Fifteen Writers and Their Digital Imprints in Numbers, Images and Words

UDC 811.163.41'322.2
DOI 10.18485/infotheca.2026.26.1.1

ABSTRACT: In this paper we present the corpus 15AUTHORS which contains 49 works of fifteen authors that wrote in the Serbian language at the end of the 19th and the beginning of the 20th century. This corpus was derived from the SrpELTeC corpus built within the framework of the COST Action "Distant Reading for European Literary History." We used existing annotations (sentences, phrases in foreign languages, part-of-speech (POS) tags, lemmas and named entities) and conducted additional analyses with open-code software Unitex and TXM, in order to reveal digital imprints left by the selected authors in their works.

KEYWORDS: literary corpus, textometry, corpus linguistics, distant reading, Serbian language, Unitex, TXM.

PAPER SUBMITTED: 16 March 2026
PAPER ACCEPTED: 01 April 2026

Cvetana Krstev
ORCID 0000-0003-3328-9392
cvetana@jerteh.rs

Language Resources and Technologies Society – JeRTeh Belgrade, Serbia

1 Introduction

The research presented in this paper is based on the ELTeC corpus, specifically its Serbian part, SrpELTeC, which was developed within the framework of the COST Action "Distant Reading for European Literary History" (CA16204).[1] Goals and results of this project, as well as challenges involved in developing the Serbian part of the corpus are presented in previous studies (Trtovac, Milnović, and Krstev 2021; Krstev 2021a).

The ELTeC corpus has already been used in various linguistic, literary, ethnological and natural language processing research. We will mention only a few here.[2] The ELTeC collection, which comprises 12 complete subcollections of 100 novels each, provides an excellent basis for different multilingual

1. European Literary Text Collection (ELTeC)
2. Detailed list

and comparative studies. Topics of interest include titling practices in novels from the period covered by the corpus (Patras et al. 2021) and the reconsideration of the common thesis in literary history that the inner life of novel characters became the central focus of literary modernism, examined using computational methods (Radak et al. 2024). Byszuk et al. (2020) address the detection of direct speech in nine ELTeC subcollections using rule-based methods and transformer architecture. The study (Škorić et al. 2022) focuses on authorship attribution using stylometric methods based on parallel embeddings and deep learning.

Studies focusing on the Serbian language that used the SrpELTeC subcollection are also numerous. SrpELTeC has been used to compile a repertoire of simile rhetoric figures in old literary texts, which was subsequently tested on a collection of contemporary literary works (Krstev 2021b). Vitas (2022) demonstrates that a carefully prepared literary corpus can serve as a basis for studying various aspects of private life during a given time period. Nešić et al. (2022) present the results of the recognition of basic classes of named entities, their annotation and their linking to knowledge bases. The results of sentiment analysis conducted on Serbian novels from the late 19th and early 20th century are given in (Stanković, Košprdić, et al. 2022). Furthermore, a comparison of topic modeling methods, illustrated with examples from the same text collection, is given in (Mihajlov et al. 2024).

2 Corpus, methods and tools

The basic Serbian subcollection (SrpELTeC) contains 100 novels, each having at least 10,000 words, originally written in Serbian and first published between 1840 and 1920. The selection of novels for SrpELTeC was balanced as much as possible, following guidelines applied to all language subcollections developed within the COST Action. The time span had to be evenly covered (concerning the first editions), novels of various lengths had to be proportionally represented, and an equal number of male and female authors had to be included. Each collection was also designed to include well-known novels (representing the canon) and lesser-known novels published only once and forgotten. In addition, each collection should contain 9 to 11 authors represented by three novels each, while all other authors had to be presented by a single novel. The extended Serbian subcollection (SrpELTeC-ext) includes 20 additional novels that meet the same criteria. The SrpELTeC-108 collection comprises all novels from the basic SrpELTeC subcollection, along with eight novels from the SrpELTeC-ext subcollection.

In the basic SrpELTeC subcollection, eleven authors are represented by three novels each, while one author is represented by five novels. The extended SrpELTeC-108 subcollection comprises works of fifteen authors, each represented by at least three works. These authors are: Jakov Ignjatović (5), Vladan Đorđević (3), Đura Jakšić (3), Milan Đ. Milićević (3+2), Lazar Komarčić (3), Draga Gavrilović (3), Janko Veselinović (3), Pera Todorović (3), Čedomilj Mijatović (2+1), Steva Sremac (3), Svetolik Ranković (3), Borisav Stanković (2+1), Jelena Dimitrijević (3), Svetozar Ćorović (3), and Milutin Uskoković (2+1). The list of all these works is provided in Appendix 4. The 49 works by these fifteen authors, collected in the 15AUTHORS subcollection, serve as the basis for our digital research.

According to the agreement reached within the D-reading COST Action, the SrpELTeC corpus was prepared in accordance with the TEI Guidelines.[3] This means that structural and textual elements such as chapters, paragraphs, highlighted passages (e.g., italics), foreign-language expressions, titles of artistic or professional works, and footnotes were annotated. The annotation was carried out by volunteer readers during the correction of texts obtained through optical character recognition. In addition, several layers of annotation were introduced automatically. Sentences were segmented, and each token was assigned a part-of-speech (POS) tag and a lemma. Furthermore, seven categories of named entities were annotated: PERS, ROLE, LOC, ORG, DEMO, EVENT, and WORK, the last of which was annotated manually by the readers. Further details on the tools used and the annotation results can be found in (Stanković, Krstev, et al. 2022).

The results presented in this paper were obtained by using existing annotations, as well as by conducting additional data searches and analyses with two open-source tools. The first one is Unitex (Paumier, Nakamura, and Voyatzi 2009), which uses electronic dictionaries to define complex queries and transformations based on finite-state transducers.[4] We used this software in combination with electronic dictionaries for Serbian and a set of general and specialized transducers (Krstev 2008) to obtain the results presented in Sections 3.1, 3.2, 3.4, 3.9, 3.10, and 3.12. The second tool is TXM, which was used to generate various statistical outputs and to visualize the results (Pincemin, Heiden, and Mazuet 2022) (see Sections 3.3, 3.5, 3.6, 3.7, 3.8, 3.12, and 3.13).[5] No tool can perform text-processing tasks "perfectly"; a certain degree of error and omission is unavoidable. This issue is not ad-

3. TEI Guidelines.
4. Unitex – corpus processing suite.
5. TXM – textometry platform.

dressed in the present paper. However, information on the performance of the tools used can be found in the cited literature.

3 Digital imprints

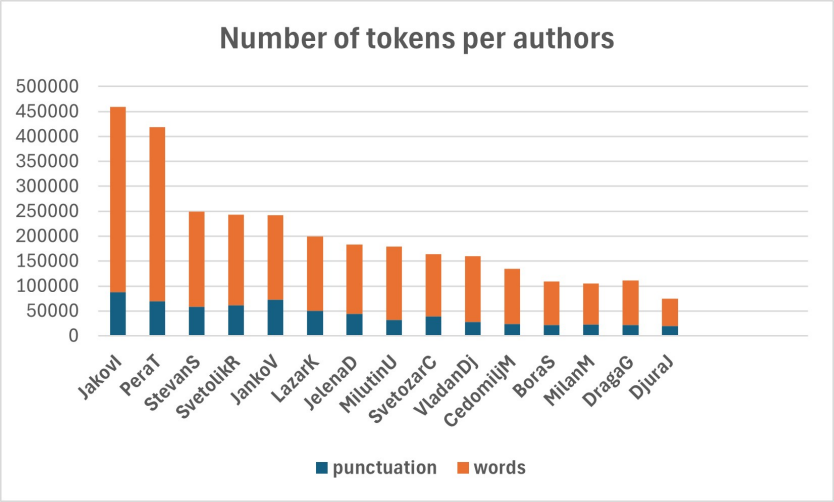

Figure 1. The length of writers' corpora measured by the number of tokens: punctuation and special characters, and words.

3.1 Corpus size

The chart shown in Figure 1 indicates that J. Ignjatović is the most represented author in the 15AUTHORS corpus, with his works accounting for 15.14% of the total corpus. The next most represented author is P. Todorović, whose works comprise 13.81% of the corpus. In contrast, M. Đ. Milićević and Đ. Jakšić are represented by the shortest works, contributing 3.45% and 2.47%, respectively. Overall, the corpus consists of 21.53% punctuation and special characters, compared to 78.47% words. The highest proportions of punctuation and special characters are found in the works of J. Veselinović (30.03%) and Đ. Jakšić (25.81%), whereas P. Todorović (16.56%) and

V. Đorđević (17.23%) use them considerably less. No firm conclusions can be drawn regarding individual authors' tendencies to produce longer or shorter works, as the selection of texts for the SrpELTeC corpus was guided by various criteria, as discussed in Section 2.

3.2 Sentence length

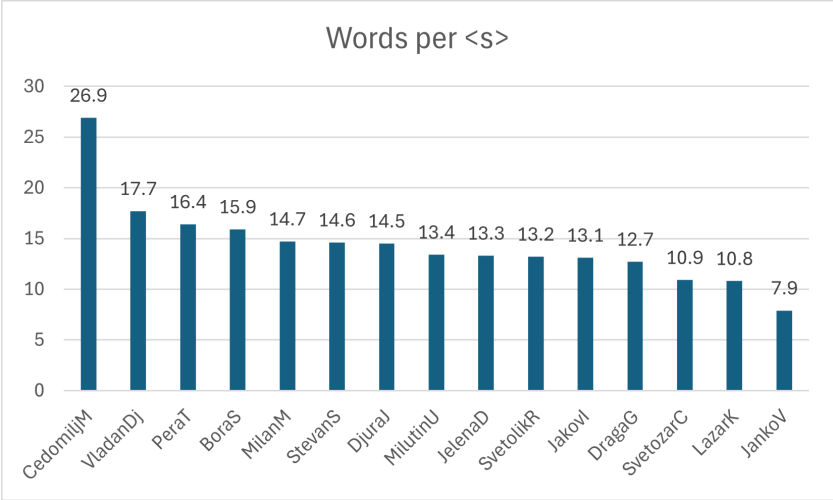

Figure 2. Sentence length measured by the number of words.

