Annotation of the Serbian ELTeC Collection

UDC 811.163.41'322.2
DOI 10.18485/infotheca.2021.21.2.3

ABSTRACT: This paper presents the so-called level-2 edition of SrpELTeC collection developed within the activities of Working Group 2 - Methods and Tools of the COST Action CA 16204 (Distant Reading for European Literary History), and its schema specification. The level-2 edition is a follow-up of the level-1 edition, which is used as input for morphosyntactic and NER annotation of novels. The Serbian level-2 pipeline outlines steps required for production of level-2, including methods and tools used in the process. Some statistics drawn from the Serbian ELTeC level-2 sub-collection brings an interesting insight into collection content.
KEYWORDS: distant reading, literary corpus, tagging, NER, lemmatization, ELTeC.

PAPER SUBMITTED: 17 November 2021
PAPER ACCEPTED: 25 December 2021

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

Cvetana Krstev
cvetana@matf.bg.ac.rs
Branislava Šandrih Todorović
branislava.sandrih@fil.bg.ac.rs
University of Belgrade
Faculty of Philology
Belgrade, Serbia

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

1 Introduction

Working group "Methods and Tools" (WG2) of the COST action "Distant Reading for European Literary History" (CA16204) is concerned with text analytic techniques and tools. WG2 coordinates activities related to sharing, evaluating, adaptation and improving methods and tools for Distant Reading research, and establishing best practices across Europe. It has a large range of activities, from the creation of manual reference annotations for the evaluation of automatic annotation tools, to the annotations' integration strategy. Namely, one of the problems WG2 tackled is integration of results of tokenization, lemmatization, part-of-speech tagging, and Named Entity Recognition (NER) into one document conforming to the ELTeC XML/TEI format. The existing tools are analysed and some guidelines for their application are published while others are still under development. Members

of WG2 are active in development of NLP resources and tools, information extraction, computational linguistics, text mining, computational stylistics, and digital literary studies.

Two main problems encountered in producing Serbian ELTeC level-2 were similar to those encountered for other languages: 1) majority of morphosyntactic taggers do not work well with XML format and 2) harmonization of NER and morphosyntactic annotations, which are performed separately with different tools. A solution was found in the TXM tool[1] (Heiden 2010; Heiden, Magué, and Pincemin 2010), an environment that enables tagging of XML files, which solves the problem of alignment of NER and morphosyntactic tags. TXM also enables the construction of a sub-corpora or partitions based on metadata (date, author, genre, etc.) or corpus structural units (like text, chapter, paragraph), querying (using the CQP browser), and the processing of more complex query results using quantitative methods (supported by the R statistical package), as well as the export of results in a tabular or graphical form (Jaćimović 2019).

The second section of this paper "Level-2 specification" will introduce concepts and current state of the schema used for morphosyntactic and NER annotation of level-1 form of novels. The third section "Serbian level-2 pipeline" will introduce steps required for the production of level-2, including methods and tools used in the process. The fourth section "SrpELTeC level-2 statistics" will bring some numerical insights from the developed dataset.

2 Level-2 specification

The encoding of novels is produced in incremental levels, each validated by the appropriate RELAXNG schemas.[2] Description of level-2 schema is given in Encoding Guidelines for the ELTeC: level 2 (distantreading.github.io).[3] At the time of writing this paper, schema for level-2 was not yet finalized, but it is expected to be done soon. ELTeC level-2 includes all elements existing in level-1 and introduces some new ones: `<s>` as the sentence tag, used for segmentation of text into sentences, and `<w>` and `<pc>`, used for tokenization of text into tokens, and their annotation. Individual words are marked using the `<w>` element and mandatory linguistic attributes `@pos`, `@lemma`, and

1. TXM is using the CQP (Corpus Query Processor) browser build on IMS Open Corpus Workbench and the R statistical package
2. ELTeC Schemas
3. Encoding Guidelines for the ELTeC: level 2

`@join`, as well as some optional attributes like the general XML attribute `@xml:id` for unique identification and `@msd` for more detailed morphosyntactic description. As tokens can be both words and punctuation marks, as well as other special characters, TEI recommends that these two cases should be distinguished by using two different elements: `<w>` for words and `<pc>` for punctuation and special characters.

The proposal is to eliminate any content within a <ref> element at level 2. The elements `<p>`, `<head>`, `<note>` and `<l>` can contain a sequence of `<s>` elements, while elements `<gap>`, `<milestone>`, `<pb>`, and `<ref>` are also permitted within text content at any point, but are disregarded in segmentation (Burnard, Schöch, and Odebrecht 2021). The element `<s>` can contain a sequence of `<w>` elements, either directly or in the sub-paragraph elements `<corr>`, `<emph>`, `<foreign>`, `<hi>`, `<label>`, `<title>`. The TEI element `<rs>` (referring string) has a special purpose in the level-2 format: it is used for the encoding of named entities, such as people, their roles, locations, organisations, works, events, and demonyms (Frontini et al. 2020; Šandrih Todorović et al. 2021).

