Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection

Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, Duško Vitas, Mihailo Škorić, Milica Ikonić Nešić

Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection | Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, Duško Vitas, Mihailo Škorić, Milica Ikonić Nešić | Proceedings of the Language Resources and Evaluation Conference, June 2022, Marseille, France | 2022 | |

http://dr.rgf.bg.ac.rs/s/repo/item/0006279

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3337–3345
Marseille, 20-25 June 2022

Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection

Ranka Stanković*, Cvetana Krstev†, Branislava Šandrih Todorović†, Duško Vitas‡, Mihailo Škorić*, Milica Ikonić Nešić†
*University of Belgrade, Faculty of Mining and Geology, Serbia
{ranka, mihailo.skoric}@rgf.bg.ac.rs
†University of Belgrade, Faculty of Philology, Serbia
cvetana@matf.bg.ac.rs, {branislava.sandrih, milica.ikonic.nesic}@fil.bg.ac.rs
‡University of Belgrade, Faculty of Mathematics, Serbia
vitas@matf.bg.ac.rs

Abstract

In this paper we present the Serbian part of the ELTeC multilingual corpus of novels written in the time period 1840–1920. The corpus is being built in order to test various distant reading methods and tools with the aim of re-thinking the European literary history. We present the various steps that led to the production of the Serbian sub-collection: the novel selection and retrieval, text preparation, structural annotation, POS-tagging, lemmatization and named entity recognition. The Serbian sub-collection was published on different platforms in order to make it freely available to various users. Several use examples show that this sub-collection is useful for both close and distant reading approaches.

Keywords: Corpus, Distant Reading, Digital Humanities, Linked Data, Named Entity Recognition, Text Analytics

1. Introduction

The term "distant reading" was first mentioned in (Moretti, 2000) for the use of quantitative text analysis methods in literary studies for the exploration of big text collections "at a distance". Observing particular features within the texts should help in discovery of new information and patterns in these collections "more objectively". The hypothesis of the current distant reading research is that useful (even if imperfect) formal and quantifiable textual features can be used as indicators or proxies for relevant literary phenomena (Schöch et al., 2020, p.1). Today, more than twenty years after its emergence, literary scholars are still discussing whether "distant reading" stands in opposition to the "close reading" ("patient, slow, word-by-word reading of literary texts that clarified the nuances and ambiguities of meaning...") and renders it obsolete (Glaubitz, 2018). For Underwood (2019) distant reading is simply a new scale of description that does not displace previous scales of literary description, but has the potential to expand the discipline, while Ciotti (2021) claims that the computational literary and cultural studies must find proper theoretical frameworks to take full advantage of the most advanced methods and analytical techniques, like text mining and machine learning.

In this paper Distant Reading (DR) refers to the currently ongoing COST action *Distant Reading for European Literary History* (CA16204) (2017–2022),[1] aiming at creating a network of scholars of different background that would produce resources and tools that can help in writing the European literary history from the new standpoint. Its main objective is the production of an unified, uniform, multilingual, digital novel collection dubbed ELTeC[2] (Odebrecht et al., 2021).

Our focus of this paper is the Serbian part of ELTeC, SrpELTeC sub-collection, and challenges that we had to overcome in order to produce it. This paper is organized as follows: Section 2 brings an overview of the ELTeC text collection and the production of its Serbian sub-collection, including the novel selection, text preparation, structural annotation, POS-tagging, lemmatization and named entity recognition (NER); Section 3 presents examples of the publication of SrpELTeC in digital libraries, corpus management systems and in Wikidata; some research tasks in digital humanities involving SrpELTeC in their solution are presented in Section 4; Section 5 presents some ongoing and future activities.

2. ELTeC and its Serbian Sub-Collection

In order to make ELTeC a solid basis for the implementation of distant reading methods it had to be meticulously prepared. First of all the eligibility criteria were defined that state that each language sub-collection should contain novels originally written in that language and first published, preferably as a book, in the period 1840–1920. For this purpose, a "novel" is defined as a fictional narrative text at least 10,000 words long. The choice of novels cannot be random, since some balancing criteria have also to be met:

- A sub-collection should contain 100 works that qualify as 'novels';

[1] Distant Reading COST action

[2] ELTeC: European Literary Text Collection

- Male and female authors should be equally represented, or at least by 30% of novels in a sub-collection;
- The whole time period should be covered, that is each of four twenty year spans should optimally contain 25% of all novels in a sub-collection;
- A sub-collection should contain novels of different sizes: short (up to 50,000 words), medium sized (having more than 50,000 words but less than 100,000) and long (more than 100,000 words); a sub-collection should contain in each of these subsets at least 20% but not more than 40% of novels;
- A sub-collection should contain both well-known canonical works and less-known and forgotten novels, the latter representing at least 30% but not more than 70% of the whole sub-collection;
- As to the authorship, optimally 9–11 authors should be represented in a sub-collection by three novels, while all other authors should be represented by one novel only.

