Scientific paper

Analysis of Family Relationships of the Characters in the SrpELTeC Corpus Based on Semantic Web Techniques and Textometry

UDC 811.163.41’322.2
DOI 10.18485/infotheca.2025.25.1_2.1

Milica Ikonić Nešić
milica.ikonic.nesic@fil.bg.ac.rs
ORCID: 0000-0002-0835-8889

University of Belgrade
Faculty of Philology
Belgrade, Serbia

ABSTRACT: This paper presents an analysis of family relationships of characters in the SrpELTeC collection of Serbian novels published between 1840 and 1920, using a textometric approach and semantic web techniques. The study explores the appearance of family relationships between literary characters in the corpus in regards to the periods in which the novels were written, as well as to the authors of the novels. Taking into account the overlap between the literary movements of Romanticism and Realism during the aforementioned periods, the analysis examines in which direction family relationships are more prevalent, and whether male characters are more represented than female characters in the collection. Visualizations of the relationships between characters are presented through linking the characters with novels in Wikidata and by expressing SPARQL queries, as well as through the network analysis using the Gephi tool.

KEYWORDS: SrpELTeC, family relationships, literary characters, Wikidata, Gephi, SPARQL

PAPER SUBMITTED: 17 Jun 2024
PAPER ACCEPTED: 22 Jun 2024

1 Introduction

The sub-collection of Serbian novels (referred to as SrpELTeC[1]) (Krstev and Stanković 2022, 2020; Stanković et al. 2022; Patras et al. 2020) was developed under the auspices of *the COST Action CA16204 Distant Reading for European Literary History*. One of the main objectives of this Action was to form a multilingual corpus (named the *European Literary Text Collection - ELTeC*), consisting of 100 novels[2], first published between 1840 and 1920, in twelve European languages, which comprises its language sub-collections[3] (Odebrecht, Burnard, and Schöch 2021). The Serbian sub-collection of novels contains 5,886,528 tokens and 4,769,262 words. All words are annotated with part-of-speech tags and lemmas (Stanković et al. 2020; Frontini et al. 2020). Also, all words contain information about named entities (Stanković et al. 2019), thus enabling the application of advanced text analysis methods in accordance with the distant reading paradigm.

Several studies have already been conducted on the SrpELTeC corpus through textometric analyses of noun and verb occurrences (Jaćimović 2019), stylometric analysis aimed at identifying the author based on the writing style (Škorić et al. 2022), as well as deep learning and model training for recognizing named entities (Šandrih Todorović et al. 2021; Ikonić Nešić et al. 2024). The occurrence of the rhetorical figures of comparison in the corpus has been studied through their usage according to various criteria: authors, time periods, an author's gender, size of the novel, and popularity of the novel. Additionally, adjectives, nouns, and markers used in the comparisons found in the SrpELTeC corpus have been analyzed, along with the entities to which they apply (Krstev 2021). Vitas (2020) utilizes 104 novels from the SrpELTeC corpus in his study and analyzes various aspects of the language regarding food and dietary habits in the Serbian language. Furthermore, he analyzes the position of women and cultural practices within the Serbian population during the second half of the 19th and the early 20th centuries using Unitex[4]. A system of electronic morphological dictionaries for the Serbian language facilitated the research since various types of food, meals and

[1] SrpELTeC, https://github.com/COST-ELTeC/ELTeC-srp
SrpELTeC, https://jerteh.github.io/srpELTeC/

[2] A narrative text that exceeds 10,000 words, in addition to a novel, can also be classified as a novella or a long short story.

[3] https://distantreading.github.io/ELTeC/

[4] Multilingual Corpus Processing Suite Unitex/GramLab

drinks are described in it in detail. However, until now, the analysis of characters in the novels and their mutual relationships has remained unexplored.

The subject of this paper is the extraction of characters and analysis of family relationships in the corpus of Serbian novels, with the aim of determining the patterns of their occurrence in relation to the authorship and the period in which the works were written. The study will employ advanced analyses of the sub-collection of Serbian novels, based on the processing of metadata, text structure, and content, as well as on semantic web techniques. One limitation of this approach is the need for contextual analysis, as certain nouns can be polysemous. For instance, the noun *žena* (woman/female/wife) has multiple meanings. In the SrpELTeC corpus, the noun *žena* appears 5,335 times, while *supruga* (wife) appears 132 times, suggesting that *žena* is often used not only to denote a female person, but also in the sense of a spouse. One method of addressing this issue could involve analyzing adjectives used as attributes of the nouns *muž* (husband) and *žena* (wife), as demonstrated in Panić's study (2018), where adjectives in the SrpKor2013 corpus (Utvić 2013) are classified into 10 groups. The most relevant group for this study is the one referring to social and personal status, such as marital status (*former*, *future*, *married*, wed, etc.) and emotional status (*beloved*, *loved*, *cherished* etc.), however, this is not the focus of this research.

The goal of this paper is to understand the relationships between characters in the SrpELTeC corpus novels through the concepts of family, romantic, or social relations, using textometric and distant reading techniques, and semantic web tools. Specifically, the paper aims to answer the following questions:

1. Can we observe similarities or differences in the portrayal of family relationships by the authors of the selected novels, through the analysis of words such as: *majka* (mother), *otac* (father), *sin* (son), *ćerka* (daughter), *žena* (wife), *muž* (husband), *brat* (brother), *sestra* (sister), and their synonyms, and do these portrayals depend on the authors themselves?
2. Taking into account that the novels were written between 1840 and 1920, during the period of overlapping literary movements of Romanticism and Realism, and knowing that one of the main themes of Realism is the concept of family, can we argue that family relationships are more prevalent in Realism than in Romanticism?
3. Gender stereotypes and the under representation of female characters in novels have been documented in the past. Can we claim that the SrpELTeC corpus also shows a greater number of male characters compared to female ones?

A review of previous related research will be presented in Section 2. Section 3 provides a brief overview of the SrpELTeC corpus on which the research was conducted. The methodology, including textometric analysis and linking characters from novels with Wikidata, is presented in Section 4. Results obtained from various methods are discussed in Section 5, while Section 6 is dedicated to the conclusions drawn from the results.

2 Related Research

Historically, written works were primarily studied from the perspective of the plot, which was considered the most important part of the story. However, more modern approaches focus on characters (Bamman, O'Connor, and Smith 2013), viewing them as the driving forces behind narrative progression (Min and Park 2016). (Woloch 2009) examines characters through their position in relation to other elements of the plot (such as the place, time, and other characters), specifically how characters are described within the narrative. The concept of the character system expands this idea to encompass the entire narrative, reflecting the unified space of all the characters within the story. Several researchers have employed this approach in order to better understand and study the methods by which writers construct their narratives.

In addition to characters themselves, researchers have also studied character interactions, which are considered the backbone of narration (Cipresso and Riva 2016; Prado et al. 2016). One approach to analyzing character interactions is through the use of graphs, which enable the representation and study of systems by mapping the interactions between their constituent elements. A character network is a graph that represents the narrative by treating characters as nodes and their interactions as edges. (Moretti 2011) has demonstrated that this approach allows for more formal handling of Woloch's concepts, as Woloch emphasizes that character spaces must be examined together-precisely what graphs, as relational spaces for modeling, enable.

However, there are alternative approaches to analyzing character relationships. (Santos, Mamede, and Baptista 2010) and a team of researchers developed a rule-based method for extracting family relationships from Portuguese narratives. The authors devised 99 rules, including some propagation rules, and combined them with named entity recognition and anaphora resolution systems. (Kokkinakis and Malm 2011) developed an unsupervised approach for extracting interpersonal relationships, including family ties.

Their method relies on the co-occurrence of the names of characters within the same sentences. These co-occurrences are marked as relationships by using context similarity and online lists of common relationships (Borin et al. 2009; Borin and Forsberg 2010). The authors tested their method on three volumes of 19th-century Swedish fiction. The extraction of family relationships from literary narratives has further evolved, including techniques like speech attribution combined with vocative detection—explicit forms of addressing used by characters in the novel (Makazhanov, Barbosa, and Kondrak 2014). The authors noted that certain vocatives indicated familial relationships between speakers, and the extracted relationships were propagated using a set of rules.

In recent studies, not only family relationships, but also romantic relationships and the role of women in the early 20th century have been analyzed. Some of these studies focus on main characters by extracting and analyzing the narratives around them. For example, Qizi Sayitqulova (2021) analyzed the novel *Kecha va Kunduz* (Night and Day) (*Kecha va Kunduz*) by Uzbek author A. Chulpan, while Firat (2018) used content analysis techniques to examine specific nouns such as grandmother, grandfather, love, and play to study relationships between grandparents and grandchildren. The relationship between sisters in 19th-century Hungarian novels (Kucserka 2020) was analyzed by examining the narrative dynamics between sisters.

3 The Serbian Novel Sub-Collection (SrpELTeC)

The process of creating the Serbian novel sub-collection (SrpELTeC) as part of the COST Action project (CA16204) took place from 2017 to 2021. The SrpELTeC[5] collection contains 100 novels, while the extended collection, SrpELTeC-ext[6] (SrpELTeC-extended), includes an additional 20 novels. All works included in any language sub-collection of the ELTeC corpus meet the same pre-established criteria: (1) only narrative prose (novels, novellas, or longer short stories) with a minimum length of 10,000 words[7]; (2) the first edition of the work must be published between 1840 and 1920, inclusive; (3) the work must be originally written in Serbian (or the respective language of the sub-collection), with translations being excluded; (4) the work must

[5] SrpELTeC, https://github.com/COST-ELTeC/ELTeC-srp

[6] SrpELTeC-extended, https://github.com/COST-ELTeC/ELTeC-srp-ext

[7] It is implied that the words are counted automatically, for example, as done by the MS Word program.

be published in Europe no more than ten years after its first edition; and (5) preference is given to works published as books during the designated period, rather than those serialized in periodicals.

To ensure diversity among the selected texts and enable comparative analysis of the sub-collection, as well as the application of statistical methods for text analysis, additional selection criteria were applied: (1) each sub-collection contains 100 novels; (2) approximately equal representation of male and female authors; (3) balanced inclusion of works with many reprints and those published only once or twice; (4) even coverage of the period 1840–1920; (5) proportional representation of short and long works; (6) ideally, 10 authors should be represented by three works each, while the rest should be represented by one work (Крстев et al. 2023). To ensure an even distribution over the time period from 1840 to 1920, the publication dates of the first editions were divided into four time periods (T1, T2, T3, and T4). The works were also categorized by length into three groups: short (up to 50,000 words), medium (50,000 to 100,000 words), and long (over 100,000 words).

