# 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 2021PAPER ACCEPTED:2 September 2021

## 1 Introduction

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

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 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. KorSAnq) 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 ( $\Pi y h y p 2020$ ).

#### 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.<sup>1</sup>
- Subcorpus of English-studies students' translations into English (KorPSAng2), where the initial texts are from the genre of literature

<sup>1. &</sup>quot;[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<sup>2</sup> 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;<sup>3</sup> 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 timeconsuming 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

<sup>3.</sup> JeRTeh

parallelization tools to obtain texts in TMX (Translation Memory eXchange) documents.<sup>4</sup> They received instruction in Notepad++ XML editor,<sup>5</sup> Unitex/GramLab,<sup>6</sup> 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.

5. Notepad++

<sup>4.</sup> 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).

<sup>6.</sup> 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. Kako замишљам доброг родитеља), and From marriage to divorce today (Sr. Od 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.

Ø	IZABERITE KORPUS type to search Q		ST. ANT	
55	Corpus or language 5 orpora 2 languages Language	Name 🛧	Words	
≡	English	KorPSAng1_en	51,524	
$\odot$	Srpski	KorPSAng1_sr	46,785	
8	English	KorPSAng2_en	50,128	
•=	Srpski	KorPSAng2_sr	41,688	
=•=	Srpski	KorSSAng	40,012	

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).

Ø	DASHBOARD Korssang Q 🛈	
	KORSSANG	CORPUS INFO MANAGE CORPUS
##	Word Sketch     Collocations and word combinations	O Word Sketch Difference Compare collocations of two words
•	• Thesaurus Synonyms and similar words	Examples of use in context
© •∎	Parallel Concordance     Translation search	↓≡ Wordlist Frequency list
•• 33	NE N-grams Multiword expressions (MWEs)	SE Keywords Terminology extraction
≡•≡ ↓≡ NE	Partition analysis, neologisms	One-Click Dictionary Automatic dictionary drafting
∿≡ δ≣		

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,<sup>7</sup> the tagset used,<sup>8</sup> 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-

<sup>7.</sup> KorSSang

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

Language Srpski		COUNTS 0		LEXICON SIZES	1
		Tokens	45,241	word	10,912
Corpus description DESC	RIPTION	words	40,012	tag	14
Tagset DESC	RIPTION	Documents	70	lemma	4,936
				Ic 🛈	10,298
				lemma Ic 🛈	4.924
		_			
	TEXT	TYPES			
	doc (19)			70 ~	
		k_englesko_podrucje, doc.boravak	_englesko_podrucje	4 🛃	
		nih_jezika, doc.br_stranih_jezika		5 🚍	
		<pre>irseva_srpskog_jezika, doc.broj_ku itrani_jezik, doc.drugi_strani_jezik</pre>	rseva_srpskog_jezika	3 🚍	
	file, do			70	
		, doc.godine		7	
		_studija, doc.gosina_studija		3 🚍	
	ispit, d			3 🚍	
	izvori,	doc.izvori		2 🖪	
	jezici_k	tuca, doc.jezici_kuca		2 🚍	
	maternj	ji_jezik, doc.maternji_jezik		2 🔳	
	n, doc.	n		70 🚍	
	naslov,	doc.naslov		22 🚍	
	nastava	a_na_univerzitetu, doc.nastava_na	univerzitetu	2 📑	
	pol, de	pc.pol		2 🛃	
	predisp	bitna_obaveza, doc.predispitna_oba	veza	2 🚍	
	sifra_sa	astava, doc.sifra_sastava		70 🚍	

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')