WG2 had several physical meetings, first in Prague (Czech Republic), Antwerp (Belgium), Lisbon (Portugal), Budapest (Hungary) and in Malaga (Spain), and several online meetings for smaller teams focused on special topics, such as: morphosyntactic tagging, NER, direct speech, semantic analysis. Some resources developed by WG2 are available in the github repository.[4]

3 The Serbian Level-2 Pipeline

The Serbian level-2 novels are produced from the level-1 edition, as proposed by the Action plan and similarly to the way it was done for some other languages. Each language has its own pipeline, since the best tools for specific languages are developed within different frameworks. For the majority of languages, the integration of morphosyntactic tagging, lemmatization and named entity annotation was not a trivial task. In this section we present the Serbian language pipeline, which comprises several steps of annotations and transformations, outlined in Figure 1, with an example of a short sentence form the well-known novel *Nečista krv* (Impure blood) (SRP19101) by Borisav Stanković.

The TEI document level-1 has elements `<teiHeader>` and `<text>` on the first level, but annotation is performed only on the content of the `<text>`

4. WG2 data repository

element. For processing purposes the `<teiheader>` element is removed in this phase, only to be updated and merged with the `<text>` element after all annotations are done.

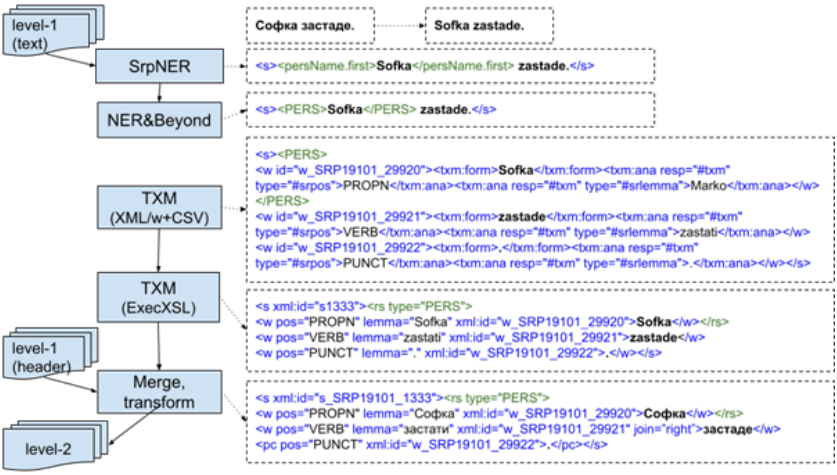

Figure 1. The Serbian SrpELTeC Level-2 pipeline.

Sentence splitting is performed by a Unitex transducer (Krstev 2008; Paumier 2021) that is adapted for this purpose, to take in consideration tags introduced in level-1. This transducer outputs the start tag `<s>` at the beginning of a sentence and the end tag `</s>` at its end.

The next step is named entity recognition performed by the rule-based system SrpNER (Krstev et al. 2014), based on large-scale lexical resources for Serbian (Krstev 2008), coupled with local grammars in the form of finite-state transducers (Vitas and Krstev 2012). Since SrpNER works on Latin texts, it is necessary to transliterate Cyrillic texts to Latin. SrpNER recognises 11 classes of NEs: dates and time (moments and periods), money and measurement expressions, geopolitical names (countries, settlements, oronyms and hydronyms), and personal names (one or more last names with or without first names and nicknames, names of church and state dignitaries). Here are some examples of SrpNER output:

1. `<pers.spec><role>`carica`</role>` `<persName.full>`Marija Terezija`</persName.full>` `</pers.spec>`
   Empress Maria Theresa

2. <pers><role>Sekretar</role> <persName.last>Živanović</persName.last></pers>
   Secretary Živanović
3. <org>Saborna crkva u <top.gr>Beogradu</top.gr></org>
   Cathedral in Belgrade
4. <org>manastir <pers.spec>Sv. Marka</pers.spec></org>
   St. Marc's monastery
5. <pers.spec><role>veliki vezir</role>
   <persName.first>Ahmed</persName.first>-
   <role>paša</role></pers.spec>
   Grand Vizier Ahmed-Pasha

Since level-2 does not allow embedded NER tags, the first step was to apply a semi-automatic procedure to remove them from the SrpNER output. Previous examples would be transformed to:

1. <role>carica</role>
   <persName.full>Marija Terezija</persName.full>
2. <role>Sekretar</role> <persName.last>Živanović</persName.last>
3. <org>Saborna crkva u Beogradu</org>
4. <org>manastir Sv. Marka</org>
5. <role>veliki vezir</role>
   <persName.first>Ahmed</persName.first><role>paša</role>

As it can be seen, besides the removal of embedded tags, the remaining SrpNER tags have to be mapped into a more simplified level-2 tagset: PERS, ROLE, LOC, ORG, DEMO, EVENT, WORK. An automatic procedure implemented as part of the NER&Beyond portal (Stanković et al. 2019; Šandrih Todorović et al. 2021) was developed and used to map SrpNER tags into the 7-categories ELTeC NER schema. Figure 2 presents the part of the NER&Beyond portal used for tagsets mapping.

