?person rdfs:label ?person_name;
  ?property ?hasValue;
          dbo:abstract ?abstract.
  FILTER(LANG(?abstract) = „en“).
  FILTER (?person_name = „%s“@en).
}
```

Listing 2. The SPARQL query used to retrieve the English abstract of a specified person from DBpedia. The variable %s will be replaced with the given name.

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?comment
WHERE {
  ?work rdf:type dbo:Book;
  dbo:releaseDate ?date.
  FILTER (?date >= „%s“^^xsd:dateTime &&
```

```

?date <= „%s”^^xsd:dateTime).
?work rdfs:label?workLabel;
rdfs:comment?comment;
dbp:country?country.
FILTER (LANG(?comment) = „en”).
}

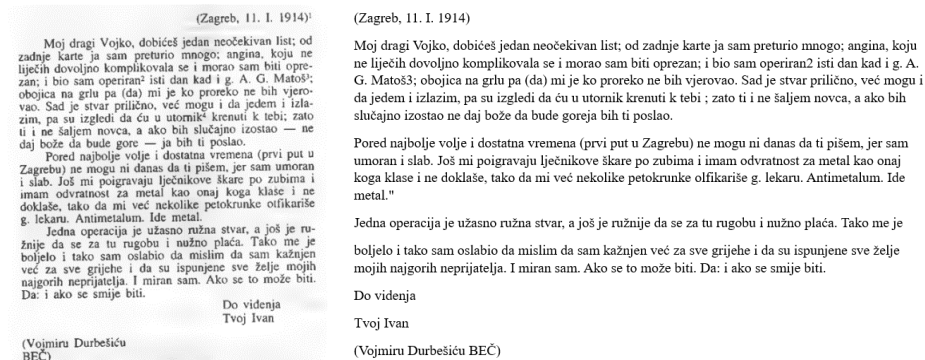
```

Listing 3. The SPARQL query used to retrieve the comments of literary works from DBpedia, published within a specified date range, while ensuring that comments are in English. The variables %s will be replaced with the given years.

5. Methodology in Action

This section demonstrates the practical application of our methodology on a test dataset comprising 11 randomly selected letters from the book (Andrić 2000). The dataset includes letters written in both Serbian Cyrillic and Serbian Latin scripts.

First, we employed OCR to digitize the letters, thereby converting the textual data into a format suitable for machine processing (Figure 2). It is important to note that OCR accuracy was affected by various factors, including shadows or smudges on the scanned letters. Such imperfections led to OCR errors, which were corrected manually.



a) The scanned letter from the book (Andrić 2000)

b) The letter converted to text format

Figure 2. The result of applying the OCR technique to one of Andrić's letters.

The subsequent application of the NER technique was crucial for the systematic identification and categorization of entities within these letters. This process effectively extracted specific personal names, locations, dates, demographic information, events, roles, and organizations. The comprehensive results of applying the NER model (Krstev et al. 2014) to the test dataset are illustrated in Figure 3, while the visualization of the NER output for one of Andrić's letters is presented in Figure 4. The NER model used for this methodology

is developed as a part of project *It-Sr-NER: CLARIN compatible NER and geoparsing web services for parallel texts: case study Italian and Serbian* (It-Sr-NER 2024-www). Although the NER model demonstrated a high F1 score, however, certain issues were identified in its performance. For instance, the verb *ide* (meaning *go*) and the noun *snaja* (meaning *some kind of relative*) were incorrectly classified as personal names. Additionally, the organization *Generale* was erroneously categorized as a role.