simple Mass 103 (2278.7 per miller)								0	+		a .		= 9		13	KWIC		0
Details Left context						KWIC	Right contex		-									0
1 🔲 () docit2 -popularna .	više kvalitetna u odnos	u na opšte pritwaćenu i	"muziku". Zabr	iniavaluće je kolik	i postotak , što	mladih	, to liudi svih	dobi , bez	: pogovos	a prihvab	a muzik	u koja ir	1 se pute	n svih me	dia serv	ra . Muzik	ku jeftinih i	melo
2 🔲 🛈 dooli2 melodija i jo	š jefinijih , zaglupljujuči	n tekstova koje pišu i izv	vode osobe upitr	nih morala . Iz vike	anda u vikend ,	mladi	su bombardo	vani plitko	om i nekv	alitetnom	muzik	om koja	raje onol	ko koliko	e puštaj	u klubovi	i katići , ko	ia po
3 🗌 🛈 doc#3 🛛 li da one	više uopšte ne predstav	ljaju kvalitet proizvoda	Sada se samo	prikazuju srećni tr	renuci u kojima	mladi	, lijepi , " savi	Seni ljudi	, koriste	proizvoc	l koji se	reklami	ra Uprar	o psiholo	ška istra	Ivarija za	ikļučila su	dajo
4 🔲 🛈 doc#7 igodina. Nav	ešću primjer sa tetovaži	ama . Koliko su se ljudi	zgrožavali prije r	na to , a danas je i	normalno kako	mladima	, tako i stariji	ma , žena	ma , muš	karcima .	Dalis	e mi ope	risali ili te	tovirali , je	dno je s	gumo . Č	ovjek se n	ajter
s 🔲 🛈 docili8 e osjeća u svi	joj koži . Od sklapanja t	iraka do razvoda danas	Danas , u 21 . v	vijeku , koliko god	je moderno da	mladi	prvo žele da	završe do	bre škole	, da ned	u posa	o i budu	Inansijsk	nezavisn	i ; u istoj	mjeri je "	moderno	okuši
s 🔲 🛈 doc#8 +, da nadu po	sao i budu finansijski ne	zavisni ; u istoj mjeri je	" moderno okuša	ati svoju sreću u b	raku", postati	mladi	bračni par , a	veoma b	rzo i mlac	I roditelj	Svako	od nas	se nada o	la ĉe stup	li u brak	iz ijubavi	, i mislim o	da je
7 🔲 🛈 docită ki nezavisni ;	u istoj mjeri je * moderno	okušati svoju sreću u l	oraku", postati r	mladi bračni par ,	a veoma brzo i	mladi	roditelj . Sval	to od nas	se nada i	ta će stu	piti u br	ak iz ljut	evi, i mi	lim da je l	bilo kom	nezamia	sivo da se	stup
a 🔲 🛈 dooli10 sporedna stv	r ili se pek zanemaruje	i ne pridaje joj se neki z	načaj . To je pita	inje kojim se dana	is najviše bave	mladi	kada dođu u	dodir sa i	stim, kad	a dođu u	situaci	u da su	u dilemi c	a li se upi	isti u br	ačni život	. Da li ćer	laj pa
9 🔲 🛈 doc#10 xačen u sjorik	u nekih drugih vrijednos	ti i ciljeva kojima kroz ži	vot težimo . Kao	takav , on se gler	da među nama	mladima	kao običan p	apir koji je	a sasvim s	uvišan ,	danas.	kada je	svakom	od nas slo	boda i 16	ni integrit	let od nepr	ocjer
10 🔲 🛈 doci/10 životu borime	. Ipak ja lično se vodim	svojim razmišljanjem i	za mene je brak	jedna prekretnica	u životu dvoje	mladih	, koja je obilje	stje , peča	at njihovit	života .	Ona vo	di ka zaj	edničkon	boljem s	utra . Tak	oda, iak	o je ovo 2	1.19
11 🔲 🕕 doci/11 go ljepši i lak	i . Mi černo samim tim b	iti jednog dana mnogo	srečniji i ponosn	iji što smo stekli s	voju porodicu .	Mladi	treba da se b	eve politi	kom Čest	o sebe ul	watim	u mislim	a o tome	sako mi je	2ao 5to	2 Nim u ov	vom danas	injerr
12 🔲 🕕 docitit li su više poši	ovali jedni druge . Više s	e radovalo , družilo , a	okupljali su se n	a zajedničkim sku	ipovima i stari i	mladi	. Radnici su i	mali svoje	a prava a	nastavnie	i dosto	janstvo	Živim u r	remenu u	i kom je :	urovi kap	italizam u	200 s
13 🔲 🛈 docif11 vaju bar nešti	da mijenjaju na bolje , l	koji imaju tračak nade i	vizije o mogućno	osti promjene . Pri	i tom mislim na	mlade	ljude koji bilo	kroz neko	akvo znar	ije , iskus	tvo ili k	roz nep	isredan r	ifiloq u be	i učestv	ju u stva	ranju bolje	
14 🔲 🛈 dooli'11 ziiku od njih	stariji imaju uglavnom d	rugačije načine razmiši	janja , oprezniji s	su i sporije prihval	taju promjene .	Mladi	zato imaju ve	de Sanse	da neŝto	promijer	e , bas	zbog sv	oje neust	ašivosti i	energije	koja je tal	kođe dopu	njen
is 🔲 🕕 docititi og svoje neus	trašivosti i energije koja	je takođe dopunjena ve	ilkom iskrenošči	u . Pri tom ne misl	lim da bi svaka	miada	osoba trebali	a da se be	ni politika	m, voć s	amo o	ni pojedi	nci koji po	sjeduju m	inogo zn	anja iz po	(BCkh i pr	avnit
16 🔲 🛈 doci/11 vimijene . Mo	raju da pored loga imaju	čvrst karakter i visoka	moralna načela	Pored toga , smi	atram da jedan	mlad	čovjek , bio 6	o muškara	ac ili žena	koji se b	avi poli	tikom tre	ba da ted	no govori	bar jeda	n strani je	zik da bi b	io ko
17 🔲 🛈 doolii11 stranih držav	a kao i korištenje stručni	a literature u smislu nov	ih saznarija . Mit	slim da je prijeko j	potrebno da se	mladi	bave politiko	m u danai	šnje vrijer	ne , konk	retno u	mojoj d	žavi jer s	natram d	a je to je	lini način	da se pok	rener
to 🔲 🕕 doc#11 renomo u pra	vcu ekonomskog razvoj	a . Moramo da idemo n	aprijed , a svi zn	amo dobro pozna	tu izreku da na	mladima	svijet ostaje	Koračajn	no neustr	ašivo u p	ravcu b	udučnos	ti . Na ov	m prostor	ima post	oji pogreš	ina prodra	suda
		ostoji pogrešna predra				mlad	čovjek po pra											