The mapping procedure allows mapping, ignore or removal of XML elements. In this case, the following XML elements are ignored: <back>, <body>, , <foreign>, <front>, <gap>, <head>, <hi>, <l>, <milestone>, <note>, <p>, <pb>, <quote>, <ref>, <s>, <text>, while the mapping is defined as follows:

- <persName.first>, <persName.full>, <persName.last>, <persName.name>, <pers.spec> → PERS
- <top.deoGr>, <top.dr>, <top.geo>, <top.gr>, <top.hyd>, <top.oro>, <top.reg>, <top.supReg>, <top.ul> → LOC

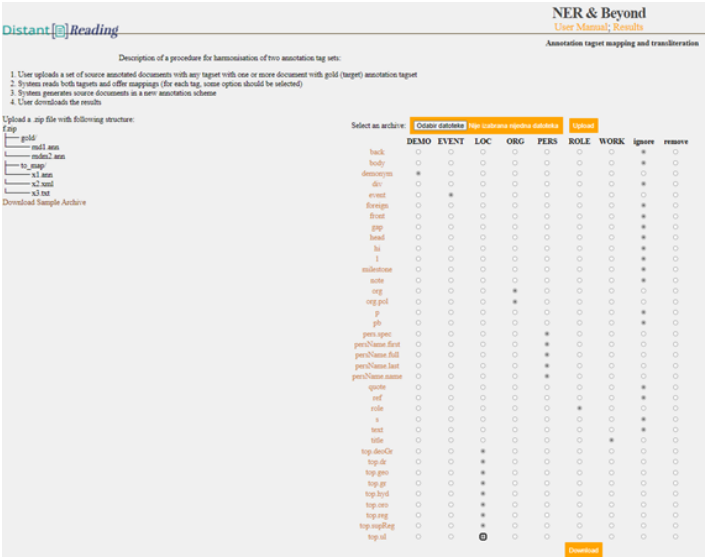

Figure 2. The tagsets mapping in the NER&Beyond portal.

- `<demonym>` → DEMO
- `<event>` → EVENT
- `<org>`, `<org.pol>` → ORG
- `<role>` → ROLE
- `<title>` → WORK.

The previous examples would be mapped as follows:

1. `<ROLE>`carica`</ROLE>` `<PERS>`Marija Terezija`</PERS>`
2. `<ROLE>`Sekretar`</ROLE>` `<PERS>`Živanović`</PERS>`
3. `<ORG>`Saborna crkva u Beogradu`</ORG>`
4. `<ORG>`manastir Sv. Marka`</ORG>`
5. `<ROLE>`veliki vezir`</ROLE>` `<PERS>`Ahmed`</PERS>`-`<ROLE>`paša`</ROLE>`

The next step was the preparation of a CSV file with metadata for 100 novels to be used for the import of the whole collection in the TXM tool (Heiden 2010).[5] The TXM import option "XML/w+CSV" was used and the required data supplied: a path to the text collection and metadata, as well as language selection. Namely, depending on language selection, TXM is

5. Textométrie//TXM

using an appropriate parameter file for TreeTagger, which is used for the part of speech tagging and lemmatization. Tokenization was applied by a set of rules. The Treetagger model[6] was trained using a dataset[7] created from several merged annotated Serbian texts, with over half a million tagged tokens (Table 1). The dataset was balanced with four literary (1-5), and three non-literary (6-8) texts, the former including one complete SrpELTeC novel (3) and a set of excerpts from SrpELTeC (5). Tokens were pre-tagged for Universal POS tagset[8] and lemma with the Unitex system,[9] using Serbian morphological dictionaries, and disambiguated manually. TreeTagger also requires a lexicon and a list of open classes for the training procedure. Serbian morphological dictionaries were used as a lexicon[10] for training, while a list of open classes was used as suggested by the Universal dependencies.

The selection of 11 full novels and excerpts from 15 novels from SrpELTeC, have been automatically labelled with SrpNER system for Serbian in the first stage of the gold standard preparation. Based on the specifically tailored guidelines, different evaluators performed careful checks and corrections, yielding a gold standard (SrpELTeC-gold) that is publicly available on European Language Grid (ELG) platform[11]. Corpus is annotated with 7 different named entity types: PERS, ROLE, LOC, DEMO, ORG, WORK, EVENT, as specified by Distant Reading for European Literary History (COST Action CA16204). Total number of text files is 242 with stend-off annotation in 242 .ann files. Total number of annotations is 330119, where PERS has 14788, ROLE has 10405, LOC has 1979, DEMO 1568, ORG 323, WORK 198, EVENT 149.

A Named Entity Recognizer (SrpCNNER) is trained using SrpELTeC-gold to recognize 7 previously mentioned named entity types, with a Convolutional Neural Network (CNN) architecture, having F1 score of approx 91% on the test dataset. Model trained for spaCy is publicly available on ELG[12].