From the very beginning it was clear that for two main reasons the development of language sub-collections would not be the equally demanding task for all languages. First, for some languages most of the novels from the chosen time period were not yet digitized or their digitized versions were not available. Second, due to the literary history of some languages, novels as a literary form were just emerging in the chosen time period (at least in its first half), leading to the difficulty to fulfill the demanding balance criteria. We had to deal with both issues in preparing the Serbian sub-collection and in the following subsections, we will dive into details of each milestone from the Serbian sub-collection creation pipeline.

2.1. The Novel Selection and Text Preparation

The development of the Serbian sub-collection proceeded in several steps. First of all, the list of candidates was prepared with the help of reference literary history books that, however, mention mostly canonical works. For more candidates we consulted the Mutual library catalog of the Republic of Serbia on the Cobiss+ platform (IZUM, 2022). Next, the already digitized versions of 157 candidates were searched for. It turned out that mostly canonical works were already digitized and even that was not done in the proper way, since they were not provided with any metadata. For that reason, we had to prepare the whole sub-collection from scratch, meaning we had to retrieve the hard copies, scan them and do the OCR, proofread and correct them and do the basic annotations, the so-called level-1 annotation. The copies were found and scanned in three largest Serbian public libraries and in the private library of authors of this paper. The result of the OCR was in most cases rather poor and the extensive correction was required. After OCR, all texts were corrected in several phases: the OCR errors were first corrected automatically using the procedure specially developed for this purpose (Krstev and Stanković, 2020), after that volunteers read and corrected texts, and in the final step the remaining potential errors were retrieved by e-dictionaries and corrected. Reading and annotation was done by volunteers, members of the Society for Language Resources and Technologies JeRTEH (`www.jerteh.rs`).

Among 157 candidates, 100 were chosen for the SrpELTeC sub-collection that best satisfy the balance criteria. The remaining novels are being prepared in the same way, and they are put in the extended sub-collection SrpELTeC-ext.[3] Still, not all balancing criteria could be satisfied in the best way: there are not enough female authors (8% instead of at least 30%), there are only two novels first published 1840-1859, and there are only 5 "long" novels (instead of at least 20%).

2.2. Novels as XML/TEI Documents

Each novel from the ELTeC collection is at level-1 prepared as an XML/TEI document. Besides some basic TEI structural elements: `<front>`, `<body>`, `<back>`, `<div>`, `<head>`, and `<p>`, some textual elements are allowed as well: `<hi>`, `<foreign>`, `<title>`, `<milestone>`, `<pb>`. For the SrpELTeC, all tags, except `<p>`, were added manually during text reading and correction. Beside that, each novel contains TEI header[4] with the following obligatory XML elements:

- `<fileDesc>`, which describes the document file: ELTeC edition (element `<titleStmt>`), its size (element `<extent>`), availability and licensing (element `<publicationStmt>`), and source(s) from which it was derived (element `<sourceDesc>`), with obligatory data about the first edition;
- `<profileDesc>`, which gives additional information about the ELTeC edition, the language the text was written in (element `<langUsage>`), and text characteristics that serve to check balance criteria of the whole sub-collection (element `<textDesc>`);
- `<revisionDesc>`, which records all changes made to the file.

Having consistent sub-collection headers enabled various further analysis: titling practices of Serbian nar-

[3] SrpELTeC is still under preparation: some novels that better fit balance criteria are being added while others are moved to the extended collection. SrpELTeC-ext presently has 14 novels.

[4] ELTeC TEI header

rative literature, authors' gender and age, publication places, modes of publication, etc.

2.3. The Annotation Pipeline

ELTeC collection is supposed to be multi-layered. The level-2 is built on the basis of level-1 presented in the previous subsection and it is more complex and informative, containing besides sentence segmentation tags <s>, token tags <w> for words and <pc> for punctuation belonging to the TEI module *analysis*, with attributes from the attribute class *att.linguistic*: mandatory attributes (@pos for word's part-of-speech, for word's lemma @lemma, and @join that specifies whether there is a space before and/or after a word), as well as optional attributes (general XML attribute @xml:id for the unique identification and @msd for the more detailed morphosyntactic description).

In order to provide such annotations for SrpELTeC, annotation pipeline built upon various language resources and tools was designed and developed. Steps included in the procedure are shown by the order of application in Figure 1.

Before entering the pipeline, some preparatory steps were necessary. All texts in SrpELTeC use the alphabet of the source used for digitization: Cyrillic or Latin. In order to unify the process, all Cyrillic texts were transformed to the Latin script. Next, the sentence boundaries were recognized and sentences were accordingly delimited between </s> and </s> tags using the Unitex transducer (Krstev, 2008).