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 " $_{JYJH}$ "<sup>9</sup> ('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 Ad-vanced<sup>10</sup> (Figure 10).

<sup>9.</sup> 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.

<sup>10.</sup> 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).

Simple MILL 103 (2.2767 per million)		Q + n @ 4 X :	E E 8 B B KWC .	1.0
Details Left context	KWIG	Right context		ΨΨ
1 O docif2 - popularna , a više kvalitetna u odnor	u na opšte pritvačenu "muziku". Zabrinjavajuče je koliki postotak, što mladih	to liudi svih dobi , bez pogovora prihvata muziku ko	ia im se outem svih medija servira . Muziku jef	nin meloc
2 O doch2 melodija i još jefinijih , zaglupljujuć	h tekstova koje pišu i izvode osobe upitnih morala . Iz vikenda u vikend , mladi	su bombardovani plitkom i nekvalitetnom muzikom k	oja traje onoliko koliko je puštaju klubovi i kafić	ći , koja po
I doci/3 li da one više uopšte ne predsta	disiu kvalitet proizvoda. Sada se samo prikazuju srećni trenuci u kojima mladi	lijepi , * savršeni ljudi * , koriste proizvod koji se rekl	amira . Upravo psihološka istraživanja zaključi	ila su da je
4 🔲 🛈 doc#7 i godina. Navešču primjer sa tetovaž	ama . Koliko su se ljudi zgrožavali prije na to , a danas je normalno kako mladima	, tako i starijima , ženama , muškarcima . Da li se mi	operisali ili tetovirali , jedno je sigurno . Čovjek	k se najljep
s 🔲 🛈 doc#8 e osjeća u svojoj koži. Od sklapanja	oraka do razvoda danas Danas , u 21. vijeku , koliko god je moderno da mladi	prvo žele da završe dobre škole , da nadu posao i bu	adu finansijski nezavisni ; u istoj mjeri je * mode	erno okuš
a 🔲 🛈 dochš II, da nađu posao i budu finansijski n	zavisni ; u istoj mjeri je " moderno okušati svoju sreću u braku " , postati mladi	bračni par , a veoma brzo i mladi roditelj . Svako od r	nas se nada da će stupiti u brak iz ljubavi , i mi	islim da je
7 🔲 🛈 doc#8 ki nezavisni ; u istoj mjeri je " modern	o okušati svoju sreću u braku ", postati mladi bračni par, a veoma brzo i mladi	roditelj . Svako od nas se nada da će stupiti u brak iz	ljubavi , i mislim da je bilo kome nezamislivo o	da se stup
s 🔲 🛈 doci/10 sporedna stvar ili se pak zanemaruje	i ne pridaje joj se neki značaj . To je pitanje kojim se danas najviše bave mladi	kada dođu u dodir sa istim , kada dođu u situaciju da	su u dilemi da li se upustiti u bračni život . Da	li će taj pr
o 🔲 🛈 dociř10 sačen u sjenku nekih drugih vrijednos	ti i ciljeva kojima kroz život težimo . Kao takav , on se gleda među nama i mladima	kao običan papir koji je sasvim suvišan , danas , kad	a je svakom od nas sloboda i lični integritet od	i neprocje
10 🔲 🕕 doci/10. životu borimo. Ipak ja lično se vodim	svojim razmišljanjem i za mene je brak jedna prekretnica u životu dvoje mladih	, koja je obilježje , pečat njihovih života . Ona vodi ka	a zajedničkom boljem sutra . Tako da , iako je o	ovo 21 . v
n 🔲 🛈 doc#11 go ljepši i lakši . Mi ćemo samim tim l	viti jednog dana mnogo srećniji i ponosniji što smo stekli svoju porodicu. Mladi	treba da se bave politikom Često sebe uhvatim u mis	slima o tome kako mi je žao što živim u ovom o	današnjer
12 🔲 🛈 doci/11 li su više poštovali jedni druge . Više	se radovalo , družilo , a okupljali su se na zajedničkim skupovima i stari i 👘 🖬 🖬 🖬	. Radnici su imali svoja prava a nastavnici dostojanst	tvo . Živim u vremenu u kom je surovi kapitaliz	am uzeo
🕫 🔲 🛈 doc#11 vaju bar nešto da mijenjaju na bolje ,	koji imaju tračak nade i vizije o mogučnosti promjene . Pri tom mislim na mlade	ljude koji bilo kroz nekakvo znanje , iskustvo ili kroz r	reposredan rad u politici učestvuju u stvaranju	bolje
, 14 🔲 🛈 doch11 izliku od njih , stariji imaju uglavnom i	и		svoje neustrašivosti i energije koja je takođe	dopunjen
is 🔲 🛈 doci/11 xg svoje neustrašivosti i energije koje	,		dinci koji posjeduju mnogo znanja iz politički	h i pravnil
16 🔲 🛈 doc#11 vimijene . Moraju da pored toga imaj	vijek slobode , nezavisnosti i samostalnosti , mislim da se sve to može postići	dogovorom , zajedničkim životom u dvoje . Na taj	reba da tečno govori bar jedan strani jezik di	a bi bio ko
🛛 🔲 🕕 docif11. stranih država kao i korištenje stručn	način će nam životi biti mnogo ljepši i lakši . Mi ćemo samim tim biti jednog da porodicu . Madi treba da se bave politikom Često sebe utvatim u mislima o to		državi jer smatram da je to jedini način da se	a pokrene
🐵 🔲 🛈 doc#11 renemo u pravcu ekonomskog razvo	vremenu . Zašto ? Iz razloga što je u današnje vrijeme jako teško biti učenik . r		osti . Na ovim prostorima postoji pogrešna p	redrasuda
/ 19 🔲 🕕 doc#11 ku budućnosti . Na ovim prostorima	x skoro		vezno naivan . Smatram da je to jako pogreš	lno i naza
🛛 20 🔲 🛈 docii11 "zelen "i obavezno naivan . Smatra			lo uključeni u politiku . Time ne doprinose ra	izvoju grad