The benefit of using TXM for tagging is that it retains XML structure elements and adds new information to each token. For example, the sentence *Sofka zastade.* (Sofka paused.):

```
<s><PERS>Sofka</PERS> zastade.</s>
```

6. SrpKor4Tagging-TreeTagger
7. SrpKor4Tagging
8. Universal POS tags
9. Unitex/GramLab Grammar-based Corpus Processing Suite
10. SrpMD4Tagging
11. SrpELTeC-gold - Named Entity Recognition Training corpus for Serbian
12. SrpCNNER - Named Entity Recognizer for Serbian

Id	Texts	Tokens	Words	Unique
1	Orwell's *1984* (Serbian translation)	108,137	96,026	18,050
2	Vern's *Around the World in Eighty Days* (Serbian translation)	68,697	62,769	12,799
3	Dragutin Ilić's *Hadži Đera* (SRP19040)	65,262	61,217	12,276
4	Excerpt from Jaroslav Hašek's *The Good Soldier Švejk*	4,122	3,347	1,475
5	Excerpts from *SrpELTeC (1840-1920)*	5,118	4,236	2,093
6	Corpus of newspaper articles on 2014 floods in Serbia	4,672	3,813	1,741
7	Excerpts from the Serbian history textbook	6,596	5,287	2,622
8	A collection of Serbian texts from Law, Finance, Education and Health domain	239,614	204,643	31,470
	Total	502,213	441,338	

Table 1. Annotated texts used for TreeTagger training, as well as the number of tokens, words and unique words for each of them.

becomes:

```
<s><PERS>
  <w id="w_SRP19101_29920"><txm:form>Sofka</txm:form>
    <txm:ana resp="#txm" type="#srpos">PROPN</txm:ana>
    <txm:ana resp="#txm" type="#srlemma">Sofka</txm:ana>
  </w></PERS>
  <w id="w_SRP19101_29921"><txm:form>zastade</txm:form>
    <txm:ana resp="#txm" type="#srpos">VERB</txm:ana>
    <txm:ana resp="#txm" type="#srlemma">zastati</txm:ana>
  </w>
  <w id="w_SRP19101_29922"><txm:form>.</txm:form>
    <txm:ana resp="#txm" type="#srpos">PUNCT</txm:ana>
    <txm:ana resp="#txm" type="#srlemma">.</txm:ana></w>
</s>
```

The obtained result is not yet level-2 compliant, which means that some additional transformations are necessary. Within the TXM tool there is an execXSL macro (in the View→Macro menu within xml macros), which performs transformations. It requires the path to the XSL file, and input and output directory with corpus files that need to be transformed (Figure 3). The initial, general purpose macro *txm-front-teitxm2xmlw.xsl* had to

be adapted for the level-2 requirements, and this new version is published on github repository.[13]

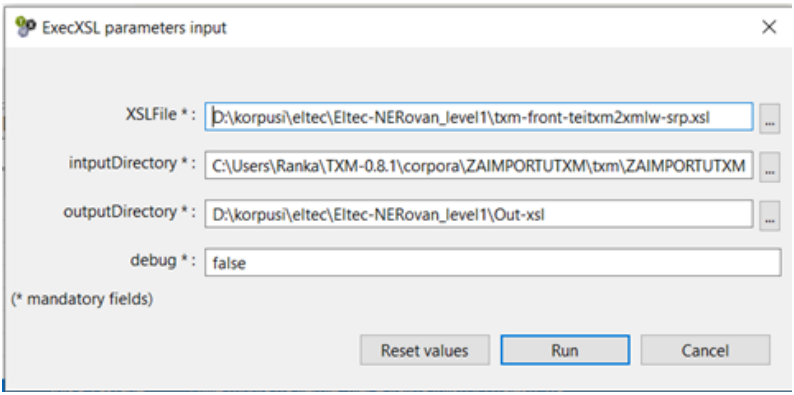

Figure 3. Invocation of the TXM macro for XSL transformation.

The adaptation of the XSL transformation macro included sentence counting, the use of required namespaces for the attributes `xml:id`, `xml:lang`, removing some attributes, and mapping XML elements for NER tags – `<PERS>`, `<LOC>`, `<ORG>`, `<DEMO>`, `<ROLE>`, `<WORK>`, `<EVENT>` – into the referring string TEI element `<rs>`, with the value of its attribute `@type` set to the appropriate value from the set {PERS, LOC, ORG, DEMO, ROLE, WORK, EVENT}. The part of this XSL code is:

```
<xsl:template match= "tei:PERS|tei:LOC|tei:ORG|tei:DEMO|
   tei:ROLE|tei:WORK|tei:EVENT">
   <!-- produce a referring string element -->
   <xsl:element name="rs"
      namespace="http://www.tei-c.org/ns/1.0">
      <xsl:attribute name="type">
         <xsl:value-of select="local-name()"/>
      </xsl:attribute>
      <xsl:apply-templates select="tei:w"/>
      <xsl:apply-templates select="tei:foreign"/>
```

