On *Corpus of English-Studies Students (KorSAng)* and Possibilities of Its Software Exploitation

UDC 81'322.2
DOI 10.18485/infotheca.2021.21.1.2

ABSTRACT: Corpus linguistics in Bosnia and Herzegovina and Republic of Serbia is not used enough. The reasons for this are many: the lack of language corpora, the fear of computer methods, and the still present traditional approach to data processing that is qualitative or does not go beyond descriptive statistics. We will often hear arguments against the computer method, such as that a large number of examples that are the result of a search query can impair the quality of the analysis and be misleading. However, the development of technology has also influenced the development of information literacy in all human activities, including the academic community. In this paper, we will try to explain how several departments of English studies in Bosnia and Herzegovina and Serbia in cooperation with The Society for Language Resources and Technologies – *JeRTeh* in Belgrade offered a solution to the student corpus, describe the process of collecting corpus until its final form and demonstrate what search options the corpus offers to its end users.
KEYWORDS: corpus linguistics, learner corpus, parallel corpora, corpus annotation, Corpus of English-studies students.

PAPER SUBMITTED: 16 August 2021
PAPER ACCEPTED: 2 September 2021

Minja S. Radonja
minja.radonja@ff.ues.rs.ba

Srđan R. Šućur
srdjan.sucur@ff.ues.rs.ba

University of East Sarajevo
Faculty of Philosophy Pale
Sarajevo
Bosnia and Herzegovina

1 Introduction

"Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular

SLA/FLT purpose. They are encoded in a standardized and homogeneous way and documented as to their origin and provenance" ((Granger 2002, 7), as cited in (Марковић 2019, 13)). *Corpus of English-studies students* (Sr. *KorSAng*) is a learner corpora of Serbian students of English studies from 4 universities – The University of East Sarajevo, The University of Banja Luka, The University of Belgrade and the University of Novi Sad. The corpus is the product of two important national projects: *Phraseological competence of Serbian speakers of English through the prism of contrastive analysis of interlanguage*, 19/6-020/961-46/18, which was realized in the period from December, 31, 2018, and *Scientific potentials of annotated student corpora in applied linguistics*, 19.032/961-135/19, realized in the period from December, 31, 2019. The KorSAng corpus is inspired by the first student corpus of the English language of Serbophone speakers – *ICLE-SE*, which was formed in the period from September 2015 to September 2016 and consists of argumentative essays of Serbophone English-studies students written in English. Previous research results based on the *ICLE-SE* corpus are presented in scientific papers (Марковић 2017, 2018, 2019, 2020; Radić-Bojanić 2019; Radonja 2019; Šućur 2019; Spajić and Suknović 2019; Tomović and Stefanović 2019). See more about the *ICLE-SE* corpus and the methods of its search in (Шуњур 2020).

1.1 Corpus structure

KorSAng includes 327 texts and has a total of about 140,000 words. It consists of three subcorpora:

- Subcorpus of English-studies students' translations from English (KorPSAng1), which consists of translations of literary texts (3 texts) and newspaper articles (3 texts) from English into Serbian. This part of the corpus has 127 translations and 46,785 words. This practically means that each translation can be paired with the original text which is in this subcorpus in English, which is together called bitext.[1]
- Subcorpus of English-studies students' translations into English (KorPSAng2), where the initial texts are from the genre of literature

1. "[B]itext is a text and its translation, i.e. translations, presented in such a way that an explicit connection is established between the elements of their logical statement, for example, at the level of paragraphs or sentences" ((Vitas 2010, 273), as cited in (Андоновски 2019, 17).

(1 text) and the newspaper genre (5 texts). The final number of translations of this subcorpus is 130, and the number of words is 50,128. As the previous subcorpus, *KorPSAng2* is also a parallel corpus.
– The subcorpus of argumentative essays[2] of English-studies students written in Serbian as a mother tongue (KorSSAng) includes 70 essays and 40,012 words.

The first two components are parallel student corpora, while the third component is designed as a reference corpus to the ICLE-SE corpus. "Parallel corpora are understood to be corpora that contain one or more original texts and their translations into one or more languages" ((Töny 2016, 11), as cited in (Андоновски 2019, 16)). On the importance and application of parallel corpora see more details in (Андоновски 2019, 2021; Ристовић 2012, 2016).

Next, in the second section of this paper, we will describe the stages of corpus preparation, from collecting and selecting texts, through processing and parallelization of texts, to creating metadata and forming the corpus. The third section shows how the corpus can be searched. Finally, in conclusion, we summarize the results so far and present further plans.

2 *KorSAng* corpus preparation

A total of 194 English-studies students participated in the formation of the corpus. *KorSSAng* subcorpus was formed in the period from 2016 to 2019, and *KorPSAng1* and *KorPSAng2* from 2017 to 2019. Of these three components, the most demanding one to collect was a corpus of essays in the Serbian language, due to students' reluctance to participate in essay writing in their mother tongue, so that the formation of this part of the corpus took the longest. The project of creating a parallel *Corpus of English-studies students* consisted of several stages:

– preparatory phase,
– collecting texts,
– text processing,
– parallelization,
– keeping records of metadata,
– publishing texts and metadata on the platform.

