

КОРПУС КАО ОРУЂЕ ЗА ПРОНИЦАЊЕ ТАЈНИ МЕЂУЈЕЗИКА

(International Corpus of Learner English:

<https://corpora.uclouvain.be/cecl/icle/home>; Sylviane Granger, Maité Dupont, Fanny Meunier, Hubert Naets, Magali Paquot (Eds.).

International Corpus of Learner English. Version 3. Louvain-La-Neuve, Belgium: Presses universitaires de Louvain, 2020, 259 pp.)

Већ дуже од пола вијека корпусна лингвистика представља изузетно живу област лингвистичких истраживања. Разнолики електронски корпуси, који наглу експанзију доживљавају захваљујући развоју модерних информационих технологија, пружају могућност квалитативних и квантитативних истраживања обиља језичких феномена, коришћењем трансверзалних и/или (псеудо)лонгитудиналних метода. Шездесетих година прошлог вијека настају први корпуси писаног и говорног енглеског језика. Касније долази до формирања мноштва других и другачијих корпуса, како на енглеском тако и на осталим језицима. Временом постаје јасан значај корпуса и корпусне лингвистике у настави и учењу страних језика, те се јавља потреба за успостављањем ученичких корпуса, тј. корпуса писмене и усмене продукције неизворних говорника на страном језику. Тиме постаје могуће проучавање комплексне природе међујезика, те међујезичког и унутарјезичког утицаја (односно трансфера у учењу страних језика) и сличних појава. Управо такав је и корпус *ICLEv3*¹ (енг. *International Corpus of Learner English – Version 3*), који ћемо, кроз пратећу публикацију наведену у парентези наслова и платформу на којој функционише, представити у овом приказу.

Како се и из самог назива може претпоставити, ради се о трећој верзији овог корпуса². Прва и друга верзија комплетирани су 2002, односно 2009, а чинили су их аргументативни састави на енглеском

¹ Корпус *ICLEv3* доступан је на следећој веб-адреси: <https://corpora.uclouvain.be/cecl/icle/home>.

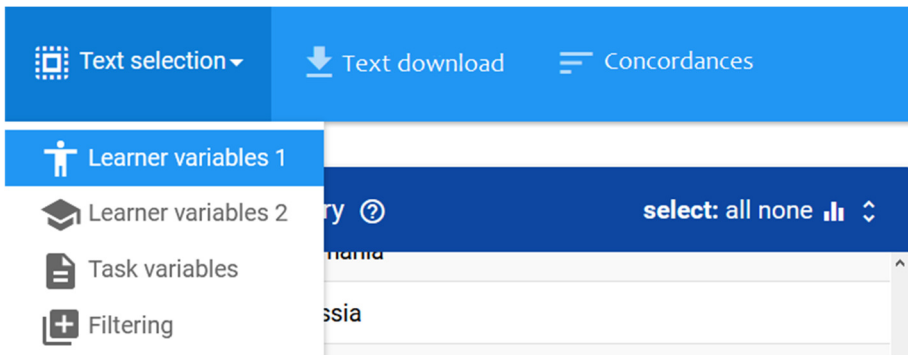
² 1995. године белгијски Центар за англистичку корпусну лингвистику започиње развој корпуса усмене продукције неизворних говорника енглеског језика (*LINDSEI – Louvain International Database of Spoken English Interlanguage*). Касније је формиран и референтни корпус усмене продукције изворних говорника енглеског језика (*LOCNEC – Louvain Corpus of Native English Conversation*).

чији су аутори неизворни говорници на напредном нивоу учења енглеског језика. У том периоду прикупљени су састави изворних говорника бугарског, кинеског, чешког, холандског, финског, француског, њемачког, италијанског, јапанског, норвешког, пољског, руског, шпанског, шведског, цванског и турског језика. Прва верзија корпуса, као резултат пројекта који је, на Католичком универзитету у Левену (фр. *Université catholique de Louvain*) у Белгији, још 1990. године покренула Силвијана Грејнцер [*Sylvianne Granger*], састојала се од 2,5 милиона ријечи, прикупљених из аргументативних састава изворних говорника 11 различитих језика. Друга верзија представљала је значајан помак у односу на прву, како у броју прикупљених ријечи (3,7 милиона) тако и у броју различитих матерњих језика ученика (16). Трећа верзија корпуса (*ICLEv3*), доступна јавности од јануара 2020, представља кулминацију дугогодишњег рада белгијског Центра за англистичку корпусну лингвистику (енг. *CECL – Centre for English Corpus Linguistics*) и значајна је из два разлога: 1) броји рекордних 5,5 милиона ријечи, а број различитих матерњих језика повећан је на 25 (од нових језика, ту су бразилски португалски, грчки, мађарски, персијски [Иран], корејски, литвански, македонски и пакистански, а значај за наше говорно подручје огледа се у томе што ова верзија први пут садржи и српску компоненту – *ICLE-SE*) и 2) корпус је доступан на потпуно новој интернетској платформи, која се одликује флексибилношћу у погледу лакоће и интуитивности претраживања самог корпуса, као и могућностима проширивања корпуса поткорпусима који накнадно буду комплетирани. И трећа верзија, као и претходне двије, врло је хомогена, пошто су сви тимови задужени за прикупљање корпуса усвојили исте смјернице у погледу: врсте састава који се пишу, нивоа познавања енглеског језика ученика и прављења разлике између енглеског језика као страног, односно другог језика.

