SrpELTeC on Platforms: Udaljeno čitanje, Aurora, NoSketch

Ranka Stanković, Mihailo Škorić, Petar Popović

10.18485/infotheca.2021.21.2.7

SrpELTeC on Platforms: *Udaljeno čitanje*, Aurora, noSketch

UDC 004.738.5:027
DOI 10.18485/infotheca.2021.21.2.7

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs

Petar Popović
petar.popovic@rgf.bg.ac.rs

University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

ABSTRACT: Serbian ELTeC collection (100 novels and extended) developed within COST action CA16204 Distant Reading for European Literary History comprises at this moment 111 novels published in the period 1840-1920. Such a valuable resource is and will be used for various lexical and linguistic research, by using different tools and methodologies. In this paper, three platforms on which these novels are published will be presented: "Udaljeno čitanje", Aurora and Sketch Engine.

KEYWORDS: distant reading, literary corpus, digital library, concordances, ELTeC.

PAPER SUBMITTED: 2 November 2021
PAPER ACCEPTED: 24 November 2021

1 Introduction

In the past a printed book was the most reliable way to store information and share it with others. Modern digital technology has made it possible to copy, store and share information even from rare, antique and fragile books. The majority of books in the Serbian ELTeC collection (SrpELTeC) were not well known and accessible to the public. Having in mind the effort invested and the importance of the whole collection, we wanted to make it available to as many people as possible. The second aim was to make it available through various channels, in order to meet the needs of different types of users. The second section will present one of the platforms where these novels are published, "Udaljeno čitanje", intended for readers who would like to see the original print as a picture while reading the digitized version. The Aurora portal, which will be elaborated in the third section, is 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 Sketch Engine, 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, which is usually explored by linguists, lexicographers, translators, students and teachers, will be described in the fourth section.

2 udaljenocitanje.unilib.rs

The platform "Udaljeno čitanje"[1] developed at the University library "Svetozar Marković", in cooperation with the University of Belgrade, Faculty of Philology, and the Society for language resources and technologies Jerteh, was supported by national projects[2] in the field of digitization of cultural heritage and contemporary creativity for 2019.

The platform is currently populated with 34 Serbian ELTeC novels, and addition of other Serbian ELTeC novels is planned for the near future. One can browse the novels, select one and read it page by page. A user gets two parallel versions of a chosen text, a picture of the original scanned page on the right and a digitized, machine readable text on the left. Figure 1 presents, in the upper part of the screen, the tenth page of the novel "Rajko od rasine" by Čedomilj Mijatović (SRP18920), with the OCR-ed and corrected text on the left and the original scan of the same page on the right - and at the lower part of the screen, metadata for the same novel. Apart from novel's title, author, publication place, the names of persons responsible for text preparation are given, as well as links to Wikidata, Wikipedia, and Cobiss.[3]

Footnotes in the original text were appropriately encoded and referenced in its digitized version. A small "information" sign in a digitized text signals the existence of a note (a footnote in the original), which, upon a click, appears in the form of a popup window containing the note's text, as can be seen in Figure 2.

1. Udaljeno čitanje

2. Project call of the Ministry of Culture and Information of Serbia - Sector for Digitization of Cultural Heritage and Contemporary Creativity for 2019, project number 119-01-00127 / 2019-09 and 401-01-00182 / 2019-09.

3. Cobiss+

Figure 1. The reading layout at udaljenocitanje.unilib.rs

Figure 2. Popup window containing text from an original footnote.

3 Aurora

The name of this portal was chosen to honour the memory of the AURORA[4] (AUtomatska Rutina za Obradu RečnikA – The Automathic 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. At the home page of this portal, a user can find more information about AURORA program and how it was used to solve problems for which today the corpus processor Unitex/GramLab,[5] a software system that integrates many initial ideas for processing input text in Serbian is being used (Vitas 1980; Krstev 1997, 2008; Vitas and Krstev 2012).

The purpose of the portal is to provide researchers of Serbian literature and other interested users with "microscopic" insight into the vocabulary of a number of works of Serbian literature, both written in prose and verse, offering a user, not only to browse the texts, their concordances and frequency lists, but also to navigate between a text and a list of words extracted from it.