13. TXM related scripts

```
   </xsl:element>
</xsl:template>
```

As a result, for our example sentence the following would be obtained:

```
<s xml:id="s1333"><rs type="PERS">
<w pos="PROPN" lemma="Sofka" xml:id="w_SRP19101_29920">
  Sofka</w></rs>
<w pos="VERB" lemma="zastati" xml:id="w_SRP19101_29921">
  zastade</w>
<w pos="PUNCT" lemma="." xml:id="w_SRP19101_29922">.
</w></s>
```

At the end, some last transformations had to be done. First, the text has to be transformed back to Cyrillic, if that was the script used in level-1, taking care about the content of the `<foreign>` element, which has to be treated in a special way. Since values of all `xml:id` attributes have to be unique for the whole ELTeC collection, the ID of a novel (value of the `xml:id` attribute of the <text> element) needs to be integrated into the sentence ID. Since TEI uses the `<pc>` element, rather than `<w>`, for punctuation, special characters `<w>` elements had to be replaced with `<pc>` and the lemma attribute removed. After this final transformation, our example sentence in the correct level-2 form is:

```
<s xml:id="s1333"><rs type="PERS">
<w pos="PROPN" lemma="Софка" xml:id="w_SRP19101_29920">
  Софка</w></rs>
<w pos="VERB" lemma="застати" xml:id="w_SRP19101_29921">
  застаде</w>
<w pos="PUNCT" lemma="." xml:id="w_SRP19101_29922">.
</w></s>
```

4 Statistical overview of level-2

SrpELTeC level-2 corpus has 100 novels annotated with part of speech tags and lemmas, while 65 novels have also named entity annotation. SrpELTeC has 5,886,528 tokens according to TXM calculation, with the four word properties: word, n, srpos, srlemma and 30 XML tags for structural elements (`back`, `body`, `front`, `div`, `div1`, `div2`, `gap`, `head`, `l`, `milestone`, `note`, `p`, `pb`,

`quote`, `ref`, `s`, `text`, `title`, `trailer`), for NER elements (PERS, LOC, ORG, DEMO, ROLE, WORK, EVENT) and other textual elements (`foreign`, `hi`).

Element `<div>` occurs at three levels. At the first level it occurs 1,763 times with the following values of the attribute `@type`: CHAPTER, GROUP, LIMINAL, NOTES, TITLEPAGE. At the second level, it occurs 463 times with chapter or group as values of the attribute @type, while at the third level it occurs 99 times as a chapter. The number of occurrences of other elements are represented in Figure 4. The distribution of named entities is represented in Figure 5.

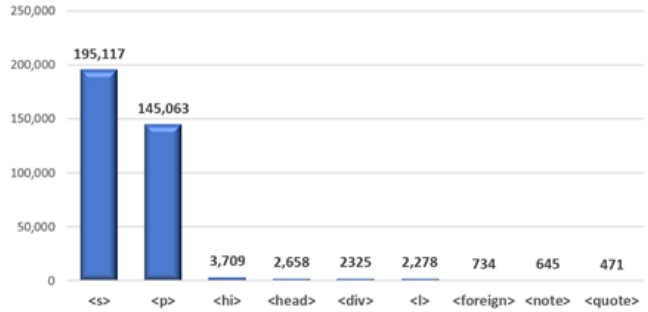

Figure 4. The number of occurrences of elements other than `<div>`.

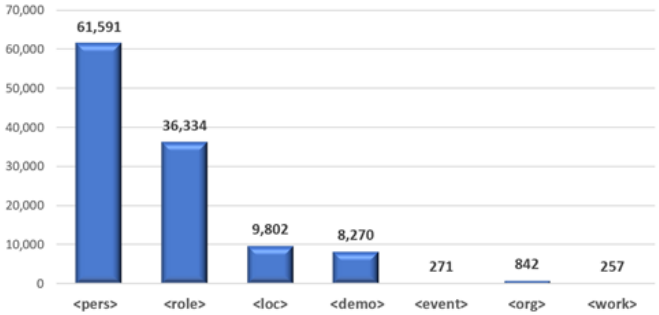

Figure 5. The number of occurrences of NE elements by class.

The average number of words per paragraph is 40, while the average number of words per sentence is 14 (Figure 6). The novels with the longest average sentences are: *Zločin jedne svekrve* (The crime of one mother in law) (SRP19062) (26), *Bespuće* (Wasteland) (SRP19121) (25), *Gmundensko jezero* (Gmunden Lake) (SRP18690) (22). The shortest sentences were used in novels: *Hajduk Stanko* (Haiduk Stanko) (SRP18963) (7), *Radetića Mara* (Radetić's Mara) (SRP18940) (8), *Srbin i Hrvatica* (A Serb and a Croat woman) (SRP18921) (9), *Seljanka* (A peasant woman) (SRP18932) (9). It is interesting to note that *Hajduk Stanko* and *Seljanka* were written by the same author.