2. "An argumentative text (discussion) is a text in which the author starts from some doubt or dilemma, and then, by stating the arguments, a solution is reached. The author of such a text strives to convince the interlocutor or the reader of the correctness of his opinion with the evidence" (CEO 2015, 79).

2.1 Preparatory phase

In the preparatory phase, several key issues related to the project were defined: main and secondary objectives of the project; determining the scope of the project; specifying cooperation and distribution of tasks among team members from the University of East Sarajevo, the University of Banja Luka, the University of Belgrade and the University of Novi Sad; specifying cooperation with team members from JeRTeh;[3] determining the time frame for the execution of individual stages. Jelena Marković, the project manager, was in charge of establishing contacts with the team members.

2.2 Collection and selection of texts

In the next phase, the team members, together with the project manager, selected several texts for translation from English and into English and made a preliminary list of the essay topics in Serbian. During the text selection, care was taken to ensure that the texts were from the literary and newspaper genres. Newspaper articles cover various topics, such as politics, economics, language, health, social networks, and show business. In the end, 6 texts in English, 6 texts in Serbian and 24 essay topics in Serbian were selected. Then ensued the collection of translations and essays, which was the most time-consuming task. It consisted of organizing the conditions for translating texts and writing the essays by the students. Team members organized this task at their faculties. The translation was limited in time, with the exception of a few works, while the writing of the essay was unlimited in at least a third of the works. In addition to the given written assignments, students were required to fill in the accompanying document profile of the participants in order to give consent for their translations and essays to be used for research purposes, while the papers themselves are anonymous and recorded under a code.

2.3 Processing and parallelization of texts

Work tasks related to text processing and parallelization were performed by research assistants, under the leadership of Jelena Marković and the team from JeRTeh. The assistants received online training in the use of

3. JeRTeh

parallelization tools to obtain texts in TMX (Translation Memory eXchange) documents.[4] They received instruction in Notepad++ XML editor,[5] Unitex/GramLab,[6] Unitex module for Serbian (Krstev 2008) and ACIDE (Aligned Corpora Integrated Development Environment) (Utvić, Stanković, and Obradović 2008), which enabled adequate text processing by prescribed rules. In Notepad++, the source and translated texts were prepared, which involved pairing paragraphs of two documents, so that each paragraph of the source text corresponded to a paragraph of the translation. The files prepared in this way were further processed in the *Unitex Visual IDE*, which enables the segmentation of paragraphs into sentences, according to the language in which the text is written. After segmentation of the original and translated text, the files were prepared for parallelization, i.e. "the process of establishing links between the appropriate variants of translation units, i.e. the formation of a set of translation units" (Андоновски 2019, 17). Parallelization was enabled using the *ACIDE* application developed by the Language Technology Group of the University of Belgrade (more on the *ACIDE* integrated development environment for parallelized corpora can be found in (Utvić, Stanković, and Obradović 2008). This application enables automatic parallelization with the possibility of checking and correcting errors, which helped the team members who had this task to successfully prepare texts for the parallel corpus. In the texts prepared for the parallel corpus within this project, we came across examples that some segments were missing in the translated text, or one segment in the original text corresponded to several segments in the translated text, which can occur in translation. The training itself was a challenge to a team of linguists who had not encountered this type of work tasks before, and the skills acquired during the creation of the parallel corpus gave a new perspective on corpus linguistics. The final result after word processing in *ACIDE* was a *TMX* document that was then entered into the corpus.

4. TMX is an ISO standard (ISO24616 2012) for the storage of so-called translation memories and their exchange between different software translation tools, as well as between different companies that deal with the maintenance of translation memories (TMX 2005). Translation memories are collections of determinants in which the text of the source language is associated with the equivalent translation of the text of the target language, i.e. the produced TMX document is composed of the obtained translation units (Андоновски 2019, 22).

5. Notepad++

6. Unitex/Gramlab, Cross-platform Corpus Processing Suite

2.4 Keeping records of metadata and corpus formation

In parallel with the preparation of texts for the corpus, the researchers kept records of metadata containing information on variables such as data about the participant (age, gender, mother tongue, education data, years of learning English, knowledge of other foreign languages, stay in English speaking country) and variables concerning the context of language production, i.e. data on the context in which the students performed the tasks (whether there was a time limit for the tasks of writing the essay or translation, use of dictionaries and manuals, whether the task was an exam obligation or written outside the exam, what is the genre, etc.). Metadata tables were made separately for translations and essays in separate Excel documents. Since the papers entering the corpus were anonymous, it was necessary to assign to each document the appropriate metadata registered under the same code. Team members from JeRTeh prepared the documents for publication: assigned metadata to each document, did part-of-speech annotation and lemmatization (for more details on the resources that made this possible, see (Stanković et al. 2020) Stanković et al. 2020).