Трећу верзију корпуса прати и публикација *International Corpus of Learner English. Version 3*, објављена 2020. од стране Католичког универзитета у Левену (*Louvain-la-Neuve*). Ова публикација броји 277 страница, а састоји се од пет поглавља и четири прилога. Прво поглавље (3–22), које слиједи након увода, садржи опис корпуса, критеријуме који су примијењени приликом његовог скупљања, ученичке варијабле, типове задатака чијој су изради ученици приступили, означавање и језичку анотацију. У другом поглављу (23–33) представљени су тимови који су учествовали у прикупљању и формирању корпуса, и то: координациони тим, тим информатичких стручњака, те национални тимови. Треће поглавље (23–33) садржи опис претходно споменутих 25 националних поткорпуса. Четврто

поглавље (53–83) представља свеобухватан приручник за коришћење саме платформе ICLEv3. Пето, завршно и најдуже поглавље (83–236), састоји се од 25 појединачних извјештаја³ различитог обима о положају енглеског језика, сачињених од стране националних тимова. Ти извјештаји, између осталог, садрже информације о језичкој ситуацији и политици учења енглеског и других страних језика у основном, средњем и високом образовању, те приступима у учењу енглеског језика и пратећим трендовима у земљама из којих национални тимови потичу. Прилоге чине: 1) списак кодова додијелих високошколским установама које су учествовале у прикупљању националних поткорпуса, 2) списак предложених тема за писање аргументативних састава, 3) списак вишечланих јединица и 4) списак ознака помоћу којих је могуће вршити претрагу.

У наставку ћемо представити платформу на којој је ова верзија корпуса заснована. Како би се корпус могао претраживати, најприје је неопходно извршити одабир текстова на основу мноштва понуђених параметара.



Слика 1. Одабир текстова (енг. *Text selection*)

Испрва су доступна три падајућа менија: 1. *Text selection* (нуди читав низ варијабли за одабир текстова, и то: *Learner variables 1*, гдје се могу одабрати матерњи језик, земља из које студенти долазе, пол, узраст, страни језици које познају, те језици који се користе код куће, затим *Learner variables 2*, гдје се бира образовна институција коју студенти похађају, те *Task variable*, гдје су понуђени аргументативни, литерарни и остали састави, услови у којима је састав писан, тј. да ли је

³ Извјештај о положају енглеског језика у Републици Србији и Републици Српској сачинили су проф. др Ненад Томовић и проф. др Јелена Марковић.

израда била временски (не)ограничена, као и да ли је састав настао у оквиру (пред)испитних обавеза, те да ли су коришћени неки извори, као и могућност груписања текстова на основу осталих претходно описаних параметара); 2. *Text download* (служи за преузимање одабраних текстова) и 3. *Concordances* (служи за претраживање конкорданци).

За потребе овог приказа бирамо цјелокупан корпус, тј. 9,529 аргументативних састава који броје преко 5,5 милиона ријечи.

selected texts: 9529 (5,766,522 words)

Native language select: all none

<input type="checkbox"/> Albanian	6
<input type="checkbox"/> Arabic	2
<input type="checkbox"/> Aromanian (Vlach)	1
<input type="checkbox"/> Bosnian	2
<input type="checkbox"/> Bulgarian	300
<input type="checkbox"/> Chinese	160
<input type="checkbox"/> Chinese-Cantonese	814
<input type="checkbox"/> Chinese-Mandarin	8
<input type="checkbox"/> Czech	241

Country select: all none

<input type="checkbox"/> Austria	70
<input type="checkbox"/> Belgium	473
<input type="checkbox"/> Bosnia Herzegovina	144
<input type="checkbox"/> Botswana	161
<input type="checkbox"/> Brazil	412
<input type="checkbox"/> Bulgaria	302
<input type="checkbox"/> China-Hong Kong	800
<input type="checkbox"/> China-Mainland	179
<input type="checkbox"/> Czech Republic	241