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.

Lemma	↓ Frequency	Frequency per million	
1 млад	103	2,276.70	 •••

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

pie mitag 103 (2.2	Oktpermilizer	2					9. ± 10 0 4		8 🖬 🕶	E KW	• + ()
ollocatio	CHANGE CRITERIA	BACK TO CONCORDA	NCE								
Word	Cooccurrences 7	Candidates ?	T-score	М	+ LogDice	Word	Cooccurrences 7	Candidates ?	T-score	М	+ LogDice
пьуди	10	39	3.13	6.82	11.17	11 да	8	250	2.63	3.81	9.54
ljudi	15	162	3.78	5.35	10.88	12 koji	7	208	2.47	3.89	9.53
треба	6	37	2.42	6.15	10.46	13 osoba	3	33	1.69	5.32	9.50
Covjek	5	28	2.21	6.29	10.29	14 su	8	262	2.62	3.75	9.49
cy	5	81	2.15	4.76	9.80	15 CB	5	139	2.09	3.98	9.40
među	3	8	1.72	7.36	9.79	15 10	6	235	2.23	3.49	9.18
svaka	3	12	1.72	6.78	9.74	17 140	3	69	1.64	4.26	9.16
док	3	13	1.71	6.66	9.73	18 .	34	1,851	5.11	3.01	9.16
тога	3	13	1.71	6.66	9.73	19 50	13	747	3.13	2.93	8.97
danas	4	66	1.92	4.73	9.60	22 10	6	297	2.17	3.15	8.94

Figure 9. Words which the lemma [MЛАД] ('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*).

ର୍ୟ CHANGE CRITERIA	
BASIC ADVANCED	ABOUT
Query type ③ simple	COL [lemma="book"][](1,3)[tag="V.*"]
lemma phrase word character	teart [] () (> *** & \ I ~ + TAGS COLUMNER :) Default athous 1 lemma •
CQL	
Sutcorpus () none (the whole corp • +	
Filter context 💿 🐱	
Text types ? 🗸	60

Figure 10. Advanced query selection

Then we select the *Parallel Concordance* tool and perform a basic search<sup>11</sup> 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:  $\kappa on anc n pojekma$  (4 examples), nponacm (uenor) npojekma (4 examples), pacnad npojekma (3 examples), nad npojekma, rauene n

We will next perform an advanced CQL query<sup>12</sup> 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.