3 Search of the *KorSAng* corpus

We briefly described the process of collecting the corpus from the preparatory phase to its final electronic form and the challenges we encountered along the way. *KorSAng* contains 136,925 words: *KorPSAng1* (46,785 words, 127 translations into Serbian), *KorPSAng2* (50,128 words, 130 translations into English), *KorSSAng* (40,012 words, 70 essays in Serbian). One of the major challenges was to motivate students to write essays in their mother tongue, which raises doubts about the decline in the competence of writing in their mother tongue (Šućur 2020, 144). We will also mention that the essay topics limit the search possibilities. Of the 24 topics offered, more than 40% of the essays were written on four topics: *Family or work: what is more important* (Sr. *Породица или посао: шта је важније*), *How music affects life* (Sr. *Како музика утиче на живот*), *How I imagine a good parent* (Sr. *Како замишљам доброг родитеља*), and *From marriage to divorce today* (Sr. *Од склапања брака до развода данас*). Therefore, we expect that the corpus of the essays offers a rich vocabulary that belongs to the semantic field of family relations, but poorer when it comes to some other offered topics. Next, after a few introductory remarks about the *Sketch Engine* platform, we shall first demonstrate how it can be used to perform simple and advanced searches

of the *KorSSAng* corpus, and then how it can be used to perform parallel corpora searches of the *KorPSAng1_ en* and the *KorPSAng1_ sr* corpus (the former consists of the source texts in English, and the latter consists of the Serbian translations of the source texts) corpus, and the *KorPSAng2_ sr* and the *KorPSAng2_ en* corpus (the former consists of the source texts in Serbian, and the latter consists of the English translations of the source texts) respectively.

3.1 *Sketch Engine*

The Sketch Engine is an online platform that comprises a series of robust electronic corpora query tools. This platform is the result of a collaboration between the British lexicographer and corpus linguist Adam Kilgarriff, and the Czech computer scientist Pavel Rychlý. Its commercial version has been available since 2004 (Kunilovskaya and Koviazina 2017, 503). As Kilgarriff et al. (2014, 15-16) describes it "the Sketch Engine has come out of the academic research world", and it is today used in linguistics, and languages departments (teaching and research), and in Computational Linguistics and Natural Language Processing. In addition, it is used in language teaching (especially in English language teaching), lexicography, translation and teaching translation, discourse analysis, etc. (Kilgarriff et al. 2014).

Below we will present a widely available non-commercial version of this platform (adapted by JeRTeh) which, in spite of a limited number of tools available, offers a multitude of possibilities for querying and researching (parallel) electronic corpora that consist of learners' production in their mother tongue, and English.

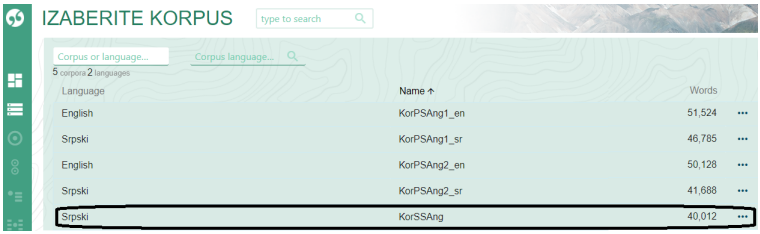

Figure 1. Corpus selection

3.2 Example of *KorSSAng* subcorpus search

First we select one, among the five corpora available, KorSSAng, which consists of argumentative essays written in Serbian, by Serbian students of the English language (Figure 1).

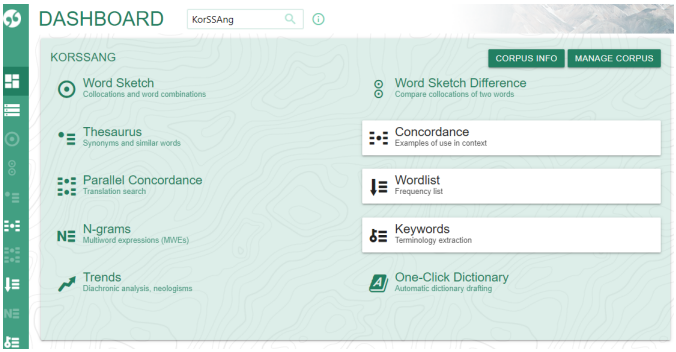

Figure 2. Dashboard

This takes us to the Dashboard (Figure 2) which offers three tools: Concordance, which provides examples of the use (of a term) in context, Wordlist, which generates a frequency list, and Keywords, which can be useful in lexicography (for terminology extraction, etc.).

Corpus Info (Figure 3) can be accessed from the Dashboard as well.

This tab offers the general information about the selected corpus, and provides its description,[7] the tagset used,[8] and the numbers of tokens, words, lemmas, and documents that comprise the corpus. By selecting the *Manage corpus* button (the upper right-hand corner), the corpus can be expanded and modified, by adding texts to the corpus, or by forming new, or altering the existing subcorpora, etc.

Next, having selected the *Concordance* tool, we perform a *Basic* query for the lemma [млад] ('young') (Figure 4), given that the essays have pre-

7. KorSSang

8. 1. N (Noun), 2. A (Adjective), 3. V (Verb), 4. PRO (Pronoun), 5. NUM (Number), 6. PREP (Preposition), 7. CONJ (Conjunction), 8. INT (Interjection), 9. PAR (Particle), 10. ADV (Adverb), 11. PREF (Prefix), 12. ABB (Abbreviation), 13. RN (Roman numeral), 14. PUNCT (Punctuation), 15. SENT (Sentence end marker), 16. ? (Non-Serbian words or suffixes in compounds).