Gender select: all none

<input type="checkbox"/> Female	7356
<input type="checkbox"/> Male	2133
<input type="checkbox"/> Unknown	40

Age reset

min: 16

max: 71

Unknown (366)

Слика 2. Ученичке варијабле 1 (енг. *Learner variables 1*)

У тзв. поједностављеном претраживању конкорданци (енг. *Simple search*) затим дајемо упит за облик који се може очекивати у међујезику неких неизворних говорника – **advices*, како бисмо провјерили његово присуство у одабраном корпусу.

selected variables: [RESET VARIABLES](#) selected texts: 9529 (5,766,522 words)

Simple search [Advanced search](#)

advices

advices

Results: 26 hit(s) [DOWNLOAD CONCORDANCE IN .XLSX \(MAXIMUM: 10,000 HITS\)](#) [DOWNLOAD CONCORDANCE IN .CSV \(MAXIMUM: 10,000 HITS\)](#)

> 1	BGSU1129	From one side is the simple matter of theory and studied advices and learned
> 2	BGSU1129	Graduates are always challenged to making deep observations and taking out the most essential advices .
> 3	BRMG1138	In addition , some advices for you to p:
> 4	BRMG1180	I 'm not saying that people repel their advices and the con
> 5	CNUK2005	During doing this , they do not just listen to their parents advices , they also c
> 6	CZUN1021	Why should n't gardeners watch their favourite programme full of clever advices and innovat

Слика 3. Поједностављена претрага конкорданци (енг. *Simple search*) за упит **advices*

Претрага за упит **advices* даје 26 резултата издвојених из састава изворних говорника бугарског, португалског, кинеског, чешког, холандског, француског, грчког, иранског, италијанског, јапанског, корејског, македонског, српског, шпанског и цванског језика.

У наставку, на цјелокупном узорку *ICLEv3* вршимо упит на основу облика који је уочен у усменој и писменој продукцији неких студената кинеског и енглеског језика, који су изворни говорници српског, и то: **no any*.

selected variables: [RESET VARIABLES](#) selected texts: 9529 (5,766,522 words)

Simple search [Advanced search](#)

no any

no any

Results: 17 hit(s) [DOWNLOAD CONCORDANCE IN .XLSX \(MAXIMUM: 10,000 HITS\)](#) [DOWNLOAD CONCORDANCE IN .CSV \(MAXIMUM: 10,000 HITS\)](#)

> 1	BGSU1196	There is no any other reasonable explanation why an experienced teacher would give up his , otherwise noble , profession
> 2	CNUK1173	There is no any good effect to our economy .
> 3	CNUK1855	and administrators and managerial personnels , but no any low-skilled workers such as physical labour workers .
> 4	CNUK1725	Ocean Park have no any accommodation facilities .
> 5	CNUK1772	There is no any benefit to either teenagers or children .
> 6	CNUK1872	written by the schoolar who have no any real practice or experience of the subject they are talking .
> 7	CNUK4810	We have no any right to decide what kind of people it will be .
> 8	BRMA2002	There is no any doubt about it that , prison is a dangerous place , especially for those who enter
> 9	KNSP1856	the same school even though there are no any solid evidence for that .
> 10	KSHY1846	watch tv after their daily routine , there is no any conversation or heat .

Слика 4. Поједностављена претрага конкорданци за упит **no any*

Претрага даје 17 резултата издвојених из састава изворних говорника бугарског, кинеског, корејског, иранског, литванског, пакистанског, руског и турског. Овакав корпус, дакле, даје могућност примјене напредних модела изучавања трансфера у учењу страних језика, који подразумевају вишеструко поређење међујезика изворних говорника различитих језика (а могуће је поређење и с контролним корпусом аргументативних састава изворних говорника британске и америчке варијанте енглеског језика – *LOCNESS – The Louvain Corpus of Native English Essays*, који броји 323 хиљаде ријечи), као и све типове усмјерености трансфера, тј. комплексног узајамног утицаја свих језика које неки ученик познаје.

У наставку ћемо описати и могућности напредне претраге корпуса (енг. *Advanced search*), те на цјелокупном корпусу *ICLEv3* приказати начин на који упити функционишу.