In the Serbian SrpELTeC sub-collection, there are 2 novels from the first period T1: [1840–1859], 18 novels from the second period T2: [1860–1879], and 40 novels from both the third T3: [1880–1899] and fourth T4: [1900–1920] periods. The core collection features works by 66 male and 4 female authors,[8] with Jakov Ignjatović being the most represented author with 5 works (Крстев et al. 2023).

3.1 Named Entity Recognition and Linking to Wikidata

Named Entity Recognition (NER) refers to the identification and classification of named entities in a text. These entities can include personal names, roles, locations, organizations, and numerical expressions such as dates, times, and monetary values. NER plays a crucial role in information extraction from texts. Within the scope of the COST Action, it was agreed that the novels should be annotated with only seven categories of entities: **PERS** (proper names of real or fictional people: first names, surnames, nicknames, or combinations of these), **ROLE** (occupations, titles, and roles of people), **DEMO** (denoting inhabitants of countries, cities, regions), **ORG** (names of organizations, political parties, educational institutions, sports clubs, hospitals, libraries, hotels, museums, taverns, and churches), **LOC** (continents,

[8] SPARQL query for visualizing the number of authors, https://w.wiki/5YSy

countries, regions, populated places, mountains, islands, caves, rivers, celestial bodies, city locations), **WORK** (titles of books, plays, poems, musical works, paintings, sculptures, newspapers), and **EVENT** (names of events, natural disasters, revolutions, battles, wars, demonstrations, concerts, sports events, holidays). These categories were determined to be the most significant for further literary studies (Frontini et al. 2020). The NER system for the Serbian language is based on manually crafted rules that rely on comprehensive lexical resources for Serbian (Krstev et al. 2014).

As a result, all words in the 100 novels from the SrpELTeC collection have been annotated with part-of-speech tags, lemmas, and named entities, according to the recommendations of the COST Action. Additionally, NER annotations have been completed for 8 more novels from the extended collection of 20 novels, bringing the total to 108 novels available for further textometric analysis in this research.

4 Methodology

This section presents the methods used to investigate and answer the research questions. To address the first and the second question, a textometric approach will be employed. The third question, along with the possibilities for visualizing relationships between characters, will be explored through linking characters with novels in Wikidata and formulating appropriate SPARQL queries (Ikonić Nešić, Stanković, and Rujević 2021; Ikonić Nešić et al. 2022).

4.1 Textometry

Textometry is a computational method of textual analysis based on word frequencies (or the counting of other linguistic features), combining statistical processing with data referencing and contextual comparison (Jaćimović 2019). Previous studies have shown that textometric analysis is highly effective and useful across various corpora, ranging from French news articles (Carrive et al. 2021) to medical abstracts (Gledhill et al. 2019) and forensic investigations (Longhi 2021). Jaćimović (2019) demonstrated the significant results of textometric analysis in her work on a sample of 21 novels from the corpus of old Serbian novels. She presented frequency lists of lemmas and specific word types (i.e. nouns and verbs) through various textometric visualizations. Considering all this, and given that 120 novels have been digitized within the extended ELTeC corpus, it is assumed that textometry will also be effective in this research, showcasing both author-specific features and novel-specific characteristics.

One of the primary tools for textometric analysis is TXM[9] (Heiden 2011). TXM is an open-source program used in various research fields within the social and human sciences. Its graphical user interface is based on the Corpus Query Processor (CQP[10]) search engine and the R[11] statistical package.

The first and fundamental corpus method applied in textometric analysis is the creation of frequency lists. Comparing the absolute frequencies of linguistic units (the exact number of occurrences in the corpus) can be useful and provides an initial impression of the contrasts within parts of the corpus. However, to compare frequencies in parts of different sizes, it is necessary to use normalization, i.e. to express frequencies using a common factor—relative frequency.

In TXM, calculating the specificity index based on the hypergeometric distribution will show the probability of a linguistic unit appearing in a particular part of the corpus. TXM also allows for graphical representation of the distribution of specificity for selected units. Specificity index values that are higher (positive values) or lower (negative values) than expected represent over- or under-represented linguistic units. However, results between -2 and +2 (probability $\geq 1\%$) are considered statistically insignificant. This method helps identify significantly frequent (positive keywords) or significantly rare (negative keywords) occurrences of linguistic units in parts of the corpus, in regard to the entire corpus. This is a useful starting point for drawing conclusions about key words, text domains, authorship, and other features.

Additionally, to more precisely describe the change in frequency of certain words in the corpus, TXM provides the option of using progression analysis. By setting up a search pattern expressed in a CQL query, a graph is generated. The graph represents the „cumulative“ frequency curve for word occurrences or search patterns. Normalization is not required since the slope of the graph directly measures and visually demonstrates the relationship between the number of occurrences and the length of the text, which is effectively a normalized value. While the absolute frequencies of words may vary, the slopes can be compared across texts of different lengths.

[9] TXM, https://sourceforge.net/projects/txm/

[10] The IMS Open Corpus Workbench , https://cwb.sourceforge.io/

[11] R: The R Project for Statistical Computing, https://www.r-project.org/

Cluster analysis in the TXM tool for textometric analysis enables the grouping of similar units (e.g., words, lemmas, paragraphs, chapters, or entire documents) based on their shared characteristics. In the context of text analysis, this could mean grouping words that appear in similar contexts or grouping documents that share similar thematic characteristics.

Using the described approaches, along with the visualization of the results, the progression of noun frequency will be displayed for the lemmatized corpus of 108 novels[12] and their synonyms: ***majka* (mother)** (мајка, мати, помајка, родитељка, маћеха, мама, дада, матер, матера, маћа); ***otac* (father)** (отац, тата, бабо, поочим, очух, бабајко, ћаћа, тајо, татко); ***muž* (husband)** (муж, супруг, супружник); ***žena* (wife)** (жена, супруга, супружница, љуба); ***ćerka* (daughter)** (ћерка, кћи, кћер, ћер, кћерка); ***sin* (son)** (син, синак, синчић); ***brat* (brother)** (брат, братић, бата, братац, браца); ***sestra* (sister)** (сестра, дада, сека, сеја, сестрица, секица). The frequency of these nouns will be examined not only in relation to the novels, but also by publication date and the author. By calculating the specificity index of the aforementioned nouns, the results will be presented based on the authors and the gender of the authors.

For cluster analysis, a sub-corpus of texts by authors with more than 120,000 words in the corpus was selected, including: Vladimir Đorđević (155,706), Ivko Ćipiko (121,469), Svetozar Ćorović (143,600), Janko M. Veselinović (225,275), Jelena Dimitrijević (177,103), Jakov Ignjatović (457,434), Dragutin J. Ilić (141,401), Lazar Komarčić (191,034), Jovan Miodragović (147,144), Čedomil Mijatović (133,392), Branislav Nušić (258,006), Svetolik Ranković (228,046), Stevan Sremac (240,862), Kamenko Subotić (192,212), Pera Todorović (415,024), and Milutin Usković (177,127). This analysis will be used to examine how authors can be grouped together based on their depiction of family relationships in their works.

For a deeper analysis of character relationships, using CQL queries within the TXM environment, relationships such as a *majka od* (mother of), *otac od* (father of), *sin od* (son of), etc., will be explored in the form of an ordered triple (X, ρ, Y), where X and Y represent the arguments of the relationship, and ρ denotes the relationship itself, etc.

$$\text{X and Y in relation } \rho \Leftrightarrow \text{a family relationship between } X \text{ and } Y \text{ exists}$$

The search will be used to identify relevant text, from which the information required to extract the following relations can then be obtained, such as:

[12] Serbian ELTeC Corpus TXM Edition (108 NER), https://live.european-language-grid.eu/catalogue/corpus/23621

$$(X,\ \text{majka/otac od},\ Y) \Rightarrow (Y,\ \text{sin/ćerka od},\ X)$$
$$(X,\ \text{sin/ćerka od},\ Y) \Rightarrow (Y,\ \text{majka/otac od},\ X)$$
$$(A,\ \text{majka/otac od},\ B)\ \&\ (B,\ \text{brat/sestra od},\ C) \Rightarrow (A,\ \text{majka/otac od},\ C)$$

For these relationships, it will be required that both X and Y are named entities labeled as <PERS>, which in this case refers to personal names. Thus, in the TXM environment, by issuing CQL queries, relationships will be searched for in the form:

$$(\text{<PERS>X</PERS>}\ \rho?\ \text{<PERS>Y</PERS>})$$

Here, ρ represents the relationship between two persons, such as a *majka od* (mother of), *otac od* (father of), *brat od* (brother of), etc., while ? indicates that one additional token may appear between the relationship and the second person. This approach will then be used to link characters in Wikidata and more precisely display the relationships between them. An example of a CQL query and the relationships extracted from the text using it is shown in Table1.

Table 1: Example of extracting relationships from the text

CQL упит	**екстрахован текст**	**екстрахован однос**
`<pers>[]*</pers> []{0,1} [srlemma = "majka|otac|kći |ćerka|sin|muž |žena"][]{0,1} <pers>[]*</pers>`	Anđe i oca joj Bogosava Stevanovića;	(**Bogosav Stevanović**, otac od, **Anđa**);
	Arslan-Nuri-pašina sina Murad;	(**Arslan-Nuri-paša**, otac od, **Murad**);
	Darinka ćerka Sretena Cvetića;	(**Sreten Cvetić**, otac od, **Darinka**);
	Rosa žena Milutinova;	(**Milutin**, muž od, **Rosa**);
	Velju sina gazda Damnjana Smiljanića;	(**Damnjan Smiljanić**, otac od, **Velja**);

4.2 Linking Characters from Novels to Wikidata

Wikidata[13] is an open knowledge base where users can create new items and edit existing ones. The process of entering the SrpELTeC collection novels into Wikidata, along with the first, digital, printed, or ELTeC edition of each novel, was automated through the synergy of two tools: *OpenRefine*[14]

[13] Wikidata, https://www.wikidata.org

[14] OpenRefine, https://openrefine.org/

and *QuickStatements*.[15] However, the entry of characters and settings was done manually, as character data could not be extracted from metadata. Instead, based on the frequency of named entities of the PERS (persons) and LOC (locations) classes, a selection of characters and settings was made based on their occurrence in the novels—those with the highest frequency of appearances were chosen (Ikonić Nešić, Stanković, and Rujević 2021; Ikonić Nešić et al. 2022).

