An Italian-Serbian Sentence Aligned Parallel Literary Corpus

Saša Moderc, Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić

An Italian-Serbian Sentence Aligned Parallel Literary Corpus | Saša Moderc, Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić | Review of the National Center for Digitization | 2023 | |

10.5281/zenodo.11203388

http://dr.rgf.bg.ac.rs/s/repo/item/0008582

Review of the NCD 43 (2023), 78–91

Saša Moderc
University of Belgrade - Faculty of Philology
Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić
University of Belgrade - Faculty of Mining and Geology

AN ITALIAN-SERBIAN SENTENCE ALIGNED PARALLEL LITERARY CORPUS

Abstract. This article presents the construction and relevance of an Italian-Serbian sentence-aligned parallel corpus, delving into the aligned sentences in order to facilitate effective translation between the two languages. The parallel corpus serves as a valuable resource for language experts, researchers, and language enthusiasts, fostering a deeper understanding of linguistic nuances and cultural expressions. By bridging the gap between Serbian and Italian, this corpus opens new avenues for cross-cultural communication and collaboration, and ultimately contributes to the improvement of language-related applications and studies. In this paper we present the methodology of corpus building, statistics on corpus content, availability, sample queries and plans for future research.
Keywords. Aligned corpus, parallel corpus, Serbian, Italian, literature.

1. Introduction

The project originated from the exigency for resources dedicated to the investigation and examination of bilingual Italian-Serbian texts. The lack of corpus resources within the domain of foreign language teaching in Serbia, as well as the limited digital resources for teaching Serbian and Italian as foreign languages, inspired this project. The complex morphology of Serbian, such as declension issues, poses challenges for foreign students. In foreign language teaching, parallel corpora (texts existing in both languages and maintaining a translation relationship) are essential for acquiring morphosyntax and lexis, allowing contrastive analysis of language structures. This approach, as advocated by Sinclair [1] in the field of corpus linguistics, provides a comprehensive view of language. In translation teaching, parallel corpora aid in word sense disambiguation and defining polysemic lexicon, thereby mitigating the limitations inherent in bilingual dictionaries. Nevertheless, the scarcity of representative parallel corpora for major languages is noted. Krstev and Vitas [2] outline the sequential stages involved in constructing parallel corpus, encompassing text selection, pre-processing, alignment procedures and result verification. They delve into potential applications of such corpus, emphasizing the significance of utilization of corpora processing tools supported by language-specific modules.

Existing corpora It-Sr-NER[1] [3], although limited in size, have been proven valuable in language teaching, stimulating collaborative work and independent research. This bilingual Italian-Serbian corpus and supporting web services for annotating named entities in text, linking them to Wikidata and geoparsing, were developed by a team from the University of Turin and the Society for Language Resources and Technologies (JeRTeh) within the It-Sr-NER project and the CLARIN infrastructure's "Bridging Gaps" initiative. Professor Olja Perišić, from the University of Turin, initiated and led the project, and JeRTeh played a crucial role in service development. The first phase of

[1] https://jerteh.rs/index.php/it-sr-ner-3/

the project involved creating an Italian-Serbian corpus with 10,000 aligned segments from classic literature based on ten novels, covering works by Italian authors like Umberto Eco and Serbian authors including Ivo Andrić. This corpus, prepared in TMX (Translation Memory eXchange) format [4] using the ACIDE [5] application, served the project's main goal of annotating named entities, namely persons, places, organizations, demonyms, events, and works of art. The aim was to create web applications for monolingual and bilingual texts, particularly Italian-Serbian, within both CLARIN and JeRTeh platforms. The project's impact extends beyond Serbian-Italian, as the services process text in twenty-four different languages [6].

The successful use of It-Sr-NER corpus was a motive to prepare a much larger corpus that would include more novels. This valuable resource was conveniently prepared by Prof. Saša Moderc, who has been aligning novels for a long period of time, with the goal of making them available to the wider professional and scientific audience. Section 2 presents the methodology of sentence aligning, metadata description, annotation description and formats used for corpus preparation. Section 3 brings corpus overview with statistical information about corpus size and content, authors and translators statistics. Typical queries and frequency statistics are presented in Section 4. At the end, we make concluding remarks and present plans for future improvements.

2. Methodology

The methodology adopted in this project will be presented from two points of view: linguistic and technical. In the first part we will give the overview of text selection, digitization and text alignment. In the second part we will present transformation of word/excel tables into TMX and other formats required for platforms on which the corpus was made available.

2.1 Text selection and corpus preparation. The interest in collecting texts of Serbian and Italian literary works and their translations into Italian and Serbian, respectively, emerged about a decade ago, around 2014. The impetus for this endeavor can be traced on multiple levels. On one hand, the collection of parallel literary texts was recognized as a tool to overcome the limitations of traditional Serbian-Italian dictionaries (through searching the collection of texts and finding applied translation solutions for lexemes and structures lacking adequate lexicographic treatment). This collection was also seen as a means to introduce documented materials into the teaching of the Italian language (in terms of contrastive analysis, syntax, lexicology, phraseology, etc.) and to provide reliable material for linguistic research. In this context, we will mention just three published works: Moderc [7], which is based on the analysis of approximately 300 Italian equivalents of the Serbian adverb 'inače'; followed by Moderc [8], presenting the initial phase of text parallelization and its potential for evaluating the translations themselves; and finally, Ceković-Janićijević [9], an article in which Italian and Serbian textual markers are analyzed, starting from the parallel texts made available to them.

On the other hand, a significant incentive for collecting texts was the fact that within researches in Italian studies there was no systematic record of published translations of Serbian and Italian authors, translator names were not listed, consequently, without specific linguistic material in digital form, it was challenging to document the criteria governing literary exchange between Serbia (and former Yugoslavia, in the realm of Serbo-Croatian) and Italy. The absence of these parameters made it impossible to systematically and quantitatively analyze the quality of existing translations. The outcome of future qualitative analyses of translations would primarily

have didactic application, without any intention to diminish the value of the effort made by the translators whose works are already included in the existing parallel texts, despite potential inaccuracies or errors in text interpretation.