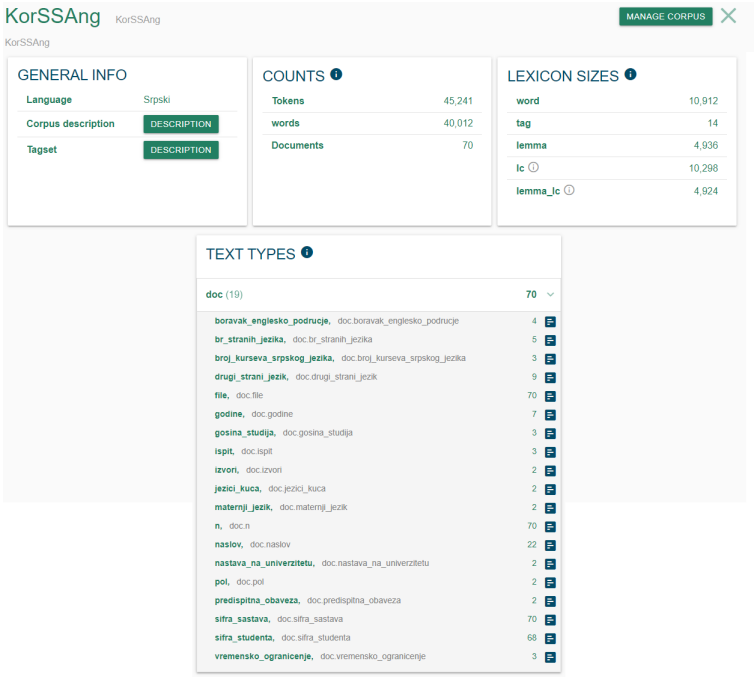

Figure 3. General info about the *KorSSAng* corpus

dominantly been written by the young, hence a significant distribution of this lemma can be expected in this corpus.

The query yields 103 results (such as the uninflected forms of the adjective "млад" ('young'), as well as the use of this adjective for a generic reference denoting young people - "млади" ('the young'), arranged in the order in which they occur in the numbered documents which make up the *KorSSAng* corpus (Figure 5).

The selection of any of the concordance lines offers a wider context of use of the lemma (Figure 6).

The search results can be sorted according to the left, or the right context (i.e. words that appear to the left, or to the right of the lemma), or according to the *Key Word in Context* (KWIC), which is the most common way of sorting concordance lines (Figure 7). The results can be exported in several formats, such as *TXT*, *CSV*, *XLS*, *XML*, whereas the current view can be saved as a *PDF*.

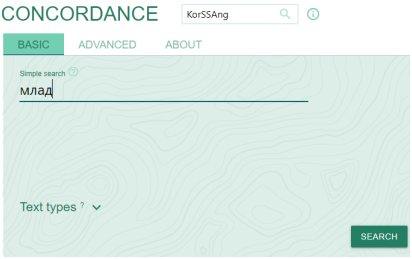

Figure 4. Basic query for the lemma [млад] (‘young’)

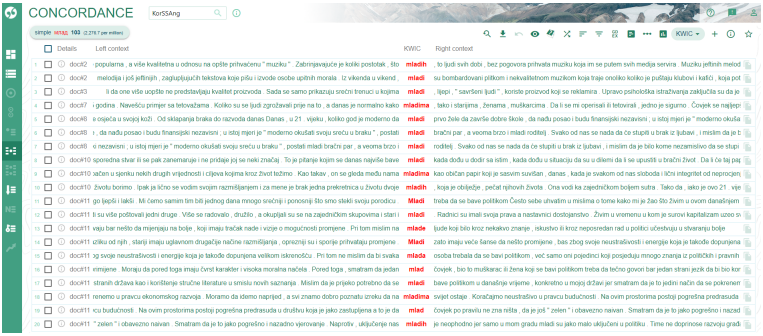

Figure 5. Search results for the basic query of the lemma [млад] (‘young’)

The frequency (Figure 8) and the distribution of the lemma in the entire corpus can be checked as well, along with the words it most frequently collocates with (Figure 9); 1-5 words to the left of the lemma, and 1-5 words to the right (in this case it is the word “људи”[9] (‘people’) with 25 co-occurrences).

Next, we perform an advanced *CQL* query (Context Query Language), which is accessed through the menu in Figure 4, by selecting the tab *Advanced*[10] (Figure 10).

9. By contrast, the query for a noun phrase [млади људи] (‘young people’) in the comparable corpus *LOCNESS* (i.e. its American component, which consists of approximately 150,000 words) yields 12 concordance lines.

10. In addition to CQL queries, advanced searches can be performed for the simple form of the word (which retrieves all the instances of the word that do not include a capital (initial) letter, or for lemmas, phrases, words (regardless of whether they contain capital letters or not), or for special characters (such as punctuation marks, numbers, etc).