Personal name	Location	Date	Demo	Event	Role	Organization
Vojko	Zagreb	11.I 1914	<i>rimskog</i>	Uskrs	g.	Hotel Geneve
A. G. Matoš	Zagreb	isti dan		Uskrs	g. lekaru	Ministrastvo Inostranih Poslova
ide	Beč	dan			MINISTAR POMOĆNIK	Hotel Evropa
Ivan	Beograd	27 maja 1938			Dr.	
Voimiru Durbešiću	Zeneva	4 o. m.			odvjetnik	
I. Andrić	Jurjevska ul, 19/1	14 oktobra 1930 g.			vojnik	
Ivo	Trst	5. VI 1926.			gospodice	
Voimir Durbešić	Trstu	15 dana			Gdica	
Tugomira Alaupovića	Višegrada	10. IX. 1945 g.			Generale	
Ivo Andrić	Beogradu	16 O. m.			drugu	
Zdenka Marković	Sarajevo	20 oktobru			pisca	
Tugomire	Evropa	8 – 10 dana			pesnik	
Anto	Vlajkovićeve ul	11. II 1925			pisac	
Omeru	5. rue des Alpu	11. II 25			Gdi.	
Tugomir Alaupović	Zagreba	6 aprila 1932 g.				
Vuču	Frankfurta	11. aprila 1939 g.				
Ivo Andrić	Berlin	7. V 1921				
Iv. Andrić	Dežmanova ul.	Tri dana				
Vera Stojić	Napulj	Jutros				
Zdenka	Rim	Večeras				
Beba	Akademski trg 7. II	Sutra				
Beba Alaupović	NAPOLI	veče				
snajku	Sinaja	7. V. (19)21				
Milutin Popović	Jugoslavija	16. VIII 1941.				
Milevu	Francuska	ovih nekoliko dana				

Figure 3. Results of applying NER model over the test dataset of Andrić's letters. Text marked in lighter tones is labelled wrong.

PERS ¹	LOC ²	ORG	DEMO	ROLE ³	EVENT	DATE ⁴
(Zagreb ² 11.I 1914 ⁴)						
Moj dragi Vojko¹, dobićeš jedan neočekivan list; od zadnje karte ja sam preturio mnogo: angina, koju ne liječi dovoljno komplikovala se i morao sam biti oprezan; i bio sam operiran isti dan⁴ kad i g.³ A. G. Matoš¹ obojica na grlu pa (da) mi je ko proteko ne bih vjerovao. Sad je stvar prilično, već mogu i da jedem i izlazim, pa su izgledi da ću u utomik krenuti k tebi: Zato ti i ne šaljem novca, a ako bih slučajno izostao — ne daj bože da bude gore — ja bih ti poslao. Pored najbolje volje i dostatna vremena (prvi put u Zagrebu²) ne mogu ni dan⁴ da ti pišem, jer sam umoran i slab. Još mi poigravaju liječnikove škarice po zubima i imam odvratnost za metal kao onaj koga klaše i ne doklaše, tako da mi već nekoliko petokrunka oftkariše g.³ lekaru³. Antimetulum. ide¹ metal. Jedna operacija je užasno ružna stvar, a još je ružnije da se za tu rugobu i užasno plaća. Tako me je boljelo i tako sam oslabio da mislim da sam kažnjen već za SVE grijeha i da su ispunjene sve želje mojih najgorih neprijatelja. I miran sam. Ako se to može biti. Da: i ako se smije biti. Do viđenja Tvoj Ivan¹ (Voimiru Durbešiću¹) BEČ²						

Figure 4. Visualization of the NER result for one of Andrić's letters.

Further, by executing the SPARQL queries outlined in Listings 1, 2, and 3 using the NER results as input variables, we gained valuable insights into the societal and political context surrounding Andrić. This facilitated the

establishment of correlations between these insights and his literary creativity. The details of these correlations are further elucidated in the following subsections.

5.1 The Influence of Historical Events on Andrić Work

In literary analysis, the particular year in which a writer produces a work often serves as a backdrop, absorbing the narrative of that socio-political and cultural time. This temporal context is not merely a passive backdrop but an active element that can influence the writer's themes, narrative structure, and character development. For instance, a writer writing during a war or political turmoil may infuse their work with themes of conflict, societal upheaval, or moral ambiguity. The historical context thus acts as a lens through which the writer views and interprets the world.