The average sentence length in the 15AUTHORS corpus is 13.4 words per sentence. Sentence length varies considerably across authors, ranging from very short sentences — an average of 7.9 words per sentence in the works of J. Veselinović — to notably long ones, with an average of 26.9 words per sentence in the works of Č. Mijatović. These two authors stand out in terms of sentence length. Among the remaining authors, the shortest sentences are found in the works of L. Komarčić (10.8 words per sentence), while the longest are observed in the works of V. Đorđević (17.7 words per sentence) (Figure 2). The average sentence length in the works of M. Uskoković is close to the overall corpus average.

3.3 Direct speech

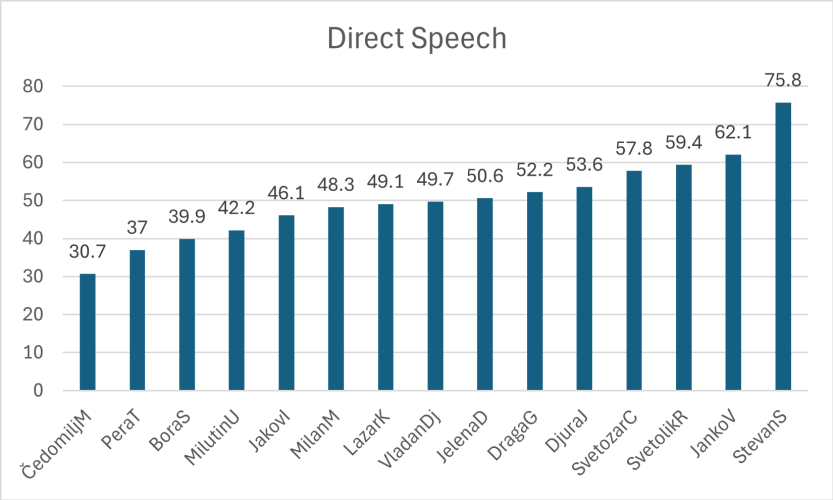

Figure 3. Percentage of paragraphs containing direct speech by author.

Some tools for the recognition and annotation of direct speech in the ELTeC corpus have already been developed, as mentioned in Section 1. In this study, however, our goal is not the precise identification of direct speech in the 15AUTHORS corpus, but rather an estimation of the proportion of paragraphs containing direct speech. We assume that each paragraph contains the speech of a single speaker and that direct speech appears at the beginning of the paragraph, marked by punctuation such as a hyphen or a quotation mark. An example of such a paragraph is given below:

> — Farewell, farewell! — tells one to another — Good night.

Although this approach is not highly precise, it provides useful insight into how different authors employ direct speech.

The results indicate that direct speech is least frequently used by Č. Mijatović and P. Todorović, whereas it is most prevalent in the works of S. Sremac and J. Veselinović (Figure 3). It is also worth noting that Č. Mijatović and P. Todorović tend to use the longest sentences, while J. Veselinović uses the

shortest ones. The sentence length in the works of S. Sremac is close to the corpus average (Section 3.2).

3.4 Lexical parameters

The lexical parameters of the authors in the 15AUTHORS corpus are presented in Table 1, including the number of words (*forms*), the number of distinct words (*types*), and the number of distinct lemmas. All data shown in Table 1 were computed using the Unitex software, except for the number of lemmas, which was calculated using TXM, as Unitex does not support automatic lemmatization. Due to differences in tokenization between these two tools, the counts of forms, types, and lemmas are not directly comparable.[6]

author	forms	types	lemmas	L	hapax	hapax%
JakovI	371243	38387	17545	0.8053	20958	5.65
PeraT	349721	43165	1804	0.7765	23820	6.81
JankoV	169685	23020	10953	0.7754	12989	7.65
StevanS	191094	27323	14431	0.7735	16125	8.44
SvetolikR	181467	26038	12165	0.7646	14859	8.19
BoraS	87601	14251	6806	0.7620	8492	9.69
DragaG	88733	14590	7188	0.7531	8522	9.6
JelenaD	139444	20906	10501	0.7514	11554	8.29
LazarK	149847	23812	11493	0.7504	14025	9.36
CedomiljM	110162	18312	8336	0.7396	10475	9.51
VladanDj	132100	23235	11002	0.7326	14027	10.62
SvetozarC	124390	23894	9816	0.7193	14670	11.79
DjuraJ	55579	12249	6573	0.7026	7791	14.02
MilanM	82356	17935	9459	0.6929	11200	13.6
MilutinU	147418	26354	12180	0.6836	15705	10.65

Table 1. Lexical parameters of writers from the 15AUTHORS corpus.

6. The TXM platform relies on TreeTagger (Schmid 2013) for automatic part-of-speech tagging and lemmatization; its Serbian model is described in (Utvić 2011; Stanković, Škorić, and Šandrih Todorović 2022).

Lexical index L is computed according to (Smith 1973):

$$L = \frac{-\sum_X f_x \frac{X}{N} log \frac{X}{N}}{logN}$$

where N represents the text size measured by the number of words, while X denotes the number of words with frequency f_x. In this way, the distribution of vocabulary is captured while normalizing for text length, allowing comparisons between texts of different sizes. The formula can be better understood through two extreme cases. If a text consists of only one word repeated N times, then $X = N$ and $L = 0$. Conversely, if no word is repeated, then $X = 1$ and $f_x = N$, resulting in $L = 1$.

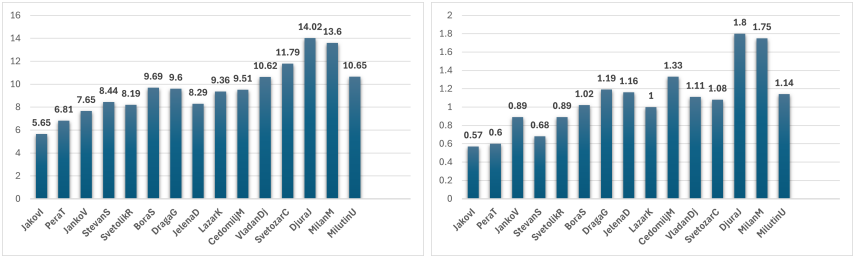

Figure 4. Percentage of word forms with frequency: one (left); greater than the median (right). Writers are sorted by the value of the index L, from highest to lowest.

Figure 4 indicates that the use of *hapax legomena* in a text is not directly connected with the value of L: the percentage of word forms with frequency one is highest in the works of Đ. Jakšić and M. Đ. Milićević (whose contribution to the size of the whole corpus is the lowest), while the percentage of these words in the works of J. Ignjatović and P. Todorović is the lowest (while their works contribute the most to the size of the corpus). This suggests that the index L is not entirely successful in neutralizing the length of the compared texts, and that a more accurate comparison could be obtained only by comparing samples of the same size.

If we arrange all words by frequency, from the lowest ($f = 1$) to the highest, the median corresponds to the frequency at which words with lower frequencies account for half of the text, while words with higher frequencies account for the other half. As shown in Figure 4, the percentage of words

with frequencies above the median is again highest in the works of Đ. Jakšić and M. Đ. Milićević, and lowest in the works of J. Ignjatović and P. Todorović. However, when considering all authors, there is no clear relationship between the percentage of *hapax legomena* and the proportion of words with frequencies above the median.

3.5 Frequency of word classes

Word classes are distributed in the 15AUTHORS corpus into 16 classes according to the guidelines of the *Universal Dependencies* project (Marneffe et al. 2021). The frequency of occurrences of these classes in the 15AUTHORS is as follows:

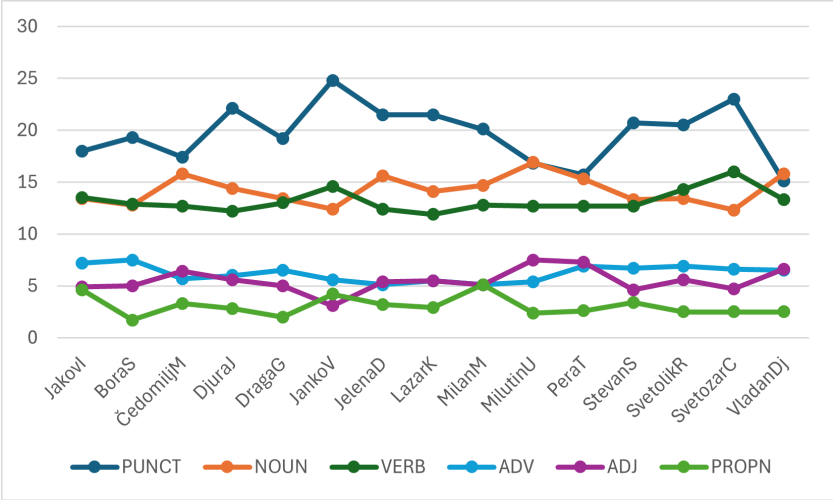

Figure 5. The participation of punctuation marks and meaningful words in the texts of individual authors.

- punctuation – PUNCT, $f = 567\,364$ (19.3%),
- nouns – NOUN, $f = 417\,261$ (14.2%);
- verbs – VERB, $f = 88\,330$ (13.2%);
- adpositions (prepositions) – ADP, $f = 188\,284$ (6.4%);

- adverbs – ADV, $f = 186\,921$ (6.4%);
- auxiliaries – AUX, $f = 179\,460$ (6.1%);
- coordinating conjunctions – CCONJ, $f = 177\,297$ (6.0%);
- adjectives – ADJ, $f = 162\,709$ (5.5%);
- pronouns – PRON, $f = 147\,878$ (5.0%);
- particles – PART, $f = 138\,947$ (4.7%);
- determiners – $f = 134\,734$ (4.6%);
- subordinating conjunctions – SCONJ, $f = 115\,004$ (3.9%);
- proper nouns – PROPN, $f = 93\,807$ (3.2%);
- numerals (cardinal) – NUM, $f = 22\,764$ (0.8%);
- interjections – INTJ, $f = 8\,953$ (0.3%);
- other (foreign words, abbreviations, etc.) – X, $f = 7\,969$ (0.3%).

Figures 5 and 6 show the percentage of occurrences of these classes by author. Punctuation marks, nouns, and verbs stand out clearly from all other word classes, being the most frequently used by all authors. The use of punctuation varies considerably: in the works of J. Veselinović, as many as 24.8% of all tokens are punctuation marks, while in the works of V. Đorđević and P. Todorović they account for just over 15% of tokens. The proportion of verbs varies less across authors—the lowest percentage is 11.9% (in the works of L. Komarčić), and the highest is 16.0% (in the works of S. Ćorović). M. Uskoković uses more nouns (16.9%), compared to an average of 14.2% across all authors.