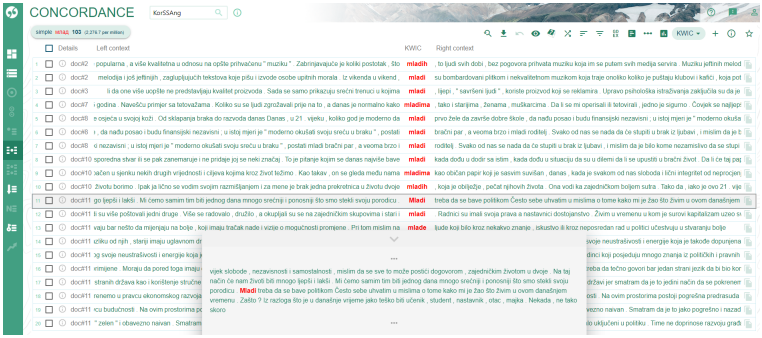

Figure 6. The wider context of use of the lemma [млад] (‘young’)

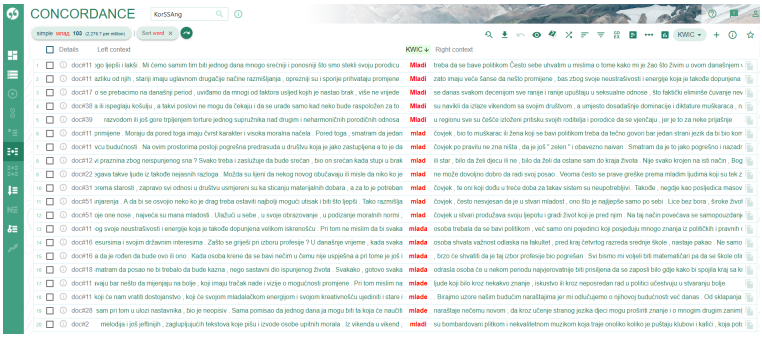

Figure 7. KWIC for the lemma [млад] (‘young’)

For the purpose of this paper we use the CQL Builder, a tool integrated in the Sketch Engine (Figure 11), aimed at generating specific syntax queries for advanced corpus search with the set of tags listed in Footnote 8. In this case we will search for verbs that appear three words to the right of the lemma [млад] (‘young’).

This query yields 94 results (Figure 12). However, before exporting, the concordance lines have to be selected manually in order to single out unique examples, so as to exclude those that are repeated several times due to the specific search range applied {0,3}.

Next we shall present concordance queries in parallel corpora: *KorPSAng1_en* and *KorPSAng1_sr*, and *KorPSAng2_en* and *KorPSAng2_sr* respectively. These corpora can be accessed through the menu in Figure 1 as well.

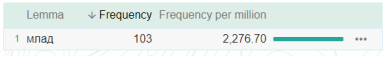

Figure 8. The frequency of the lemma [млад] ('young') in the entire *KorSSAng* corpus

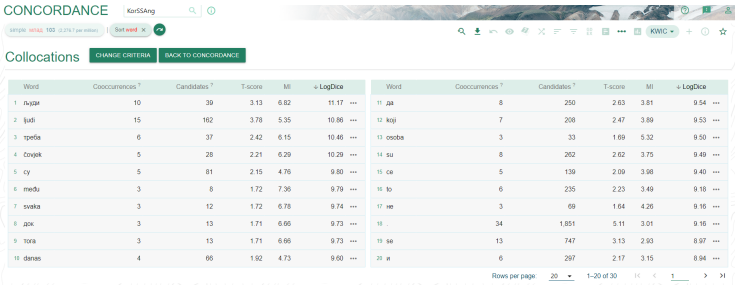

Figure 9. Words which the lemma [млад] ('young') most frequently collocates with

First we select the *KorPSAng1_en* corpus and access the Dashboard (Figure13). This Dashboard differs from the one in Figure 2 2 in that it has an additional tool; Parallel Concordance, since there is a Serbian corpus that *KorPSAng1_en* is paired with (*KorPSAng1_sr*).

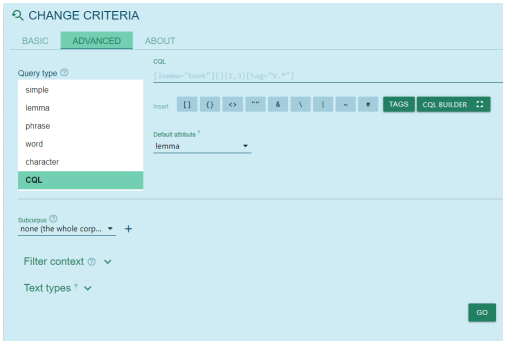

Figure 10. Advanced query selection

Then we select the *Parallel Concordance* tool and perform a basic search[11] for the noun phrase "project's collapse" (which is a part of a larger relative clause - *which then led to the project's collapse*) present in one of the source texts in English, to see how it is translated into Serbian 14.

Figure 11. Generating a CQL query `[lemma="млад"] [ ] {0,3}[tag="V.*"]`

The search for the noun phrase "project's collapse" yields 19 parallel concordance lines (each line consisting of an aligned text in English and its translation into Serbian) (Figure 15)

A detailed overview of the results yielded shows that the noun phrase is translated into Serbian in several different ways, predominantly as a noun phrase: *колапс пројекта* (4 examples), *пропаст (целог) пројекта* (4 examples), *распад пројекта* (3 examples), *пад пројекта*, *гашење пројекта*, *колапс (самог) филма*, *колапс*, *крај првобитне верзије*, *крах пројекта*, *пропадање пројекта*, and, in one case, it is translated as a part of a clause: *које су довеле до тога* [...] *да цео пројекат пропадне.*

We will next perform an advanced CQL query[12] for perfect verb forms in the same parallel corpora; `[lemma="have"] [tag="V.*"]` (Figure 16), since complex verb phrases, more frequently than not, present a stumbling block in the learning of a foreign language.