The default preview on the portal's main page shows all titles in two big groups: prose works and poetry. In each of these groups works are listed by authors. The toolbar at the top of the page offers filtering: only names of authors, only female authors and their works, or only works from the ELTeC text collection. Filtering can also be done using the search box, by starting to type an author's name or a work's title, either in Cyrillic or Latin script.

Several authors and titles are linked with Wikidata, while further linking is an ongoing activity and expected to be finished soon. Linking of Wikidata and ELTeC collection is supported by Wikimedia Serbia[6] within the project "wikiELTeC – Wikidata about old Serbian novels from collection ELTeC

4. Aurora
5. Unitex Corpus Processing Suite
6. Wikimedia Serbia

(input, linking of named entities, visualization and analysis)". All 100 novels from Serbian ELTeC sub-collection[7] and 11 from the extended sub-collection are available through the AURORA platform.

Figure 3. The home page of the AURORA platform.

Each text in the AURORA collection is initially processed in order to obtain its inverted version, an alphabetically ordered index in which each entry containing a word and its frequency in a text points to a list of all occurrences of that word. This index can be presented to a user for browsing, ordered either alphabetically or by frequencies. The list can be complete (using button ▤ on the page represented in Figure 3) or filtered, so that the most frequently used words such as conjunctions, prepositions etc. are eliminated (using button ▤). This representation of texts enables the construction of concordances directly and linking of each concordance keyword with a broader context in the full text preview. Namely, using the Unitex-Gramlab locate module with the following regular expression, concordances are generated for all words in a text (except XML elements and their attributes).

```
<WORD><<[^li|div|head|n|p|lg|p\srend=\\\"Tekst\\\"|text|appInfo|
application|encodingDesc|fileDesc|item|encoding|author|body|
```

7. SrpELTeC

```
document|label|meTypesetSize|publicationStmt|sourceDesc|type|
unknown]>>
```

Another option deletes the so-called stopwords from the list of words:

```
<WORD><<[^li|div|head|n|p|lg|p\srend=\\\"Tekst\\\"text|appInfo|
application|encodingDesc|fileDesc|item|encoding|author|body|
document|label|meTypesetSize|publicationStmt|sourceDesc|type|
unknown|и|у|да|на|за|од|а|са|о|из|али|до|што|као|или|по|како|с|када|
јер|због|према|па|после|ако|без|пре|док|око|код|против|него|кад|
уз|већ|где|између|под|пред|преко|међу|иако|кроз|ни]>>"
```

The process of concordance generation is integrated into Leximir (Stanković et al. 2011; Stanković et al. 2012) and for each novel in SrpELTeC that is in level-1 TEI form the following is produced:

- a separate header for metadata extraction;
- a separate body of the text for production of concordances;
- index files (full and reduced);
- html files with concordances (full and reduced);
- html form of the novel.

The full use of the system is illustrated in Figure 4. In this way, AURORA provides insight into the vocabulary of a literary work and is the initial step in creating a dictionary of words used by individual writers. Future versions of this portal, which will use the full content of the system of electronic morphological dictionaries for the Serbian language, will give an even more elaborate insight into literary works.

Let us mention here some directions for future development, one of which will be lemmatizing concordances and associating words in the index with semantic attributes contained in electronic dictionaries. We also plan to integrate named entities, extracted from level-2 version of texts annotated with names of persons and their roles (professions, positions and titles), locations, organisations, events, work titles, and demonyms (Frontini et al. 2020; Šandrih Todorović et al. 2021). This would enable users to browse and search for concordances for a particular named entity class or a particular named entity. Named entities linking with Wikidata and integration with other knowledge bases is also envisaged.