The pipeline starts with the named entity recognition, for which the rule- and lexicon-based SrpNER system was used that recognizes various classes of NEs, such as dates, time, money and measurement expressions, geopolitical, personal names (<persName.first> in the example) and organizations (Krstev et al., 2014). Scholars involved in the D-reading COST action agreed that the level-2 tagset should contain the following: PERS, ROLE, LOC, ORG, DEMO, EVENT and WORK; therefore, we had to map SrpNER tags to the corresponding ones from this tagset. For this purpose, we joined various existing and newly developed NER-related tools into a NER&Beyond (JeRTeh, 2021) online platform (Šandrih et al., 2019; Šandrih Todorović et al., 2021). For the purpose of lemmatization and POS-tagging we have used the TXM tool (Heiden, 2010) that keeps existing XML structure and adds new information to each token.

TXM is using an appropriate parameter file for TreeTagger (Schmid, 1999), used for the part-of-speech tagging and lemmatization. In order to train the TreeTagger model for Serbian (Škorić and Stanković, 2021) with the Universal Dependencies tagset, a dataset created from several annotated Serbian texts (Vitas et al., 2021) was used. TreeTagger also requires a lexicon and a list of open classes for the training procedure. For this purpose, Serbian morphological dictionaries (Krstev and Vitas, 2006) were used to produce a lexicon (Krstev et al., 2021a) in required format.

After processing done in TXM, text was converted back to the original alphabet, and after merging the transformed document with the existing TEI header from the level-1, the level-2 edition was produced.

The outcome of this process is not only the publicly available level-2 annotated SrpELTeC, but also the additional resource and tool that were made publicly available on the European Language Grid (ELG) platform.[5] The first one is the named-entity (NE) annotated corpus SrpELTeC-gold (Krstev et al., 2021b) that contains full texts of 11 novels and excerpts from additional 15 novels from the SrpELTeC. In the first stage of the gold standard preparation, this corpus was automatically labelled with the already mentioned SrpNER system, after which different evaluators following specifically tailored guidelines performed careful checks and corrections producing the gold standard – SrpELTeC-gold. The corpus is divided in separate files with stand-off annotations. Total number of annotations is 330,119, distributed as follows: PERS - 14,788, ROLE - 10,405, LOC - 1,979, DEMO - 1,568, ORG - 323, WORK - 198, EVENT - 149.

The second one is the NE Recognizer (SrpCNNER) (Šandrih Todorović et al., 2021) that was trained for spaCy[6] on SrpELTeC-gold to recognize seven previously mentioned NE types with a Convolutional Neural Network (CNN) architecture. The model achieved F_1 score of $\approx 91\%$ on the test dataset.

3. SrpELTeC in Various Settings and for Different Uses

3.1. Digital Libraries and Corpus Querying

SrpELTeC as a valuable resource is and will be used for various lexical and linguistic research, by using different tools and methodologies. As the majority of books in SrpELTeC were not well known and easily accessible to the public, our goal was to make them available and accessible in different environments in order to meet the needs of different types of users. In this section, three platforms on which these novels are published will be presented: "Udaljeno čitanje", Aurora and Sketch Engine.

The platform "Udaljeno čitanje" (UBSM, 2019) is intended for readers who would like to see original pages as pictures while reading in parallel a digitized version. It was developed at the University Library "Svetozar Marković" in which the majority of novels for SrpELTeC were scanned, in cooperation with the University of Belgrade and JeRTeH members, qnd supported by a national project.[7] The initial version has 34 Serbian ELTeC novels, but further preparation and

[5] European Language Grid

[6] spaCy, Python module for advanced NLP

[7] The project was financed by the Ministry of Culture and Information of Serbia - Sector for Digitization of Cultural Heritage and Contemporary Creativity in 2019.

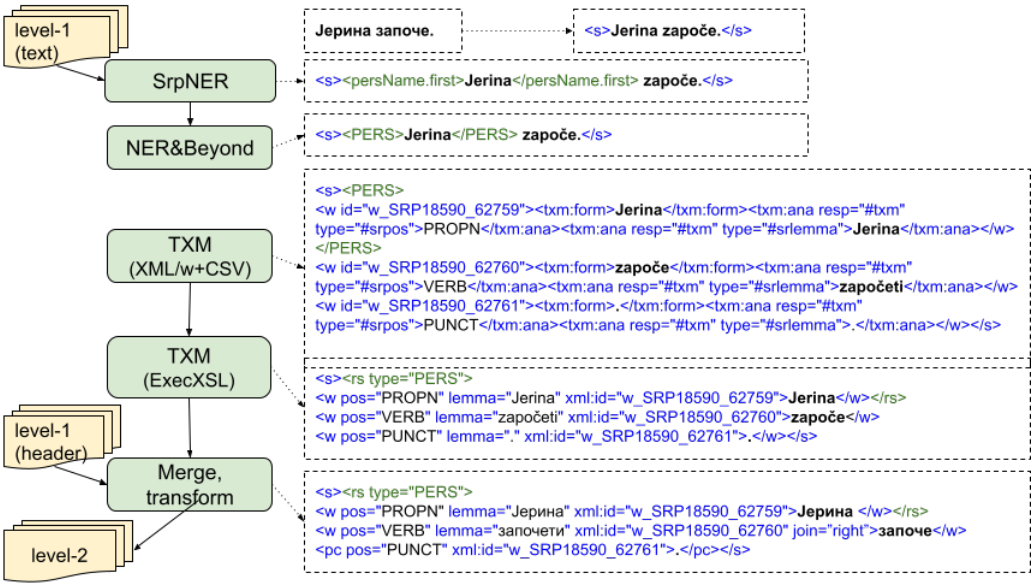

Figure 1: SrpELTeC level-2 pipeline illustrated by the sentence *Jerina započe.* (Jerina began.)