11. An advanced search for this noun phrase can be performed via the following CQL query as well: `[lemma="project"][word="'"][word="s"][lemma="collapse"]`.

12. CQL queries of the English components of the corpus are generated with the Penn Treebank Tagset.

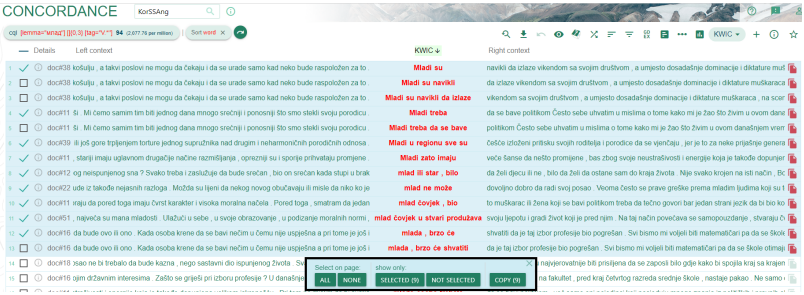

Figure 12. Search results for the CQL query `[lemma="млад"] [ ]{0,3}[tag="V.*"]`

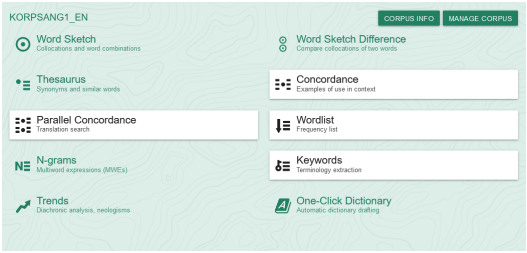

Figure 13. Parallel corpora Dashboard

This search yields 432 results (Figure 17) that include examples of the past, and the present perfect tense, perfect infinitives, and participles.

The search can be further narrowed down using a fine tagset,[13] and, if we want to reduce it to the forms of the past perfect tense, we can do so with the following CQL query: `[word="had"][tag="VVN.*"]`, (where *VVN*

13. VB → verb BE, base form (be), VBD → verb BE, past tense (was, were), VBG → verb BE, gerund/present participle (being), VBN → verb BE, past participle (been), VBP → verb BE, sing. present, non-3rd (am, are), VBZ → verb BE, 3rd person sing. present (is), VH → verb HAVE, base form (have), VHD → verb HAVE, past tense (had), VHG → verb HAVE, gerund/present participle (having), VHN → verb HAVE, past participle (had), VHP → verb HAVE, sing. present, non-3rd (have), VHZ → verb HAVE, 3rd person sing. present (has), VV → verb, base form (take), VVD → verb, past tense (took), VVG → verb, gerund/present participle (taking), VVN → verb, past participle (taken), VVP → verb, sing. present, non-3rd (take), VVZ → 3rd person sing. present (takes).

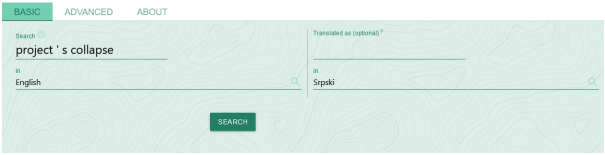

Figure 14. Basic search for the noun phrase "project's collapse"

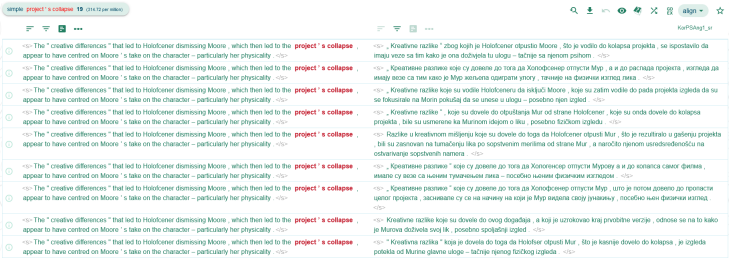

Figure 15. Basic search results for the noun phrase "project's collapse"

is the tag for the past participle of the main verb). This query yields 168 results (Figure 18).

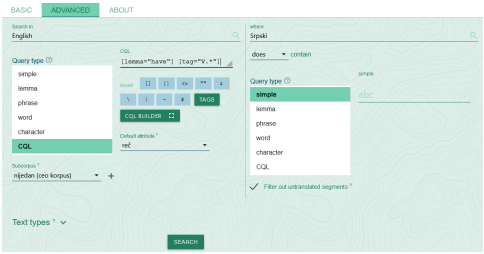

Figure 16. The advanced CQL query for perfect verb forms; `[lemma="have"] [tag="V.*"]`

Finally, through the menu in Figure 1, we select the *KorPSAng2_sr* corpus and perform an advanced CQL query for the noun phrase "istrošene svađalačke snage" `[lemma="истрошен"] [lemma="свађалачки"] [lemma="снага"]` to search for parallel concordances (Figure 19).