Adverbs, adjectives, and proper nouns are used considerably less frequently than nouns and verbs. The use of adverbs is fairly consistent, close to the corpus average of 6.4%, with a few deviations: 7.5% in the works of B. Stanković, and 5.1% in the works of J. Dimitrijević and M. Đ. Milićević. In J. Veselinović's works, adjectives account for only 3.1% of the text, compared to the overall corpus average of 5.5%. Higher-than-average use of adjectives is observed in the works of M. Uskoković (7.5%) and P. Todorović (7.3%). For proper nouns, the average is 3.2%; M. Đ. Milićević uses more (5.1%), while B. Stanković uses the fewest (1.7%).

Cardinal numbers, interjections, and "other" words are used considerably less frequently than other functional words, ranging from zero to 1.2%, which corresponds to the frequency of cardinal numbers in the works of Č. Mijatović. The use of other functional words varies among authors. For instance, the average frequency of nominal pronouns is 5%: P. Todorović and M. Đ. Milićević use them less frequently (4.3%), while J. Veselinović uses them more often (6.2%). S. Ranković uses fewer auxiliary verbs (4.3%) compared to the average (6.1%), whereas J. Ignjatović and D. Gavrilović use them more frequently (7.4%).

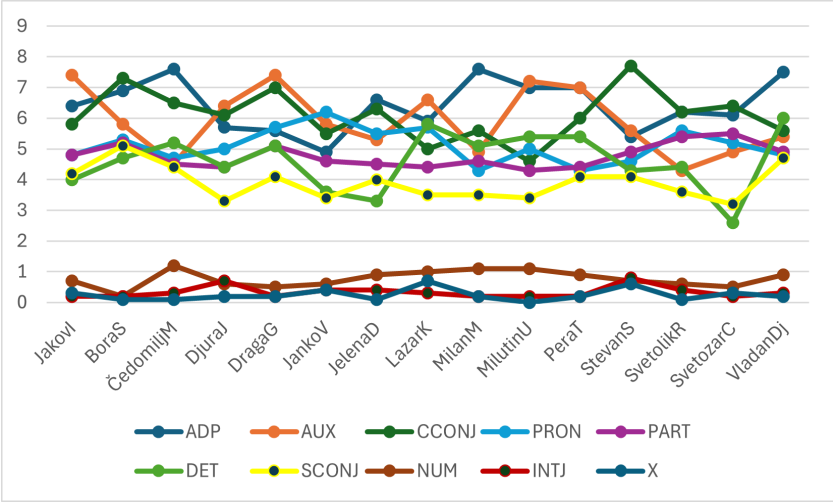

Figure 6. The participation of functional words in the texts of individual authors.

3.6 Nouns

The twenty most frequent nouns in the 15AUTHORS corpus, in the descending order of frequency, are: *ruka* 'hand', *kuća* 'house', *glava* 'head', *dan* 'day', *oči* 'eyes', *bog* 'god', *čov(j)ek* 'man', *ljudi* 'men', *put* 'road', *žena* 'woman/wife', *srce* 'heart', *r(ij)ečˇ* 'word', *lice* 'face', *otac* 'father', *glas* 'voice', *godina* 'year', *majka* 'mother', *strana* 'side', *vreme* 'time/weather', and *sv(ij)et* 'world'. Three of these nouns appear among the twenty most frequent nouns for all authors: *ruka* 'hand' (ranks 1–11), *dan* 'day' (ranks 1–12), and *oči* 'eyes' (ranks 1–15). The noun *vreme* 'time/weather' appears among the top twenty nouns for only four authors. In L. Komarčić's list of the twenty most frequent nouns, sixteen also occur among the twenty most frequent nouns in the entire 15AUTHORS corpus, whereas in Č. Mijatović's list, ten nouns are "new," i.e., they do not appear among the twenty most frequent nouns in the corpus as a whole.

Most authors (11 in total) most frequently use a noun that belongs to the set of the twenty most frequent nouns in the 15AUTHORS corpus as a whole. Exceptions include Đ. Jakšić, whose most frequent noun is *gospodin* 'mister/sir', D. Gavrilović, whose most frequent noun is *ljubav* 'love', and P. Todorović and Č. Mijatović, who most frequently use the noun *knez* 'prince'.

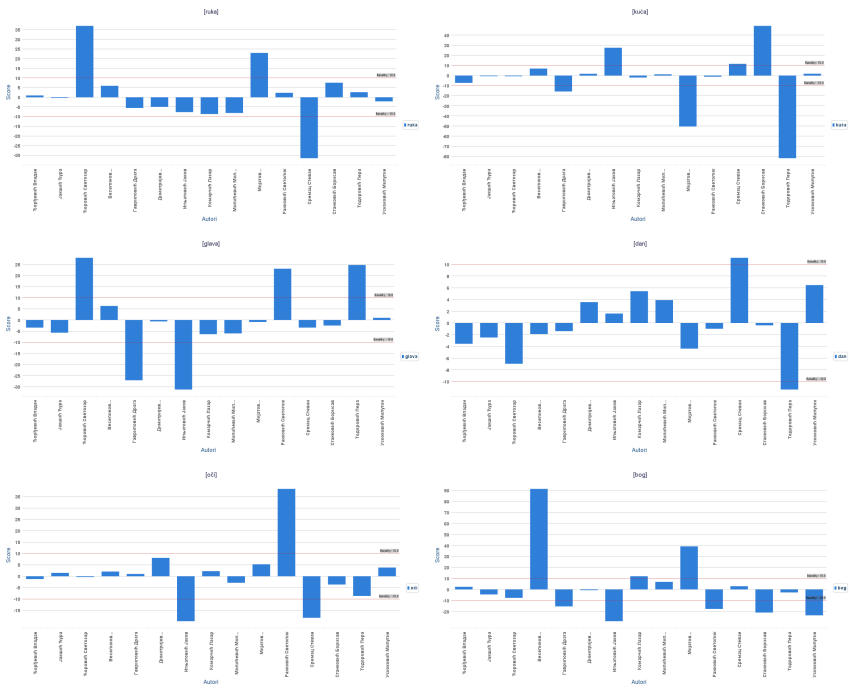

Figure 7. The specificity of use of the nouns *ruka* ‘hand’, *kuće* ‘house’, *glava* ‘head’, *dan* ‘day’, *oči* ‘eyes’, *bog* ‘god’ by 15 authors.

Two authors have as their second most frequent noun a word that does not appear among the twenty most frequent nouns in the 15AUTHORS corpus: *čiča* ‘old man/uncle’ in the works of Đ. Jakšić and *pop* ‘priest’ in the works of S. Sremac. Nouns that are frequent in individual authors’ lists but do not belong to the overall top twenty include *duša* ‘soul’ (appearing in the lists of six authors) and *zemlja* ‘land/country’ (in the lists of five authors). It is noteworthy that in the female author D. Gavrilović’s list of the twenty most frequent nouns, in addition to *ljubav* ‘love’, *srce* ‘heart’, and *duša* ‘soul’, there are also several nouns denoting family relations: *žena* ‘woman/wife’, *muž* ‘husband’, *d(ij)ete* ‘child’, *d(j)eca* ‘children’, and *sestra* ‘sister’. A similar pattern is observed in the works of J. Dimitrijević, the other female author, who, alongside *duša* ‘soul’ and *srce* ‘heart’, frequently uses *majka* ‘mother’, *otac* ‘father’, *sin* ‘son’, *kćer* ‘daughter’, and *hanuma* ‘khanum’. Among nouns

denoting emotions, only two appear in the top 20 lists: *ljubav* 'love' (used frequently by D. Gavrilović and M. Uskoković) and *strah* 'fear' (used by B. Stanković).

The *specificity* index enables the comparison of the use of selected linguistic units (words, phrases, constructions) across individual authors. This index reflects the degree of deviation from the expected frequency of a given unit in the works of an author and is computed on the basis of the following parameters: f — the number of occurrences of a selected unit in the works of a given author; F — the total number of occurrences of that unit in the corpus as a whole; t — the total number of units in the works of the given author; and T — the total number of units in the entire corpus. In order to compare the frequencies of linguistic units across corpus partitions of different sizes, normalization is required. A common approach is to calculate relative frequency as the ratio between the absolute frequency of a unit in a given partition and the size of that partition, which corresponds to a *normal distribution*. However, it has been observed that the distribution of linguistic units more closely follows a *hypergeometric distribution*. The probability that a selected unit occurs f times in a given author's subcorpus (i.e., partition) is therefore:

$$Pr(X = f) = \frac{\binom{F}{f}\binom{T-F}{t-f}}{\binom{T}{t}}$$

Positive values of the specificity index indicate that a linguistic unit occurs more frequently than expected, whereas negative values indicate lower-than-expected frequency. Further details on the specificity index can be found in (Jaćimović 2019).

Specificity indices, as illustrated in Figure 7, show that the most frequent nouns — *ruka* 'hand', *kuća* 'house', *glava* 'head', *dan* 'day', *oči* 'eyes', and *bog* 'god' — occur within expected ranges in the works of V. Đorđević, Đ. Jakšić, J. Dimitrijević, L. Komarčić, and M. Đ. Milićević. The only notable deviation concerns the noun *bog* 'god' in the works of J. Veselinović and M. Uskoković: the former uses it considerably more often than expected, while the latter uses it considerably less. S. Ćorović uses the nouns *ruka* 'hand' and *glava* 'head', and S. Ranković *glava* 'head' and *oči* 'eyes', significantly above expected levels. In contrast, in the works of D. Gavrilović, *kuća* 'house', *glava* 'head', and *bog* 'god' occur considerably below expectation. In Č. Mijatović's works, *ruka* 'hand' and *bog* 'god' occur more often than expected, whereas *kuća* 'house' occurs less frequently. In S. Sremac's works, *dan* 'day' occurs more often, while *ruka* 'hand' and *oči* 'eyes' occur less often than expected.

In B. Stanković's works, *kuća* 'house' occurs above expected levels and *bog* 'god' below them, whereas in P. Todorović's works, *glava* 'head' occurs more often, and *kuća* 'house' and *dan* 'day' less often than expected. The greatest deviations are observed in the works of J. Ignjatović: *kuća* 'house' occurs more often than expected, while *glava* 'head', *oči* 'eyes', and *bog* 'god' occur less often.