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

<sup>11.</sup> An advanced search for this noun phrase  $\operatorname{can}$ be performed via the following CQL query aswell: [lemma="project"] [word="''] [word="s"] [lemma="collapse"].

Jointan-Younzi (1903) (1997**) 194 dott za provinski ji ji listi viniti X.          Image: Statistic statisti statistatis statistic statistic statisti statisti statisti stati	Q, ± ∞ Q 4 X = = 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
🗸 🛈 doci/38 koŝulju ; a takvi poslovi ne mogu da čekaju i da se urade samo kad neko bude raspoložen za to . 🛛 Miadi su	naviki da iztaze vikendom sa svojim društvom , a umjesto dosadašnje dominacije i diktature muš 🛽
O doc#38 kośulju , a takvi poslovi ne mogu da čekaju i da se urade samo kad neko bude raspoložen za to . Mladi su navikli	
	da izlaze vikendom sa svojim društvom , a umjesto dosadašnje dominacije i diktature muškaraca
O doc#38 košulju, a takvi poslovi ne mogu da čekaju i da se urade samo kad neko bude raspoložen za to. Mladi su navikli da izla	vikendom sa svojim društvom , a umjesto dosadašnje dominacije i diktature muškaraca , na scer
🗸 🛈 doc#11 ši . Mi čemo samim tim bili jednog dana mnogo srečniji i ponosniji što smo slekli svoju porodicu . 🛛 Mladi treba 👘	da se bave politikom Često sebe uhvatim u mislima o tome kako mi je žao što živim u ovom dane
📋 🕕 doc#11 ši . Mi čemo samim tim bili jednog dana mnogo srečniji i ponosniji što smo slekli svoju porodicu . 👘 Mladi treba da se bav	e politikom Često sebe utivatim u mislima o tome kako mi je žao što živim u ovom današnjem vrem (
🗸 🛈 doc#39. ili još gore trpljenjem torture jednog supružnika nad drugim i neharmoničnih porodičnih odnosa. 👘 Mladi u regionu sve s	u češče izloženi pritisku svojih roditelja i porodice da se vjenčaju , jer je to za neke prijašnje genera 🛾
🗸 🛈 doc#11., stariji imaju uglavnom drugačije načine razmišljanja , oprezniji su i sporije pritvataju promjene . 👘 Mladi zato imaju	veće šanse da nešto promijene , bas zbog svoje neustrašivosti i energije koja je takođe dopunjer
🗸 🛈 doof12 og neispunjenog sna ? Svako treba i zaslužuje da bude srečan , bio on srečan kada stupi u brak 👘 🛛 mlad 🖩 star , bilo	da želi djecu ili ne , bilo da želi da ostane sam do kraja života . Nije svako krojen na isti način , Bc
🗸 🛈 doci/22 ude iz takođe nejasnih razloga . Možda su lijeni da nekog novog obučavaju ili misle da niko ko je 👘 🛛 mlad ne može 👘	dovoljno dobro da radi svoj posao . Veoma često se prave greške prema mladim ljudima koji su t
🗸 🛈 doolf11 iraju da pored toga imaju čvrst karakter i visoka moralna načela . Pored toga , smatram da jedan 👘 🖬 mlad čovijek , bio	to muškarac ili žena koji se bavi politikom treba da tečno govori bar jedan strani jezik da bi bio ko
🗸 🛈 doci/51 , najveća su mana mladosti . Ulažući u sebe , u svoje obrazovanje , u podizanje moralnih normi , mlad čovjek u stvari prod	lužava svoju ljepotu i gradi život koji je pred njim . Na taj način povećava se samopouzdanje , stvaraju či l
🗸 🛈 doolf16 da bude ovo ili ono . Kada osoba krene da se bavi nečim u čemu nije uspješna a pri tome je još i 👘 👘 mlada , brzo će 👘	shvatiti da je taj izbor profesije bio pogrešan . Svi bismo mi voljeli biti matematičari pa da se školé
O dooi/16 da bude ovo il ono. Kada osoba krene da se bavi nečim u čemu nije uspješna a pri tome je još i mlada , brzo će shvat	da je taj izbor profesije bio pogrešan . Svi bismo mi voljeli biti matematičari pa da se škole otimaji 🖡
O doci118 jsao ne bi trebalo da bude kazna , nego sastavni dio ispunjenog života . Sv Select on page: show only.	X i najvjerovatnije bili prisiljena da se zaposli bilo gdje kako bi spojila kraj sa krajen