publishing is planned. Apart from novel's title, author, publication place and the names of persons responsible for text preparation, links to Wikidata, Wikipedia, and Cobiss are given as well.

The Aurora portal is being developed to provide researchers of Serbian literature and other interested users with a detailed insight into the vocabulary of novels, offering them to browse texts, concordances and frequency lists. The name of this portal was chosen to honour the memory of the AURORA (JeRTeh, 2022) (The Automatic Routine for Dictionary Processing) software system for the production of concordances (Vitas, 1979), which was the first step in the automatic processing of written texts in the Serbian language. The content of Aurora portal can be browsed by a text type (prose or poetry), author's name or by collections, SrpELTeC being one of them. Several authors and titles are linked with Wikidata, while further linking is an ongoing activity.[8] All 100 novels from the Serbian ELTeC sub-collection and 11 from the extended sub-collection are available through the AURORA platform. An alphabetically ordered index of all words used in a text is pre-prepared for each novel using Unitex-Gramlab (Maurice Gross, 2022) each word (and its frequency) in this list is linked to a list of all its occurrences in the context and presented in the form of concordances, while a broader context in the full text preview can also be seen. The list of all words can be browsed, or a word can be searched for through the regular search field.

The third platform SrpELTeC is published on is the Sketch Engine (CZ, 2022). It is a platform for corpora management and exploration, as well as for analyzing texts to identify what is typical in a language and what is rare, unusual or emerging usage. It also enables text analysis and text mining applications through API features.[9] With the Sketch Engine, a user can search for a word, phrase or pattern, and results can be presented in the form of word sketches, concordances, word lists, frequency graphs, sketch differences etc (Kilgarriff et al., 2004; Kilgarriff et al., 2014).

3.2. The SrpELTeC in Wikidata

Preparing data about Serbian novels from the SrpELTeC for Wikidata and linking Wikidata to various applications started as a manual process. The opportunity for speeding up this process was seen in using information already encoded in the header of each novel, as explained in Subsection 2.2.

After the preliminary manual Wikidata population with SrpELTeC novels, the automation of preparing and importing information was implemented using OpenRefine (David Huynh, 2022) and QuickStatements,[10] along with the purpose-designed extraction procedure. Namely, after the extraction of metadata from TEI headers, the mapping with Wikidata schema was defined in OpenRefine and predicates (properties) that connected subjects and objects in RDF triples were specified in the table header. Each statement for a subject has a property and a value that can be a Wikidata item, an external URL, or a literal (string). After consolidation in OpenRefine, RDF triplets were imported in Wikidata using QuickStatements.

Several groups of data were added or improved: au-

[8]Linking of Wikidata and ELTeC collection is supported by Wikimedia Serbia within the project "wikiELTeC – Wikidata about old Serbian novels from collection ELTeC (input, linking of named entities, visualization and analysis)"

[9]Sketch Engine API features

[10]QuickStatements

thors, publishers, metadata about novels, their printed and electronic editions, including SrpELTeC, main characters and their relations, places in which novels are settled. As a results, 66 authors and 110 novels from the core and the extended SrpELTeC collection are represented in Wikidata, comprising with associated items for first editions and digital SrpELTeC editions approximately 2,500 statements. Just to mention that novels' metadata were automatically imported, while main characters and their relations were added manually by volunteers, mostly students and members of the JeRTeH Society.

The SPARQL queries for retrieval of authors, novel titles, publication places and other metadata with different visualization options were implemented on top of SrpELTeC Wikidata. The visualization as interactive graphs of authors and ELTeC editions can be retrieved by the following query:

```
#defaultView:Graph
SELECT DISTINCT ?author
?authorLabel ?edition
?editionLabel
WHERE {
  # published in (P1433)
ELTeC collection (Q106927517)
  ?edition wdt:P1433
wd:Q106927517;
  # instance of (P31)
  # edition (Q3331189)
  wdt:P31 wd:Q3331189.
  # optional author (P50)
  OPTIONAL
  {?edition wdt:P50 ?author}
SERVICE wikibase:label
{bd:serviceParam
wikibase:language
"sr,[AUTO_LANGUAGE],en".}}
```