In our case, applying the query from Listing 1 for the timeframe 1914–1945 and where variable `comments` contains the term *Yugoslavia*, we obtain the results presented in Figure 5. The historical events in this timeframe had profound implications for the Balkan region which was a focal point of military conflict, and many battles significantly affected the geopolitical landscape. These events not only altered the course of history in the Balkans but also left a memorable mark on the collective psyche of the people, including Andrić. During the mentioned timeframe, Andrić was jailed for his anti-Austrian activities at the war. This experience influenced his literary perspective, as seen in his books *Nemiri* and *Ex Ponto* which reflect his imprisonment. The profound impact of these historical events provided Andrić with a rich tapestry of experiences and insights that shaped his literary voice and topics. The themes of loss, identity, and the quest for meaning amidst war chaos could be prevalent in his later works and could have their roots in the experiences and observations of this period. The backdrop of Andrić's stories, mainly set in the 19th century but also spanning earlier and later periods, is crucial to understanding the complexity of the Balkans as they approached the Second World War. His stories are set in Bosnia, a region he portrays as a cultural and geopolitical borderland between East and West. This characterization is particularly resonant in the context of the impending World War II, as the Balkans were once again becoming a focal point of significant power interests and conflicts. Andrić's narratives, soaked in themes of irrationalism and emotional eruptions, reflect the turbulent history of the Balkans, where religious and ethnic identities often intersected and clashed. This historical context is crucial to understanding, as the region's complex ethnic and religious tapestry played a significant role in the time's broader political and military dynamics. The year 1925 can be considered the year when Andrić mainly or exclusively adopted the *ekavica*. This linguistic choice is noteworthy

as it reflects a deeper cultural and political alignment within Yugoslav, particularly in a region marked by diverse languages and dialects.

"Bombing of Belgrade (1944)"@en	"The Allies carried out an aerial bombing campaign against the Axis in Belgrade during the Allied bombing of Yugoslavia in World War II. The air strikes lasted from 16 April 1944 to 6 September 1944. Belgrade was bombed eleven times by the Anglo-American air force. Infrastructure in Belgrade was bombed three times in April, twice in May, once in June and July and four times in September 1944. The heaviest casualties were recorded during the April bombing on 16 April and 17 April 1944, which coincided with the first and second days of Orthodox Easter that year. The main unit in this action was the American Fifteenth Air Force, with a base in Foggia in the south of Italy. 600 bombers took part, dropping carpet bombs from 3,000 - 5,000 metres. There was no anti-aircraft defense. Several people"@en
"Uprising in Serbia (1941)"@en	"The uprising in Serbia was initiated in July 1941 by the Communist Party of Yugoslavia against the German occupation forces and their Serbian quisling auxiliaries in the Territory of the Military Commander in Serbia. At first the Yugoslav Partisans had mounted diversions and conducted sabotage and had attacked representatives of Aćimović's quisling administration. In late August some Chetniks joined the uprising and liberated Lončica. The uprising soon reached mass proportions. Partisans and Chetniks captured towns that weak German garrisons had abandoned. The armed uprising soon engulfed great parts of the occupied territory. The largest liberated territory in occupied Europe was created by the Partisans in western Serbia, and was known as the Republic of Užice. Rebels shared power on the"@en
"Battle of Višegrad"@en	"The Battle of Višegrad was the battle between Chetnik forces and Axis, and part of an Chetnik offensive in Eastern Bosnia in autumn of 1943, in Axis occupied Yugoslavia during World War II. The Chetnik forces of 2,500 captured Višegrad, destroyed big railway bridge across river Drina and continued their advances toward Rogatica and Sokolac. The German and Ustaše garrison in Višegrad and garrison that protected the bridge of total 1,100 soldiers had 350 dead and 400 wounded. The Chetniks had 21 dead and 30 wounded. In subsequent battle for Rogatica waged ten days later, the Chetniks captured Rogatica and killed more than 200 Axis soldiers."@en

Figure 5. The portion of the result of SPARQL query from Listing 1, for the timeframe 1914–1945, and location *Yugoslavia*. The complete result is available at the link.