3.7 Verbs

The twenty most frequent verbs in the 15AUTHORS corpus, listed in the descending order of frequency, are: *moći* 'be able', *reći* 'say', *znati* 'know', *vid(j)eti* 'see', *imati* 'have', *doći* 'come', *kazati* 'tell', *ići* 'go', *гледати* 'watch', *govoriti* 'speak', *misliti* 'think', *čuti* 'hear', *морати* 'have to', *početi* 'begin', *dati* 'give', *trebati* 'need', *uzeti* 'take', *ostati* 'stay', *nemati* 'not have', and *otići* 'leave'. Five of these verbs rank among the twenty most frequent verbs for all authors: *moći* (ranks 1–5), *znati* (1–6), *vid(j)eti* (1–11), *imati* (2–11), and *doći* (2–14). The verb *reći* 'say' appears among the top twenty verbs for all authors (ranks 1–9), except for B. Stanković. The set of the twenty most frequent verbs in the works of J. Ignjatović nearly coincides with the overall set, with the only exception being *гледати* 'watch', which is absent. The smallest number of verbs from the above list — exactly sixteen — appears in the works of five authors: J. Veselinović, Č. Mijatović, B. Stanković, M. Uskoković, and S. Ćorović.

The most frequent verb used by each of the 15 authors appears in the list of the twenty most frequent verbs in the corpus as a whole, with the exception of S. Sremac, whose most frequent verb is *velim* 'say'. This verb ranks 21st in the overall corpus. Verbs that are frequent in the works of individual authors but do not belong to the overall top twenty include *pogledati* 'take a look' (attested in eight authors' works), as well as *odgovoriti* 'answer', *pitati* 'ask', and *ustati* 'stand up' (each attested in five authors' works). It is also worth noting that both the overall top twenty list and the top twenty lists of individual authors include verbs derived by prefixation, such as *gledati* 'watch' and *pogledati* 'take a look', *govoriti* 'speak' and *odgovoriti* 'answer', as well as *pitati*, *upitati* 'ask', and *zapitati* 'wonder'. In addition, aspectual verb pairs are present, including *osetiti* / *osećati* 'feel', *stati* / *stajati* 'stand', and *doći* / *dolaziti* 'come'.

No verb denoting emotion appears in the overall top twenty list. However, the verb *voleti* 'love' is among the twenty most frequent verbs in the works of D. Gavrilović (rank 10), J. Dimitrijević (rank 9) — both female authors

— and M. Uskoković (rank 5). In addition, D. Gavrilović's list also includes *ljubiti* 'love/kiss' (rank 11). The verb *osećati* 'feel' appears among the top twenty verbs in the works of B. Stanković (rank 12) and M. Uskoković (rank 13), while B. Stanković's list also contains its aspectual counterpart *osetiti* (rank 19). The top twenty list of J. Dimitrijević includes the verb *plakati* 'cry' in the final position.

3.8 Adjectives

The twenty most frequent adjectives in the 15AUTHORS corpus, listed in the descending order of frequency, are: *drugi* 'second/other', *velik* 'big', *star* 'old', *dobar* 'good', *l(ij)ep* 'beautiful', *mlad* 'young', *ceo/cijel* 'whole/entire', *prvi* 'first', *crn* 'black', *nov* 'new', *mali* 'small', *pun* 'full/plump', *beo/bijel* 'white', *živ* 'alive/vivid', *isti* 'same', *srpski* 'Serbian', *po(s/š)(l/lj)ednji* 'last', *turski* 'Turkish', *sre(ć/t)an* 'happy', and *čitav* 'whole/entire'. Possessive adjectives derived from proper names (e.g. *Kočin* 'belonging to Koča' and *Darinčin* 'belonging to Darinka') were excluded from the top twenty lists of individual authors. As a result, five of the above adjectives appear among the top twenty adjectives for all authors: *drugi*, *velik*, *star*, *l(ij)ep*, and *mlad*. In addition, three adjectives occur among the top twenty for all but one author: *dobar* (ranked 27th in the works of B. Stanković), *prvi* (25th in the works of Đ. Jakšić), and *mali* (25th in the works of P. Todorović). The adjective *čitav* 'whole/entire' appears in the top twenty lists of only three authors; however, in the works of S. Ćorović it ranks as high as third, in contrast to its 20th position in the whole corpus.

The most frequent adjective used by each of the 15 authors appears in the list of the twenty most frequent adjectives in the corpus as a whole. For the majority of authors—eleven in total—this adjective is *drugi* 'second/other', which occupies the top position in the overall list. Adjectives that are frequent in the works of individual authors but do not belong to the overall top twenty include *dug* 'long' (attested in six authors' works), as well as *strašan* 'horrible/scary', *težak* 'heavy', *veseo* 'cheerful', and *ženski* 'female' (each attested in five authors' works). The adjectives *crn* 'black' and *beo/bijel* 'white', which are among the twenty most frequent adjectives in the corpus overall, appear in the top twenty lists of most individual authors — specifically, nine of them. Neither adjective appears in the lists of J. Ignjatović and D. Gavrilović; only *crn* 'black' appears in the lists of P. Todorović, S. Ranković, and S. Ćorović; while only *beo/bijel* 'white' appears in S. Sremac's list.

Two demonyms appear among the twenty most frequent adjectives in the 15AUTHORS corpus: *srpski* 'Serbian' (rank 16) and *turski* 'Turkish' (rank 18). Both adjectives are included in the top twenty lists of only two authors, J. Ignjatović and P. Todorović; in the latter's list, *balkanski* 'related to Balkan' also appears, ranked 20th. The adjective *srpski* 'Serbian', but not *turski* 'Turkish', occurs in the top twenty lists of M. Đ. Milićević, L. Komarčić, and Č. Mijatović. Conversely, *turski* 'Turkish', but not *srpski* 'Serbian', appears in the lists of J. Veselinović and J. Dimitrijević; the latter's list additionally includes the demonyms *francuski* 'French' and *evropski* 'European'.

Other colors, apart from the adjectives *crn* 'black' and *beo/bijel* 'white', do not appear in the top twenty lists of individual authors, with the exception of Č. Mijatović's list, which includes *crven* 'red' and *zelen* 'green', as well as *srebrn* 'silver' and *zlatan* 'golden', which can also denote colors (see Section 3.9). Adjectives that appear in the top twenty lists of specific authors but not necessarily in the overall top twenty may reflect the particular style or thematic focus of those authors. For example:

- V. Đorđević's list includes *strašan* 'awesome', *силан* 'powerful', *диван* 'wonderful', *cjajan* 'brilliant', and *красан* 'beautiful'.
- Đ. Jakšić's list contains *sirot* 'poor', *bled* 'pale', *gladan* 'hungry', *sam* 'alone', *nesrećan* 'unhappy', and *tužan* 'sad'.
- D. Gavrilović's list features *sirot* 'poor', *bogat* 'rich', *valjan* 'worthy', *čist* 'clean/pure', *veseo* 'cheerful', and *pošten* 'honest'.
- S. Ćorović's list includes *širok* 'wide', *golem* 'big', *dug* 'long', *tanak* 'thin', *симан* 'tiny', *debeo* 'thick', and *јак* 'strong'.

3.9 Adjectives denoting colors

A specific marker used in Serbian electronic dictionaries for adjectives denoting colors allows us to examine their use in the works of fifteen selected authors. A total of 3,057 occurrences of these adjectives were retrieved from the 15AUTHORS corpus, representing 109 different lemmas. The adjectives *crn* 'black' and *beo/bijel* 'white' were not taken into account. Other commonly used color terms, in the descending order of frequency, include: *crven* 'red', *zelen* 'green', *zlatan* 'golden', *plav* 'blue', *rumen* 'light red', *srebrn* 'silver', *žut* 'yellow', *siv* 'gray', *modar* 'dark blue', *ružičast* 'pink', *smeđ* 'brown', *sur* 'lead gray', *beličast* 'whitish', *žućkast* 'yellowish', *riđ* 'russet', *zelenkast* 'greenish', *ljubičast* 'purple', *crvenkast* 'reddish', *plavetan*, *plavičast* 'bluish'. We did not disambiguate senses, so adjectives such as *zlatan* 'golden' and *srebrn* 'silver' may refer to material rather than color.

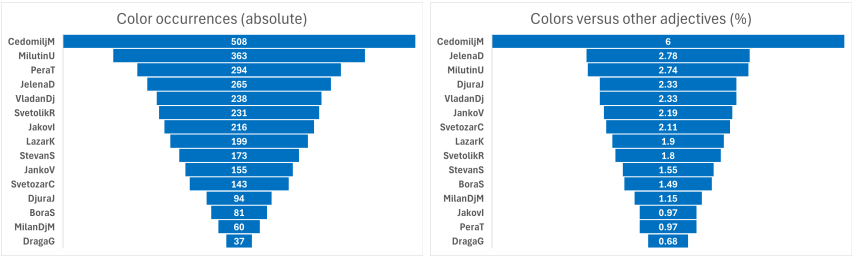

Figure 8. The use of adjectives denoting colors by individual authors.

In absolute terms, Č. Mijatović uses the highest number of color adjectives (508), as well as the greatest number of different ones (35). D. Gavrilović uses these adjectives the least — 37 occurrences and only 13 different (the same number of different is also found in the works of M. Đ. Milićević) (Figure 8, left). When compared with the total number of adjectives used by each author, the proportion of color adjectives is highest in Č. Mijatović's works and lowest in D. Gavrilović's works (Figure 8, right).

Among 109 different adjectives denoting colors, there are 17 adjectives that indicate that something is a bit of a certain color or tends towards that color, and they occur in various forms: *beličast* 'whitish', *žu(t/ć)kast* 'yellowish', *zagasitorumenkast* 'dark reddish', *zelenkast* 'greenish', *zlatkast* 'a bit golden', *morast* 'darkish', *mrkajast* 'darkish', *otvorenozelenkast* (light greenish', *plavetnikast*, *plavičast*, *plav(k/č)ast*, *plavuš(k)ast* 'bluish', *purpurast* 'purplish', *siv(k)ast* 'grayish', *srebrnkast* 'a bit silver', *crvenkast* 'reddish', *crnkast* 'blackish'.