Using SPARQL queries, it can be determined that there are currently 978 characters from novels entered into Wikidata (https://w.wiki/5Yuz), with the largest number of entries from the novel *Đurađ Branković: istorijski roman*. Wikidata contains 100 novels from the core SrpELTeC collection and 20 novels from the expanded SrpELTeC-extended collection (https://w.wiki/5Ydd). A full list of all novels in these two collections is provided in the works by (Krstev and Stanković 2022).

All created characters are linked to novels in Wikidata using the characters (`P674`) property. The search capabilities within the TXM environment allowed for the extraction of information about the characters and the determination of their relationships with other characters by issuing CQL queries as shown in Table 1. Afterward, each character was linked to other characters in Wikidata using the properties *father* (`P22`), *mother* (`P25`), *spouse* (`P26`), and *child* (`P40`). An example of the entry for the character *Omer Vidajić* (Q109613221) from the novel *Omer Čelebija: pripovijetka iz života srpskoga naroda*, and his connections with other characters is shown in Figure 1, where it can be seen that his **father** is Hamza Vidajić (`Q109654569`), his **mother** is Melećа (`Q109654771`), and his **spouse** is Pava Tešnjak (`Q109659439`). Based on these entries, it is possible to query the relationships between characters using SPARQL.

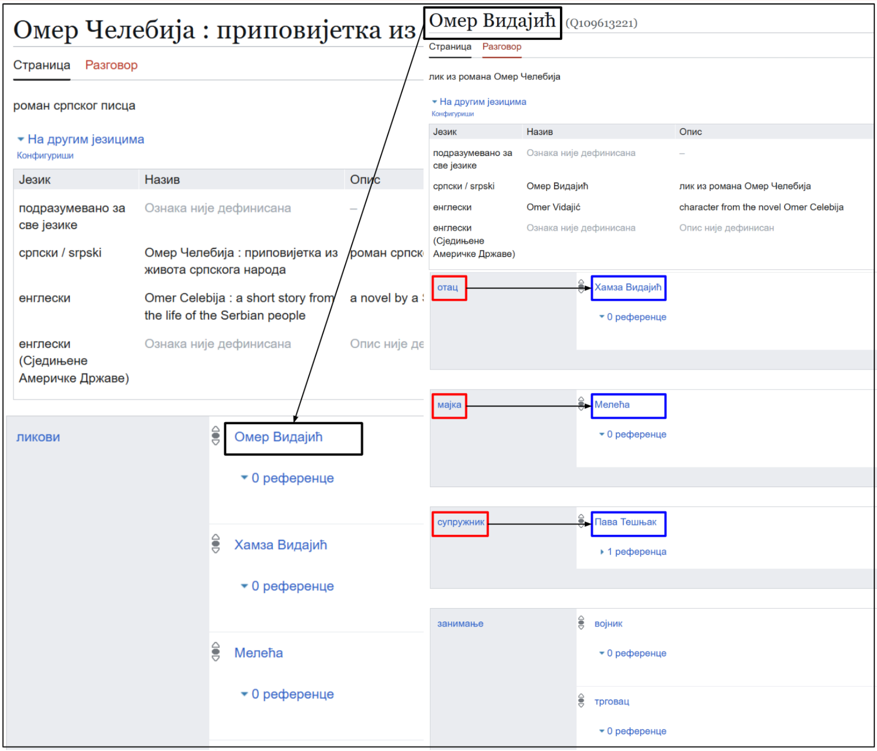
4.3 Network Analysis of Characters

After manually entering the main characters into Wikidata, the tool INCEpTION was used to explore the potential of network analysis for uncovering character relationships in novels. INCEpTION[16] connects entities in the text with corresponding items in the open knowledge base Wikidata. Following this, the open-source tool Gephi[17] (Bastian, Heymann, and Jacomy 2009) was employed for the visualization and network analysis of characters, using

[15] QuickStatements, https://quickstatements.toolforge.org/#/batch

[16] INCEpTION, https://inception-project.github.io

[17] Gephi, https://gephi.org/



Омер Челебија : приповијетка из

Страница Разговор

роман српског писца

▾ На другим језицима
Конфигуриши

Језик	Назив	Опис
подразумевано за све језике	Ознака није дефинисана	–
српски / srpski	Омер Челебија : приповијетка из живота српскога народа	роман српск
енглески	Omer Celebija : a short story from the life of the Serbian people	a novel by a
енглески (Сједињене Америчке Државе)	Ознака није дефинисана	Опис није де

ликови

Омер Видајић
▾ 0 референце

Хамза Видајић
▾ 0 референце

Мелећа
▾ 0 референце

Омер Видајић (Q109613221)

Страница Разговор

лик из романа Омер Челебија

▾ На другим језицима
Конфигуриши

Језик	Назив	Опис
подразумевано за све језике	Ознака није дефинисана	–
српски / srpski	Омер Видајић	лик из романа Омер Челебија
енглески	Omer Vidajić	character from the novel Omer Celebija
енглески (Сједињене Америчке Државе)	Ознака није дефинисана	Опис није дефинисан

отац → Хамза Видајић
▾ 0 референце

мајка → Мелећа
▾ 0 референце

супружник → Пава Тешњак
▸ 1 референца

занимање
војник
▾ 0 референце
трговац
▾ 0 референце

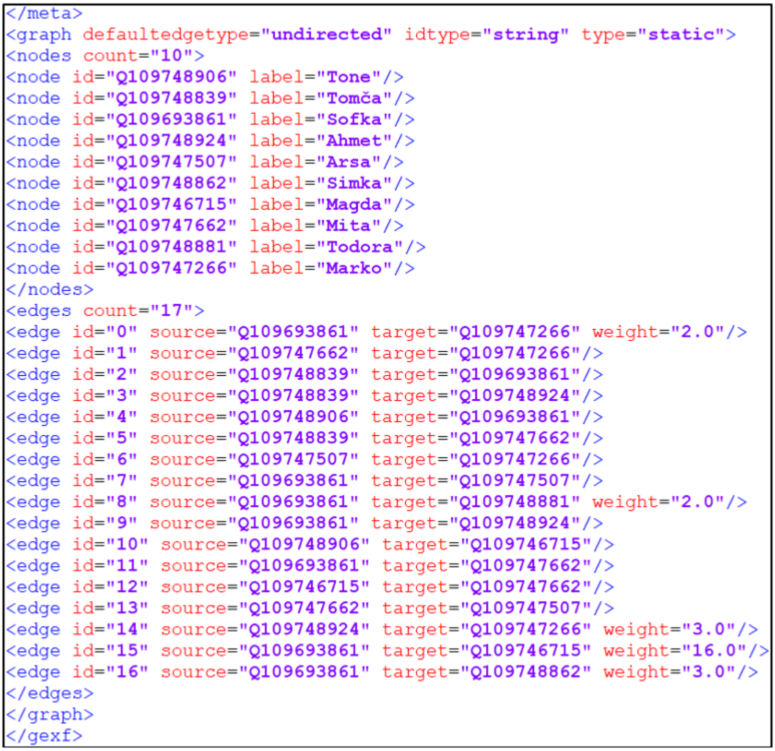
Figure 1: Example of character relationships from the novel *Omer Čelebija: Pripovijetka iz života srpskoga naroda* (Omer Celebija : a short story from the life of the Serbian people).

a passage from the novel *Nečista krv* (Impure Blood) by Borisav Stanković. INCEpTION is a web environment that enables interactive text annotation and entity linking with external knowledge bases, including Wikidata. The annotation process involves identifying entity mentions in the text and linking them to Wikidata items. Linking the text to an item (class or instance) starts by selecting the entity span in the text, followed by searching for the appropriate item using an identifier to automatically search and connect to Wikidata items (Ikonić Nešić, Stanković, and Rujević 2021; Ikonić Nešić et al. 2022). Currently, 13 characters from the novel *Nečista krv* (Impure Blood) are listed in Wikidata (https://w.wiki/ApXd). After the characters were linked to Wikidata items, the annotated sections of the novel were exported as tab-separated value (*tsv*) files, according to the WebAnno TSV 3.3 format.[18]

The next step was preparing the input file for the Gephi tool. In the example of the analyzed network, the Force Atlas layout algorithm was used in the Gephi environment, along with Network Diameter and Modularity statistics. These statistics provide insights into the network structure, which will be presented in Section 5 in the analysis of the research results, offering a better understanding of the dynamics between characters. Force Atlas is a suitable choice for the character network analysis in novels, as it provides visualizations that reveal the structure and dynamics among characters. In novels, characters often form groups or clusters based on their interactions. For example, characters who frequently communicate or appear in the same scene can form clusters. Force Atlas positions characters with relationships closer to each other in the network, making it easier to identify these groups. The algorithm also highlights characters with intense interactions, such as main protagonists or significant supporting characters. Central characters become prominent, while less significant characters, who do not form many connections, tend to occupy the periphery of the network. Given the number of nodes (fewer than 100), Force Atlas was deemed appropriate. Force Atlas 2 is recommended for larger node counts. Furthermore, it allows for flexible adjustments in order to meet specific analysis needs, such as focusing on particular characters, scenes, or themes (Moretti 2011).

For the network analysis, a text sample from novel *Нечиста крв* (Impure blood) containing 3,912 sentences is used, with 1,195 named entities of the PERS class, where Sofka is mentioned in approximately 50% of the cases. The main characteristics of the network analyzed from this passage are: *number of nodes* 10, *number of edges* 17 (each edge represents the appearance of characters in the same sentence), *graph type* `undirected`, for ranking, the **`weighted degree`** module is used, where „weighted“ indicates how many times two characters are mentioned in the same sentence. In this research, an interface is created that generates a *gexf* file from the input *tsv* file, which is then used as input for Gephi, as shown in Figure 2. It can be observed that Sofka (`Q109693861`) and Magda (`Q109746715`) have the most co-occurrences in the same sentences, with 16 occurrences (Figure 2).