The work on collecting texts was not conducted within a particular formal project, and in that regard, the criteria for selection and processing of texts did not follow a precise, predetermined procedure. There were no strict deadlines set for processing the texts. Initially, the focus was on Serbian authors translated into Italian, with sources including the library of Italian Studies at the Faculty of Philology, followed by the National Library of Serbia, the University Library "Svetozar Marković," and finally, some libraries in Italy, which allowed access to translations not available in Serbia (specifically, the translation of Bora Stanković's "Nečista krv"). Considering that the number of translations of Serbian works into Italian was relatively limited until the 1980s, the decision was made to encompass the majority of existing translations, setting aside chronological criteria, which are, however, dominant in the didactics of contemporary Italian language (therefore, works by Dositej Obradović, Ljubomir Nenadović, and Svetozar Marković were included among the processed authors). Strict language criteria for translations into Italian were also set aside, and works by Croatian authors, like Miroslav Krleža and August Šenoa, were also included in the collection of text. The intention was to gather existing translations into Italian and preserve them in electronic form as part of the Serbian and Yugoslav cultural heritage in Italian language.

When it comes to translations of Italian works into Serbian, the initial criterion was to digitize translations of established Italian authors of the twentieth century (specifically, those already included in Italian literary histories), regardless of the linguistic variations of the translations: in former Yugoslavia, some relevant Italian novels were translated only once, either in Croatian or in Serbian, and regardless of the perception of diversity among these two languages (or variations), in our estimation the translation labeled as “Croatian”, deserve also to be included in this selection of translations.

From the 1980s onwards, works by Serbian authors began to be translated into Italian in larger volumes. Consequently, the selection of available texts expanded, and the criterion of exceptionality or representativeness of translated literary works gave way to titles where, in our opinion, commercial reasons may coexist alongside literary and artistic criteria. This observation is particularly relevant when it comes to contemporary translations from the Italian language, where genres such as romance and crime novels are present, and some of these works may not leave a significant literary impact; nevertheless, they too contribute to the representativeness and comprehensiveness of the linguistic corpus.

The process of acquiring texts is based on scanning paper editions or downloading existing electronic editions, converting them into Word format, technical text editing, spell checking, and finally, parallelization using the wrapper LF Aligner[2] (developed by Andras Farkas) that helps translators create translation memories from texts and their translations. It relies on HunAlign[3] [10] for automatic sentence pairing and works for various input formats: txt, doc, docx, rtf, pdf, html. The primary objective of Aligner is to produce interchangeable translation memories (TMX), utilizing an open XML-based standard designed to facilitate the seamless exchange of translation memory data, specifically aligned parallel texts, among various tools and translation service providers. Parallel texts are generated in Excel table format and then transferred

[2] https://sourceforge.net/projects/aligner/
[3] http://mokk.bme.hu/resources/hunalign/

to Word as tables. From the perspective of a language teacher, this approach is justified by easier access and use of parallel texts in a familiar text processing program environment.

For linguistic needs and simultaneous consultation of multiple texts, the free program *DocFetcher*[4] was used. *DocFetcher* indexes many texts, enabling the search for a sequence of characters in selected texts. The program provides a rudimentary but quite satisfactory insight into both the original text and the translated text, meeting the needs of language study and analysis.

After some of the Serbian-Italian parallel texts were processed and made available to the public as a corpus as It-Sr-NER, the intention was to make available in a new online public corpus all the already parallelized texts, adding to this number those texts whose processing was started but was left uncompleted.

After processing and releasing certain Serbian-Italian parallel texts as the It-Sr-NER corpus, the goal was to create a new online public corpus comprising all previously parallelized texts. This would include texts that had undergone processing, as well as those whose processing had been initiated but remained incomplete.

2.2 Corpus processing, annotation and publication. The metadata for corpus comprise information about author (name and QID in Wikidata), title of the works, publisher and year of publishing of the work that was used for digitization, translator(s) names, number of translation (1-4), source language code, title of translation, publisher and year of translation (used for digitization). Since the ultimate goal is to make this collection accessible and visible as much as possible, digital versions of the corpus are prepared for several working environments in order to reach different types of users and solve different tasks.

One of the platforms is Biblisha, a tool that offers various possibilities of enhancing queries submitted to large collections of TMX documents generated from aligned parallel articles, novels or other documents. The queries, initiated by a keyword, can be expanded, both semantically and morphologically, using different supporting monolingual and multilingual resources, such as wordnets and electronic dictionaries. The tool is accessible on the web, both for authorized and unauthorized users, with different possibilities (Stanković at al. 2012) [11]. New collection number (15) has been created for SerbItaCor3 and two subcollections, 15.1 for Italian novels and 15.2 for Serbian novels, and a code was assigned for each author: 15.1.01-15.1.43 for Italian authors and 15.2.01-15.2.31 for Serbian authors. Additionally, a code for each (parallel) novel is assigned to each work of the author. For example: code 15.1.11 is assigned to Calvino Italo and his works are assigned codes 15.1.11.001-15.1.11.014. File name is also recorded for word version, excel version (with translation equivalents) and generated TMX xml format. If a novel, as “Il barone rampante” of Calvino Italo, has three translations, they are coded as 15.1.11.001.1-15.1.11.001.3. We can see in Figure 1 an example of an excel file containing multiple translations of the “Il barone rampante”.