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

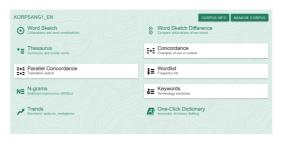


Figure 13. Parallel corpora Dashboard

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,<sup>13</sup> 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

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



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

samp	a project's collegies 10 (cm/cz.perimice)	의 🛓 🗠 🥹 🖉 😫 💩 🕫
	F F 8	🖅 🐺 📴 🚥 KaPSAagt_ar
	<>> The " creative differences" that led to Holotener dismissing Moore , which then led to the project ' s collapse , appear to have centred on Moore 's take on the character – particularly her physicality ,	<c> "Kreativne razlike " zbog kojih je Hototoener objustio Moore , što je vodilo do kolapsa projekta , se isposlavilo da imaju veze sa tim kako je ona doživjela tu ulogu – tačnje sa njenom psihom . <b></b></c>
	$\ll$ The "creative differences " that led to Holotoener dismissing Moore , which then led to the <b>project's collapse</b> , appear to have centred on Moore 's take on the character – particularly her physicality. $\ll$	«с»- "Креатиене разлике које су довеле до тога да Холофсенер отпусти Мур , а и до распада пројекта , когледа да имају возо са тим како је Мур жаљела одиграти улогу , тачније на физички коглед лика .
	$\ll$ The "creative differences " that led to Holofoener dismissing Moore , which then led to the <b>project's collapse</b> , appear to have centred on Moore 's take on the character – particularly her physicality. $\ll$	-co- " Kreativne razlike koje su vodile Holofceneru da isključi Mocre , koje su zatim vodile do pada projekta izgleda da su se fokusirale na Morin pokušaj da se unese u ulogu – posebno njen izgled
	It creative differences " that led to Holdcener dismissing Moore , which then led to the project 's collapse , appear to have centred on Moore 's take on the character – particularly her physicality.	
	$\ll$ The "creative differences" that led to Holoicener dismissing Moore, which then led to the <b>project's collapse</b> , appear to have contried on Moore's take on the character – particularly her physically, $\sim t_{e^{-1}}$	Rackie u kreatanom milijenju koje su dovele do toga da Hotočener objusti Mur, što je rezultnalo u gašenju projekta , bil su zasnovan na tambćinju lika po soprtvinim minilima od strane Mur , a natočito njinom usrodendenošću na ostvartvarje ospotvanih nameca
	$<\!\!>$ The "creative differences " that led to Holofoener dismissing Moore , which then led to the project " a collapse , appear to have centred on Moore 's take on the character – particularly her physicality. $<\!\!>\!\!>$	«»- "Кревтнене реалике" које су довеле до тога да Хологинсер отпусти Мурову а и до копалса самог филма , имало су возе са њеним тумачењем лика – посебно њеним физичали исгладом ; «!»>
	$\ll$ The "creative differences" that led to Hotoicener dismissing Moore, which then led to the <b>project</b> 's <b>collapse</b> , appear to have contried on Moore"'s take on the character – particularly her physically. $\ll$	«> "Креализне разлика" које су довеле до тога да Холофсенер отпусти Мур. "што је потом довело до прогасти цалот пројаста "заснивала су са на начину на који је Мур видала своју јунакину "поскбно њен филичан изглед. <>>>
	$\ll$ The " creative differences " that led to Hotolcener dismissing Moore , which then led to the <b>project</b> " <b>s collapse</b> , appear to have centred on Moore ' s take on the character – particularly her physicality. $\ll$	-s>- Kreativne razlike koje su dovele do ovog događaja , a koji je uzrokovao kraj prvobitne verzije , odnose se na to kako- je Murova doživela svoj lik , posebno spojažnji izgled . <l>&gt;</l>
	-S>The "creative differences " that led to Hotoloener dismissing Moore , which then led to the project 's collapse , appear to have centred on Moore 's take on the character – particularly her physicality.	-cs> "Kriestivna razlika " knje je dovela do toga da Holofser otpusti Mar., što je kasnije dovelo do kolapsa , je izgleda potekla od Murine glavne uloge - tačnije njenog fizičkog izgleda .