[18] WebAnno TSV 3.3, https://inception-project.github.io/releases/26.8/docs/user-guide.html#sect_formats_webannotsv3

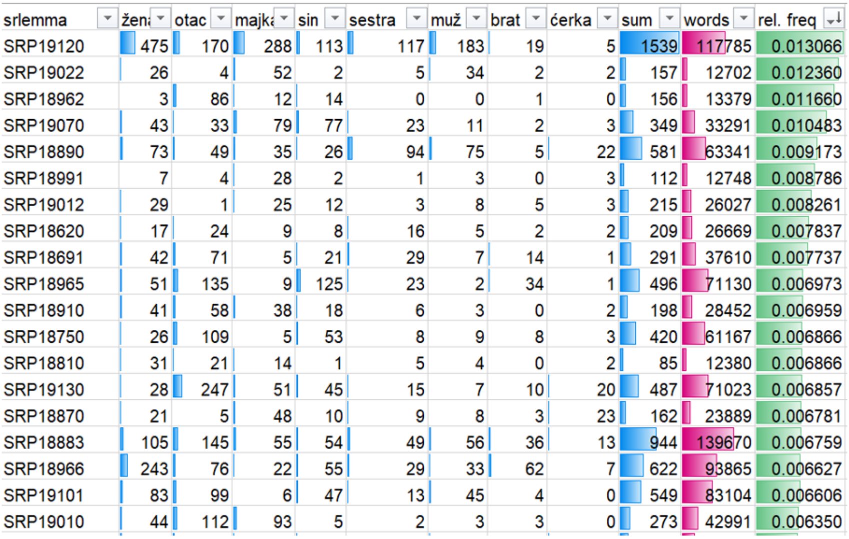


```
</meta>
<graph defaultedgetype="undirected" idtype="string" type="static">
<nodes count="10">
<node id="Q109748906" label="Tone"/>
<node id="Q109748839" label="Tomča"/>
<node id="Q109693861" label="Sofka"/>
<node id="Q109748924" label="Ahmet"/>
<node id="Q109747507" label="Arsa"/>
<node id="Q109748862" label="Simka"/>
<node id="Q109746715" label="Magda"/>
<node id="Q109747662" label="Mita"/>
<node id="Q109748881" label="Todora"/>
<node id="Q109747266" label="Marko"/>
</nodes>
<edges count="17">
<edge id="0" source="Q109693861" target="Q109747266" weight="2.0"/>
<edge id="1" source="Q109747662" target="Q109747266"/>
<edge id="2" source="Q109748839" target="Q109693861"/>
<edge id="3" source="Q109748839" target="Q109748924"/>
<edge id="4" source="Q109748906" target="Q109693861"/>
<edge id="5" source="Q109748839" target="Q109747662"/>
<edge id="6" source="Q109747507" target="Q109747266"/>
<edge id="7" source="Q109693861" target="Q109747507"/>
<edge id="8" source="Q109693861" target="Q109748881" weight="2.0"/>
<edge id="9" source="Q109693861" target="Q109748924"/>
<edge id="10" source="Q109748906" target="Q109746715"/>
<edge id="11" source="Q109693861" target="Q109747662"/>
<edge id="12" source="Q109746715" target="Q109747662"/>
<edge id="13" source="Q109747662" target="Q109747507"/>
<edge id="14" source="Q109748924" target="Q109747266" weight="3.0"/>
<edge id="15" source="Q109693861" target="Q109746715" weight="16.0"/>
<edge id="16" source="Q109693861" target="Q109748862" weight="3.0"/>
</edges>
</graph>
</gexf>
```

Figure 2: Input *gexf* file for the novel *Nečista krv* (Impure blood) (SRP19101).

5 Results and Discussion

Calculating the relative frequencies of occurrences of the lemmas *majka* (mother), *otac* (father), *sin* (son), *ćerka* (daughter), *muž* (husband), *žena* (wife) and their synonyms within the corpus, revealed that *Glava šećera* (Sugar loaf) (SRP18751) by Milovan Glišić had the lowest relative frequency for this set of lemmas at 0.000553, while the novel *Nove* (New ones) by Jelena Dimitrijević (SRP19120) had the highest relative frequency at 0.013066 (Figure 3). These results confirm the assumption that the relative frequencies of the selected lemmas can somewhat indicate whether the theme of family is a central one in a novel. The relative frequencies in the short story *Glava šećera* (Sugar loaf) suggest that the theme of family is not present. This aligns with the content of the story that portrays the position of peasants in relation to the political atmosphere of that era. On the other hand, Jelena Dimitrijević's novel focuses on the tragic fate of a woman and describes her position within a family that adheres to ancient customs. In order to present results more clearly, only the lemmas *žena* (wife), *majka* (mother), *otac* (father), *sin* (son), *ćerka* (daughter), *muž* (husband), *brat* (brother), *sestra* (sister), are shown in Figure 3, without their synonyms.



srlemma	žen	otac	majk	sin	sestra	muž	brat	ćerka	sum	words	rel. freq
SRP19120	475	170	288	113	117	183	19	5	1539	117785	0.013066
SRP19022	26	4	52	2	5	34	2	2	157	12702	0.012360
SRP18962	3	86	12	14	0	0	1	0	156	13379	0.011660
SRP19070	43	33	79	77	23	11	2	3	349	33291	0.010483
SRP18890	73	49	35	26	94	75	5	22	581	63341	0.009173
SRP18991	7	4	28	2	1	3	0	3	112	12748	0.008786
SRP19012	29	1	25	12	3	8	5	3	215	26027	0.008261
SRP18620	17	24	9	8	16	5	2	2	209	26669	0.007837
SRP18691	42	71	5	21	29	7	14	1	291	37610	0.007737
SRP18965	51	135	9	125	23	2	34	1	496	71130	0.006973
SRP18910	41	58	38	18	6	3	0	2	198	28452	0.006959
SRP18750	26	109	5	53	8	9	8	3	420	61167	0.006866
SRP18810	31	21	14	1	5	4	0	2	85	12380	0.006866
SRP19130	28	247	51	45	15	7	10	20	487	71023	0.006857
SRP18870	21	5	48	10	9	8	3	23	162	23889	0.006781
SRP18883	105	145	55	54	49	56	36	13	944	139670	0.006759
SRP18966	243	76	22	55	29	33	62	7	622	93865	0.006627
SRP19101	83	99	6	47	13	45	4	0	549	83104	0.006606
SRP19010	44	112	93	5	2	3	3	0	273	42991	0.006350

Figure 3: Top 20 novels sorted by the relative frequency of occurrences of the lemmas *majka* (mother), *otac* (father), *sin* (son), *ćerka* (daughter), *muž* (husband), *žena* (wife).

Observing the progression diagram (Figure 4), for the lemmas *majka* (mother), *otac* (father), *brat* (brother), *sestra* (sister), *sin* (son), *ćerka* (daughter), *muž* (husband), *žena* (wife), and their synonyms in the work *Patnica* (Patnica)(`SRP8883`) by Jakov Ignjatović, a notable increase can be observed for the lemma *majka* (mother) and a slightly smaller one for *otac* (father). Hence, that fact supports the idea that the novel is about family and the position of women in society. The largest increase for the lemma *žena* (wife) (and its synonyms *supružnica* (spouse), *ljuba* (beloved)), as well as *muž* (husband) and *ćerka* (daughter), occurs in *Nove* (New ones) (`SRP19120`) by Jelena Dimitrijević. A smaller increase for *žena* (wife) was seen earlier in the novel *Nazareni* (Nazarenes) (`SRP18966`) by Jaša Tomić. Both novels discuss the clash between traditional and modern values, with family as a central theme, particularly focusing on women in Dimitrijević's case. In the novel

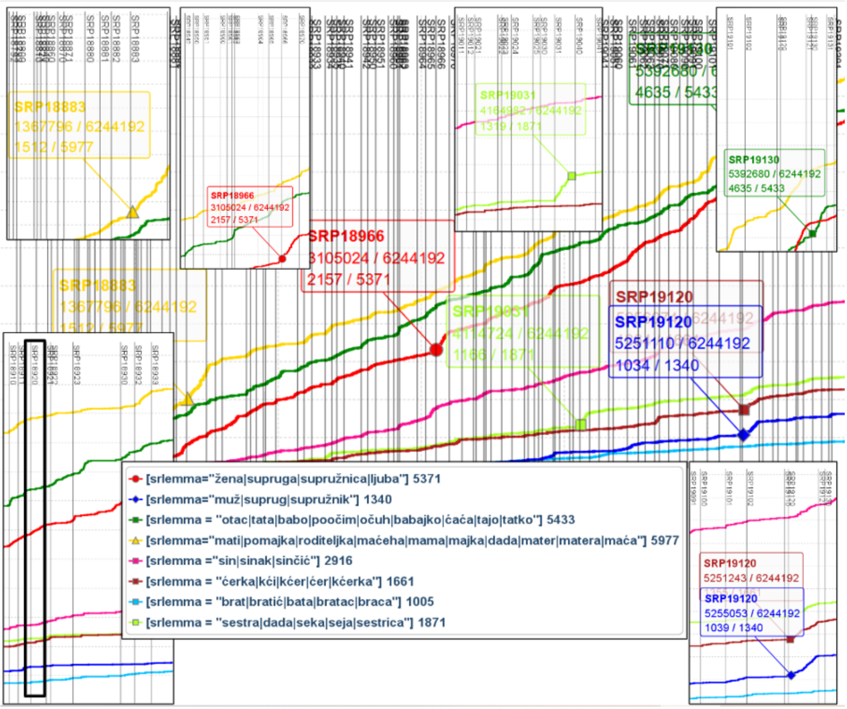


Figure 4: Progression diagram of lemmas *majka* (mother), *otac* (father), *sin* (son), *ćerka* (daughter), *muž* (husband), *žena* (wife) and their synonyms within the corpus.