[4] https://docfetcher.sourceforge.io/en/index.html

ID	original	prevod1	prevod2	prevod3
1	Italo Calvino	Italo Calvino	Italo Kalvino	Italo Kalvino
2	Il barone rampante	Barun penjač	Baron na drvetu	Baron na drvetu
3		Prevela Karmen Milačić	Prevela Jugana Stojanović	Prevela s italijanskog Ana Srbinović
4	I Fu il 15 di giugno del 1767 che Cosimo Piovasco di Rondò, mio fratello, sedette per l'ultima volta in mezzo a	I Bio je 15. lipnja 1767. kad je Cosimo Piovasco di Rondò, moj brat, posljednji put sjeo između nas.	I Petnaestog juna 1767. Kozimo Piovasko di Rondo, moj brat, poslednji put je s nama seo za sto.	I Bio je 15. juni 1767, kada je Kozimo Pjovasko di Rondo, moj brat, poslednji put seo sa nama za sto.
5	Ricordo come fosse oggi.	Sjećam se kao danas.	To mi se živo urezalo u sećanje kao da se danas zbilo.	Sećam se toga kao da je bilo juče.
6	Eravamo nella sala da pranzo della nostra villa d'Ombrosa, le finestre inquadravano i folti rami del grande elce del parco.	Bili smo u blagovaonici naše vile u Ombrosi, prozore su uokvirivale guste grane velikih vrtnih česmina.	Sedeli smo u trpezariji našeg letnjikovca u Ombrozi, u okvirima prozora videle su se brsnate grane velikog cera u parku.	Sedeli smo u trpezariji naše kuće u Ombrozi, prozori su uokvirivali gustu krošnju velike crnike koja je rasla u vrtu.
7	Era mezzogiorno, e la nostra famiglia per vecchia tradizione sedeva a tavola a quell'ora, nonostante fosse già invalsa tra i nobili la moda, venuta dalla poco mattiniera Corte di Francia, d'andare a desinare a metà del pomeriggio.	Bilo je podne, a naša je obitelj po starom običaju sjedala za stol u to vrijeme, iako je do plemića već bila doprla moda, prenesena s francuskog dvora gdje baš nisu bili ranoranioci, da se ruča sredinom poslijepodneva.	Bilo je podne i naša porodica se po starom običaju u to vreme okupila oko stola, iako je plemstvo već prihvatilo modu francuskog dvora, na kojemu se nije rano ustajalo, da se ruča sredinom popodneva.	Bilo je podne i naša porodica je po davnašnjem običaju u to doba sedela za trpezom, uprkos tome što se među plemstvom već raširila moda, preuzeta s Francuskog Dvora koji se baš nije odlikovao ranoraniocima, da se obeduje tek u kasno popodne.
8	Tirava vento dal mare, ricordo, e si muovevano le foglie.	Sjećam se, vjetar je puhao s mora i lišće se pomicalo.	Sećam se, duvao je vetar s mora i treperilo je lišće.	Duvao je vetar s mora, sećam se, i lišće se njihalo.
9	Cosimo disse: - Ho detto che non voglio e non voglio! - e respinse il piatto di lumache.	Cosimo je kazao: - Rekao sam da neću i neću! - i odgurnuo tanjur puževa.	Kozimo reče: - Kad sam kazao da neću, neću! - i odgurnu tanjir kuvanih puževa.	Kozimo reče: - Rekao sam da neću i tačka! - i odgurnu tanjir s puževima.
10	Mai s'era vista disubbidienza più grave.	Veće neposlušnosti još nije bilo.	To je bilo nezapamćeno otkazivanje poslušnosti.	Nikada pre u našoj kući nije ispoljena tako krupna neposlušnost.

Figure 1. Multiple parallel translation example

The development of previous parallel corpora was based on use of ACIDE application that has as output TMX file, but it requires file by file processing. Having large number of files, of which some have multiple translations, the approach with ACIDE would be time consuming. That motivated us to prepare a new application for metadata and aligned text management. First part of the application implemented conversion of metadata from excel with providing also statistical analysis of metadata. The second part was the conversion of excel files with aligned text into TMX format. It should be noted that in case or multiple translation, for each translation a separate TMX file is generated.

Since the starting text collection was in word format, organized as tables with aligned segments, the first step was to produce excel files that have a more versatile format for automatic processing. Just to mention that aligned segments are most of the time sentences, but due to differences in translations, sometimes one sentence can be translated as two or more sentences can be translated into one. Sometimes, for very long sentences, aligned segments can be just parts of the sentences. This step has been performed manually unifying headers in all excel documents. The python notebook is written for conversion of excel files into TMX. For multiple translations, a separate file is generated for each translation. Figure 2 presents two translation equivalents from the novel “Anikina Vremena” of Ivo Andrić, where it can be seen that TMX uses ISO

standards for country and language codes, as well as for date and time. A TMX document consists of a header (metadata describing the aligned texts) and a body, containing a set of translation units (TU) composed of two or more semantically equivalent translation unit variants (TUVs). Each TUV contains the text (one or more sentences or segments) in one of the TMX document languages, where the text in the first TUV is usually in the source language, and the texts in the remaining TUVs are in one or more target languages. In this example one translation unit variant is Italian (it) and the other is Serbian (sr). A TUV attribute xml:lang denotes the language of the text within the TUV.

```
<tu>
    <prop type="Domain"/>
    <tuv creationid="n5" xml:lang="it" creationdate="20231128T103638Z">
        <seg>Negli anni sessanta dello scorso secolo penetrò anche nelle più sperdute
        regioni di questo paese un indefinito ma forte desiderio di apprendere e di
        conseguire quella esistenza migliore che l'istruzione e la cultura comportano.
        </seg>
    </tuv>
    <tuv creationid="n5" xml:lang="sr" creationdate="20231128T103638Z">
        <seg>Šezdesetih godina prošloga stoleća prodirala je i u najudaljenije krajeve
        ove zemlje neodređena ali jaka želja za znanjem i boljim životom koji znanje i
        prosvećenost donose sobom.</seg>
    </tuv>
</tu>
<tu>
    <prop type="Domain"/>
    <tuv creationid="n6" xml:lang="it" creationdate="20231128T103638Z">
        <seg>Né la Romanija [Massiccio montano della Bosnia orientale] né la Drina
        poterono impedire che questo desiderio raggiungesse pure Dobrun [Cittadina
        della Bosnia orientale, nei pressi del confine con la Serbia] illuminando anche
        pop [Sacerdote cristiano ortodosso] Kosta Porubović.</seg>
    </tuv>
    <tuv creationid="n6" xml:lang="sr" creationdate="20231128T103638Z">
        <seg>Ni Romanija ni Drina nisu mogle sprečiti da ta želja ne prodre i u Dobrun
        i ne ozari i popa Kostu Porubovića.</seg>
    </tuv>
</tu>
```

Figure 2. An example of the TMX document for the novel “Anikina vremena”

The TMX format can be used for direct upload in Biblisha digital library [12] and for the m tool for textometrie [13]. The use of these two platforms will be illustrated in Section 4. For the third version of the corpus, the noSketch engine, additional preparation was required and also a custom application in python was prepared[5]. Namely, noSketch engine requires separate files for two languages, with the same number of segments in both files. Also, metadata are added as XML attributes of the root element for each novel.

The parallel corpus was annotated with part-of-speech (POS) tags and lemmas. The Serbian part of the corpus was annotated using a specialized, multi-model tagger for Serbian language[6] [14], which uses both *TreeTagger* [15] and *spaCy*[7] models trained on part-of-speech tagging task using the manually annotated, publicly available corpus *SrpKor4Tagging*[8] [16]. This system provides two tagsets: a traditional classification for Serbian part of speech tags and a Universal Dependency tagset[9] [17]. The lemmatization is performed after the POS-tagging step, using electronic morphological dictionaries for Serbian *SrpMD* [18, 19], incorporated through the aforementioned TreeTagger model.