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).

inglish		stev Srpski	
Query type (1)	CGL [lemma="have"] [tag="V.*"]	does • contain	
simple Iemma	inter II () 🗢 ** 4	Query type ③	since
	1 I - 2 TAGS	simple	abc
phrase		lemma	
word	cdr anirbia 🖸	phrase	
character	Default attribute 7	word	
CQL	<u>reč •</u>	character	
Autocompus 1		CQL	
nijedan (ceo korpus)	<u>+</u> +		
		<ul> <li>Filter out untranslated</li> </ul>	segments?
Fext types ? 🗸			
lext types			

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.

simple	e Dernr	na="hav	re"[]tage"V."] 432 (7.155.75 per million)			۹. ₹	10	0	<b>a</b> >	< 8	align	• \$
	F.	$\overline{\tau}$	··· 🛙		F 🔻 🖬 🚥						KorPSAn	g1_sr
	not a	travel	writer at all , as her mother had	, thinner , paler , both worldly and ethereal , as though in her heart she was said she wished to be , but simply a traveler , the purer form , someone of feeling , but does not care to record them .	<co ,="" bila="" bleda="" dana="" deniz="" godinu="" i="" je="" od="" starija="" sveti<br="" tanja="" u="" viša="">uopšte bila putopisac , kao što je njena majka govorila da ona to želi prikuplja utiske , guste anatomije osečanja ali je nije briga da ih zabe</co>	da posti	ane , ni					
	<>> SI	he was	self-possessed and thoughtful ,	had brought us hand-carved gifts from the jungles .	<ul> <li>Bila je staložena i brižljiva , donosila nam je ručno napravljene p</li> </ul>	oklone iz	: džung	ji. <td></td> <td></td> <td></td> <td></td>				
	<>*I	I could	have prepared some kind of ki	mchi dish . *	<>> * Mogla sam da napravim neko kimči jelo . *							
	not a	travel	writer at all , as her mother had	, thinner , paler , both worldly and ethereal , as though in her heart she was said she wished to be , but simply a traveler , the purer form , someone of feeling , but does not care to record them . $<\!\!\!/s\!\!>$	<so ,="" bila="" bleda="" dana="" deniz="" godinu="" i="" je="" mršavija="" od="" starija="" sv<br="" viša="">putopisac , kao što je njena majka govorila da ona želi da bude , neg zbijene anatomije osećanja , ali ne mari da ih zapisuje .</so>							
	<> SI	he was	self-possessed and thoughtful,	had brought us hand-carved gifts from the jungles .	Sila je prisebna i uviđavna , donosila nam je ručno izrađene pokl	one iz d	žungli .					
	<>*I	I could	have prepared some kind of ki	mchi dish . *	s , Mogla sam ja napraviti neko kimči jelo . *							
	not a	travel	writer at all , as her mother had	, thinner , paker , both worldly and ethereal , as though in her heart she was said she wished to be , but simply a traveler , the purer form , someone of feeling , but does not care to record them .	Била је годину дана старија од Дениз, виша , иршавија , бле небеском свету , као да у дубини душе уопште није путујући писе , већ само путник , чистијег облика , неко ко сакуппља утиске , сло их забелези	ц, што	је њен	а мајк	а рекла	да би	желепа д	да буде
	<>> SI	he was	self-possessed and thoughtful,	had brought us hand-carved gifts from the jungles .	«» Била је сталожена и пакљива , донела нам је ручно рађене и	токлоне	из џун	юм	45>			

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

F	Ŧ	E					<b>T T</b>	E										KorPSAng1_	57
hear the p	t she	was not a iorm , som	n Denise , she was taller , travel writer at all , as her teone who collects impress	nother had said she	wished to be , but sir	nply a traveler ,	Godini njenom sro putnik , čist	u nije uopš	Ste bila puto	itopisac ,	kao što j	e njena m	ajka go	worila c	ta ona	to želi d	la posta	ine , no obi	ičan
<<>>	She w	as self-pos	ssessed and thoughtful, F	d brought us hand-c	arved gifts from the j	ungles .	<>> Bila je	staložena	i brižljiva , i	, donosila	a nam je	ručno nap	ravljen	e poklo	ne iz d	žungli .			
hear the p	t she	was not a orm , som	n Denise , she was taller , travel writer at all , as her teone who collects impress	nother had said she	wished to be , but sir	nply a traveler ,	<s> Godini nije bila pu ko sakuplja</s>	opisac , ka	ao što je nje	njena maji	ijka govor	ila da ona	2eli da	bude ,	nego :	samo pu			
<>>	She w	as self-por	ssessed and thoughtful , F	ad brought us hand-c	arved gifts from the j	ungles .	<>> Bila je	prisebna i	uvidavna ,	, donosila	la nam je	ručno izra	dene p	oklone	iz džu	ngli . <td>&gt;</td> <td></td> <td></td>	>		
hear the p	t she	was not a orm , som	n Denise , she was taller , travel writer at all , as her seone who collects impress	nother had said she	wished to be , but sir	nply a traveler ,	Била ј овоземаљи рекла да б анатомија	жом и неб и желепа	беском све да буде , е	юту, као већ сам	о да у дуб ю путник	іини души , чистијег	уопш облия	те није а, нею	путују	ћи писа	ц, што	је њена м	