5.2 Effects of Personal Relationships on Andrić Writing

Exploring personal relationships and their influence on a writer's work is a nuanced and multifaceted endeavour. The interaction between a writer and a specific individual can profoundly impact the oeuvre's thematic content, emotional tone, and stylistic characteristics. This influence is particularly discernible when the individual in question plays a significant role in the writer's personal or professional life. Analysing these relationships through a biographical lens allows a deeper understanding of the writer's artistic choices. It provides a richer interpretation of their literary creations, unveiling the interplay between personal experiences and creative expression. Applying the query for Listing 2, for input variable *Tugomir Alaupović*, we get results as in Figure 6. This helps us to understand to whom Andrić is writing. For the uninitiated, explaining how Tugomir Alaupović of the same social milieu influenced Andrić's work is ungrateful because DBpedia does not mention their direct connection. However, excellent connoisseurs of Andrić can assume whether and how his gymnasium professor Alaupović influenced his work.

"Tugomir Marko Alaupović (Dolac, Travnik, 18 August 1870 – Zagreb, 9 April 1958) was a Yugoslav professor at "First Grammar School ,Sarajevo, poet, storyteller and politician. In addition to his rich political biography, he was also minister of religion in the government of the Kingdom of Serbs, Croats and Slovenes. He has written several literary works that have been translated into French, German, Czech and Italian. He was one of the initiators of the Croatian Society for the "Setting up of Children in Crafts and Trade" in Sarajevo and later initiated the change of the society name to Napredak. He was a member of the Main Board of the Serbian St. Sava Society in Belgrade. On 16 January 1934, after a serious operation, in a letter to Tihomir Djordjevic, a prominent Serbian ethnologist, he said: "Unfortunately, my hopes have not been fulfilled and I will have to stay long or maybe even definitely in Zagreb. It hurts and I'm sorry that for these reasons, I have to resign as a member of the Main Board of the St. Sava Society. But rest assured that for the rest of my life, I will remain faithful to that beautiful and noble saying: 'Everyone is my dear brother be he any religion'".">@en

Figure 6. The results of the query from Listing 2 for variable *Tugomir Alaupović*.

The absence of information about specific persons such as *Vera Stojić*, *Vojmir Durbešić*, *Zdenka Marković*, and others in the DBpedia, can be attributed to several factors inherent in the structure and scope of this resource. DBpedia is primarily constructed by extracting structured Wikipedia content, so the inclusion of individuals in DBpedia is contingent on their notability and representation within Wikipedia. Additionally, personal names such as *Vučo*, *Mileva*, *Anto*, and *Omer* discovered with NER, provide limited utility, as their identities cannot be definitively determined without consulting the corresponding footnotes in the book (Andrić 2000).

5.3 Influence of Contemporary Literary Works on Andrić Writing

The early 20th century saw significant literary innovation. Andrić was part of this broader literary context and the international literary milieu may have influenced on his narrative technique or thematic choices. Andrić, whose work often reflects on the social and political dynamics of the Balkans, might have found inspiration in how other authors dealt with similar themes in their respective cultural and historical contexts.

Executing the SPARQL query presented in Listing 3 for the time frame 1914–1945 for the output labelled in *English*, yielded the results depicted in Figure 7. The result shows a wide range of literary genres, from mystery and adventure novels to essays and historical fiction. During this period, literature was witnessing various movements and trends globally. Many of these works reflect the societal and historical contexts of their time. For instance, *Native Son* deals with racial issues in America, *The Gone with the Wind* deals with the American Civil War, the Reconstruction Era, and personal destinies, and *The Road to Wigan Pier* by George Orwell deals with the political phenomena at that time. Also, the query was tested for other languages, such as *German*, *French* and *Italian*, where there was no difference in the output result. For the *Serbian* and *Croatian* languages, there were no results. Andrić, whose work is deeply rooted in the historical and cultural milieu of the Balkans, might have been influenced by how world authors integrated contemporary social and political issues into

their narratives. The exposure to such a wide array of literary works might have influenced Andrić in terms of thematic choices, narrative techniques, and the portrayal of historical and cultural realities.