3.3. Analysis of the SrpELTeC with the TXM Tool

Except for the purpose of adding different morphosyntactic annotations, the TXM tool (Serge, 2020) can be used to calculate various text statistics (Krstev et al., 2019) including textometry as a powerful technique for the analysis of large body of texts (Heiden, 2010). TXM provides the following qualitative tools: 1) KWIC (KeyWord In Context) concordances of word patterns based on CQP (Corpus Query Processor) full text search engine and CQL query language; 2) word pattern frequency lists based on tokens, lemmas, POS, or structural annotations including NEs; 3) word pattern progression graphics; 4) rich HTML-based text edition navigation with links from all other software modules (e.g. from concordances). TXM provides quantitative analysis tools, based on R packages: factorial correspondence analysis, cluster analysis, specific word patterns analysis, collocations analysis. It also helps to build various sub-corpora types or partitions (for contrastive analysis between text structures or word selections) (Heiden et al., 2015).

SrpELTeC presently consists of 5,886,528 tokens, 4,769,262 words, 223,727 types and 89,977 distinct lemmas. The average number of words per paragraph is 40, while the average number of words per sentence is 14. The novel with the longest average sentences is *Zločin jedne svekrve* (The crime of one mother in law) with 26 words, while the shortest sentences were used in the novel *Hajduk Stanko* (Haiduk Stanko), the average length being 7 words.

The textometric approach to the corpus research provides the possibility of recognizing some entities or characteristics specific to some texts, their remarkably high or low representation in certain parts. The calculation of the specificity score based on the hypergeometric distribution in the TXM environment shows the probability of a lexical unit occurring in a particular part of the corpus (Heiden et al., 2015). The TXM also provides a graphical representation of the specificity distribution of the selected units. Specificity score values higher (positive) or lower (negative) than expected express a more or less represented lexical unit or pattern (Heiden, 2010).

Novels written by the same authors have been joined together for more detailed stylometry analysis. Results presented here focus on the use of nouns, verbs and adjectives by different authors. The distribution of these classes is shown in Figure 2 for the authors whose novels, when joined together, have at least 50,000 tokens.

The specificity score on corpus partitions based on authorship reveal that the adjectives are specific for novels written by Milutin Uskoković which is presented by the high positive value of the specificity score (a grey bar over the axis), whereas in works of Janko Veselinović the use of adjectives is extremely low which is presented by the high negative value of the specificity score (a grey bar below the axis). Branislav Nušić and Milutin Uskoković use nouns more often than other authors whose works are included in this corpus partition (orange bars over the axis), while Janko Veselinović and Jakov Ignjatović used them less (orange bars below the axis).

On the other hand, the verbs are less used by Lazar Komarčić compared to their degree of use throughout the corpus, which is presented by the high negative value of the specificity score (a blue bar below the axis). Janko Veselinović and Svetolik Ranković use more verbs than the others (a blue bar above the axis). The prevalence of the use of verbs in novels by Janko Veselinović, the author of *Hajduk Stanko*, can be explained by his very short sentences.

Some of these findings correspond to previous research of literary criticism, although more detailed analysis is necessary. Uskoković is known for his liric style that corresponds to more frequent use of adjectives. Mladenov (1980) wrote that Nušić is characterized by the dy-

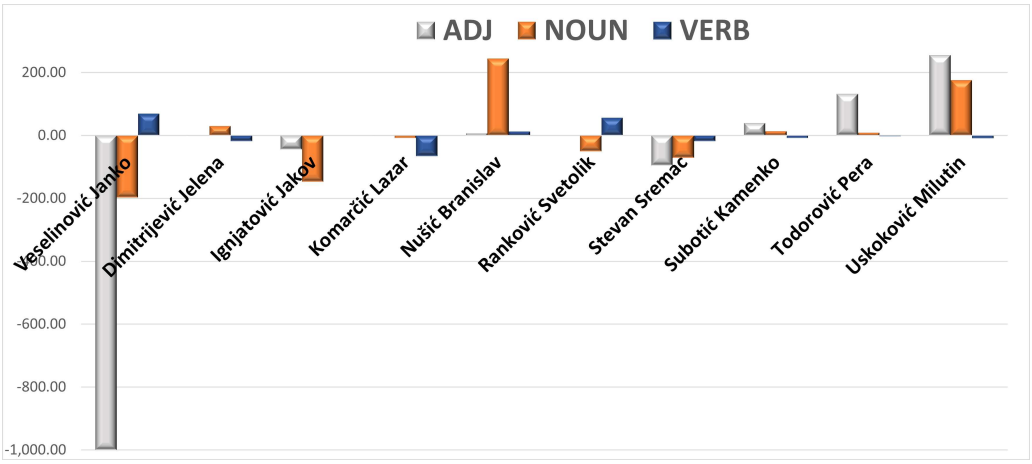

Figure 2: The specificity of nouns (NOUN), verbs (VERB) and adjectives (ADJ) use in the SrpELTeC corpus by 10 authors with largest share (contribution)

namism of the narrative and the rhythm of the sentence that is characteristic of the conversational style.