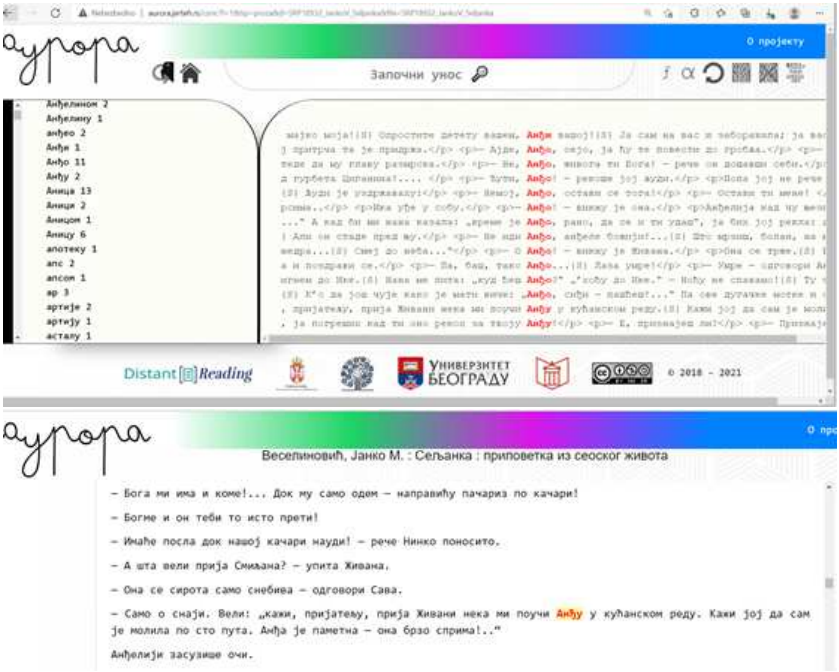

Figure 4. The novel *Seljanka* (The Peasant Woman) (SRP18932) by Janko Veselinović: a list of word forms with their frequencies (top left); b) selected concordances for forms of the name *Anđa* (top right); c) one of the chosen forms displayed in full context (bottom).

4 Sketch Engine

Sketch Engine[8] is a widespread tool to explore how language works, based on analysis of corpora compiled from authentic texts of billions of words. It can promptly identify what is typical in language and what is rare, unusual or emerging usage, and enables text analysis and text mining applications through API features.[9] Main end users of Sketch Engine are linguists, lexicographers, translators, students and teachers.

The Sketch Engine contains 500 ready-to-use corpora in 90+ languages, each having a size of up to 60 billion words to provide a truly representative sample of a language. With the Sketch Engine the 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).

A reduced version of the Sketch Engine is available as an open source edition under the name NoSketch Engine. It offers core corpus processing and search features, but it does not support word sketches, preinstalled corpora, term extraction and other more advanced features. A NoSketch Engine node[10] is installed and maintained by the Society for Language Resources and Technologies JeRTeh, offering access to several monolingual and bilingual corpora. For some of them, access is granted to authorized users only, while a number of them are available without authorisation. The SrpELTeC corpus can be freely accessed and searched using CQL (Corpus Query Language).

The SrpELTeC corpus in NoSketch is part of speech annotated and lemmatized using TreeTagger (Schmid 1999) with a tagging model, located in the parametric language parameter file, trained on the harmonized resources, which have been manually annotated within different projects (Stanković et al. 2020). The vocabulary that TreeTagger consults when lemmatizing is the system of morphological electronic dictionaries of the Serbian language authored by Cvetana Krstev and Duško Vitas (Krstev 2008; Vitas and Krstev 2012).

Figure 5 presents a page with a simple CQL query `[tag="A.*"][lemma="život"]`, which retrieves concordances with bigrams, where the first word is an adjective and the second is any form of