"Gone with the Wind is a novel by American writer Margaret Mitchell, first published in 1936. The story is set in Clayton County and Atlanta, both in Georgia, during the American Civil War and Reconstruction Era. It depicts the struggles of young Scarlett O'Hara, the spoiled daughter of a well-to-do plantation owner, who must use every means at her disposal to claw her way out of poverty following Sherman's destructive "March to the Sea". This historical novel features a coming-of-age story, with the title taken from the poem "Non Sum Qualis eram Bonae Sub Regno Cynarae", written by Ernest Dowson."@en

"Native Son (1940) is a novel written by the American author Richard Wright. It tells the story of 20-year-old Bigger Thomas, a black youth living in utter poverty in a poor area on Chicago's South Side in the 1930s. While not apologizing for Bigger's crimes, Wright portrays a systemic causation behind them. Bigger's lawyer, Boris Max, makes the case that there is no escape from this destiny for his client or any other black American since they are the necessary product of the society that formed them and told them since birth who exactly they were supposed to be."@en

"The Road to Wigan Pier is a book by the English writer George Orwell, first published in 1937. The first half of this work documents his sociological investigations of the bleak living conditions among the working class in Lancashire and Yorkshire in the industrial north of England before World War II. The second half is a long essay on his middle-class upbringing, and the development of his political conscience, questioning British attitudes towards socialism. Orwell states plainly that he himself is in favour of socialism, but feels it necessary to point out reasons why many people who would benefit from socialism, and should logically support it, are in practice likely to be strong opponents."@en

Figure 7. The portion of the result of SPARQL query from Listing 3, for the timeframe 1914–1945. The complete result is available at the link.

6. Limitations

Our methodology has uncovered potential influences on Andrić's creativity work by integrating key technologies – OCR, NER, and SPARQL queries. However, this methodology is subject to some limitations.

Applying OCR to digitize letters while effectively making the text available for further computer analysis faced inherent challenges. Due to the variability in text quality and font styles in the original documents, some text parts are converted incorrectly, resulting in gaps or errors in the digitized output.

Using NER was instrumental in extracting key entities and effectively identifying person names, dates, and locations, providing a structured approach to unstructured text. However, the limitations of NER became apparent in instances of ambiguous or context-specific words, which occasionally led to misclassifications or omissions of certain entities. This could affect the accuracy of the data used for subsequent analyses.

Using SPARQL queries to extract contextual information from DBpedia introduced a novel dimension of analysis of Andrić's letters. However, it also revealed the dependency of such an approach on the availability and comprehensiveness of data within DBpedia. The queries sometimes yielded limited results, especially for less documented or obscure entities. This limitation is particularly relevant when dealing with personal names that may not be extensively or at all documented in DBpedia.

7. Conclusions and Future work

The methodology presented in this paper applies OCR, NER, and SPARQL for analyzing the personal letters of Ivo Andrić and denotes an interesting method for exploring the literary legacies of Ivo Andrić. By applying NER to Andrić's letters, we can systematically extract entities like persons, dates, locations, and others, thereby uncovering the intricate web of personal, historical, and cultural references that influenced his writing. These extracted entities serve as inputs for predefined SPARQL queries targeted at DBpedia – a rich repository of structured data. In this way we obtain contextual information about the identified entities, providing a deeper understanding of the historical moments, world literary legacies, and figures mentioned or alluded to in Andrić's correspondence. This methodology offers a nuanced view of how Andrić's personal experiences and the socio-political climate of his era may have seeped into his literary works. This approach also exemplifies the potential of DH to augment traditional literary analysis, offering new insights and perspectives that were previously difficult to access or quantify.

In light of the identified limitations, future work could focus on several key steps to enhance the robustness and accuracy of the methodology. Firstly, OCR accuracy can be improved, including machine learning approaches tailored to texts. Secondly, further development in the NER process, such as training custom models on domain-specific datasets or incorporating more sophisticated context analysis, could reduce misclassification and improve entity recognition. Thirdly, expanding the database sources beyond DBpedia to include other SW resources could provide more comprehensive datasets for SPARQL queries, especially for less documented individuals or events. Collaborative efforts with linguists or historians might enrich the analysis, adding expert interpretation and validation layers.

Authors contribution statement

M. Đ. Petrović: Conceptualization, Investigation, Methodology, Validation, Writing; D. Bečejski Vujaklija: Conceptualization, Conceptualization; J. Đorđević: Conceptualization, Methodology, Supervision; J. Mitrović: Validation, Supervision.