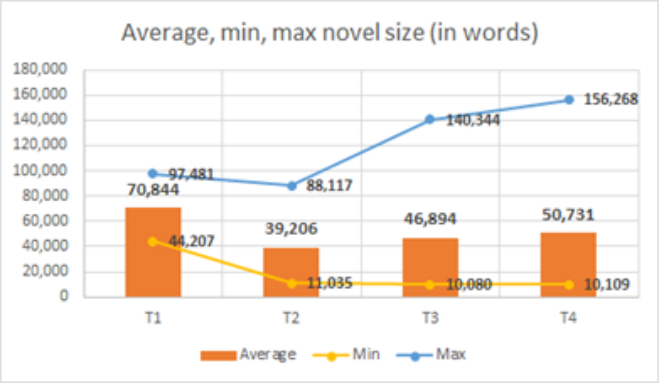

Figure 6. The average size, shortest and longest novels counted by the number of words per time period.

The frequencies for part of speech tags are given in Figure 7, which shows that nouns are the most frequent, followed by verbs and other parts of speech.

The lexicon statistics retrieved by TXM gives insight into most frequently used nouns, verbs, adjectives and pronouns. The 12 most frequent words from each of these groups are given in Table 2. One can see that the most frequent nouns are *ruka* (hand), *kuća* (house) and *dan* (day); the most frequent verbs (apart from auxiliaries) are *moći* (to can), *reći* (to say) and *znati* (to know); the most frequent adjectives are *drugi* (other or second), *velik* (big) and *star* (old); the most frequent pronouns are personal pronouns *on* (he), *ja* (I) and *ona* (she).

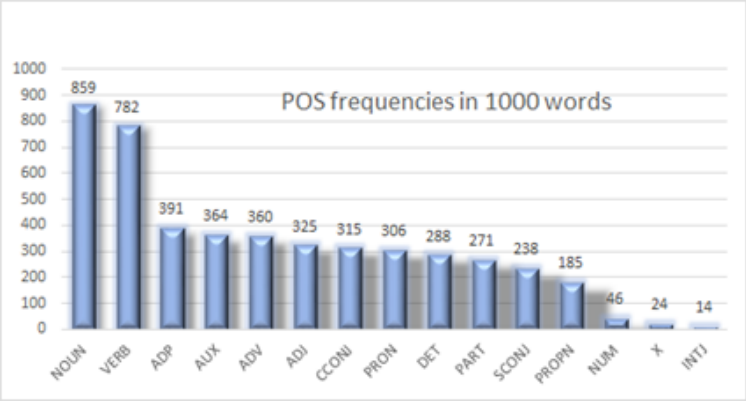

Figure 7. The frequencies of POS in SrpELTeC.

For each novels are calculated absolute frequency (F_i), where (i) represents specific novel and normalized length (Len_i) as the integer division of number of words in novel and 10000, so Len_i are numbers of values that fall in the interval $[1, 15]$. Figure 8 illustrates the most frequent named entities for four categories, using their lemmatized forms. Three frequency values are given for each category: absolute frequency in the whole corpus (green) (F_a), the number of novels in which a NE occurs (divided by 100) (blue) (F_n) and the relative frequency, taking into account both the length of a novel and the number of novels in which a particular NE occurs (yellow) (F_r) calculated using Equation (1).

$$F_r = F_n \cdot \sum_{i=1}^{65} \frac{F_i}{Len_i} \tag{1}$$

The most frequent PERS named entity, both measured by absolute and relative frequency is *Bog* (God). Apart from it the highest absolute frequency have the masculine personal name *Miloš* and feminine name *Darinka*. Measured by number of novels in which they occur, the most used are the masculine personal name *Pera* and feminine name *Mara*. The most frequent ROLE entities are *gospodin* (mister), *pop* (priest) and *gospođa* (missis), measured both by absolute and relative frequency. The roles that appear in most of the novels, besides *gospodin*, are *seljak* (peasant) and *gazda* (landlord). Other frequent roles are *kapetan* (captain), *učitelj* (teacher) and *kmet* (farmer).

As for DEMO entities, the entities referring to inhabitants or ethnic groups having the highest relative frequency are *Turci* (Turks), *Srbi* (Serbians) and

NOUN	FREQ	VERB	FREQ	ADJ	FREQ	PRON	FREQ
ruka	10,802	moći	20,353	drugi	12,553	on	89,345
kuća	10,461	reći	18,215	velik	8,341	ja	60,830
dan	9,238	znati	15,399	star	5197	ona	56,925
glava	8,046	imati	13,432	dobar	4,774	ti	23,599
oči	7,841	doći	9,106	prvi	4,773	vi	13,707
bog	6,998	viditi	8,252	mlad	4,474	šta	12,547
put	6,370	ići	7247	ceo	4,314	mi	8,299
čovek	6,019	kazati	7,065	lep	4,151	ko	8,274
ljudi	5,586	početi	6,780	nov	3,706	sebe	8058
reč	5,141	govoriti	6,462	crn	2,961	oni	5,262
žena	4,923	gledati	6,410	srpski	2,908	ništa	3,036
vreme	4,825	misliti	6,206	mali	2,887	ono	2,764