8. Sketch Engine
9. API features
10. SrpELTeC at JeRTeh

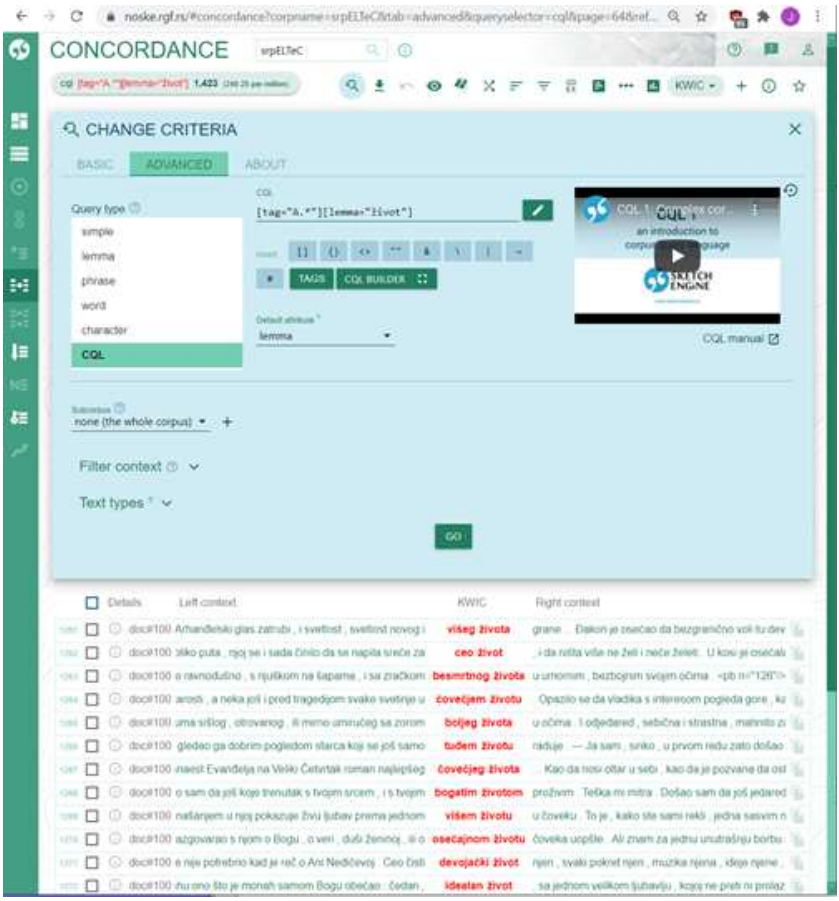

Figure 5. Concordances of the query `[tag="A.*"][lemma="život"]` in SrpELTeC.

the lemma *život* (life), e.g. *višeg života* 'higher life', *ceo život* 'whole life', *besmrtnog života* 'immortal life', *boljeg života* 'better life', etc.

The statistics of retrieved KWIC (key words in context) from concordances are presented to the user in the form of a table with absolute and relative (per million) frequencies for lemmatized forms, as presented in Figure 6. The most frequent adjectives that precede the lemma život are: *ceo* 'whole', *nov* 'new', *bračni* 'marital', *društven* 'social', *dug* 'long', etc.

The Faculty of Mining and Geology obtained access to the Sketch engine through the ELEXIS project.[11] Also, the Serbian ELTeC sub-collection is available on this platform for authorized users.[12] As already mentioned, there are additional features available in this environment, such as: word sketches, word clouds, thesaurus, sketch differences etc.

The word sketch feature processes the collocates of a word and other words in its neighborhood (McCarthy et al. 2015; Thomas 2014). Figure 7 presents a word sketch in the form of a set of collocations, grouped by grammatical patterns organized into categories, called grammatical relations, such as words that serve as an object of a verb, words that serve as a subject of a verb, words that modify a word etc. This one-page summary of a word's grammatical and collocational behavior allows for further browsing of concordances for a selected collocate. The sketch grammar is a set of rules written in CQL, based on part of speech tags and regular expressions defining which tokens should be included in the grammatical relation. For example, a subject may be defined as a noun preceding a verb, with additional requirements specified for both components, such as relative positions of the noun and the verb. Also, patterns can include required and optional words between specified components (in this case a noun and a verb).

The visualisation of the word sketch for lemma *život* in SrpELTeC in the form of a diagram is given in Figure 8. Distance from the centre of the big circle in which *život* is located reflects typicality (score): *ceo život* is more typical than *bračan život*. Circle size is related to the frequency: *ceo život* is more frequent than *dug život*. Circle colour indicates which segment (grammatical relation) collocations belong to, because circles may be positioned out of their segments for better visualization. Segment size indicates the size of the grammatical relation relative to other visualized relations, i.e. the number of collocations it contains in total, not just the number of collocations that are visualized.

11. European lexicographic infrastructure
12. SrpELTeC at Sketch Engine

	Lemma	↓ Frequency	Frequency per million
1	ceo život	89	15.03
2	nov život	78	13.17
3	bračni život	37	6.25
4	društven život	24	4.05
5	drugi život	23	3.88
6	nem život	20	3.38
7	javan život	19	3.21
8	lep život	17	2.87
9	seoski život	16	2.70
10	dobar život	16	2.70
11	zajednički život	14	2.36
12	pun život	14	2.36
13	miran život	14	2.36
14	đački život	13	2.19
15	čovečji život	13	2.19
16	prav život	12	2.03
17	običan život	12	2.03
18	narodni život	12	2.03
19	domaći život	12	2.03
20	porodičan život	11	1.86

Figure 6. The frequency results of the query `[tag="A.*"][lemma="život"]` on SrpELTeC.