References

- Andrić 2000: Andrić, Ivo. *Pisma (1912–1973): privatna pošta*. Novi Sad: Matica srpska.
- Auer et al. 2007: Sören, Auer; Christian, Bizer; Georgi, Kobilarov; Jens, Lehmann; Richard, Cyganiak; Zachary, Ives, Z. *DBpedia: a Nucleus for a Web of Open Data*. In: Aberer, K. et al. (ed.). ISWC ASWC 2007. Lecture Notes in Computer Science, 4825. Berlin – Heidelberg: Springer-Verlag. Pp. 722–735. DOI: https://doi.org/10.1007/978-3-540-76298-0_52. 15. 12. 2024.
- Awel & Abidi 2019: Awel, Muna Ahmed, and Ali Imam, Abidi. Review on Optical Character Recognition. In: *International Research Journal of Engineering and Technology (IRJET)*, 6(6). Pp. 3666–3669.
- Biographical Criticism 2024-www: *Biographical Criticism in Literature & Theory*. Retrieved from: English Studies, <https://english-studies.net/biographical-criticism-in-literature-theory/>. 18. 12. 2024.
- Dekker et al. 2019: Dekker, Niels; Kuhn, Tobias; van Erp, Marieke. Evaluating Named Entity Recognition Tools for Extracting Social Networks from Novels. In: *PeerJ Computer Science*. 5, e189. DOI: 10.7717/peerj-cs.189.
- Hawkesworth 2000: Hawkesworth, Celia. *Ivo Andrić: Bridge Between East and West*. London: Bloomsbury Publishing. <https://www.bloomsbury.com/us/ivo-andric-9781847140890/>. 15. 12. 2024.
- Helmi 2021: Helmi, Laila C.A. Digital humanities: A paradigm for the 21st century. In: *BAU Journal-Society, Culture and Human Behavior*. 2(2), S. 5. DOI: 10.54729/2789-8296.1057. 12. 12. 2024.
- Ikonić Nešić et al. 2022: Ikonić Nešić, Milica; Stanković, Ranka; Schöch, Christof; Škorić, Mihailo. From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back). In: Declerck, T. et al. (ed.). In: *Proceedings of the 8th Workshop on Linked Data in Linguistics*. Marseille: ELRA Pp. 7–16. URL: <https://aclanthology.org/2022.ldl-1.2>. 17. 12. 2024.
- It-Sr-NER 2024-www: *It-Sr-NER*. Retrieved from: <https://ners.jerteh.rs/>. 19. 12. 2024.
- Khurana et al. 2023: Diksha, Khurana; Aditya, Koli; Kiran, Khatter; Sukhdev, Singh. Natural Language Processing: State of the Art, Current Trends and Challenges. In: *Multimedia Tools and Applications*. 82(3). Pp. 3713–3744. DOI: 10.1007/s11042-022-13428-4.
- Krstev et al. 2014: Krstev, Cvetana; Obradović, Ivan; Utvić, Miloš; Vitas, Duško. A System for Named Entity Recognition Based on Local Grammars. In: *Journal of Logic and Computation*. 24(2). Pp. 473–489. DOI: 10.1093/logcom/exs079.