Table 2. The most frequent nouns, verbs, adjectives and pronouns

Cigani (Roma people). The most freqeunt adjectives referring to toponyms, inhabitants or ethnic groups are *srpski* (referring to Serbia or Serbians), *turski* (referring to Turkey or Turks) and *beogradski* (referring to Belgrade). The most frequent LOC entities both measured by absolute and relative frequency are *Beograd* (Belgrade) and *Srbija* (Serbia). Besides them, the frequently occurring countries are *Rusija* (Russia) and *Turska* (Turkey), the frequently occurring cities in Serbia are *Niš*, *Kragujevac* and *Užice*, the cities that are not in Serbia are *Beč* (Vienna), *Carigrad*/*Stambol* (Istanbul) and *Pariz* (Paris), and the most frequent rivers are *Dunav* (Danube), *Sava* and *Morava*.

Multi-word units are not annotated in the level-2 version of ELTeC collection, except for the named entities. Due to the existence of the incomplete morphological dictionaries of multi-word units we were able to retrieve the most frequently used multi-word nouns and adjectives. By far the most frequent multi-word noun is *srpski narod* (Serbian people), followed by *bojno polje* (battle field) and *vrhovna komanda* (High Command). The frequent multi-word nouns referring to education are *osnovna škola* (elementary school), *školska godina* (school year) and *učitelj muzike* (music teacher). It is interesting that adjectives *železnički* (referring to railway) and *električni* (electric) are used in numerous multi-word nouns revealing the modernization of Serbia: *železnička pruga* (railway) and *železnička stan-*

PERS	F	Numb	RelFKc
bog	2211	0.52	358.50
Boža	641	0.57	112.82
Milan	601	0.2	60.60
Pera	536	0.24	50.66
Miloš	1203	0.19	41.46
Mara	732	0.24	40.94
Milana	335	0.18	36.19
Stojan	572	0.16	34.57
Srba	257	0.32	31.59
Danica	672	0.14	31.07
Marka	382	0.31	31.00
Jov	428	0.21	30.25
Ljubica	671	0.15	28.08
Sima	634	0.19	26.97
Jelena	366	0.1	26.77
Nikola	356	0.21	23.12
Darinka	765	0.08	22.19
Milica	244	0.12	19.09
Ana	563	0.13	17.94
Steva	481	0.14	17.37
Mari	230	0.22	17.33
Sava	265	0.24	16.39
Petar	253	0.24	15.73
Jova	193	0.17	15.00
Ivan	362	0.15	14.72

srlemma	F	Num	RelFK
gospodin	2238	0.57	431.87
pop	1257	0.44	156.53
gospođa	1123	0.43	144.23
kapetan	1075	0.31	142.40
učitelj	996	0.45	135.87
gazda	655	0.51	96.91
seljak	558	0.52	95.79
g.	650	0.40	90.66
gospodar	639	0.41	70.64
kmet	810	0.30	63.31
ministar	453	0.24	59.29
doktor	407	0.35	58.37
đak	590	0.36	50.28
gospođica	426	0.33	47.37
trgovac	312	0.47	46.51
predsednik	433	0.27	45.07
činovnik	232	0.40	36.38
sluga	289	0.43	34.89
lekar	283	0.35	27.26
car	294	0.30	27.22
pandur	272	0.28	25.79
knez	416	0.24	25.16
pisar	386	0.27	23.49
poslanik	178	0.25	21.04
radnik	192	0.33	19.78

DEMO	F	NumNc	RelFKol
Turčin	1788	0.37	159.95
srpski	1047	0.52	148.41
turski	674	0.41	71.44
Srbin	273	0.33	33.20
Ciganin	139	0.26	12.54
beogradski	166	0.27	11.98
nemački	93	0.25	9.82
Srpkinja	83	0.19	8.43
ruski	128	0.21	6.13
grčki	95	0.2	5.29
francuski	67	0.17	3.82
Arnautin	137	0.07	3.76
Grkinja	158	0.04	3.14
Hrvat	58	0.06	2.75
Grk	74	0.14	2.20
Nemac	43	0.14	1.96
Madžar	67	0.1	1.94
Arapin	84	0.09	1.76
užički	35	0.09	1.72
Vlah	26	0.11	1.69
Rus	79	0.11	1.69
carigradski	35	0.11	1.31
Srba	24	0.11	1.11
Bugarin	37	0.11	1.10
Ciganka	31	0.11	1.06