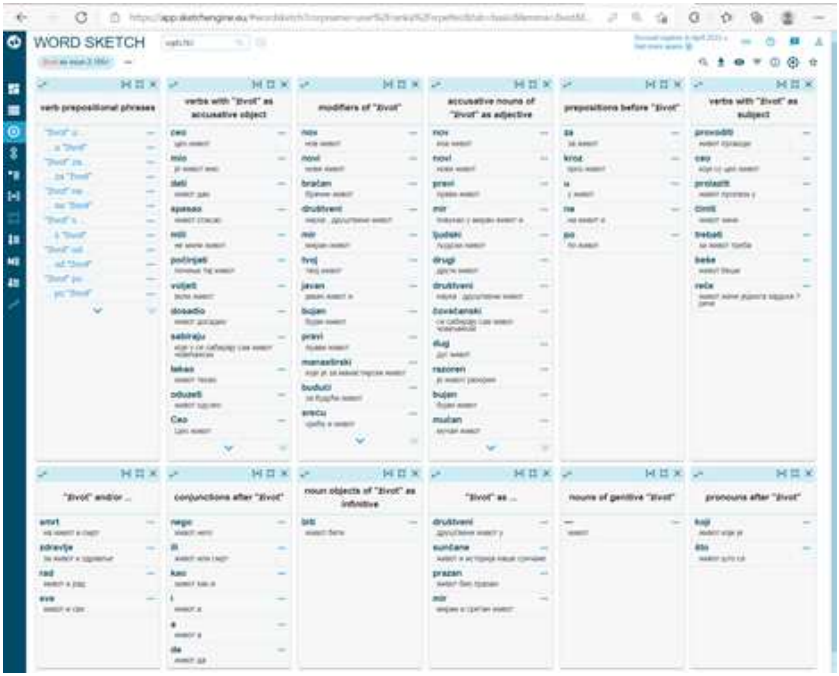

Figure 7. One-page overview of the word sketch for the lemma *život* in srpELTeC.

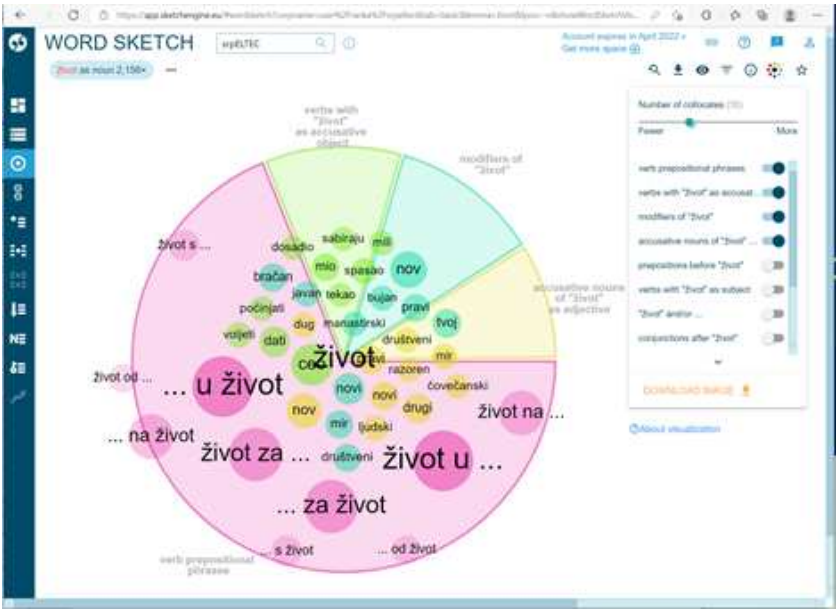

Figure 8. The visualization of word sketch for lemma *život* in SrpELTeC.

The thesaurus presented in Figure 9 is an automatically generated list of synonyms or words belonging to the same semantic field as the required word *život*. The thesaurus entries are retrieved from the context in which the word *život* appears in the srpELTeC corpus.