Dve sestre ili Samoubistvo jedne švalje: slika iz beogradskog života (Two sisters or suicide of a seamstress: a picture from Belgrade life) (SRP19031) by Boža Savić, there is a noticeable increase for the lemma *sestra* (sister), which perfectly aligns with the novel's theme. When analyzing the novel *Kaluđer i hajduk: pripovetka o poslednjim danima Srbije u XV veku* (A monk and a hajduk: a short story about the last days of Serbia in the 15th century) (SRP1930) by Stojan Novaković, one can observe a rise in the occurrence of the lemma *oca* (father). Although the novel is historical in nature and family relations are not the main theme, through the analysis of concordances and the context in which the lemma *otac* (father) appears, it is noted that in this novel, the lemma *otac* (father) refers to the title of a clerical figure. In this case it is a monk, who is also one of the main characters in the novel. In *Beogradske tajne* (Belgrade secrets) (SRP18923) by Pera Todorović, all lemmas stagnate, which confirms that the novel is written in a realistic style, with elements of social drama and adventure, without family as a central theme.

To further describe how the appearance of lemmas depends on the author, the specificity indices for the lemma usage depending on the gender of an author will first be presented. The results show that the specificity index for the lemma *otac* (father) is insignificant for both male and female authors, while the specificity indices for *ćerka* (daughter) (>110), *majka* (mother) (>60), *sin* (son) (>10), *muž* (husband) (>100), and *žena* (wife)(>7) indicate that these lemmas are more prevalent in works by female authors compared to male authors (Figure 5). A deeper qualitative analysis of the context would be necessary to further explore the concept of family in novels based on the author's gender. Here, the specificity of the lemmas home, family, and parent will be briefly examined, depending on the author's gender.

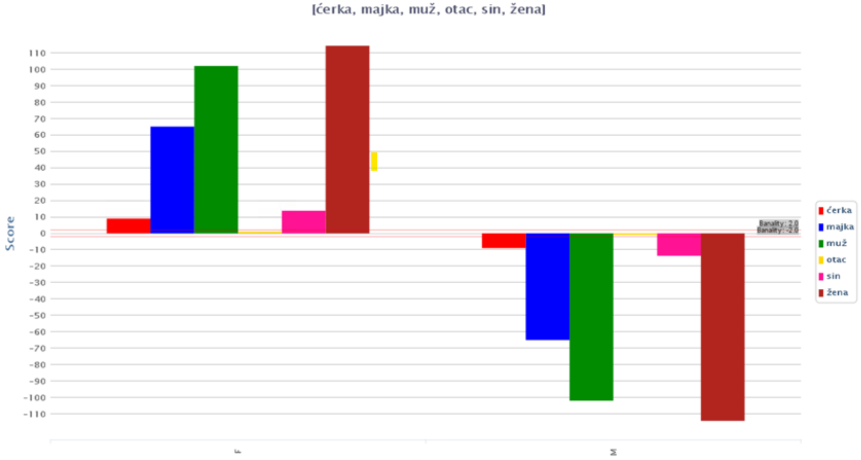


Figure 5: Specificity of the lemmas *ćerka* (daughter), *majka* (mother), *sin* (son), *muž* (husband), and *žena* (wife) based on the author's gender (left: female, right: male).

In Figure 6, the histogram illustrates the specificity of the lemmas *dom* (home), *porodica* (family), and *roditelj* (parent) depending on the author's gender. It is clear that the lemma *roditelj* (parent) is more specific to female authors, while *dom* (home) and *porodica* (family) are statistically insignificant. These results somewhat confirm the notion that female authors tend to focus on individual roles within the family, such as the roles of: a mother, wife, daughter, father, or son. This may indicate that female authors delve deeper into specific relationships between parents and children. However, male authors present a broader, perhaps a more traditional view of family relations and the family institution as a whole, and emphasize the importance of the home as both a physical and symbolic space. This also suggests gender differences in approaching the themes of family and home. Women are more likely to write about emotional and personal experiences, while men focus on the family as a social institution.

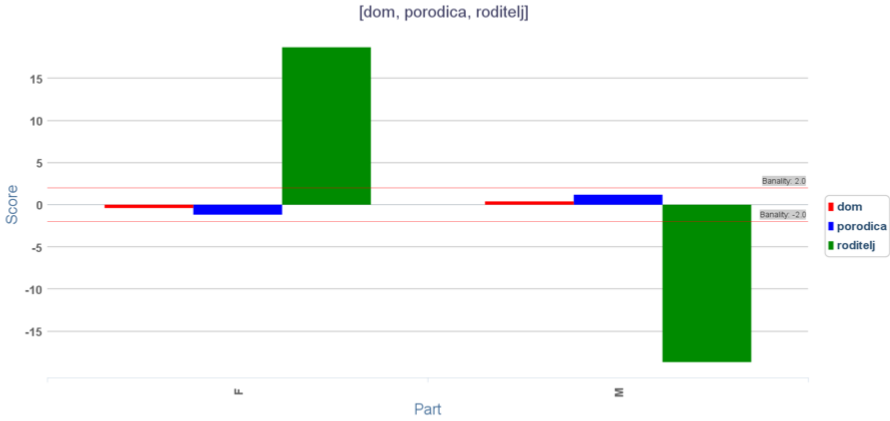


Figure 6: Specificity of lemmas *dom* (home), *porodica* (family) and *parent* based on the author's gender (left: female, right: male authors).

In order to take a closer look at the representation of the lemmas by authors, pairs of lemmas that describe specific family relationships will be examined. Figure 7 presents the specificity of the lemmas *majka* (mother) and *otac* (father) by authors. The results show that the lemma *majka* (mother), with a specificity index greater than 100, is highly represented in the works of Jelena J. Dimitrijević. This can be explained by the fact that her novels primarily focus on women. The low specificity of the lemma *otac* (father) (<10) confirms that the relationship between a mother and father is not the main theme in her works. The balanced specificity of the lemmas *majka* (mother) and *otac* (father) in Todor Lj. Popović's work, who only has one novel in this corpus *Gila: novela iz seoskog života* (Gila: a novella from rural life), suggests that parental relationships are indeed one of the themes of this novel. The low specificity for mother and father in Pera Todorović's work (<-40 and <-20, respectively) clearly indicates that parental relationships are not a central theme in his novels, which aligns with the historical genre of his works.

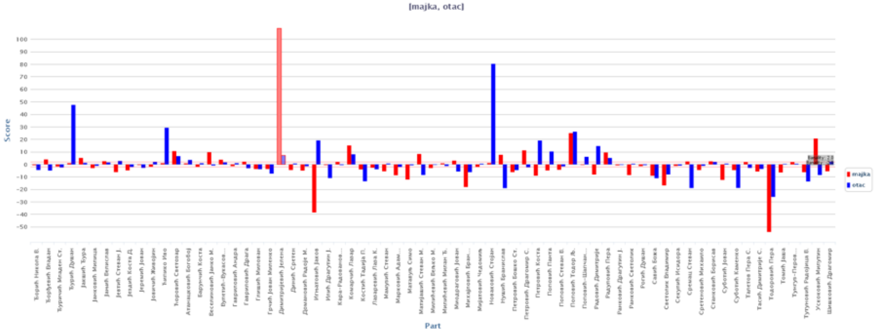


Figure 7: Specificity of the lemmas *majka* (mother) and *otac* (father) in the SrpELTeC corpus by author.

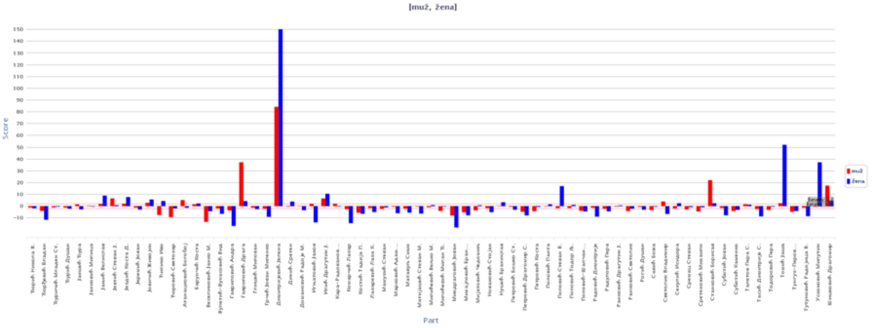


Figure 8: Specificity of the lemmas *žena* (wife) and *muž* in the SrpELTeC corpus by author.

However, if we look at the lemmas *žena* (wife) and *muž* (husband) (Figure 8), the high specificity of Jelena Dimitrijević confirms, once again, that the relationship between a wife and husband (whether current, future, or desired) is one of the main themes in her novels. On the other hand, the low specificity (<-10) for Andra Gavrilović indicates that he does not focus on women in his works. This aligns with the historical themes of his novels *Despotova vlastela: roman iz srpske prošlosti* (Despot's nobles: a novel from the Serbian past) and *Prve žrtve: pripovetka iz srpske prošlosti* (The first victims: a story from the Serbian past).

The cluster analysis shown in Figure 9 divides selected authors into three groups based on how family relationships are represented in their works. Authors such as Ivan Ćipiko, Svetolik Ranković, Lazar Komarčić, Janko Veselinović, and Svetozar Ćorović can be grouped by several key aspects regarding the theme of family in their works. These authors often explore disruptions and crises within traditional family structures, showing how social and economic changes impact family relationships and destroy old patterns and values. Čedomil Mijatović and Stevan Sremac use family relationships as a tool for critically highlighting social and societal issues. Their works often examine how family relationships reflect and influence broader societal structures and norms. Authors such as Jelena Dimitrijević, Milutin Uskoković, Branislav Nušić, Dragutin Ilić, Pera Todorović, Vladan Đorđević, and Jovan Miodragović explore family conflicts, generational clashes, and interpersonal

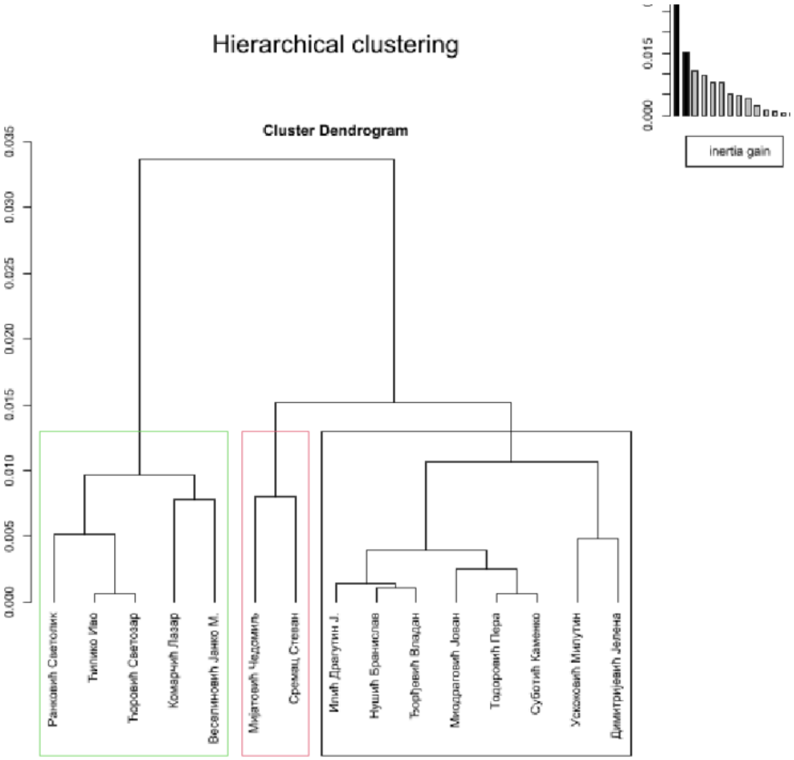


Figure 9: Cluster analysis of family relationships by author.

relationships within the family. Their works often depict dramatic conflicts and disagreements among family members. The cluster analysis successfully groups authors who approach the theme of family in similar ways. However, given that automatically generated clustering was applied, the validity of this clustering is something that should be further investigated.