[5] https://github.com/procesaur/tmx2noske
[6] https://github.com/procesaur/BEaSTagger
[7] https://spacy.io
[8] https://live.european-language-grid.eu/catalogue/corpus/9295
[9] https://universaldependencies.org/u/pos/

The Italian part of the corpus was annotated using the Italian *TreeTagger* model[10] provided by Achim Stein. The tagset used by the Italian model[11] comprises 38 tags, 13 for verbs, 8 for pronouns, 2 for determiners, 2 for prepositions and only one for all other categories.

3. Corpus overview

The corpus consists of 243 literary works, 129 written by Italian authors and 114 by Serbian, as presented in Table 1. Most of the works are included with one translation, but 8 Serbian works have 2 translations, with 3 translations there are 4 Italian and 1 Serbian novel. One Italian and one Serbian novel have 4 different translations. Luigi Pirandelo's "Il fu Mattia Pascal" (1904, we used a publication from 2001) has been translated by Filip M. Dominiković (1927), Miodrag Ristić (1940), Jugana Stojanović (1966) and Mirela Radosavljević and Aleksandar Levi (2004). Ivo Andrić's "Priča o vezirovom slonu" (published in 1947; we used a publication from 2012) has 4 translations: Luigi Salvini (1954), Bruno Meriggi (1966), Dunja Badnjević Orazi and Manuela Orazi Bašić (1995) and Dunja Badnjević (2001). The total number of translation pairs is 267, 140 for Serbian and 127 for Italian authors.

Language	Number of translations				Total works	Total translation pairs
	1	2	3	4		
Italian	124		4	1	129	140
Serbian	104	8	1	1	114	127
Total	228	8	5	2	243	267

Table 1. Number of translated works per language with number of translations

The total number of segments is 664.044, where partition with Serbian authors has 258.423 aligned segments, and Italian has 405.621 segments. The Italian part of the corpus has approximately 11.1 million words (4.9+6.2), while Serbian has 10.4 million (4.2+6.1). Figure 3 presents the number of words in the partitions where the source language is Serbian (left), per language, and where the source language is Italian (right), also per language.

[10] https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[11] https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/italian-tagset.txt

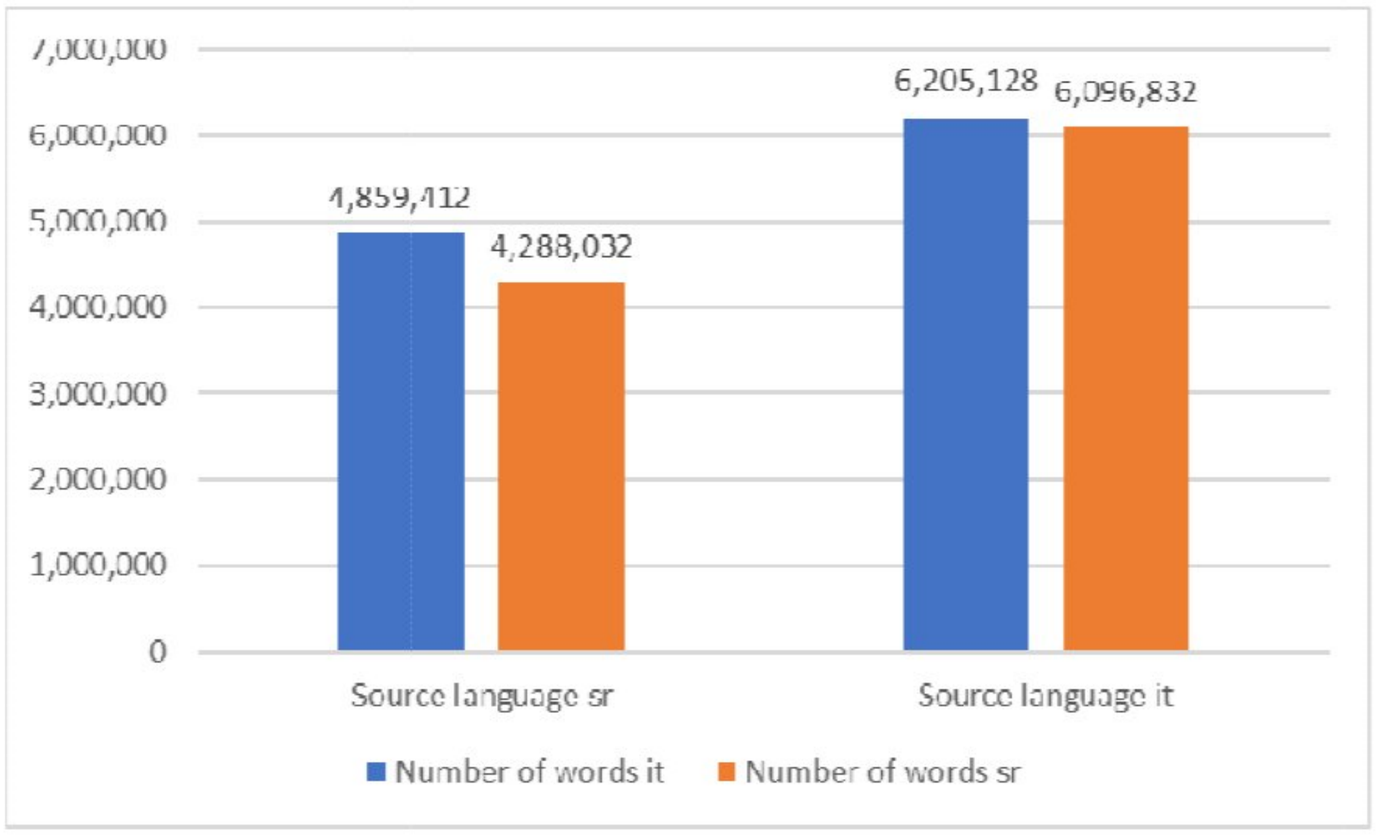

Figure 3. Number of words in language parts in partitions by source language

The total number of authors is 74, 43 Italian and 31 Serbian authors. In this corpus 24 Italian authors are represented with one work, 7 authors are represented with 2, 8 authors with 3 works, Baricco Alessandro has 4 works, Eco Umberto has 5 works, Calvino Italo has 14, and Pirandello Luigi has 43 (mainly stories, and one novel). The 19 Serbian authors are represented with one work, 3 authors with 2 works, 3 with 3 works, Crnjanski Miloš and Pavić Milorad are represented with 4 works, Kiš Danilo and Marković Svetozar with 5 works, Albahari David with 6 works and Andrić Ivo with 57 novels and short stories.