The synonym candidates are identified automatically from the context in which they occur, relying on the distributional semantic theory hypothesis according to which words that appear in the same context have similar meaning. In the Sketch Engine interpretation, it means that words with similar collocations probably have similar meaning, so the word sketch serves as the base for the calculation of the similarity score. To determine synonyms for *život*, the word sketches of all nouns are compared with its word sketch, and those that share the largest proportion of collocates are listed as similar words. The score assigned to the synonym represents the percentage of shared collocates.

Visualisation in Figure 10 contains information about the frequency and similarity scores of the lemma *život* in SrpELTeC. A circle size reflects the frequency of the encircled word: *čovek* ‘man’ is more frequent than *rad* ‘work’.

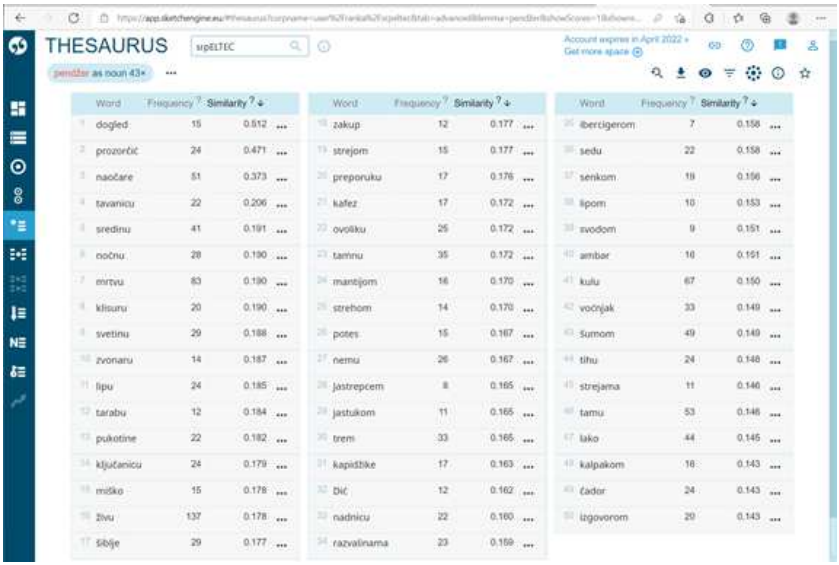

Figure 9. Thesaurus for lemma *život* in SrpELTeC.

The distance from the circle centre, where again *život* is situated, depends on the similarity score: *dan* 'day' is more similar to *život* than *posao* 'work'. This example shows that none of the candidates need actually be synonyms. To obtain more reliable results one would need to work on a much bigger corpus.

The word sketch difference provides comparisons by contrasting collocations that can be retrieved by lemmas, word forms or subcorpora. Comparing the collocations can provide a deeper understanding of the difference in use and meaning. Figure 11 presents the word sketch difference for *život* and *smrt* (death), which compares the use of the two lemmas by comparing their collocates. Different colours are assigned to the chosen words and their word sketches, and for each collocate, in each grammatical relation separately, the results for both words are compared. The colours indicate the word for which the collocate is more frequent (blue for *život*, pink for *smrt*), while the shade of the colour indicates the strength of the collocation. The words in white (for example *ja*, *sam* in the centre of the bottom boxes in Figure 11) are collocates without a preference.

An additional option is the "word forms" option, which compares the use of two different word forms of the same lemma via their collocates. A third

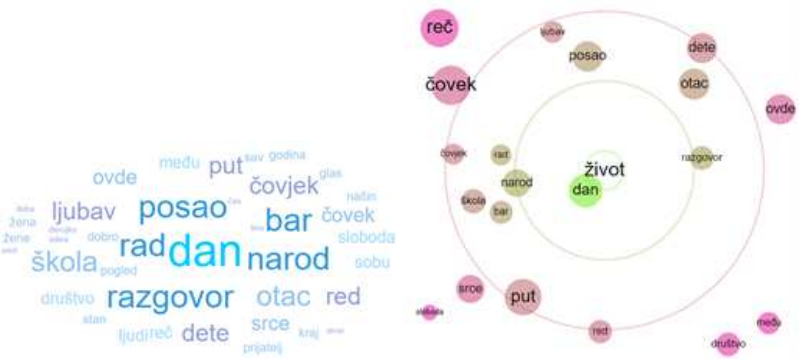

Figure 10. Visualisation of the thesaurus for lemma *život* in SrpELTeC.

option is "subcorpora", which compares the use of the same lemma in two different subcorpora of the same corpus, via their collocates.