To address the second research question, the frequency and specificity of the lemmas *ćerka* (daughter), *majka* (mother), *sin* (son), *muž* (husband), and *žena* (wife) were examined across the time periods into which the corpus is divided: T1: [1840–1859], T2: [1860–1879], T3: [1880–1899], and T4: [1900–1920] (Figure 10).

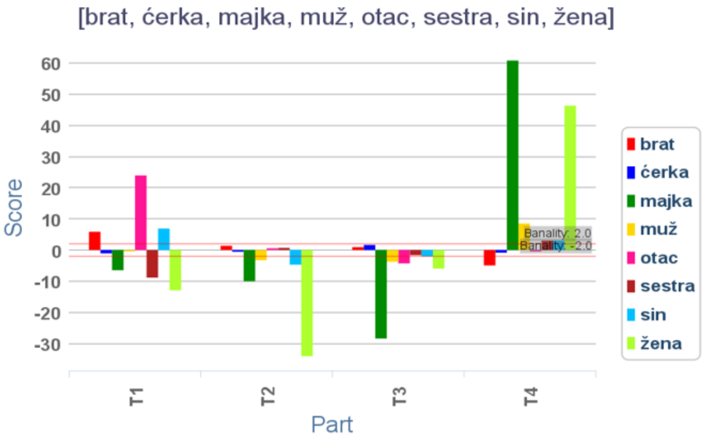


Figure 10: Specificity of the lemmas *ćerka* (daughter), *majka* (mother), *sin* (son), *muž* (husband), and *žena* (wife) by time periods.

Romanticism in Serbian literature developed from 1848 to 1870, with Branko Radičević (1824–1853) as the founder of Serbian lyrical romanticism. In prose, Bogoboj Atanacković and Ljubo Nenadović (1826–1895), the most significant travel writer, were notable figures. Realism emerged as a literary movement in Serbia in the early 1870s. Milovan Glišić was writing his stories in the 1880s, alongside Laza Lazarević, Simo Matavulj, and Janko Veselinović, who also published works during this period. Stevan Sremac and Svetolik Ranković published their most important works in the 1890s. The development of realism continued into the early 20th century (Deretić 1983). Examining the collection of Serbian novels written between 1840 and 1920, it was challenging to establish a clear distinction between these two literary movements. Based on this, it can be said that T1 is dominated by romanticism, while T2 is intertwined with both romanticism and realism. Periods T3 and T4 are dedicated to realism, with realism reaching its peak during T4.

Looking at the specificity indices (Table 2) and the histogram (Figure 10) for the lemmas *ćerka* (daughter), *majka* (mother), *sin* (son), *muž* (husband), and *žena* (wife) across the time periods, it can be observed that the lemmas wife, mother, and husband are more specific to T4 than to other periods, as can also be seen in Figure 10. The significant presence of the lemma son further supports the idea that family relationships are more prevalent in the later stages of realism (T4), compared to its earlier phase (T3).

Table 2: Specificity indices of the lemmas mother (majka), daughter (ćerka), father (otac), husband (muž), son (sin), wife (žena), brother (brat), sister (sestra) by time periods

лема	Σ_F	f_{T1}	S_{T1}	f_{T2}	S_{T2}	f_{T3}	S_{T3}	f_{T4}	S_{T4}
žena	5,108	64	-12.9	353	-34.1	2,040	-5.9	2,651	**46.3**
otac	4,870	267	**23.9**	604	0.6	1,970	-4.3	2,029	-0.4
majka	3,637	55	- 6.5	320	-10.0	1,242	-28.4	2,020	**60.8**
sin	2,366	111	**6.9**	223	-4.7	965	-2.0	1,067	3.0
muž	1,293	33	-0.4	119	-3.3	495	-3.7	646	**8.5**
ćerka	433	7	-1.1	48	-0.5	208	**1.6**	170	-0.8
brat	789	47	**5.9**	112	1.4	358	0.9	272	-4.9
sestra	1,542	10	**-8.9**	198	0.7	627	-1.6	707	3.0

For the lemma *otac* (father), which stands out in terms of specificity, it can be noted that it is highly specific in T1 with a value of +23.9 (Figure 10, Table 2). This result can be explained by the fact that only two novels from T1 are present in the corpus. Thus, that leads to a bias in the results: of the total number of occurrences of father in the entire corpus (267), 200 of them appear in the novel *Đurađ Branković: istorijski roman* by Jakov Ignjatović, where the term often refers to a religious figure or spiritual father.

However, examining the relationships between characters required a different approach, which is why data from Wikidata were used. Each novel in Wikidata represents an RDF (Resource Description Framework)[19] structure. One such structure is shown below using the example of the novel *Ivkova slava* (Ivko's patron saint's day) by Stevan Sremac

[19] RDF, https://w3c.github.io/rdf-primer/spec/

```
   wd:Q109336082      # Ivko's patron saint's day
         wdt:P31      wd:Q7725634;    # is a literary work
         wdt:P50      wd:Q559989;    # author Stevan Sremac
         wdt:P407     wd:Q9299;      # language of the work: Serbian
         wdt:P840     wd:Q3711;      # narrative place Belgrade
         wdt:P840     wd:Q129259;     # narrative place Niš
         wdt:P1433    wd:Q106927517; # published in ELTeC
         wdt:P1476    Ивкова слава : приповетка    # title
         wdt:P2408     wd:Q212829       # the plot takes place around St.
   George's Day
      # characters in the novel
         wdt:P674      wd:Q109378039
         wdt:P674      wd:Q109378481
         wdt:P674      wd:Q109554049
         …
      Q559989         # Stevan Sremac
         wdt:P31      wd:Q5;        # is a human
         wdt:P21      wd:Q6581097;    # gender is male
         wdt:P19      wd:Q571136;     # place of birth: Senta
```

Using the displayed structure, it is possible to utilize the Wikidata Query Service[20] to visualize relationships between characters in novels. The advantage of using Wikidata to analyze character relationships is that we can quickly access the data and obtain accurate information about the relationships between individual characters. By using SPARQL queries, character relationships entered into Wikidata can be visually displayed. For instance, a SPARQL query can show the relationships between all spouses in the core and extended novel collections. A visual representation of all spouses is available here, and a part of that graph is shown in Figure 11. By clicking on Lazar Pantić (`Q110271637`), additional properties and relationships of that character can be revealed (represented by dashed red arrows), such as who his father is (*father* Milađin Pantić (`Q110271589`)), who his mother is (*mother* Anđelija Pantić (`Q110271431`)), and the work in which the character appears (*present in work Seljanka: pripovetka iz seoskog života* (Peasant woman: a story from rural life) (`Q109336237`)). The property (`P31`) indicates the number 2, which in this case means that Lazar Pantić is associated with two statements: *fictional human* (`Q15632617`) and *literary character* (`Q3658341`). Currently, 121 such husband-wife relationships are detected between characters in SrpELTeC collection.

[20] Wikidata Query Service

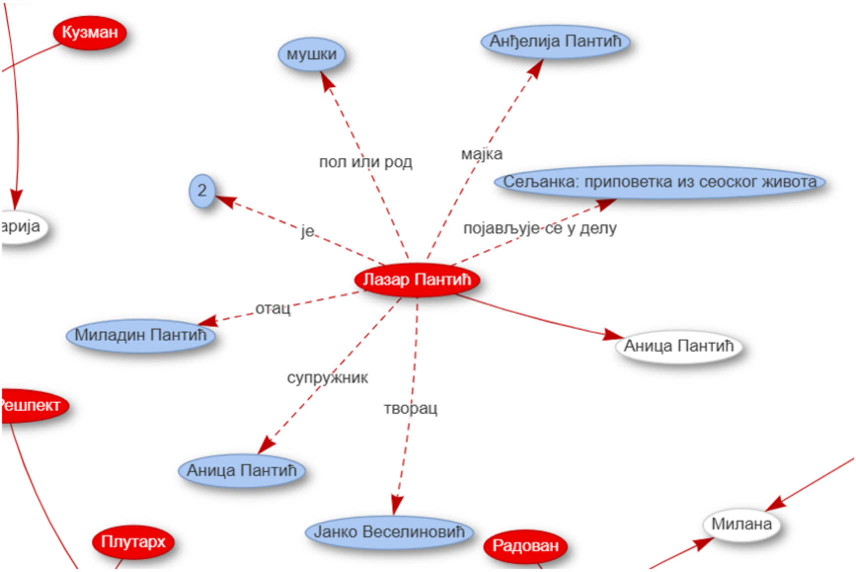


Figure 11: Husband-wife relationship graph within the collection.