4. SrpELTeC Application in Digital Humanities

4.1. The Authorship Attribution

The research on comparative stylistic and morphosyntactic analysis of ELTeC texts using *stylo R* package[11] was performed in cooperation with the Polish language institute in Krakow.

The experiments were performed with different text representations produced from level–2 morphosyntactic annotations in order to obtain their numeric comparisons. The task was to test the performance of the morphosyntactic annotations in stylometric analysis to improve classification results in various scopes (gender, time period and authorship). Tests were performed on the SrpELTeC text collection. Balanced subsets — text representations of the documents — were prepared for the research by cross sectioning two groups of variants: one was using the metadata (author, author's gender and time-period) and the other was using different levels of morphosyntactic annotation (POS trigrams, fine–grained POS trigrams, lemma, and word forms) totalling in twelve different text representations.

The research tried to find out what better defines an author's style: the use of word forms, lemmas, or certain part of speech. Based on metadata, analysis of the influence of author's sex or a time period was also investigated.

The classification was based on the cosine delta distance distribution (Evert et al., 2015), where the indicators of classification reliability are overlapping areas of calculated distributions. The higher probability of accurate classification is indicated with smaller areas of overlapping without multiple intersections. The *stylo* package (Eder et al., 2016) was used for the analysis. It was found that 1) classifications by neither gender nor time period could be obtained for Serbian ELTeC texts, 2) the document representation using lemma produces better results than the original word representation (Figure 3), 3) POS and fine grained POS yielded higher standard deviation but also the higher area of overlap. Current research is focused on the best way to combine different document representation to increase the classification accuracy on the authorship attribution task.

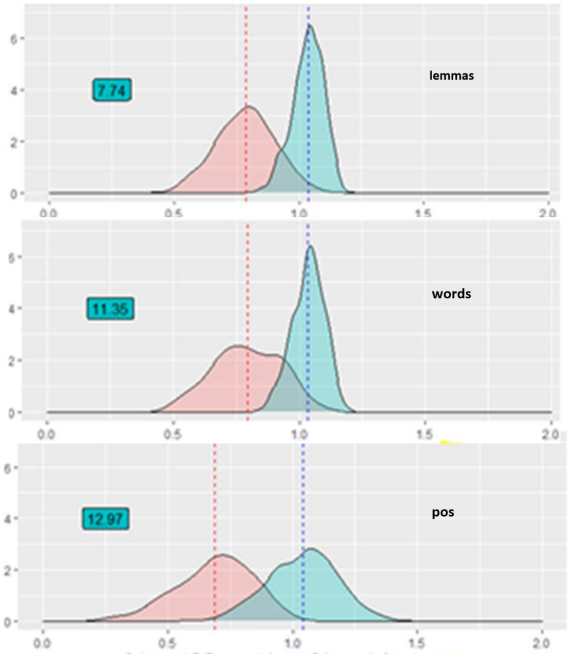

Figure 3: Delta distances distributions (same-class is blue and different-class is red) using lemma, word and POS document representations

[11] Stylometry with *stylo*

4.2. The Similes in SrpELTeC

In the previous subsections 3.3 and 4.1 we presented some successful applications of distant reading methods to SrpELTeC. In this and the following subsection we will briefly present some applications of close reading methods to SrpELTeC.

SrpELTeC corpus represents a solid basis to analyze the use of rhetorical figures in Serbian literary texts. We focused on simile figures due to their prevalence and recognizable structure. The aim of this research was manifold: (a) to establish the list of simile figures used in old Serbian literary text; (b) to analyse the use of simile figures per authors; (c) to develop means for retrieving simile figures in unseen texts.

In order to achieve the first aim we used some simple patterns relying on information from the Serbian morphological e-dictionaries, and closely looked at the results to distinguish false recognitions and metaphoric uses from similes. That enables us to establish that 556 different simile figures appeared 1,051 times in SrpELTeC, 10.5 similes per novel, and 2.18 similes figures per each 10,000 words of the collection. The most frequently used similes are *beo kao sneg* (white as snow), *bled kao krpa* (pale as a cloth), *bled kao smrt* (pale as death) and *hladan kao led* (cold as ice), which are also similes used in most of novels – *beo kao sneg* was used in 30 out of the 100 novels. Simile figures were mostly used to describe people – men (335) more than women (184), children (8) and groups of people (46), and their features (187).

The use of similes was analyzed for 13 authors represented in SrpELTeC by 3 or more novels. It was concluded that Janko Veselinović and Milutin Uskoković were authors that used simile figures the most – 5.26 and 4.08 simile figures per 10,000 words, respectively. J. Veselinović uses 10 times more simile figures than the author with the fewest simile figures from this group – Draga Gavrilović with 0.54 per 10.000 words. It is interesting to note that J. Veselinović is the author that uses few adjectives, as shown in Subsection 3.3.