5 Concluding remarks

In this paper three platforms supporting the SrpELTeC collection of novels are presented. The first is "Udaljeno čitanje", which enables browsing and reading of digitized text, with a preview of the scanned original. The platform will be further improved by publishing more novels, as well as by introducing advanced features for search and filtering. The second platform is the Aurora portal, which provides researchers with "microscopic" insight into the vocabulary of selected Serbian literary works. In addition to the text, concordances and word frequencies are also available, as well as navigation between a text and a list of words extracted from it. Apart from further expansion of resources, Aurora will be more tightly integrated with Wikidata, presenting results from predefined SPARQL queries, with authors and their novels, novels linked with main characters and their roles, in form of graphs, tables, timelines (Ikonić Nešić, Stanković, and Rujević 2021) and locations of the events on the maps etc. Browsing lists of rare words and browsing by authors, will also be enabled. The third platform is the Sketch Engine, for corpora management and exploration, as well as for analyzing large texts. Integration of the Aurora portal and the Sketch Engine is envisaged, since Aurora is not optimised for large novels. We believe that the developed platforms will contribute to raising the visibility of SrpELTeC, a valuable resource in Serbian language for linguists, but also bring a part

of literary history that was unknown or unavailable closer to to the wider community.

Figure 11. The word sketch difference for *život* (blue if preferred) and *smrt* (pink if preferred) in SrpELTeC.

Acknowledgment

The platform Udaljeno čitanje is developed with the support of projects in the field of digitization of cultural heritage and contemporary creativity for 2019: 119-01-00127 / 2019-09 and 401-01-00182 / 2019-09, supported by the Ministry of Culture and Information of Serbia – Sector for Digitization of Cultural Heritage and Contemporary Creativity.

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)".

The text collection preparation is supported by the COST Action 16204 – Distant Reading for European Literary History support. Access to SketchEngine is provided by the ELEXIS project funded by the European Union's Horizon 2020 research and innovation program under grant number 731015.

References

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.*

Ikonić Nešić, Milica, Ranka Stanković, and Biljana Rujević. 2021. "Serbian ELTeC Sub-collection in Wikidata." *Infotheca - Journal for Digital Humanities* 21 (2): 60–87. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.4.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. "The Sketch Engine: ten years on." *Lexicography* 1 (1): 7–36.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. "Itri-04-08 the sketch engine." *Information Technology* 105 (116).

Krstev, Cvetana. 1997. "Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije." PhD diss., Univerzitet u Beogradu, Matematički fakultet, September.

Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* Faculty of Philology of the University of Belgrade.

McCarthy, Diana, Adam Kilgarriff, Milos Jakubicek, and Siva Reddy. 2015. "Semantic word sketches." *Corpus Linguistics 2015.*

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 Angelova et al., 1252–1260. INCOMA Ltd. https://doi.org/10.26615/978-954-452-072-4_141.

Schmid, Helmut. 1999. "Improvements in part-of-speech tagging with an application to German." In *Natural language processing using very large corpora,* 13–25. Springer.

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac, and Miloš Utvić. 2012. "A tool for enhanced search of multilingual digital libraries of e-journals." In *Proc. of the 8th LREC),* edited by Nicoletta Calzolari et al. Istanbul, Turkey: European Language Resources Association (ELRA).

Stanković, Ranka, Ivan Obradović, Cvetana Krstev, and Duško Vitas. 2011. "Production of morphological dictionaries of multi-word units using a multipurpose tool." In *Proceedings of the Computational Linguistics-Applications Conference, October 2011, Jachranka, Poland,* 77–84.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian." In *Proc. of The 12th LREC,* 3947–3955. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.487.

Thomas, James. 2014. "Discovering English with the Sketch Engine." *Research-publishing. net,* 161–176.

Vitas, Duško. 1979. "Prikaz jednog sistema za automatsku obradu teksta." In *INFORMATICA'79, Bled,* 7101 1–5.

Vitas, Duško. 1980. "Generisanje imeničkih oblika u srpskohrvatskom." *Informatica,* no. 3, 49–55.

Vitas, Duško, and Cvetana Krstev. 2012. "Processing of Corpora of Serbian Using Electronic Dictionaries." *Prace Filologiczne* XVIII:279–292.