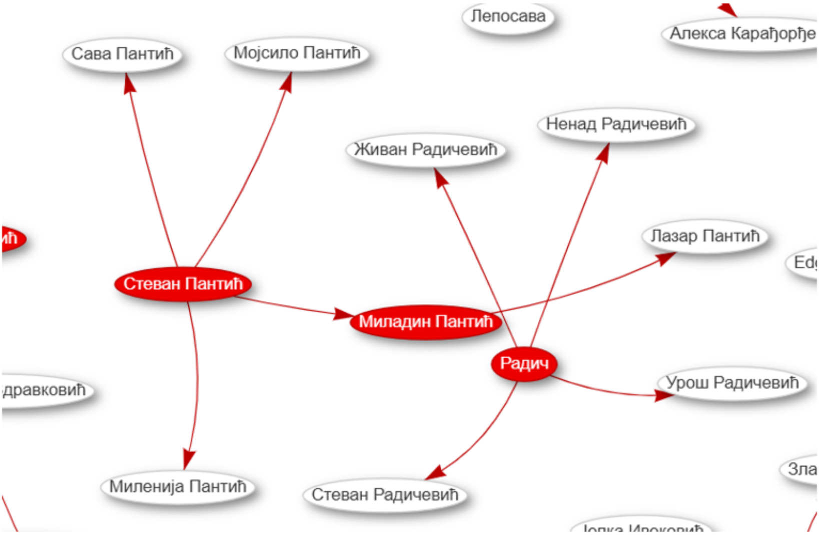


Figure 12: Father-child relationship graph within the collection.

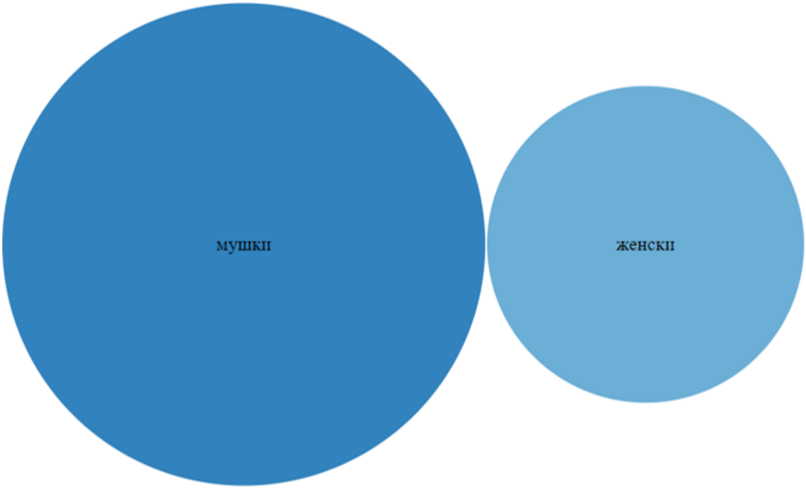


Figure 13: Appearance of characters by gender in the SrpELTeC collection.

In Figure 12, the graph of husband-wife relationships within the SrpELTeC collection is shown. The nodes representing husbands are colored red, and the solid red arrows point to their wives. A similar query can be used to analyze parent-child relationships (https://w.wiki/AqN4). Figure 12 presents a part of the visual representation of the network generated by this query. In Figure 12, it can be viewed that Stevan Pantić is the *father* of Miladin Pantić, Sava Pantić, Mojsije Pantić, and Milena Pantić, while Miladin Pantić is the *father* of Lazar Pantić. Currently, there are 175 father-child relationships in the SrpELTeC collection.

To answer the third research question, a SPARQL query (https://w.wiki/AqNF) that visualizes the results using a bubble chart is used. The sizes of the circles reflect the relative proportion of characters by gender. The results of the query, shown in Figure 13, indicate that male characters (671) are more represented than female characters (289), confirming the hypothesis that male characters are dominant in the corpus of older Serbian novels.

After conducting the textometric analysis and applying semantic web techniques, a network graph, based on a passage from the novel *Nečista krv*, is shown as an example of how network analysis can be used to determine relationships between characters. The statistical analysis of the network is shown in Table 3. The *Label* column contains the names of the characters, while *Id* represents the QID of the corresponding Wikidata item. Charac-

Table 3: Network statistics

Id	Label	MC	WD	E	CC	HCC	BC
Q109748906	Tone	0	2.0	2.0	0.56	0.61	0.0
Q109748839	Tomča	1	3.0	2.0	0.6	0.67	0.33
Q109693861	Sofka	0	28.0	1.0	1.0	1.0	23.67
Q109748924	Ahmet	1	5.0	2.0	0.6	0.67	0.33
Q109747507	Arsa	1	3.0	2.0	0.6	0.67	0.0
Q109748862	Simka	0	3.0	2.0	0.53	0.56	0.0
Q109746715	Magda	0	18.0	2.0	0.6	0.67	0.5
Q109747662	Mita	1	5.0	2.0	0.69	0.78	2.33
Q109748881	Todora	0	2.0	2.0	0.53	0.56	0.0
Q109747266	Marko	1	7.0	2.0	0.64	0.73	0.83

ters with different Modularity Class values belong to different parts of the narrative network and are divided into two clusters. One cluster (MC = 0, green nodes in Figure 14) could represent a group of characters involved in a specific scene or plot, while the other cluster (MC = 1, blue nodes in Figure 14) represents a group of characters connected to a different part of the story. The class MC = 0, which includes the characters Sofka, Magla, Simka, Todora, and Tone, represents a group dominated by the female sphere within the novel, as well as characters who mark the action taking place inside the household. Magda is also identified here as Sofka's loyal companion, which explains their close connection. Sofka's brother Tone belongs to the passive part of the family; his character appears primarily within the household, which structurally places him closer to the female characters than to the male group, thus confirming his affiliation with class MC = 0. Class MC = 1 includes Marko, Ahmet, Mita, Tomča, and Arsa. This class represents a group dominated by male characters or figures of patriarchal power and trade. Marko (WD = 7.0) and Mita (WD = 5.0) are the key figures in this group.

The nodes represent the characters in the novel, with the size of the node determined by the Betweenness Centrality (**BC**) statistic. In character analysis, a character with high BC frequently mediates communication or interaction between other characters, and is represented by a larger node in the network. Minor characters, who do not bridge different groups or scenes, are represented by smaller nodes. The Correlation Coefficient (**CC**) quantifies the degree of correlation between a character and other characters in the same sentence. A higher CC indicates greater interaction with other characters, as seen in the case of Sofka (1.0), who appears most frequently

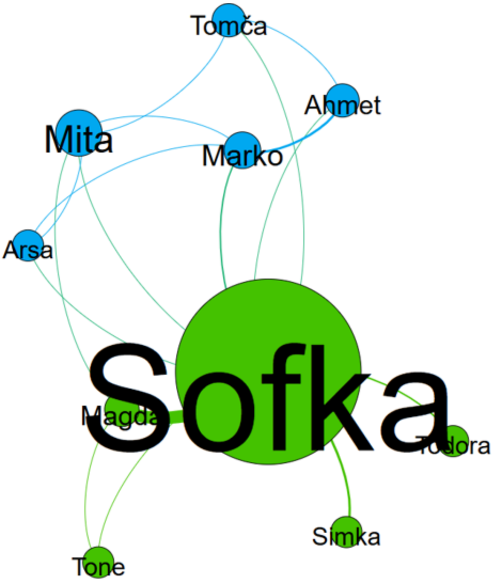


Figure 14: Character network graph in the novel *Nečista krv (Impure Blood).*

in the analyzed passage, and has the most connections with other characters. Centrality measures, such as Betweenness Centrality, indicate how central a character is in the network. The Hierarchical Clustering Coefficient (**HCC**) provides additional information on how these central characters are connected to their surroundings, especially in terms of the hierarchical structure. A higher HCC value suggests that a node is more strongly connected to its neighbors, who are also connected to each other, as in the case of Sofka (1.0). Eccentricity (**E**) describes the distance of a node from the most distant nodes in the network. Characters with low eccentricity are often crucial to the story, as they are connected to most other characters and play a central role in the social network. From Table 3, it can be noted that Sofka is the central character, as her eccentricity value is the lowest (1.0). Weighted Degree (**WD**) represents the weight (importance) calculated as the sum of the weights (number of occurrences) of all connections associated with a specific node (character in the novel).

By summarizing all the parameters from Table 3, the statistical analysis of the network confirms that Sofka occupies a central position in the narrative structure of the examined passage. Her high weighted degree (WD = 28.0) indicates her ubiquity and dominant number of interactions, while her Betweenness Centrality (BC = 23.67) defines Sofka as a key character who connects all other characters. Her strategic advantage is further emphasized by minimal eccentricity (E = 1.0), which confirms that Sofka is directly connected to all actors, establishing her as the primary character in the narrative. Although Mita has a relatively small number of interactions (WD = 5.0), he has the second-highest BC = 2.33. While he would naturally be expected to be the center of the family cluster (class 0), he appears as an external factor connecting Sofka to the buyers (class 1), which is confirmed by his high BC and affiliation with characters outside the household. This result also confirms the connection between Effendi Mita and Sofka in the father-daughter relationship. Magda (WD = 18.0, but BC = 0.5) appears frequently, but almost always alongside Sofka. Her low BC indicates that she does not have an independent role in linking different groups of characters; she is “locally” important within Sofka’s environment but does not serve as a connector between different clusters. The peripheral status of characters such as Tone, Arsa, Simka, and Todora is confirmed by their minimal centrality (BC = 0.0), and of Ahmet and Tomča (BC = 0.33), which indicates that their interactions do not affect the integration of the wider network of relationships, remaining confined to isolated narrative segments. The edges between the characters indicate their connection within the sentences of the novel. The thicker the line, the higher the number of shared appearances of the two connected characters within the same sentences in the novel. In Figure 14, it can be observed that the most connected characters are Sofka and Magda, which is expected since they are a mother and daughter, while Sofka and Arsa share only one appearance together, as indicated by the very thin line connecting them.

The network analysis clearly confirmed family connections through high clustering coefficient (CC) values (Sofka’s mother Todora, Sofka’s father Mita, Sofka’s brother Tone) and modularity (MC), which highlighted the household’s internal world as a distinct cluster. The statistical parameters reflected family dynamics, indicating that while the high frequency of connections (WD) demonstrates the daily closeness of household members (Magda, WD = 18.0, Sofka’s maid), the high BC values (Sofka 23.67 and Mita 2.33) confirm the patriarchal role of the father as a key bridge controlling the relationship between the family home and the external world of trade and power.