The information about translators is available for 255 translations. Most translators translate from either Serbian or Italian.

The most productive translator from Serbian to Italian in corpus SerbItaCor3 are Alice Parmeggiani with 25 translations, Dunja Badnjević (Orazi) with 16, Bruno Meriggi with 14. The number of translators working in pairs is 4: the most productive are Dunja Badnjević (Orazi) and Manuela Orazi (Bašić) with 13 translations, followed by Jelena Mirković and Elisabetta Boscolo Gnolo with 3 translations. From Italian to Serbian, the most productive translator is Ana Srbinović with 23 works, followed by Elizabet Vasiljević with 19 and Gordana Subotić with 10. Six translators worked together, in pairs: Mirela Radosavljević and Aleksandar Levi (7 works), Srbislava Vukov-Simentić and Snježana Marinković (3 works) and Ana Srbinović and Elizabet Vasiljević (1).

4. Show cases: practical usage

In this section we will give a few insights from a practical point of view for three different environments. The first is Biblisha, with a demonstration of query expansion capabilities [20]. Figure 4 presents the panel for search with keyword “vojnik” (soldier), with restriction to search over SerbItaCor3 collection. System expands the query semantically using Serbian WordNet [21, 22] and adds the synonym “ratnik” (warrior). Next query expansion is morphological (based on previously mentioned *SrpMD*), where all inflected forms for both words are included in the query. The part of the result can be seen in Figure 4, where aligned concordances that contain *ratnik, ratnika, vojnici, vojnicima*… are retrieved and matches are highlighted.

The automatic procedure for introducing all data into Biblisha is an on-going activity, but a few novels were manually entered into the database. Biblisha is built over

Mongo[12] database, so JSON format is required with specific procedures for data management.

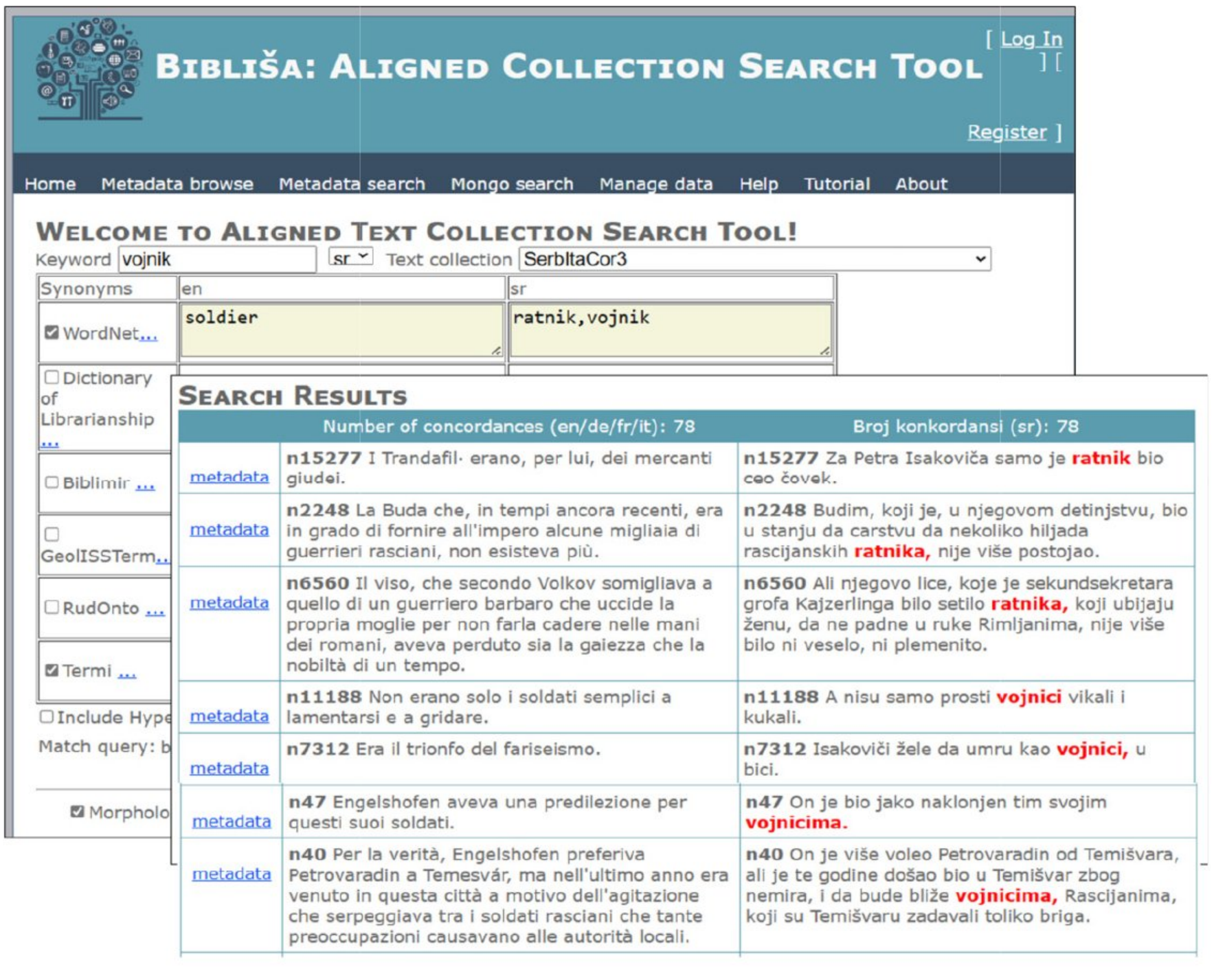

Figure 4. Example of query expansion in Biblisha

The second instance of aligned corpus SerbItaCor3 is available via desktop application TXM, in which various queries can be submitted. Namely, two parts of the corpus are ITSR23_IT0, the Italian part, and ITSR23_SR1, the Serbian part, so these two words should be used in CQL [23] queries. For example, if we want to explore translations in context for the words “guerriero” or “soldato” into “vojnik” or “ratnik”, we can submit some of the following queries:

- ➔ [itlemma="soldato"] :ITSR23_SR1 [srlemma="vojnik"]
  Occurrences of the “soldato” lemma for which we find the “vojnik” lemma in the aligned Serbian segment.
- ➔ [itlemma="soldato"] :ITSR23_SR1 ! [srlemma="vojnik"]
  Occurrences of the “soldato” lemma for which we do not find the “vojnik” lemma in the aligned Serbian segment (Figure 5).
- ➔ <seg> [itlemma!="soldato"]* </seg> :ITSR23_SR1 [srlemma="vojnik"]
  Segments not containing the “soldato” lemma and for which we find the “vojnik” lemma in the Serbian aligned segment.