In order to retrieve (and annotate) simile figures from any Serbian text, we developed a local grammar in the form of finite-state automata that relies on the Serbian morphological e-dictionaries. The system is expandable since newly discovered simile figures can easily be added to it. The next step in this research is to use this system to annotate simile figures in SrpELTeC and add annotations as a new layer to the level-2 representation in TXM, which would enable more refined analyses.

4.3. SrpELTeC as the Evidence of Eating Habits in Serbia in 1840–1920

SrpELTeC is a valuable source for studying everyday life in Serbia in the second half of the 19th and the beginning of the 20th century, and especially their eating habits. The everyday and available foodstuff, the prepared dishes depended on a wide variety of factors: rural and urban diets were different, meals prepared in the regions dominated by the Ottoman Empire differed significantly from those prepared in regions dominated by Austro-Hungarian Empire, while people's financial situation also greatly influenced what they ate and how their meals were prepared. Since characters in SrpELTeC novels, and its extended subset, can be illiterate peasants or highly educated persons, poor or rich, that live in various regions populated by Serbs - south, north, west – or abroad, in present or in some ancient time, this collection represents diversiform setting for our research.

Some of the questions posed were: (a) what were the most popular foodstuff and meals; (b) how they were named; (c) and how these changed over time. For this research the Serbian morphological e-dictionaries implemented in Unitex were indispensable since various foodstuff, meals and drinks are described in it in detail. Some of the discoveries were that *luk* (onion) and *hleb* (bread) represent the inevitable ingredients of every meal, often the only ingredients, which is consistent with evidences of previous ethnological studies. Various variant generic names were used for bread – *hleb*, *leb*, *lebac*, *ljeb* – with a number of names for a bread of a special type: *lepinja*, *somun*, *pogača*, *simit*, *peksimit*, *beškot*, *proja*, etc. The interesting result was that coffee (under different names: *kafa*, *kava*, *kahva*, *kajmaklija*, *crna čorba*) was mentioned more often than any other food. It was consumed both in villages and cities, by peasants and bishops, prepared and served in Turkish or European style. The analysis of the use of some foodstuff over time revealed that potatoes – under the names *krompir*, *krumpijer*, and *krtola* – were scarcely mentioned in novels dated at the beginning of the investigated time period, while in novels published in its second half it appeared significantly more, which corresponds with the previous knowledge about the cultivation and use of potatoes in Serbia. According to the evidences from SrpELTeC, wine was drunk more often than rakia, moreover, not only the "ordinary" wine, but champagne as well (under the names *šampanj*, *šampanjer*, *šampanjsko vino*, *penušavo vino*).

5. Conclusion and Future Plans

The COST action D-reading joined people of various backgrounds with a common motivation to join their expertise and create a powerful and a complex language resource, as well as to develop a plethora of language processing tools as a basis for many future research and cooperation to come. In this paper we presented our share in the project, namely SrpELTeC sub-collection. We have shown that although it was primarily developed to be used for distant reading methods and tools, it was equally useful for close reading approaches; moreover, it can and is already used for "real" reading by interested audience. Our experience in this project will help us build similar collections, e.g. novels from the later time period, travelogues, etc.

Linking of annotated NEs with Wikidata is currently

done manually, but we intend to automate it by training a suitable machine learning model. Expansion of the NE tagset with other entities, like drinks, or transportation means is also planned. We also intend to assign semantic attributes present in electronic dictionaries to the Aurora's concordances in order to integrate NEs from the level–2 and link them with Wikidata and other knowledge bases. Finally, the annotated level–2 corpus will be published in the Linguistic Linked Open Data.

6. Acknowledgements

The text collection preparation is supported by the COST Action 16204 – Distant Reading for European Literary History support. Linked data development was done in the scope of the project "WikiELTeC–Wikidata about old Serbian novels from collection ELTeC" supported by the Wikimedia Serbia.

7. Bibliographical References

Ciotti, F. (2021). Distant Reading in Literary s Studies: A Methodology in Quest of Theory. *Testo e Senso*, 23:195–213.

Eder, M., Rybicki, J., and Kestemont, M. (2016). Stylometry with R: a Package for Computational Text Analysis. *The R Journal*, 8(1).

Evert, S., Proisl, T., Schöch, C., Jannidis, F., Pielström, S., and Vitt, T. (2015). Explaining Delta, or: How do Distance Measures for Authorship Attribution Work? *URL: http://dx. doi. org/10.5281/zenodo*, 18308.

Glaubitz, N. (2018). Zooming in, zooming out: The debate on close and distant reading and the case for critical digital humanities. In *Anglistentag 2017*, page 21. WVT Wissenschaftlicher Verlag Trier.

Heiden, S., Pincemin, B., and Decorde, M. (2015). Manuel de TXM, July. Manuel d'utilisation du logiciel TXM.