This query yields 32 results. What follows are several examples of various complexity: dissipated energy for quarrelling, a drained argumentative spirit, spent fighting endurance, the wasted energy on fights, used up strength for

quarreling, used up strength for falling out, worn out quarrelsome energy, wasted fighting energy, wasted argumentative energy, a drained will to fight, a small fighting force, etc.

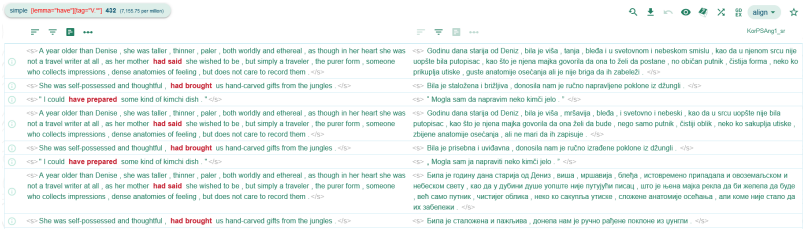

Figure 17. CQL query `[lemma="have"] [tag="V.*"]` search results

4 Conclusion

In this paper, we have briefly described the steps of the creation of the Corpus of English-studies Students (*KorSAng*), and presented the possibilities of its search. It is known that the lack of student corpora in particular is a handicap for researchers in the field of applied linguistics, and we hope that our efforts to overcome this problem will result in greater popularity of corpus linguistics as a research method in this area, especially in English studies. The results of the use of the *KorSAng* corpus so far have been presented in the following works: (Šućur 2020; Spajić and Suknović 2019; Tomović and Stefanović 2019; Марковић and Станковић, у штампи), and a doctoral dissertation, part of the corpus of which is based on *KorSAng*. The scientific potential of this corpus can be expected in many scientific and professional papers, monographs and scientific research projects. We will add that the plan is to create a second version of the *KorSAng* corpus, with an expanded number of essays and translations, and with the possibility of integrating new software tools that will provide easier search and more comfortable use of the corpus to end users.

Acknowledgment

This paper is based on research conducted within two national projects: *Scientific potentials of annotated student corpora in applied linguistics*

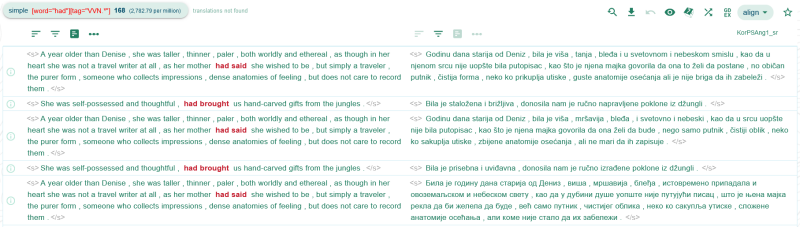

Figure 18. Search results for the past perfect tense via a CQL query `[word="have"] [tag="VVN.*"]`

[Naučni potencijali anotiranih učeničkih korpusa u primijenjenoj lingvistici], 19.032/961-135/19 and *Phraseological competence of Serbian speakers of English through the prism of contrastive analysis of interlanguage* [Frazeološka kompetencija srpskih govornika engleskog kroz prizmu kontrastivne analize međujezika], 19/6-020/961-46/18. The projects were financially supported by the Ministry for Scientific and Technological Development, Higher Education and Information Society, Banja Luka. We thank *The Society for Language Resources and Technologies – JeRTeh* for the cooperation in the project (19.032/961-135/19), especially Professor Ranka Stanković. Special thanks go to project coordinator, Professor Jelena Marković, for providing us with the opportunity to cooperate in the projects mentioned, as well as for her continuous support in the form of encouragement, advice, motivation and expertise.

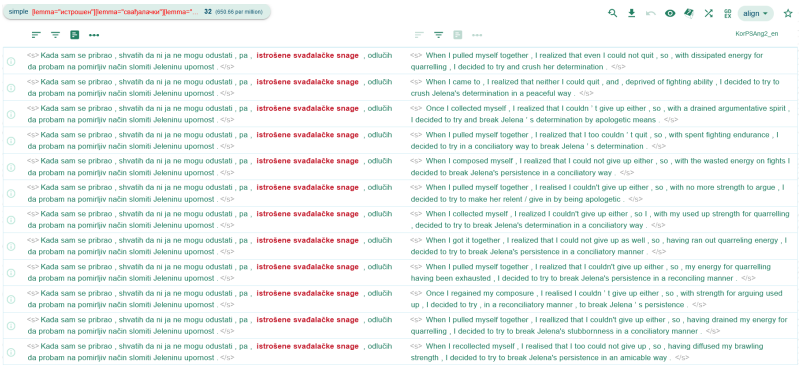

Figure 19. Basic search results for the noun phrase “istrošene svađalačke snage”

References

CEO. 2015. *Општи стандарди постигнућа за крај општег средњег и средњег стручног образовања и васпитања у делу општеобразовних предмета.* Београд: Завод за вредновање квалитета образовања и васпитања.

Granger, Sylviane. 2002. "A bird's-eye view of learner corpus research." *Computer learner corpora, second language acquisition and foreign language teaching* 6:3–33.