6 Conclusion

Upon analyzing the results for the first research question, it can be concluded that the textometric analysis provided expected results when examining the specificity of the lemmas *ćerka* (daughter), *majka* (mother), *sin* (son), *muž* (husband), *žena* (wife), *sestra* (sister) and *brat* (brother), as well as their synonyms. The results confirmed a correlation between the specificity of these lemmas and the gender of the authors, with the specificity aligning with the themes of the authors' novels. Further research related to this question could take several directions. First, additional NLP tools could be applied to detect nouns, verbs, or phrases associated with family relationships, as well as adjectival attributes that accompany the lemmas. Investigating family relationships in literary texts could also include examining their appearance concerning certain sociological characteristics of the authors (gender, age, social background, education, profession, etc.), individual stylistic features, and the themes and types of narratives.

Pointing out the higher specificity of the examined lemmas in T4, compared to other periods, partially confirmed the claim that family relationships were more prevalent in the period of realism's peak development than in the romanticism period. However, future studies on this research question could utilize data from Wikidata once each novel in Wikidata is tagged with its corresponding literary movement.

Linking the novels to Wikidata and entering characters allowed for the counting of male and female characters across the entire collection, and revealed that gender stereotypes were present in the srpELTeC collection. This also enabled the visualization of family relationships for the characters and their connections which were entered into Wikidata.

Finally, the possibility of network analysis of the characters in the excerpt from the novel *Nečista krv* (Impure Blood) is also presented, which, through statistical analysis of network parameters, confirmed the family relationships among the characters. Future research will certainly move in the direction of network analysis of characters across the entire corpus

Acknowledgment

This research was supported by the Science Fund of the Republic of Serbia, #7276,Text Embeddings – Serbian Language Applications – TESLA. The author thanks Professor Cvetana Krstev and Professor Ranka Stanković for providing the extensively prepared dataset and for their valuable suggestions that improved the paper.

References

Bamman, David, Brendan O'Connor, and Noah A. Smith. 2013. "Learning Latent Personas of Film Characters." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,* 352–361. Association for Computational Linguistics.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi: an open source software for exploring and manipulating networks." In *Proceedings of the international AAAI conference on web and social media,* 3:361–362. 1.

Borin, Lars, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2009. "Thinking green: toward swedish FrameNet++." In *FrameNet Masterclass and Workshop,* 3:1–8.

Borin, Lars, and Markus Forsberg. 2010. "Beyond the synset: Swesaurus–a fuzzy Swedish wordnet." In *Workshop on Re-thinking synonymy: Semantic sameness and similarity in languages and their description. Helsinki,* 00137–7.

Carrive, Jean, Abdelkrim Beloued, Pascale Goetschel, Serge Heiden, Antoine Laurent, Pasquale Lisena, Franck Mazuet, Sylvain Meignier, Bénédicte Pincemin, Géraldine Poels, et al. 2021. "Transdisciplinary analysis of a corpus of French newsreels: The ANTRACT Project." *Digital Humanities Quarterly* 15 (1).

Cipresso, Pietro, and Giuseppe Riva. 2016. "Computational psychometrics meets Hollywood: the complexity in emotional storytelling." *Frontiers in Psychology* 7:227145. https://doi.org/10.3389/fpsyg.2016.01753.

Deretić, Jovan. 1983. *Историја Српске Књижевности.* Нолит.

Firat, Hatice. 2018. "Grandparent-Grandchild Relationships in Turkish Children's Novels." *Universal Journal of Educational Research* 6 (10): 2047–2060. https://files.eric.ed.gov/fulltext/EJ1192746.pdf.

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. "Named entity recognition for distant reading in ELTeC." In *CLARIN Annual Conference 2020,* 37–41.

Gledhill, Christopher, Hanna Martikainen, Alexandra Mestivier, and Maria Zimina-Poirot. 2019. “Towards a linguistic definition of ‘simplified medical English’: Applying textometric analysis to cochrane medical abstracts and their plain language versions.” *LCM-La Collana/The Series,* 91–114.

Heiden, Serge. 2011. “The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme.” In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation,* 389–398.

Ikonić Nešić, Milica, Saša Petalinkar, Mihailo Škorić, and Ranka Stanković. 2024. “BERT downstream task analysis: Named Entity Recognition in Serbian.” In *Conference on Information Technology and its Applications,* 333–347. Springer.

Ikonić Nešić, Milica, Ranka Stanković, and Biljana Rujević. 2021. “Serbian ELTeC Sub-Collection in Wikidata.” *Infotheca - Journal for Digital Humanities* 21 (2): 60–87. https://doi.org/10.18485/infotheca.2021.21.2.4.

Ikonić Nešić, Milica, Ranka Stanković, Christof Schöch, and Mihailo Škorić. 2022. “From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back).” In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference,* edited by Thierry Declerck, John P. McCrae, Elena Montiel, Christian Chiarcos, and Maxim Ionov, 7–16. European Language Resources Association. https://aclanthology.org/2022.ldl-1.2/.

Jaćimović, Jelena. 2019. “Textometric methods and the TXM platform for corpus analysis and visual presentation.” *Infotheca - Journal for Digital Humanities* 19 (1): 30–54. https://doi.org/10.18485/infotheca.2019.19.1.2.

Kokkinakis, Dimitrios, and Mats Malm. 2011. “Character Profiling in 19th Century Fiction.” In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage,* edited by Cristina Vertan, Milena Slavcheva, Petya Osenova, and Stelios Piperidis, 70–77. Association for Computational Linguistics. https://aclanthology.org/W11-4111/.

Krstev, Cvetana. 2021. “White as Snow, Black as Night – Similes in Old Serbian Literary Texts.” *Infotheca - Journal for Digital Humanities* 21 (2): 119–135. https://doi.org/10.18485/infotheca.2021.21.2.6.

Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. "A system for named entity recognition based on local grammars." *Journal of Logic and Computation* 24 (2): 473–489. https://doi.org/10.1093/logcom/exs079.

Krstev, Cvetana, and Ranka Stanković. 2020. "Old or new, we repair, adjust and alter (texts)." *Infotheca - Journal for Digital Humanities* 19 (2): 61–80. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2019.19.2.3.

Krstev, Cvetana, and Ranka Stanković. 2022. "Novels and Authors of the Serbian ELTeC Collection." *Infotheca - Journal for Digital Humanities* 21 (2): 172–186.

Kucserka, Zsófia. 2020. "Friends or Enemies? Sisterhood in Nineteenth-Century Hungarian Novels and Diaries." *The Hungarian historical review: new series of Acta Historica Academiae Scientiarum Hungaricae* 9 (4): 650–666. https://doi.org/10.38145/2020.4.650.

Longhi, Julien. 2021. "Using digital humanities and linguistics to help with terrorism investigations." *Forensic science international* 318:110564.

Makazhanov, Aibek, Denilson Barbosa, and Grzegorz Kondrak. 2014. "Extracting family relationship networks from novels." *arXiv preprint arXiv:1405.0603.*

Min, Semi, and Juyong Park. 2016. "Network Science and Narratives: Basic Model and Application to Victor Hugo's Les Misérables." In *Complex Networks VII: Proceedings of the 7th Workshop on Complex Networks CompleNet 2016,* edited by Hocine Cherifi, Bruno Gonçalves, Ronaldo Menezes, and Roberta Sinatra, 257–265. Cham: Springer International Publishing.

Moretti, Franco. 2011. "Network theory, plot analysis."

Odebrecht, Carolin, Lou Burnard, and Christof Schöch. 2021. *European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels.* https://zenodo.org/records/4662444.

Patras, Roxana, Carolin Odebrecht, Ioana Galleron, Rosario Arias, J. Berenike Herrmann, Cvetana Krstev, Katja Mihurko Poniž, and Dmytro Yesypenko. 2020. *Dataset for ELTEC titles [Data set].* http://zenodo.org/records/4268669.

Prado, Sandra D., Silvio R. Dahmen, Ana LC Bazzan, Padraig Mac Carron, and Ralph Kenna. 2016. “Temporal network analysis of literary texts.” *Advances in Complex Systems* 19 (03): 1650005.

Qizi Sayitqulova, Zilola Husniddin. 2021. “The Reflection of Family Relations in the Novel of A. Chulpan “Kecha Va Kunduz”.” *International Journal of Linguistics, Literature and Culture* 7 (4): 188–193. https://doi.org/10.21744/ijllc.v7n4.1635.

Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. “Serbian NER&Beyond: The Archaic and the Modern Intertwinned.” In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021),* 1252–1260. Held Online: INCOMA Ltd. https://aclanthology.org/2021.ranlp-1.141/.

Santos, Daniel, Nuno Mamede, and Jorge Baptista. 2010. “Extraction of family relations between entities.” In *INForum,* 9–10.

Škorić, Mihailo, Ranka Stanković, Milica Ikonić Nešić, Joanna Byszuk, and Maciej Eder. 2022. “Parallel Stylometric Document Embeddings with Deep Learning Based Language Models in Literary Authorship Attribution.” *Mathematics* 10 (5): 838. https://doi.org/10.3390/math10050838.

Stanković, Ranka, Cvetana Krstev, Branislava Šandrih Todorović, Duško Vitas, Mihailo Škorić, and Milica Ikonić Nešić. 2022. “Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection.” In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 3337–3345. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.lrec-1.356/.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. “Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian.” In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 3954–3962. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.487/.

Stanković, Ranka, Diana Santos, Francesca Frontini, Tomaz Erjavec, and Carmen Brando. 2019. “Named entity recognition for distant reading in several european literatures.” *DH Budapest 2019.*

Utvić, Miloš. 2013. “Izgradnja referentnog korpusa savremenog srpskog jezika.” PhD diss., Univerzitet u Beogradu, Filološki fakultet. https://phaidrabg.bg.ac.rs/view/o:10061.

Vitas, Duško. 2020. “Food as Text.” *Infotheca - Journal for Digital Humanities* 19 (2): 139–161. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2019.19.2.7.

Woloch, Alex. 2009. *The One vs. the Many: Minor Characters and the Space of the Protagonist in the Novel.* Princeton University Press.

Крстев, Цветана, Ранка Станковић, Бранислава Шандрих Тодоровић, and Милица Иконић Нешић. 2023. “Нове технологије за оживљавање старих текстова.” In *Зборник радова Међународне научне конференције Дигитална хуманистика и словенско културно наслеђе II, Београд, 28-29 јуни 2021,* 1252–1260. Београд: Савез славистичких друштава Србије.