[12] https://www.mongodb.com/

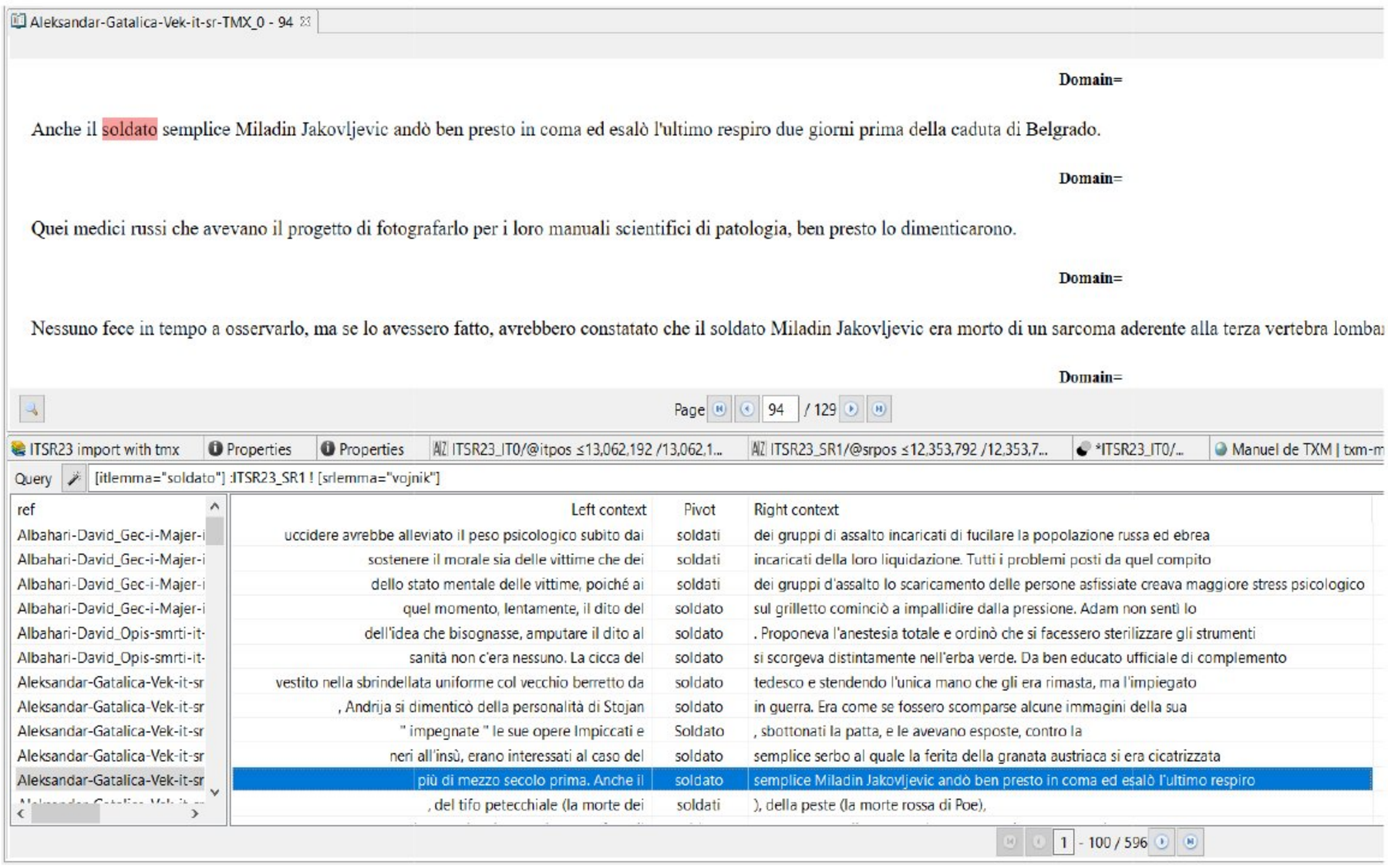

Figure 5. Example of query in TXM

Table 2. presents absolute frequencies, as the number of occurrences of lemmatized word in corpus in Italian and in Serbian part of the corpus. Having in mind that two parts have slightly different numbers of tokens, it is common to normalize frequencies, so frequency per million words (PM) is also given.

Rang	it lemma	it Frequency	it Frequency PM		sr lemma	sr Frequency	sr Frequency PM
1	il	636297	48713		jesam	697617	56470
2	e	334221	25587		i	360565	29187
3	di	330217	25280		da	332539	26918
4	essere	261181	19995		se	254675	20615
5	che	254883	19513		u	238576	19312
6	a	196080	15011		biti	181115	14661
7	del	186864	14306		na	139318	11277
8	non	161251	12345		taj	137366	11119
9	avere	157529	12060		koji	122789	9939
10	un	149741	11464		ja	95857	7759
11	in	140896	10787		on	89745	7265
12	si	126505	9685		ona	89679	7259
13	per	112116	8583		ne	85735	6940
14	al	109775	8404		a	71449	5784
15	una	95795	7334		kao	63783	5163
16	con	89809	6875		od	60013	4858