We identified eleven adjectives of Turkish origin in the 15AUTHORS corpus: *alev*, *alen*, *alov* 'bright red', *đuvez* 'dark red', *karpus* 'watermelon-colored', *kremzli* 'a type of red', *maven* 'clear blue', *mor* 'purple', *pembe* 'pink', and *skrl(a/e)tan* 'bright red'. It is interesting to note that the adjective *alev* and its variants occur twelve times in the 15AUTHORS corpus, but only twice in the expression *aleva paprika* 'cayenne pepper', namely in M. Uskoković's *Čedomir Ilić* and Đ. Jakšić's *Čiča Tima*. All other occurrences refer to garments — *čarape* 'socks', *šalvare* 'a type of women's trousers', *jagluk* 'kerchief', *libade* 'a type of women's jacket', and *fes* 'fez hat'.[7]

7. A notable difference from present-day usage: in the SrpKor21 corpus, the adjective *alev* occurs 976 times modifying a noun, and that noun is always *paprika* 'pepper'.

Ten indeclinable adjectives, typically of foreign origin, appear in several variants: *bra(o|u)n* 'brown', *violet* 'violet', *grao* 'gray', *đuvez*, *zejtini* 'olive (colored)', *karpus*, *kremzli*, *mor*, *pembe*, and *roz(a|e)* 'pink'.

Among the retrieved adjectives denoting colors, 31 are compounds. Ten of these refer to combinations of colors: *beličastosrebrn* 'white-silver', *žućkastozelen* 'yellowish-green', *zelenkastoplavetan* 'greenish-blue', *zelenožut* 'green-yellow', *plavobeo* 'blue-white', *plavosiv* 'blue-gray', *sivoplav* 'gray-blue', *crvenkastosiv* 'reddish-gray', *crnožut* 'black-yellow', and *čivitasto-modar* 'indigo blue'. Color modification is attested in 19 adjectives. These include *bledo-* 'pale-' (*-žut* '-yellow', *-zelen* '-green', *-ljubičast* '-purple', *-plav* '-blue', *-rumen* '-bright red'), *zagasitorumenkast* 'dull red', *zatvorenožut* 'dark yellow', *mrko-* 'dark-' (*-žut* '-yellow', *-siv* '-gray'), *mutnoplav* 'cloudy blue', *otvoreno-* 'clear-' (*-zelenkast* '-greenish', *-ljubičast* '-purple'), *polu-* 'half-' (*-zelen* '-green', *-zlatan*' golden'), *snežnobeo* 'snow white', and *tamno-* 'dark-' (*-zelen* '-green', *-plav* '-blue', *-siv* '-gray', *-crven* '-red'). In the remaining two cases, color functions as a modifier of another adjective: *žućkastobled* 'yellowish pale' and *rumenosjajan* 'bright red'.

3.10 The rhetorical figure of simile

In our previous work (Krstev 2021b; Krstev, Stanković, and Marković 2023), we described a tool for the recognition and annotation of the rhetorical figure of simile in Serbian texts, based on a formal description of these figures, electronic dictionaries, and finite-state transducers. Using this tool, we identified 837 instances of simile in the 15AUTHORS corpus, of which 521 are adjectival and 315 verbal.

Among the 301 distinct adjectival similes, the most frequent one is *beo kao sneg* 'white as snow', occurring 30 times. It is followed by *bled kao smrt* 'pale as death' (25 occurrences), *hladan kao led* 'cold as ice' (15), *bled kao krpa* 'pale as a rag' (10), *beo kao mleko* 'white as milk' (9), *mlad kao kaplja* 'young as a drop' (8), and *crven kao krv* 'red as blood' (7). The similes *dobar kao dan* 'good as day', *žut kao vosak* 'yellow as wax', *ljut kao ris* 'angry as a lynx', and *mek kao pamuk* 'soft as cotton' each occur six times in the corpus. In total, 223 adjectival similes appear only once. The most frequently used adjectives in adjectival similes are *beo* 'white' and *crn* 'black', each occurring in 12 similes. They are followed by *bled* 'pale', *crven* 'red', and *rumen* 'bright red', each appearing in 10 similes, and *čist* 'clean/pure', which appears in nine. The nouns *anđeo* 'angel', *jagnje* 'lamb', *ptica* 'bird', and *sunce* 'sun' are used as a base for comparison in five similes each. The nouns *vatra* 'fire',

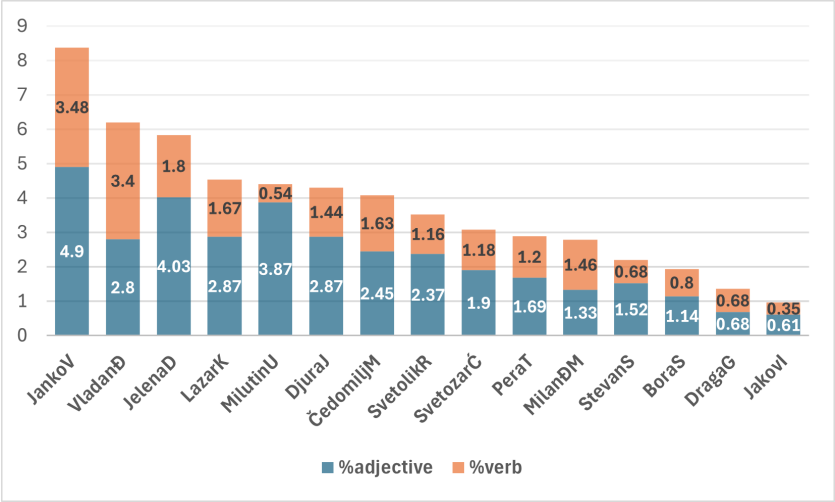

Figure 9. Relative frequency of adjectival and verbal similes in individual authors' works.

dan 'day', *žeravica* 'ember', *zemlja* 'land/soil', *nebo* 'sky', and *smrt* 'death' each occur as vehicles in four similes.

Among the 164 distinct verbal similes, the most frequent one is *drhtati kao prut* 'tremble like a stick', occurring 20 times. It is followed by *plakati kao dete* 'cry like a child' (13 occurrences); *zaspati kao zaklan* 'fall asleep as if slaughtered' and *ciknuti kao zmija* 'shriek like a snake', each with 8 occurrences; *viti (se) kao crv* 'wriggle like a worm' and *prebledeti kao krpa* 'turn pale as a rag', each with 7 occurrences; *slušati kao dete* 'obey like a child' (6); and *smejati (se) kao lud* 'laugh like a madman' (5). In total, 99 verbal similes occur only once in the corpus. With regard to verbs used in verbal similes, the most frequent is *stajati* 'stand', which appears in five different similes. It is followed by *živeti* 'live' and *izgledati* 'look (like)', each occurring in four similes. The following verbs appear in three similes each: *viti se* 'wriggle', *zaspati* 'fall asleep', *jurnuti* 'rush', *padati* 'fall', *plakati* 'cry', *sejati* 'sow', *skočiti* 'jump', and *spavati* 'sleep'. As for the nouns used, *dete* 'child' occurs in nine verbal similes; *zmija* 'snake' and *crv* 'worm' in six each; and *put* 'road' and *strela* 'arrow' in five similes each.

The largest number of similes in absolute terms — 142 in total — was retrieved from the works of J. Veselinović. This author also exhibits the highest

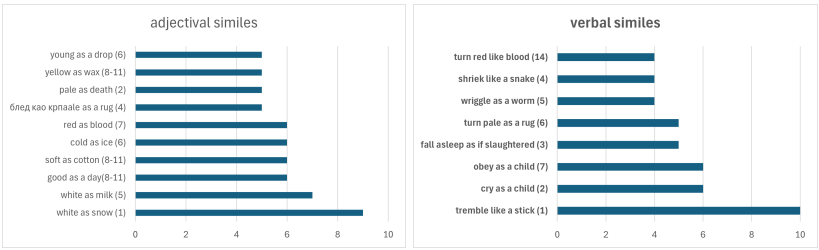

Figure 10. Similes that appear in the works of most authors (in brackets is the rank of these expressions by absolute frequency in the corpus).

relative frequency[8], as shown in Figure 9. In contrast, only 12 similes were identified in the works of D. Gavrilović, while the lowest relative frequency is observed in the works of J. Ignjatović. Furthermore, both the absolute frequencies of adjectival (83) and verbal similes (59), as well as their relative frequencies, are highest in J. Veselinović's works. Some authors show a clear preference for adjectival over verbal similes (J. Dimitrijević, and especially M. Uskoković), whereas V. Đorđević uses considerably more verbal than adjectival similes.

The most frequent adjectival simile, *beo kao sneg* 'white as snow', is used by nine authors. It is followed by *beo kao mleko* 'white as milk', which appears in the works of seven authors (see Figure 10, left). The most frequent verbal simile, *drhtati kao prut* 'tremble like a stick', is used by ten authors. It is followed by *plakati kao dete* 'cry like a child' and *slušati kao dete* 'obey like a child', each used by six authors (see Figure 10, right).

3.11 The use of foreign languages

Volunteers who proofread and corrected scanned and OCR-processed texts annotated segments in foreign languages using the XML tags `<foreign>` and `</foreign>`, assigning a two-letter language code to the `xml:lang` attribute. As most volunteers were not specialists in the history of the Serbian language, they were instructed to use the Old Church Slavonic code "cu" for all variants of Serbian not written in Vuk's Cyrillic alphabet. Occasionally, texts contained segments in foreign languages such as Turkish or German,

8. The relative frequency was calculated as the number of similes per 10,000 running words.

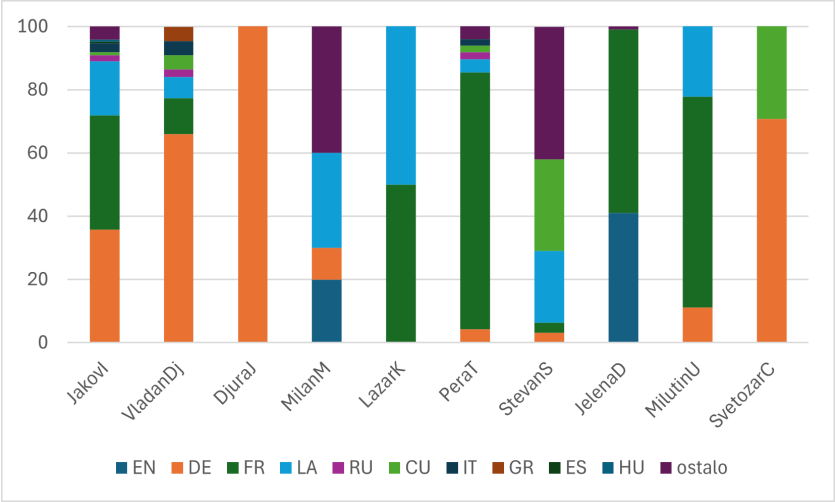

Figure 11. Use of different languages by authors expressed as a percentage of the total number of segments in a foreign language.

but written in Cyrillic and adapted to pronunciation (e.g., *мајн грус*, from the German 'Meine Grüße'). In our analysis, all such instances were classified as "other." While repetition of foreign-language phrases is generally not expected, the French phrase "Mon Dieu" appeared 17 times (always in J. Dimitrijević's novel *Nove*), and "corpus delicti" appeared four times (in S. Sremac's novels *Pop Ćira i pop Spira* and *Ivkova slava*).