- Labatut & Bost 2019: Labatut, Vincent; Bost, Xavier. Extraction and Analysis of Fictional Character Networks: A Survey. In: *ACM Computing Surveys* (CSUR). 52(5). Pp. 1–40. DOI: 10.1145/3344548.
- Literature in Context 2024-www: *Literature in Context*. Cambridge: University Press. Retrieved from: <https://www.cambridge.org/core/series/literature-in-context/7F33BBA656571804864A3787944D845C>. 17. 12. 2024.
- McGillivray et al. 2020: McGillivray, Barbara; Poibeau, Thierry; Pablo, Ruiz Fabo. Digital Humanities and Natural Language Processing: „Je t’aime... Moi non plus“. In: *Digital Humanities Quarterly*. 14(2). DOI: 10.17863/CAM.55816.
- Mihailovich & Andrić 1966: Mihailovich, D. Vasa; Andrić, Ivo. The Basic World View in the Short Stories of Ivo Andrić. In: *Slavic and East European Journal*. 10. Pp. 173. <https://api.semanticscholar.org/CorpusID:172077383>. 15. 12. 2024.
- Nasar et al. 2021: Nasar, Zara; Jaffry, Syed Waqar; Malik, Muhammad Kamran. Named Entity Recognition and Relation Extraction: State-of-the-art. In: *ACM Computing Surveys* (CSUR). 54(1). Pp. 1–39. DOI: 10.1145/3445965.
- Pan 2009: Pan, Jeff Z. Resource Description Framework. In: Staab, S., Studer, R. (ed.). *Handbook on Ontologies*. Berlin – Heidelberg: Springer. Pp. 71–90. DOI: 10.1007/978–3–540-92673–3_3.
- Polleres 2014: Polleres, Axel. SPARQL. In: Alhajj, R., Rokne, J. (ed.). *Encyclopedia of Social Network Analysis and Mining*. New York: Springer. Pp. 1960–1966. DOI: 10.1007/978–1–4614–6170-8_124.
- Retsinas et al. 2022: George, Retsinas; Giorgos, Sfikas; Basilis, Gatos; Christophoros, Nikou. Best Practices for a Handwritten Text Recognition System. In: *International Workshop on Document Analysis Systems, Springer*. Pp. 247–259. DOI: 10.1007/978–3–031–06555–2_17.
- ScanDocFlow 2024-www: *ScanDocFlow: OCR and data extraction API*. <https://scandocflow.com/>. 14. 12. 2024.
- Tambuscio and Andrews 2021: Tambuscio, Marcella; Andrews, Tara Lee. Geolocation and Named Entity Recognition in Ancient Texts: A Case Study about Ghewond’s Armenian History. In: *Computational Humanities Research*. Amsterdam: CEUR Workshop Proceedings. Pp. 136–148.
- Tedeschi et al. 2021: Tedeschi, Simone; Conia, Simone; Cecconi, Francesco; Navigli, Roberto. Named Entity Recognition for Entity Linking: What Works and What’s Next. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Pp. 2584–2596. DOI: 10.18653/v1/2021.findings-emnlp.220.
- Tošović 2012: Tošović, Branko (Hg./ur.). *Ivo Andrić – Literat und Diplomat im Schatten zweier Weltkriege (1925–1941) – Ivo Andrić – književnik i diplomata u sjeni dvaju svjetskih ratova (1925–1941)*. Graz – Beograd: Institut für Slavistik der Karl-Franzens-Universität Graz – Beogradska knjiga. 606 S./s.

[Andrić-Initiative: Ivo Andrić im europäischen Kontext – Ivo Andrić u evropskom kontekstu, tom 5]

Tošović 2014: Tošović, Branko (Hg./ur.). *Andrićeva HRONIKA. Andrićs CHRONIK*. Graz – Banjaluka – Beograd: Institut für Slawistik der Karl-Franzens-Universität Graz – Narodna i univerzitetska biblioteka Republike Srpske – Svet knjige – nmlibris. 1077 S./s. [Andrić-Initiative: Ivo Andrić im europäischen Kontext – Ivo Andrić u evropskom kontekstu, tom 7]

Tošović 2015: Tošović, Branko (Hg./ur.). Andrić ispred i iza AVLJE. In: Tošović, Branko (Hg./ur.). *Andrićeva AVLJA / Andrićs HOF*. Graz – Banjaluka – Beograd: Institut für Slawistik der Karl-Franzens-Universität Graz – Narodna i univerzitetska biblioteka Republike Srpske – Svet knjige – nmlibris. S. 22–66. [Andrić-Initiative: Ivo Andrić im europäischen Kontext – Ivo Andrić u evropskom kontekstu, tom 8]

van Dalen-Oskam et al. 2014: van Dalen-Oskam, Karina; de Does, Jesse; Marx, Maarten; Sijaranamual, Isaac; Depuydt, Katrien; Verheij, Boukje; Geirnaert, Valentijn. Named Entity Recognition and Resolution for Literary Studies. In: *Computational Linguistics in the Netherlands Journal*. 4. Pp. 121–136. URL: <https://clinjournal.org/clinj/article/view/45>.