ISO24616. 2012. *Language resources management — Multilingual information framework.* International Standard Organization.

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

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

Kunilovskaya, Maria, and Marina Koviazina. 2017. "Sketch engine: A toolbox for linguistic discovery." *Jazykovedny Casopis* 68 (3): 503.

Radić-Bojanić, Biljana B. 2019. "NEODREĐENA ZAMENICA ONE U PISANJU KOD NEIZVORNIH GOVORNIKA ENGLESKOG JEZIKA." *Годишњак Филозофског факултета у Новом Саду* 44 (2): 39–52. https://doi.org/10.19090/gff.2019.2.39-52.

Radonja, Minja S. 2019. "The use of interactive metadiscourse in Serbian students." *Радови Филозофског факултета: Часопис за хуманистичке и друштвене науке* 8 (21). https://doi.org/10.7251/FIN1921121R.

Spajić, Sonja, and Mina Suknović. 2019. "The Choice of Lexemes According to Their Frequency in Translation into L2." *Komunikacija i kultura online* 10 (10): 104–119. https://doi.org/10.18485/kkonline.2019.10.10.6.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Skoric. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian." In *Proceedings of The 12th Language Resources and Evaluation Conference,* 3954–3962.

Šućur, Srđan. 2019. "Distribucija frazalnih glagola u pisanju na engleskom kao stranom kod srbofonih govornika." *Komunikacija i kultura online* 10 (10): 120–143. https://doi.org/10.18485/kkonline.2019.10.10.7.

Šućur, Srđan. 2020. "REVERSE TRANSFER IN ADULT SERBIAN EFL LEARNERS'WRITING1 2-A CORPUS BASED STUDY." *BEYOND HERMENEUTICS,* 141.

TMX. 2005. "Translation Memory eXchange (TMX) 1.4b Specification." Accessed 1.08.2021. https://www.gala-global.org/knowledge-center/industry-development/standards/lisa-oscar-standards.

Tomović, Nenad, and Sofija Stefanović. 2019. "Uticaj L2 i leksički i leksičko-sintaksički kalkovi u prevodu. Studija slučaja." *Komunikacija i kultura online* 10 (10): 144–154. https://doi.org/kkonline.2019.10.10.8.

Töny, Luzius. 2016. "Corpora als Ressourcen für die maschinelle Übersetzung." Accessed 17.04.2016. https://swanrad.ch/downloads/mt_1.pdf.

Utvić, Miloš, Ranka Stanković, and Ivan Obradović. 2008. "Integrisano okruženje za pripremu paralelizovanog korpusa." *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen,* 563–578.

Vitas, Duško. 2010. "Resursi i metode za obradu srpskog – stanje i perspetive." In *Srpska lingvistika/Serbissche Linguistik, Eine Bestandsaufnahme, Studies of Language and Culture in Central and Eastern Europe (SLCCEE,* edited by Biljana Golubović and Cristian Voß, 7:257–277. München: Verlag Otto Sagner.

Андоновски, Јелена. 2019. "Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса." PhD diss., Универзитет у Београду, Филолошки Факултет, January.

Андоновски, Јелена. 2021. "Паралелни корпуси у Србији — могућности за паралелно проналажење информација на два или више језика" [in serbian]. 3, *Библиотекар* 63 (1): 51–74. ISSN: 0006-1816. https://doi.org/10.18485/bibliotekar.2021.63.1.3.

Марковић, Јелена. 2017. "Лични метадискурс у писању код изворних и неизворних говорника енглеског језика." *Филолог–часопис за језик, књижевност и културу,* no. 15, 44–60. https://doi.org/10.21618/fil1715044m.

Марковић, Јелена. 2018. "Употребе глагола make у писању на енглеском језику као страном код изворних говорника српског језика (корпуснолингвистичка анализа)." *Зборник матице српске за филологију и лигвистику* 61 (1): 165–180. https://www.maticasrpska.org.rs/stariSajt/casopisi/ZMSFL_61_1.pdf.

Марковић, Јелена. 2019. *Кроз призму контрастивне анализе међујезика.* Филозофски факултет.

Марковић, Јелена. 2020. "Концесивни конектори though и however у писању на енглеском језику код изворних и неизворних говорника." *Филолог–часопис за језик, књижевност и културу,* no. 21, 13–35. https://doi.org/10.21618/fil2021013m.

Марковић, Јелена, and Ранка Станковић. у штампи. "Ја/ти/ми/ви у дискурсној компетенцији у светлу контрастивне анализе међујезика." *Методички видици.*

Ристовић, Зоран. 2012. "Од корпуса до учионице: примена паралелизованих текстова у настави енглеског језика у основној школи." *ИНФОтека* 13 (2): 52–66.

Ристовић, Зоран. 2016. "Кумулативни ефекти експлоатације вишејезичних корпуса у настави страних језик." PhD diss., Универзитет у Београду, Филолошки Факултет.

Шуњур, Срђан Р. 2020. "Корпус као оруђе за проницање тајни међујезика." *Радови Филозофског факултета: Часопис за хуманистичке и друштвене науке,* no. 22, https://doi.org/10.7251/RFFP2022321S.