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In R. Otoguro, et al., editors, *24th Pacific Asia Conference on Language, Information and Computation*, volume 2, pages 389–-398, Sendai, Japan, November. Institute for Digital Enhancement of Cognitive Development, Waseda University.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). Itri-04-08 the Sketch Engine. *Information Technology*, 105(116).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P., and Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Krstev, C. and Stanković, R. (2020). Old or new, we repair, adjust and alter (texts). *Infotheca - Journal for Digital Humanities*, 19(2):61–80.

Krstev, C., Obradović, I., Utvić, M., and Vitas, D. (2014). A System for Named Entity Recognition Based on Local Grammars. *Journal of Logic and Computation*, 24(2):473–489.

Krstev, C., Jaćimović, J., Šandrih, B., and Stanković, R. (2019). Analysis of the First Serbian Literature Corpus of the Late 19th and Early 20th Century with the TXM Platform. In *DH_BUDAPEST_2019*, pages 36–37. Centre for Digital Humanities - Eötvös Loránd University.

Krstev, C. (2008). *Processing of Serbian. Automata, Texts and Electronic Dictionaries*. Faculty of Philology of the University of Belgrade.

Mladenov, M. (1980). *Novinarska stilistika [Journalistic stylistics]*. Naučna knjiga.

Moretti, F. (2000). Conjectures on World Literature. *New Left Review*, 1 (February):54–68.

Šandrih Todorović, B., Krstev, C., Stanković, R., and Ikonić Nešić, M. (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)*, pages 1252–1260, Held Online, September. INCOMA Ltd.

Schmid, H. (1999). Improvements in Part-of-Speech Tagging with an Application to German. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Schöch, C., Eder, M., Arias, R., and Pieter Francois, A. P. (2020). Foundations of Distant Reading: Historical Roots, Conceptual Development and Theoretical Assumptions around Computational Approaches to Literary Texts. In *Digital Humanities 2020*.

Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Vitas, D. (1979). Prikaz Jednog Sistema za Automatsku Analizu Teksta. *Informatica*, 79.

Šandrih, B., Krstev, C., and Stanković, R. (2019). Development and Evaluation of Three Named Entity Recognition Systems for Serbian - The Case of Personal Names. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1060–1068, Varna, Bulgaria, September. INCOMA Ltd.

8. Language Resource References

Lexical Computing CZ. (2022). *SketchEngine*. Lexical Computing CZ s.r.o., `https://www.sketchengine.eu/`.

David Huynh. (2022). *OpenRefine*. Metaweb Technologies, Inc, `https://openrefine.org/`.

IZUM. (2022). *Cobiss+ platform*. Institut informacijskih znanosti (IZUM), `https://plus.sr.cobiss.net/opac7/bib/search`.

JeRTeh. (2021). *NER&Beyond portal for NER*. JeRTeh, `http://nerbeyond.jerteh.rs/`.

JeRTeh. (2022). *Aurora – AUtomatska Rutina za*

Obradu RečnikA. JeRTeh, http://aurora.jerteh.rs/.

Cvetana Krstev and Duško Vitas. (2006). *SrpMD - Serbian morphological dictionaries*. ELG, https://live.european-language-grid.eu/catalogue/lcr/17355, 1.0.

Cvetana Krstev and Duško Vitas and Ranka Stanković and Mihailo Škorić. (2021a). *SrpMD4Tagging - Serbian Morphological Dictionaries for Tagging*. ELG, https://live.european-language-grid.eu/catalogue/lcr/9294, 1.0.

Cvetana Krstev and Branislava Šandrih Todorović and Ranka Stanković and Milica Ikonić Nešić. (2021b). *SrpELTeC-gold - Named Entity Recognition Training Corpus for Serbian*. ELG, https://live.european-language-grid.eu/catalogue/corpus/9485, 1.0.

Maurice Gross. (2022). *Unitex/GramLab*. Maurice Gross and LADL team, https://unitexgramlab.org/.

Carolin Odebrecht and Lou Burnard and Christof Schöch. (2021). *European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels.* Zenodo.

Heiden Serge. (2020). *The TXM Platform*. https://txm.gitpages.huma-num.fr/textometrie/.

UBSM. (2019). *Udaljeno čitanje*. Univerzitetska biblioteka "Svetozar Marković", Beograd (UBSM), https://udaljenocitanje.unilib.rs/.

Duško Vitas and Cvetana Krstev and Ranka Stanković and Miloš Utvić and Mihailo Škorić. (2021). *SrpKor4Tagging*. ELG, https://live.european-language-grid.eu/catalogue/corpus/9295, 1.0.

Branislava Šandrih Todorović and Cvetana Krstev and Ranka Stanković and Milica Ikonić Nešić. (2021). *SrpCNNER - Named Entity Recognizer for Serbian (7 classes)*. ELG, https://live.european-language-grid.eu/catalogue/corpus/9485, 1.0.

Mihailo Škorić and Ranka Stanković. (2021). *SrpKor4Tagging-TreeTagger*. ELG, https://live.european-language-grid.eu/catalogue/ld/9296, 1.0.