We used these tags to analyze the use of foreign languages in the 15AUTHORS corpus. Our findings show that five authors — D. Gavrilović, J. Veselinović, S. Ranković, Č. Mijatović, and B. Stanković — did not use foreign languages in their works. However, this does not allow us to conclude whether these authors were familiar with foreign languages or not; a notable example is Č. Mijatović.[9]

Some authors frequently used foreign languages — 291 occurrences were annotated in the works of J. Ignjatović and 112 in those of J. Dimitrijević — while others used them only rarely, with just one occurrence in Đ. Jakšić's works and two in those of L. Komarčić. The remaining authors used foreign languages to varying extents: P. Todorović (48 occurrences), V. Đorđević

9. Čedomilj Mijatović on Wikipedia.

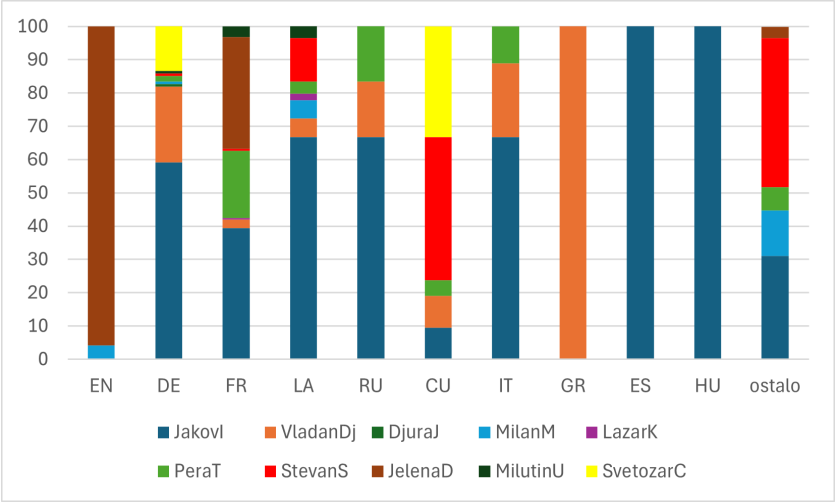

Figure 12. Language use by different authors expressed as a percentage of the number of segments in that language.

(44), S. Sremac (31), S. Ćorović (24), M. Đ. Milićević (10), and M. Uskoković (9). As shown in Figure 11, some authors incorporated segments in multiple languages. J. Ignjatović used all languages identified in the 15AUTHORS corpus except English and Greek; V. Đorđević used all except English, Spanish, and Hungarian; and P. Todorović used all except English, Greek, Spanish, and Hungarian. By contrast, J. Dimitrijević used only English and French. German appears in the works of eight authors, while French is used by six.

As shown in Figure 12, English appears only in the works of J. Dimitrijević, apart from two segments in M. Đ. Milićević's works that were erroneously annotated. In contrast, German, Latin, and French were used by several authors. Russian appears in the works of three authors, although only six instances were identified in total. The figure also indicates that S. Sremac and J. Ignjatović more frequently employed foreign-language segments written in Cyrillic script according to pronunciation than other authors.

3.12 People's roles

As noted in Section 1, the SrpELTeC corpus, as well as its subset the 15AUTHORS, includes automatic annotation of seven classes of named entities. One

of these classes refers to roles of people appearing in novels (ROLE), whether they are real or fictional. Serbian electronic morphological dictionaries, in which such roles are associated with appropriate semantic markers, made it possible to classify all roles identified in the corpus into those denoting professions, positions in society or organizations, and titles or forms of address. In addition, these e-dictionaries enabled a further distinction between roles referring to men and those referring to women. Table 2 presents the twenty most frequent roles identified in the 15AUTHORS corpus, classified according to these criteria. The data indicate that professions, positions, and titles referring to men are used far more frequently than those referring to women, confirming preliminary findings reported in (Stanković, Krstev, and Vitas 2024).

prof/M	*F*	**prof/F**	*F*	**poz/M**	*F*	**pos/F**	*F*	**title/M**	*F*	**title/F**	*F*
priest	1892	maid	157	vizier	806	landlady	173	prince	2254	misses	1163
teacher	676	woman teacher	147	hajduk	589	wife if priest	159	mister	1693	Mrs	498
doctor	558	nun	48	boss	519	wife of notary	24	Mr.	1423	miss	293
serf	523	woman cook	46	father	403	girl pupil	18	duke	800	miss	289
peasant	481	woman milliner	40	pupil	356	prioress	15	pasha	739	khanum	244
monk	476	woman cook	33	minister	295	clerk	12	captain	686	khanum	207
servant	400	woman server	31	chief	251	overseer	10	despot	605	madam	185
soldier	280	maid	28	president	225			abbot	587	ma'am	145
clerk	260	cook	28	bishop	201			czar	478	misses	142
merchant	253	peasant woman	28	hodja	121			uncle	400	czarina	141
cop	236	maid	26	colonel	119			sultan	352	aunt	131
coachman	166	midwife	25	supervisor	97			aga	330	misses	125
officer	156	danseuse	24	deputy	93			general	262	princess	89
doctor	145	teacher	22	governor	77			count	260	pashinitza	78
professor	138	woman cook	21	peasant	60			bey	247	frau	75
craftsman	134	barmaid	19	adjutant	57			bey	227	young frau	59
lawyer	121	governess	19	nobleman	57			king	223	madam	56
archpriest	119	nanny	14	secretary	56			mullah	187	gracious	55
clerk	116	maid	14	president	54			uncle	152	countess	48
guard	103	danseuse	12	manager	53			knight	106	captain's wife	37

Table 2. English translation of the most frequent professions, positions and titles of men and women in the 15AUTHORS corpus. The Serbian roles as identified in the corpus are given in the Appendix.

To establish specificity of the use of nouns across the six classes, we selected those occurring in the corpus with a frequency greater than 10. The results are presented in Figure 13. It can be observed that professions referring to men (top row, left) occur less frequently than expected — that is, relative to their distribution in the overall corpus — in the works of J. Dimitrijević

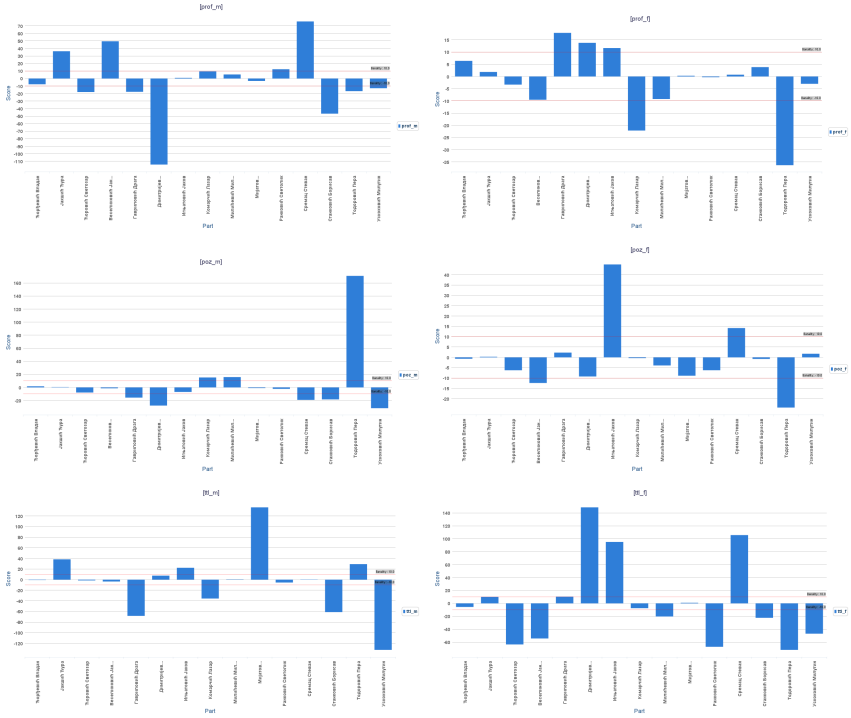

Figure 13. Specificity of the appearance of professions, positions and titles of men and women in the 15AUTHORS.

and B. Stanković, and significantly more frequently in those of J. Veselinović and S. Sremac. In the latter case, this is likely due to the frequent use of the noun *pop* 'priest', which is the second most frequent noun in Sremac's works (see Subsection 3.6). Professions referring to women (top row, right) occur significantly more often in the works of D. Gavrilović, J. Dimitrijević, and J. Ignjatović, and much less often in those of P. Todorović and L. Komarčić. As for positions referring to men (middle row, left), they are used far more frequently than expected by P. Todorović — *prince* being his most frequent noun — and much less frequently by M. Uskoković and J. Dimitrijević. Positions referring to women (middle row, right) occur more often than expected in J. Ignjatović's works, and less often in those of P. Todorović. Titles referring to men (bottom row, left) appear more frequently than expected in

the works of Č. Mijatović, and less frequently in those of M. Uskoković, J. Dimitrijević, and B. Stanković. There is also considerable deviation from expected frequencies in the use of titles referring to women (bottom row, right) across most authors: they are used more frequently by J. Dimitrijević, J. Ignjatović, and S. Sremac, and less frequently by P. Todorović, S. Ranković, S. Ćorović, J. Veselinović, B. Stanković, and M. Đ. Milićević.