17	suo	80586	6169		za	59033	4779
18	quello	75455	5777		sav	54707	4428
19	nel	74042	5668		hteti	53965	4368
20	da	73151	5600		što	52218	4227
21	come	68329	5231		ali	50639	4099
22	fare	68168	5219		sa	47975	3883
23	ma	67890	5197		ga	45964	3721
24	dire	57780	4423		jedan	45940	3719
25	tutto	56246	4306		moći	44100	3570
26	se	54311	4158		s	38555	3121
27	più	54024	4136		svoj	36517	2956
28	dal	50676	3880		kako	35160	2846
29	mi	50154	3840		neki	34418	2786
30	questo	43745	3349		iz	33935	2747
31	sul	42922	3286		reći	31119	2519
32	potere	39235	3004		ti	30438	2464
33	lo	35530	2720		o	30220	2446
34	mio	34041	2606		onaj	29615	2397
35	anche	32576	2494		samo	28862	2336
36	lui	28578	2188		imati	28469	2304
37	sapere	28393	2174		tako	27582	2233
38	o	27739	2124		ovaj	27005	2186
39	vedere	27612	2114		znati	26988	2185
40	gli	27153	2079		njegov	26287	2128
41	andare	26627	2038		kad	26028	2107
42	cosa	26237	2009		ni	25266	2045
43	solo	26070	1996		po	25178	2038
44	volere	25724	1969		drugi	25096	2031
45	perché	25487	1951		ili	24101	1951
46	stare	25456	1949		moj	23280	1884
47	così	25049	1918		još	21266	1721
48	loro	25028	1916		li	20517	1661
49	quando	24508	1876		do	20406	1652
50	dovere	24355	1865		više	20318	1645

Table 2. Most frequent words (lemma) in Italian and Serbian part of the corpus

NoSketch Engine[13] is an open-source project combining Manatee, Bonito and Crystal into a powerful and free corpus management system [24, 25]. NoSketch Engine

[13] https://nlp.fi.muni.cz/trac/noske

is a limited version of the software empowering the Sketch Engine[14], a commercial variant offering word sketches, thesaurus, keyword computation, user-friendly corpus creation and many other outstanding features. An instance of NoSketch Engine has been maintained by JeRTeh with numerous monolingual and bilingual corpora, including this new parallel corpus SerbItaCor3. Figure 6 presents an example of the CQL query that, when positioned on the Serbian part (SerbItaCor3_sr), retrieves concordances that contain lemma “ratnik” or “vojnik”, and in the Italian part (SerbItaCor3_it) looks for lemma “soldato” or lemma “guerriero”:

[lemma="ratnik|vojnik"] within SerbItaCor3_it: [lemma="soldato|guerriero"].

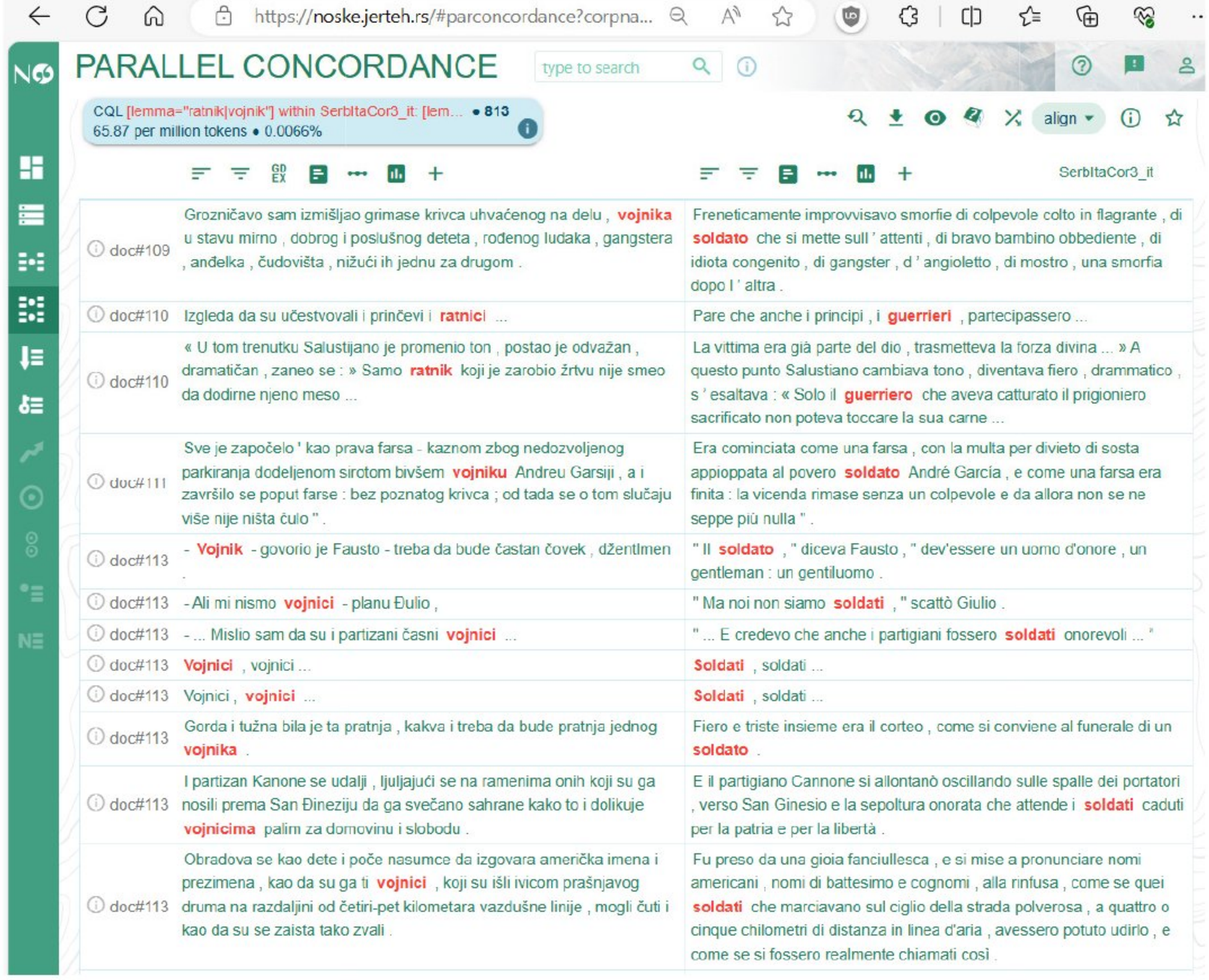

Figure 6. Example of query in noSketch engine instance

5. Conclusion and future work

The paper presents the largest Serbian-Italian parallel literary corpus so far, SerbItaCor3. This collaborative and multidisciplinary effort aimed to enhance Serbian and Italian language teaching by creating and publishing a parallel corpus and developing supporting transformation applications. Future research on the subject will concentrate on organizing additional training to promote corpus and its integration into teaching. Enhancing the annotation by adding named entities and linking them to knowledge bases are identified as one of the goals. Additionally, there is a plan to expand the existing selection of texts with new translations of Serbian literary works into Italian, and vice versa. The goal is to contribute to a deeper understanding of linguistic and cultural aspects of translation studies. Apart from using parallel Serbian-

[14] https://www.sketchengine.eu/

Italian corpus, existing resources also provide a means to explore multiple translations in the same language. Hence, there are plans to construct corpora for Italian-Italian and Serbian-Serbian language pairs as well. Such parallel corpora are especially valuable for the evaluation of text-embedding models, since they are composed of pairs of same-meaning sentences. Another goal of future activities is to enable query expansion for Italian with the introduction of Italian WordNet and other available dictionaries.