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.

sin	ple: (Jemma="wctpowew"))emma="caalyana-wx"(Jemma="	오 🛓 🗠 💿 🖑 💥 align 🕶 5								
	F F 🖬 🚥	🗁 🔻 📘 🚥 KorPSAng2_en								
	$<\!\!\!>$ Kada sam se pribrao , shvatih da ni ja ne mogu odustati , pa , istrošene svadalačke snage , odlučih da probam na pomirljiv način slomiti Jeleninu upomost , $<\!\!\!>$	$<\!\!>$ When I pulled myself together , I realized that even I could not quit , so , with dissipated energy for quarterling , I decided to try and crush her determination . $<\!\!>\!\!>\!\!>$								
	<s>Kada sam se pribrao , shvatih da ni ja ne mogu odustati , pa , istrošene svađalačke snage , odlučih da probam na pomirljiv način storniti Jeleninu upomost .</s>	<>> When I came to , I realized that neither I could quit , and , deprived of fighting ability , I decided to try to crush Jelena's determination in a peaceful way . >								
	Kada sam se pribrao , shvatih da ni ja ne mogu odustati , pa , istrošene svadalačke snage , odlučih da probam na pomirljiv način slomiti Jeleninu upomost ,	Once I collected myself, I realized that I couldn't give up either, so, with a drained argumentative spirit, I decided to try and break Jelena's determination by apologetic means.								
	$<\!$	When I pulled myself together , I realized that I too couldn't quit , so , with spent fighting endurance , I decided to try in a conclisiony way to break Jelena's determination.								
	<s>Kada sam se pribrao , shvatih da ni ja ne mogu odustati , pa , istrošene svađalačke snage , odlučih da probam na pomirljiv način slomiti Jeleninu upomost ,</s>	SWhen I composed myself , I realized that I could not give up either , so , with the wasted energy on fights I decided to break Jelena's persistence in a conciliatory way . <i>&gt;</i>								
	Kada sam se pribrao , shvatih da ni ja ne mogu odustati , pa , istrošene svadalačke snage , odlučih da probam na pomirljiv način slomiti Jeleninu upomost .	SWhen I pulled myself together , I realised I couldn't give up either , so , with no more strength to argue , I decided to try to make her relent / give in by being apologetic .								
	<s>Kada sam se pribrao , shvatih da ni ja ne mogu odustati , pa , istrošene svađalačke snage , odlučih da probam na pomirljiv način slomiti Jeleninu upomost ,</s>	When I collected myself , I realized I couldn't give up either , so I , with my used up strength for quarreling , decided to try to break Jelena's determination in a conciliatory way.								
	Kada sam se pribrao , shvatih da ni ja ne mogu odustati , pa , istrošene svadalačke snage , odlučih da probam na pomirljiv način slomiti Jeleninu upomost ,	When I got it together. I realized that I could not give up as well , so , having ran out quarreling energy , I decided to try to break Jelena's persistence in a conciliatory manner . >								
	<s>Kada sam se pribrao , shvatih da ni ja ne mogu odustati , pa , istrošene svađalačke snage , odlučih da probam na pomirljiv način slomiti Jeleninu upomost ,</s>	When I pulled myself together , I realized that I couldn't give up either , so , my energy for quarrelling having been exhausted , I decided to try to break Jelena's persistence in a reconciling manner .								
	$\ll\!\!>$ Kada sam se pribrao , shvatih da ni ja ne mogu odustati , pa , istrošene svadalačke snage , odlučih da probam na pomirljiv način slomiti Jeleninu upomost , $<\!\!<\!\!s\!\!>$	<>> Once I regained my composure , I realised I couldn't give up either , so , with strength for arguing used up , I decided to try , in a reconciliatory manner , to break Jelena's persistence .								
	$<\!$	When I pulled myself together , I realized that I couldn't give up either , so , having drained my energy for quarreling , I decided to try to break Jelena's stubbornness in a conciliatory manner . >								
	Se> Kada sam se pribrao , shvatih da ni ja ne mogu odustati , pa , istrošene svadalačke snage , odlučih da probam na pomirljiv način slomiti Jeleninu upomost ,	SWhen I recollected myself, I realised that I too could not give up, so, having diffused my brawing strength, I decided to try to break Jelena's persistence in an amicable way. >								

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

# References

- СЕО. 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 Jakubiček, 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 leksičkosintaksički 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\_l.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, по. 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, по. 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.
- Шућур, Срђан Р. 2020. "Корпус као оруђе за проницање тајни међујезика." Радови Филозофског факултета: Часопис за хуманистичке и друштвене науке, по. 22, https://doi.org/10. 7251/RFFP2022321S.