From these observations, it can be concluded that roles referring to women, as well as other designations of women, are used significantly less frequently in the works of P. Todorović, and considerably more frequently in those of J. Ignjatović. In the works of B. Stanković and M. Uskoković, neither male nor female designations appear to play a significant role, as their frequencies are generally at or below expected levels across all categories. In contrast, in V. Đorđević's works, nouns from all categories consistently fall within the expected range.

3.13 Correspondence analysis

Correspondence analysis is widely used in the analysis of textual data. It is specifically designed for contingency tables, such as those that cross-tabulate texts and words. It can be applied to word forms, lemmas, or part-of-speech tags within a corpus. In textual data analysis, correspondence analysis is commonly used for visualization through two-dimensional factorial maps.

TXM, which we used for processing, builds lexical tables in which the rows represent texts by specific authors and the columns represent selected words, while the corresponding cells contain their frequencies. Correspondence analysis reorganizes these data to facilitate visualization and provide clearer insight into their underlying structure. First, the center of gravity of the data cloud is calculated, relative to which the dispersion of the data is measured. Then, the principal axes of dispersion are determined, and the coordinates of the points are computed with respect to these axes. In a two-dimensional representation, the first two axes — those that account for the greatest dispersion — are displayed. The points in the plane derive from both sets of data, namely rows (authors) and columns (words). These axes have no intrinsic meaning; therefore, the positions of the points should be interpreted only in relation to one another.

We replicated the experiment presented in (Lavrentiev et al. 2021) by cross-tabulating authors with the 200 most frequent tokens from the 15AUTHORS corpus. The results are shown in Figure 14. During the data preparation, we normalized various quotation marks by merging them into a single

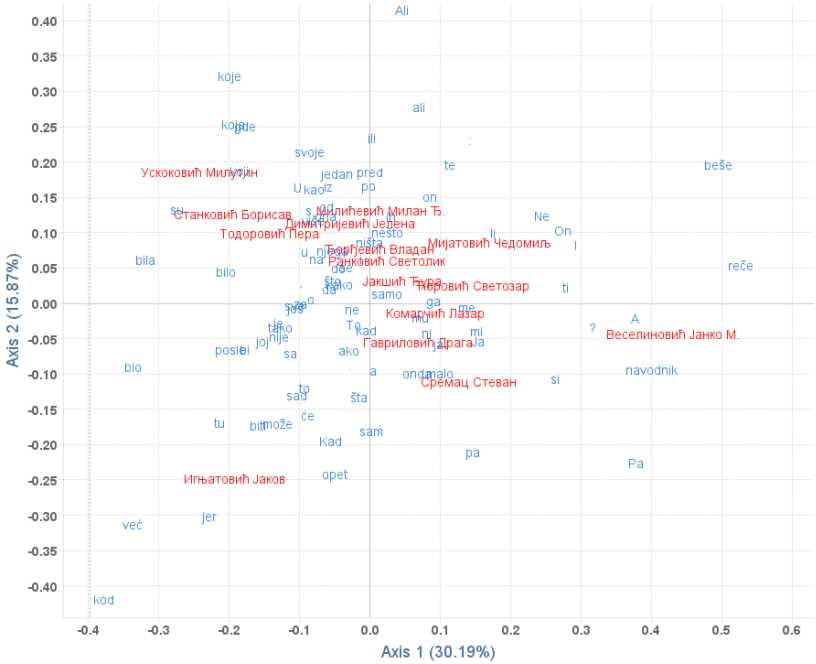

Figure 14. Correspondence analysis of authors crossed with the 200 most frequent tokens.

token—*наводник* ('quotation mark'). The figure shows that J. Veselinović is clearly separated from the other authors, while his relative proximity to S. Sremac, and to the point representing *наводник*, is consistent with the findings from Section 3.3, indicating that these two authors use direct speech more frequently than others. It can also be observed that J. Ignjatović is distinctly separated from all other authors.

A somewhat different pattern emerges when authors are cross-tabulated with the 100 most frequent nouns, verbs, and adjectives from the 15AUTHORS corpus (Figure 15). P. Todorović and Č. Mijatović are clearly separated from the other authors, as are J. Veselinović and S. Sremac, whose proximity was also observed in Figure 14. Also, M. Uskoković, D. Gavrilović, and J. Dimitrijević appear very close in this diagram; the lemma *žena* 'woman' serves as a key element linking them.

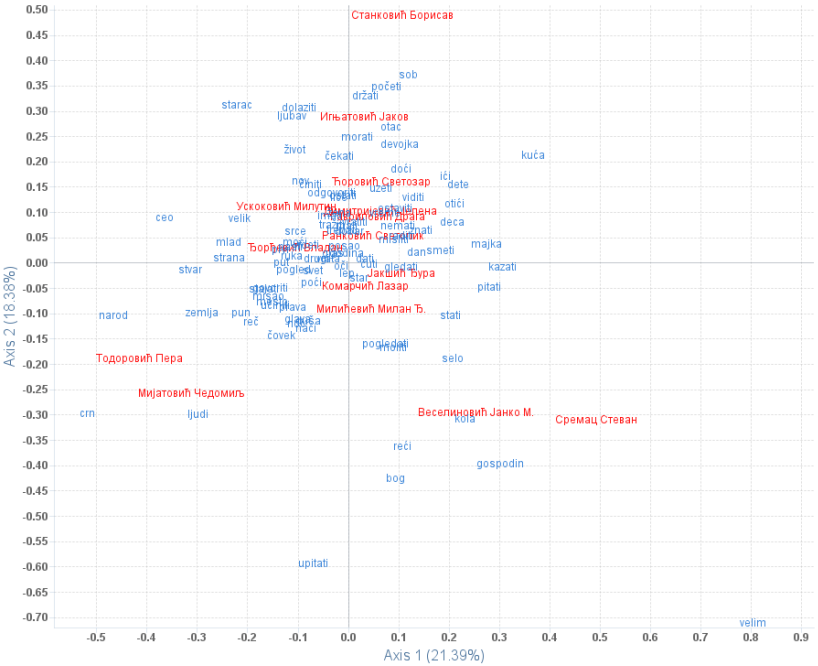

Figure 15. Correspondence analysis of authors crossed with the 100 most frequent lemmas (nouns, verbs and adjectives).

4 Conclusions

The aim of this paper is neither to evaluate the literary merits of the selected authors, nor to contrast our findings with the judgments of literary critics and readers. Rather, our goal is to present the digital imprints that authors leave in their works and to explain how we identified these imprints, which may serve as a starting point for further philological and computational research.

In our future research, we will focus on expanding our corpus. We have already compiled a collection of works by the same fifteen authors, including not only novels, but also short stories, travelogues, and memoirs. This expanded dataset will enable comparisons of digital imprints across different literary genres. Certain experiments should be repeated using balanced

samples, particularly those related to lexical richness (Section 3.4) and correspondence analysis (Section 3.13). For a more in-depth examination of individual authors' vocabularies, it would be beneficial to perform sense disambiguation. For example, is the noun *srce* 'heart', the second most frequent noun in the works of D. Gavrilović and V. Đorđević, used by these authors in the same sense?

The digital imprints presented here, as well as many others yet to be identified, can be used for the (semi-)automatic classification of selected works according to genre, themes, authors' attitudes, and other features. Such classifications could, in turn, support contrastive analyses of digital imprints across genres and other characteristics.

Acknowledgment

The author is grateful to the collaborators of the COST Action "Distant Reading for European Literary History" (CA16204), whose support provided the impetus for and facilitated the compilation of the SrpELTeC corpus. The contribution of the University Library "Svetozar Marković" was outstanding, as most of the required books were identified in its collections and subsequently digitized. Special thanks are also due to its collaborators, Dr. Aleksandra Trtovac and Dr. Vasilije Milnović, who organized this extremely demanding work. We also wish to emphasize the contributions of the many readers who proofread the texts and performed basic annotations, particularly Dr. Duško Vitas and Dr. Ivan Obradović, who personally processed dozens of books. Without the support of Prof. Dr. Ranka Stanković, the SrpELTeC corpus would not have been enriched with advanced annotations and made available to users for further free use.

References

Byszuk, Joanna, Michał Woźniak, Mike Kestemont, Albert Leśniak, Wojciech Łukasik, Artjoms Šeļa, and Maciej Eder. 2020. "Detecting direct speech in multilingual collection of 19th-century novels." In *Proceedings of lt4hala 2020-1st workshop on language technologies for historical and ancient languages,* 100–104.

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.

Krstev, Cvetana. 2008. *Processing of Serbian – Automata, Texts and Electronic Dictionaries.* Belgrade: University of Belgrade, Faculty of Philology.

Krstev, Cvetana. 2021a. "The Serbian Part of the ELTeC Collection Through the Magnifying Glass of Metadata." *Infotheca – Journal for Digital Humanities* 21 (2): 26–42. https://doi.org/10.18485/infotheca.2021.21.2.2.

Krstev, Cvetana. 2021b. "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, Ranka Stanković, and Aleksandra Marković. 2023. "Multiword Expressions-Comparative Analysis Based on Aligned Corpora." In *Book of Abstracts of the UniDive 1st general meeting, 16-17 March 2023, Paris-Saclay University, France.* COST.

Lavrentiev, Alexey M., Tatiana Yu. Sherstinova, Andrey M. Chepovskiy, and Benedict Pincemin. 2021. "Using TXM platform for research on language changes over time: The dynamics of vocabulary and punctuation in Russian Literary Texts." *Вестник Томского государственного университета. Филология,* no. 70, 69–89.

Marneffe, Marie-Catherine de, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. "Universal Dependencies." *Computational Linguistics* (Cambridge, MA) 47, no. 2 (June): 255–308. https://doi.org/10.1162/coli_a_00402.

Mihajlov, Teodora, Milica Ikonić Nešić, Ranka Stanković, and Olivera Kitanović. 2024. "Topic modeling of the SrpELTeC corpus: A comparison of NMF, LDA, and BERTopic." In *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS),* 649–653. IEEE. https://doi.org/10.15439/2024F1593.

Nešić, Milica Ikonić, 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,* 7–16.

Patras, Roxana, Carolin Odebrecht, Ioana Galleron, Rosario Arias, Berenike J. Herrmann, Cvetana Krstev, Katja Mihurko Poniž, and Dmytro Yesypenko. 2021. "Thresholds to the "Great Unread": Titling Practices in Eleven ELTeC Collections." *Interférences littéraires/Literaire interferenties* 25:163–187.