Acknowledgment

The results presented in this paper are partially supported by the TESLA project (Text Embeddings - Serbian Language Applications), Program PRIZMA, the Science Fund of the Republic of Serbia, grant number 7276.

References

[1] J. McH. Sinclair, *Corpus, Concordance, Collocation*, Oxford: Oxford University Press, 1991.

[2] C. Krstev, D. Vitas, *An Aligned English-Serbian Corpus*, In: ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality), Volume I, Belgrade, 4-6 December 2009, eds. N. Tomović & J. Vujić, pp. 495-508, Faculty of Philology, University of Belgrade, ISBN 978-86-6153-005-0, 2011.

[3] O. Perišić, R. Stanković, M. Ikonić Nešić, M. Škorić, *It-Sr-NER: Web Services for Recognizing and Linking Named Entities in Text and Displaying Them on a Web Map*, In Infotheca, Belgrade : Faculty of Philology, University of Belgrade (2023). https://doi.org/10.18485/infotheca.2023.23.1.3

[4] TMX Format Specifications Version 1.0. Open Standards for Container/Content Allowing Re-use (OSCAR). 25 November 1997.

[5] M. Utvić, R. Stanković, and I. Obradović, *Integrisano okruženje za pripremu paralelizovanog korpusa.* Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen (2008): 563-578.

[6] O. Perišić, R. Stanković, M. Ikonić Nešić, M. Škorić, *It-Sr-NER: CLARIN Compatible NER and GeoparsingWeb Services for Italian and Serbian Parallel Text*, In Selected Papers from the CLARIN Annual Conference 2022 Prague, Czechia, 2022, 10-12 October, pp. 99-110. Linköping University Electronic Press, 2023.

[7] S. Moderc, *I testi letterari paralleli e la valutazione della traduzione: il caso dell'interpunzione*. Nasleđe 29, str. 203–215, 2014.

[8] S. Moderc, *Su un modo di tradurre l'avverbio serbo "inače" in italiano: il caso dell'equivalente "altrimenti"*, Italica Belgradensia 1, 2015, str. 61-79, 2015.

[9] N. Ceković, N. Janićijević, *La traduzione dei segnali discorsivi italiani in serbo: il caso dei verbi di percezione visiva e uditiva*, Filološki pregled 45 (2), str. 93-106, 2018.

[10] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy, *Parallel corpora for medium density languages*, In Proceedings of the RANLP 2005, pages 590-596.

[11] R. Stanković, C. Krstev, I. Obradović, A. Trtovac, M. Utvić, *A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals*, in Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, May 2012, Istanbul, Turkey, Istanbul, Turkey : European Language Resources Association, 2012.

[12] R. Stanković, C. Krstev, D. Vitas, N. Vulović, and O. Kitanović, *Keyword-based search on bilingual digital libraries*. In Semantic Keyword-Based Search on Structured Data Sources: COST Action IC1302 Second International KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania, September 8–9, 2016, Revised Selected Papers 2, pp. 112-123. Springer International Publishing, 2017.

[13] S. Heiden, *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 (pp. 10 p.). Sendai, Japon, 2010. Retrieved from http://halshs.archives-ouvertes.fr/docs/00/54/97/64/PDF/paclic24_sheiden.pdf
[14] R. Stanković, M. Škorić, B. Šandrih Todorović, *Parallel Bidirectionally Pretrained Taggers as Feature Generators*, In Applied Sciences, MDPI AG (2022). https://doi.org/10.3390/app12105028
[15] H. Schmid *Probabilistic part-ofispeech tagging using decision trees*, In New methods in language processing 2013 Nov 5 (p. 154).
[16] R. Stanković, B. Šandrih, C. Krstev, M. Utvić, M. Škorić, *Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian* in Proceedings of the 12th Language Resources and Evaluation Conference, May Year: 2020, Marseille, France, European Language Resources Association, 2020.
[17] M. C. de Marneffe, C. Manning, J. Nivre, D. Zeman, *Universal Dependencies* in Computational Linguistics, ISSN 1530-9312, vol. 47, no. 2, pp. 255-308, 2021.
[18] C. Krstev, *Processing of Serbian. Automata, texts and electronic dictionaries*, Faculty of Philology of the University of Belgrade; 2008.
[19] D. Vitas, C. Krstev. *Processing of corpora of serbian using electronic dictionaries*, Prace Filologiczne. 2012; 63:279-92.
[20] R. Stanković, C. Krstev, D. Vitas, N. Vulović, O. Kitanović, *Keyword-Based Search on Bilingual Digital Libraries*, in Semantic Keyword-Based Search on Structured Data Sources - Second COST Action IC1302 International KEYSTONE Conference, IKC 2016, Springer, 2017. https://doi.org/10.1007/978-3-319-53640-8_10
[21] C. Krstev, G. Pavlović-Lažetić, I. Obradović, *Using textual and lexical resources in developing serbian wordnet*, Romanian Journal of Information Science and Technology. 2004; 7(1-2):147-61.
[22] R. Stanković, M. Mladenović, I. Obradović, M. Vitas, C. Krstev, *Resource-based WordNet augmentation and enrichment*, InProceedings of the Third International Conference Computational Linguistics in Bulgaria (CLIB 2018) 2018 May 28 (pp. 104-114).
[23] O. Christ, B. M. Schulze, A. Hofmann, and Esther König, *The IMS Corpus Workbench : Corpus Query Processor (CQP), User's Manual*, August 16, 1999 (CQP V2.2), University of Stuttgart, http://corpora.dslo.unibo.it/TCORIS/cqpman.pdf
[24] P. Rychlý, *Manatee/Bonito-A Modular Corpus Manager*, In: RASLAN. 2007. p. 65-70.
[25] A. Kilgarriff, et al., *The Sketch Engine: Ten Years on*, Lexicography, 2014, 1.1: 7-36.