Marija Đokić Petrović (Graz)
 Dragana Bečejski Vujaklija (Belgrade)
 Jasmina Đorđević (Niš)
 Jelena Mitrović (Passau)

Unveiling Literary Legacies: Integrating Named Entity Recognition and SPARQL to Analyse Andrić's Letters

This paper explores the potential of Natural Language Processing for analysing the letters of Ivo Andrić, a notable Yugoslavian writer and Nobel Prize laureate. By utilizing Named Entity Recognition (NER), the study enabled the systematic extraction of significant entities such as people, dates, and locations from Andrić's correspondence. These entities were then used as input for SPARQL queries to retrieve contextual information from the DBpedia knowledge base. The focus of these queries includes historical events, profiles of relevant individuals, and notable literary works within specific timeframes. By correlating the query results with broader historical, cultural, and personal contexts, the research uncovers the influences that may have shaped Andrić's literary output. Overall, the study findings underscore the value of applying advanced computational techniques in literary scholarship in alignment with contemporary trends in the Digital Humanities field. This research could pave the way for similar analyses of the works of other literary figures, enriching our comprehension of the interplay between personal experience and creative expression.

Marija Đokić Petrović (Graz)
Dragana Bečejski Vujaklija (Belgrade)
Jasmina Đorđević (Niš)
Jelena Mitrović (Passau)

Razotkrivanje književnih zaostavština: integrisanje prepoznavanje imenskih entiteta i SPARKL za analizu Andrićevih pisama

Ovaj rad istražuje potencijal obrade prirodnog jezika za analizu pisama Ive Andrića. Korišćenjem prepoznavanja imenovanih entiteta (NER), studija je omogućila sistematsko izdvajanje značajnih entiteta kao što su ljudi, datumi i lokacije iz Andrićeve prepiske. Ovi entiteti zatim su korišćeni kao ulaz za SPARQL upite za preuzimanje kontekstualnih informacija iz Dbpedia baze znanja. Fokus tih upita obuhvata istorijske događaje, profile relevantnih pojedinaca i značajna književna dela u određenim vremenskim okvirima. Dovođenjem u vezu rezultata upita sa širim istorijskim, kulturnim i ličnim kontekstima istraživanje otkriva uticaje koji su mogli oblikovati Andrićev književni opus. Generalno, nalazi studije naglašavaju vrednost primene naprednih računarskih tehnika u književnoj nauci u skladu sa savremenim trendovima u oblasti digitalnih humanističkih nauka. Ovo istraživanje moglo bi da otvori put sličnim analizama dela drugih književnih ličnosti, obogaćujući naše razumevanje međugre između ličnog iskustva i kreativnog izražavanja.

Marija Đokić Petrović
FH Joanneum, University of Applied
Sciences
Alte Poststraße 147,
8020 Graz, Austria
School of Computing, University Union
Dragana Bečejski Vujaklija
Serbian Society of Informatics
Kneza Miloša 9,
11000 Belgrade
Serbia
dragana.becejski-vujaklija@fon.bg.ac.rs

Jelena Mitrović
Faculty of Computer Science and Mathe-
matics, Universität Passau
Innstraße 41, 94032 Passau
Germany

Kneza Mihaila 6,
11000 Belgrade
Serbia
+43 667 784 68 68
office@marijadjokicpetrovic.com
<https://marijadjokicpetrovic.com/>
Jasmina Đorđević
Faculty of Science and Mathematics,
University of Niš, Višegradska 33
18000 Niš, Serbia
+381 64 13 09 353
jasmina.djordjevic@pmf.edu.rs
<http://jasminadjordjevic.com/>
The Institute for Artificial Intelligence
Research and Development of Serbia
Naučno-tehnološki park Novi Sad
Fruškogorska 1
21000 Novi Sad, Serbia
+49 851 509 3353
jelena.mitrovic@uni-passau.de
<https://ca-roll.github.io/>