Paumier, Sébastien, Takuya Nakamura, and Stavroula Voyatzi. 2009. "Unitex, a corpus processing system with multi-lingual linguistic resources." In *eLEX2009 – Book of Abstracts – eLexicography in 21st century: New chalanges, new applications,* 173–175.

Pincemin, Bénédicte, Serge Heiden, and Franck Mazuet. 2022. "The textometric concept of active corpus." In *JADT 2022 Proceedings of the 16th International Conference on Statistical Analysis of Textual Data,* edited by Michelangelo Misuraca, Germana Scepi, and Maria Spano, II:691–698. Naples, Italy: VADISTAT - Per Simona Balbi, Univ. of Naples Federico II.

Radak, Tamara, Lou Burnard, Pieter Francois, Agnes Hilger, Fotis Jannidis, Gábor Palkó, Roxana Patras, Michael Preminger, Diana Santos, and Christof Schöch. 2024. "Towards a computational history of modernism in European literary history: Mapping the Inner Lives of Characters in the European Novel (1840–1920)." *Open Research Europe* 3:128.

Schmid, Helmut. 2013. "Probabilistic part-of-speech tagging using decision trees." In *New methods in language processing,* 154–164. Routledge.

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.

Smith, Raoul N. 1973. *Probabilistic Performance Models of Language.* Mouton.

Stanković, Ranka, Miloš Košprdić, Milica Ikonić Nešić, and Tijana Radović. 2022. "Sentiment Analysis of Serbian Old Novels." In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data,* edited by Ilan Kernerman, Sara Carvalho, Carlos A. Iglesias, and Rachele Sprugnoli, 31–38. Marseille, France: European Language Resources Association, June.

Stanković, Ranka, Cvetana Krstev, Branislava Šandrih Todorović, and Mihailo Škorić. 2022. "Annotation of the Serbian ELTeC Collection." *Infotheca – Journal for Digital Humanities* 21 (2): 43–59. https://doi.org/10.18485/infotheca.2021.21.2.3.

Stanković, Ranka, Cvetana Krstev, and Duško Vitas. 2024. "SrpELTeC: A Serbian Literary Corpus for Distant Reading." *Primerjalna književnost* 47 (2): 45–63. https://doi.org/10.3986/pkn.v47.i2.03.

Stanković, Ranka, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. "Parallel bidirectionally pretrained taggers as feature generators." *Applied Sciences* 12 (10): 5028.

Trtovac, Aleksandra, Vasilije Milnović, and Cvetana Krstev. 2021. "The Serbian Part of the ELTeC – from the Empty List to the 100 Novels Collection." *Infotheca – Journal for Digital Humanities* 21 (2): 7–25. https://doi.org/10.18485/infotheca.2021.21.2.1.

Utvić, Miloš. 2011. "Annotating the corpus of contemporary Serbian." *Infotheca: Journal of informatics and librarianship* 12 (2): 36a–47a.

Vitas, Duško. 2022. "From Onions to Champagne – Food and Drink in the SrpELTeC Corpus." *Infotheca – Journal for Digital Humanities* 21 (2): 88–118. https://doi.org/10.18485/infotheca.2021.21.2.5.

Works included in 15AUTHORS corpus

1. Игњатовић, Јаков (1822–1889):
   - *Ђурађ Бранковић : историческиј роман* [Đurađ Branković : a historical novel], 1859;
   - *Једна женидба : слика из живота* [A wedding: pictures from life], 1862;
   - *Милан Наранџић* [Milan Narandžić], 1863;
   - *Васа Решпект* [Vasa Rešpekt], 1875;
   - *Патница : роман* [The Suffering Women : a novel], 1888.
2. Ђорђевић, Владан (1844–1930):
   - *Кочина крајина : историјски роман* [Koča's froniter: a historical novel], 1863;
   - *Гмунденско језеро : путничка новела* [Gmunden lake: a travel novel], 1869;

- *У фронт : приповетка из живота једног бившег краља* [To the front : a short story from the life of a former king], 1913.

3. Јакшић, Ђура (1832–1878):
   - *Сељаци : приповетка из сеоског живота, из године 1857* [Peasants : a short story from rural life, from the year 1857.], 1874;
   - *Сирота Банаћанка* [The poor woman from Banat], 1875;
   - *Чича Тима : приповетка из учитељског живота* [Uncle Tima : a short story from a teacher's life], 1876.
4. Милићевић, Милан Ђ. (1831–1908);
   - *Потурченица Лејла : (црте из ратова за слободу)* [Turkified Lejla : (sketches from the wars for freedom)], 1879;
   - *Јурмуса и Фатима или Турска сила сама себе једе: прича о ослобођењу шест округа 1832-1834* [Jurmusa and Fatima or Turkish empire eats itself : the story of the liberation of six districts 1832-1834], 1879;
   - Хајдуци : биљешке с пута по Рујну [Hajduks : notes from a travel across Rujno], 1879;
   - *Десет пара: прича из живота у вароши* [Ten cents : a story from life in the town], 1881;
   - *Омер Челебија : приповиједетка из живота српскога народа* [Omer Čelebija : a short story from the life of the Serbian people], 1886.
5. Комарчић, Лазар (1833–1909):
   - *Драгоцена огрлица : прича у своје време* [Precious necklace : story in its time], 1880;
   - *Мој кочијаш : слике 1883 године* [My coachman : images from the year 1883], 1887;
   - *Један разорен ум* [A ruined mind], 1893.
6. Гавриловић, Драга (1854–1917):
   - *Из учитељичког живота* [From a teacher's life], 1884;
   - *Бабадевојка* [Old maid], 1887;
   - *Девојачки роман* [A maiden novel], 1889.
7. Веселиновић, Јанко М. (1862–1905);
   - *Сељанка : приповетка из сеоског живота* [Peasant woman: a story from rural life], 1888;
   - *Борци : роман из сеоског живота* [Fighters: a novel from rural life], 1889;
   - *Хајдук Станко : историјски роман* [Hajduk Stanko: a historical novel], 1896.
8. Тодоровић, Пера (1852–1907):

- *Силазак с престола : роман / написао Карио Амурели* [Descent from the throne : a novel/written by Kario Amureli], 1889;
- *Београдске тајне : историски роман из српске прошлости с краја прошлог века!* [Belgrade secrets : a historical novel from the Serbian past, from the end of the last century!], 1892;
- *Смрт Карађорђева : историски роман из недавне прошлости* [Death of Karađorđe : a historical novel from the recent past], 1983.

9. Мијатовић, Чедомиљ (1842–1932):
   - *Иконија везирова мајка : приповетка из XVII века* [Ikonija, vizier's mother: a short story from 17th century], 1891;
   - *Рајко од Расине: приповетка с краја XVII века* [Rajko from Rasina : a short story from the end of 17th century], 1892;
   - *Кнез Градоје од Орлова Града : приповетка из времена боја на Косову* [Prince Gradoje of Eagle City : a short story from the time of the battle of Kosovo], 1899.
10. Сремац, Стеван (1855–1906):
    - *Поп Ћира и поп Спира : приповетка* [Father Ćira and father Spira : a short story], 1894;
    - *Ивкова слава : приповетка* [Ivko's patron saint's day : a short story], 1895;
    - *Зона Замфирова : приповетка* [Zamfire's Zona : a short story], 1907.
11. Ранковић, Светолик (1863–1899):
    - *Горски цар : роман* [Forest emperor : a novel], 1897;
    - *Сеоска учитељица : роман* [Village teacher : a novel], 1899;
    - *Порушени идеали : роман* [Broken ideals : a novel], 1900.
12. Станковић, Борисав (1876–1927):
    - *Увела ружа* [Withered rose], 1899;
    - *Нечиста крв* [Impure blood], 1901;
    - *Покојникова жена* [The deceased's wife], 1902.
13. Димитријевић, Јелена (1862–1945):
    - *Ђул-Марикина прикажња : приповетка* [Jul-Marikina's narrative: a short story], 1901;
    - *Фати-султан* [Fati-Sultan], 1907;
    - *Нове : роман* [New ones: a novel], 1912.
14. Ћоровић, Светозар (1875–1919):
    - *Женидба Пере Карантана* [Marriage of Pera Karantan], 1905;
    - *Јарани : приповетка* [Buddies : a story], 1913;
    - *У ћелијама* [In the cells], 1919.
15. Ускоковић, Милутин (1884–1915):
    - *Дошљаци : роман* [Newcomers : a novel], 1910;
    - *Потрошене речи* [Spent words], 1911;
    - *Чедомир Илић : роман* [Čedomir Ilić: a novel], 1914.

Roles of people in the 15AUTHORS corpus as identified in texts (in Serbian)

проф/м	*F*	**проф/ж**	*F*	**поз/м**	*F*	**поз/ж**	*F*	**ттл/м**	*F*	**ттл/ж**	*F*
поп	1892	слушкиња	157	везир	806	газдарица	173	кнез	2254	госпођа	1163
учитељ	676	учитељица	147	хајдук	589	попадија	159	господин	1693	гђа	498
доктор	558	калуђерица	48	газда	519	натарошевица	24	г .	1423	госпођица	293
кмет	523	кухарица	46	отац	403	ученица	18	војвода	800	фрајла	289
сељак	481	машамода	40	ђак	356	приорка	15	паша	739	ханума	244
калуђер	476	ашчика	33	министар	295	службеница	12	капетан	686	ханум	207
слуга	400	редара	31	начелник	251	надзорница	10	деспот	605	мадама	185
војник	280	измећарка	28	председник	225			игуман	587	баба	145
писар	260	куварица	28	владика	201			цар	478	госпоја	142
трговац	253	сељанка	28	хоџа	121			чича	400	царица	141
пандур	236	собарица	26	пуковник	119			султан	352	тетка	131
кочијаш	166	бабица	25	старешина	97			ага	330	госпа	125
официр	156	чочек	24	посланик	93			ђенерал	262	кнегиња	89
лекар	145	учитељка	22	намесник	77			гроф	260	пашиница	78
професор	138	куварка	21	паор	60			бej	247	фрау	75
мајстор	134	бирташица	19	ађутант	57			бег	227	фрајлица	59
адвокат	121	гувернанта	19	великаш	57			краљ	223	мадам	56
прота	119	дадиља	14	секретар	56			мула	187	милостива	55
чиновник	116	служавка	14	преседник	54			чика	152	графица	48
стражар	103	играчица	12	управитељ	53			витез	106	капетаница	37