LOC	Freq	Num	RelFKc
Beograd	785	0.52	120.39
Srbija	594	0.42	57.25
Niš	118	0.20	7.44
Rusija	157	0.20	6.98
Dunavo	118	0.24	6.93
Kosovo	111	0.24	6.78
Sava	105	0.20	5.24
Beč	99	0.19	5.10
Carigrad	107	0.18	4.45
Morava	81	0.17	4.39
Turska	89	0.19	3.79
Pariz	66	0.16	3.22
Kragujevac	62	0.16	3.03
Pešta	102	0.11	2.75
Dunav	60	0.18	2.72
Stambol	50	0.13	2.33
Bosna	86	0.09	2.05
Užice	62	0.08	1.98
Kruševac	65	0.08	1.80
Zemun	30	0.14	1.65
Šumadija	30	0.15	1.59
Karlovci	74	0.07	1.51
Kalemegdan	43	0.07	1.39
Srem	24	0.15	1.37
Rudnik	90	0.07	1.31

Figure 8. The frequencies of PERS, ROLE, DEMO and LOC categories in 65 novels of SrpELTeC.

ica (railway station), *električna struja* (electric current), *električna baterija* (electric battery), *električna lampa* (electric lamp), *električna sijalica* (electric bulb), *električno zvonce* (electric bell) and *električna centrala* (electric power station). Frequently occurring multi-word nouns with figurative meaning are *mrtva tišina*, (dead silence) *grobna tišina* (grave silence) and *crne misli* (black thoughts). Among multi-word adjectives, excluding similes (see (Krstev 2021) in the same issue) and demonyms are: *živ i zdrav* (alive and healthy), *go i bos* (nude and barefoot), *mrtav pijan* (deadly drunk), *mrtav umoran* (deadly tired).

5 Conclusions

In this paper we presented the results of the team work on producing the so-called level-2 edition of SrpELTeC. We gave an overview of the required schema with its main characteristics, and challenges in processing. Serbian level-2 pipeline included adaptation of SrpNER for named entity annotation, preparation of TreeTagger model for Serbian with the Universal Dependencies tagset, part of speech annotation and lemmatization within TXM tool, and preparation of several scripts for file transformations. Finally, statistics generated by TXM are supplied for several tags used as structural elements. Statistics are generated by using TXM.

Further plans include NER annotation of remaining 35 novels of SrpELTeC and adaptation of the output format to be compliant with the final level-2 schema, which is expected soon. The addition of the new layer with multi-word units annotation is also envisaged. The srpELTeC corpus will be further analysed by the quantitative and qualitative approach to researching textual corpus elements within the TXM program with the textometric approach (Heiden 2010; Jaćimović 2019) and visual presentation of the obtained results, as well as Latent semantic analysis. Various other analyses will be possible with this valuable resource, like authorship attribution, the lexical attraction between words (co-occurrence analysis), text specificity analysis, MWE and collocation extraction, dictionary example extractions, named entity linking, sentiment analysis, direct speech, word embeddings.

Acknowledgment

We would like to thank Prof. Serge Heiden and his team from the IHRIM, École normale supérieure de Lyon for helping us adapt TXM platform for the purpose of this work. We are also grateful to Prof. Diana Santos from the Department of Literature, Area Studies and European Languages at University of Oslo for many fruitful discussions that helped in finding best solutions to various problems.

References

Burnard, Lou, Christof Schöch, and Carolin Odebrecht. 2021. "In search of comity: TEI for distant reading." *Journal of the Text Encoding Initiative,* no. 14.

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

Heiden, Serge. 2010. "The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme." In *24th Pacific Asia conference on language, information and computation,* 2:389–398. 3. Institute for Digital Enhancement of Cognitive Development, Waseda University.

Heiden, Serge, Jean-Philippe Magué, and Bénédicte Pincemin. 2010. "TXM: Une plateforme logicielle open-source pour la textométrie-conception et développement." In *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010,* 2:1021–1032. 3. Edizioni Universitarie di Lettere Economia Diritto.

Jaćimović, Jelena. 2019. "Textometric methods and the TXM platform for corpus analysis and visual presentation." *Infotheca* 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.* Faculty of Philology of the University of Belgrade.

Krstev, Cvetana. 2021. "White as Snow, Black as Night – Similes in Old Serbian Literary Texts." *Infotheca - Journal for Digital Humanities* 21 (2): 119–135. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.6.

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

Paumier, Sebastian. 2021. *Unitex 3.3 User Manual.* Université Paris-Est Marne-la-Vallée. https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-en.pdf.

Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. "Serbian NER& Beyond: The Archaic and the Modern Intertwined." In *Deep Learning Natural Language Processing Methods and Applications – Proc. of the Int. Conf. Recent Advances in Natural Language Processing (RANLP 2021),* edited by Galia et al. Angelova, 1252–1260. INCOMA Ltd. https://doi.org/10.26615/978-954-452-072-4_141.

Stanković, Ranka, Diana Santos, Francesca Frontini, Tomaž Erjavec, and Carmen Brando. 2019. "Named Entity Recognition for Distant Reading in Several European Literatures." *DH Budapest 2019.*

Vitas, Duško, and Cvetana Krstev. 2012. "Processing of corpora of serbian using electronic dictionaries." *Prace Filologiczne* 63:279–292.