The screenshot shows the 'Advanced search' interface. At the top, there are two tabs: 'Simple search' and 'Advanced search'. Below the tabs is a search bar with a magnifying glass icon and a search button. To the right of the search bar are three filter icons and a close button. Below the search bar is a list of search filters, each with a dropdown menu:

- Form (case insensitive) equals Form (case insensitive)
- Form (case sensitive) equals Form (case sensitive)
- Lemma equals Lemma
- Part of speech equals Part of speech
- Simplified Part of speech equals Simplified part of speech
- Any word

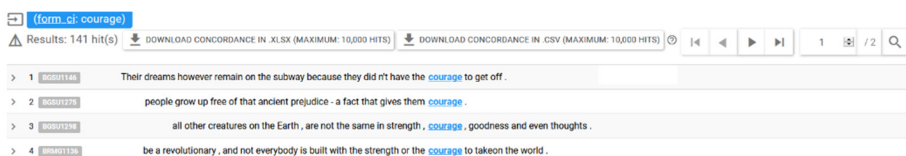
At the bottom of the interface, there are five buttons: '← MOVE TO THE LEFT', '→ MOVE TO THE RIGHT', 'DELETED THIS WORD', 'DELETED THE QUERY', and 'SEARCH'.

Слика 5. Напредна претрага (енг. *Advanced search*)

Напредна претрага корпуса нуди могућност задавања шест различитих типова упита, почевши од: 1) претраге било ког облика ријечи, независно од тога да ли садржи велика слова (*Form: case insensitive*), као и 2) претраге која издваја ријечи које не садрже велика слова (*Form: case sensitive*), затим 3) претраге по лемама (*Lemma*), односно свих могућих облика неке леме (тако се нпр. претрагом помоћног глагола *do* добијају и резултати који, поред самог глагола који се претражује, садрже и облике *does, doesn't, did, didn't, done, doing*). Могућа је и 4) претрага помоћу врсте ријечи (*Part of Speech*) (засно-

вана на детаљном скупу етикета које предвиђа CLAWS7⁴, попут: упитних детерминатора DDQ (*wh-determiner*), приједлога (IF – *preposition for*, IO – *preposition of*), затим вишечланих приједлога (II21, II22, II31), заједничких именица (NN – *common noun*), разних типова сложеница (NN121, NN132, NN141), глагола (VB0 – *be*, VDD – *did*, VHZ – *has*, VV0 – *lexical verb base form*, VVD – *lexical verb past tense*), прилога (RR – *general adverb*), придјева (JJR – *general comparative adjective*), итд. Доступна је и 5) претрага помоћу поједностављених етикета за врсте ријечи (*Simplified Part of Speech*), такође заснованих на CLAWS7, попут ADJ (придјеви), ADV (прилози), DET (детерминатори), N (именице), NEG (одрични облици), PREP (приједлози), GE (генитиви), итд. Поред свих наведених типова претраге, могуће је вршити и 6) претрагу по „било којој ријечи“ (*any word*), гдје је астериск (*) ознака за такву ријеч. На тај начин могуће је у низовима узастопних ријечи посматрати једну позицију означену астериском, која је промјенљива. Тако претрага упитом *on the (*) hand* даје и резултат *on the one hand*, као и *on the other hand*. Сваки од описаних типова упита нуди могућност одабира да ли издвојени резултати почињу траженом ријечју, завршавају њом, или је садрже и сл., те извоза резултата у документе формата .XLSX и .CSV.

Даље вршимо претрагу именице *courage* помоћу првог и трећег типа упита, те по једног типа лексичких (V+N) и граматичких колокација (*Adj+Prep*), заснованих на поједностављеним етикетама за врсте ријечи.



Слика 6. Напредна претрага помоћу првог типа упита, за именицу *courage*

Претрага именице *courage* извршена на цијелом корпусу ICLEv3 даје 141 резултат.

⁴ <http://ucrel.lancs.ac.uk/claws7tags.html>.

Уколико обавимо трећи тип упита, тј. извршимо претрагу по ле-
мама, добијамо и садашње и прошле партиципе ових глагола и пер-
фекатске облике (*discouraging, encouraged*), као и именице
(*discouragement* и *encouragement*).

> 14	By encouraging them to think critically , organise and express their thoughts orally or in written form .
> 15	They encourage and win the student 's interest thus facilitating the educational process .
> 16	Their dreams however remain on the subway because they did nt have the courage to get off .
> 17	cycle functions perfectly in our century The modern scientific , technical and industrial development encourages us to dream and strive to have our dreams fulfilled (in simulative or real ways
> 18	And this might be much discouraging .
> 19	Nonetheless , it should be encouraged adequately by the government - for instance , by tax concession for the firms which steadily
> 20	of higher learning where the brightest were brought together in an environment that would encourage wild ideas , creativity , debate and in the process expand the horizons of basic science
> 21	who study and work ? " said , " We can neither forbid nor encourage them to do that but statistics show that students who have been working during these 4
> 22	The idea of teamwork may be encouraged in theory , but in practice , teamwork , if it exists , is usually optional
> 23	Furthermore , the absence of practical skills has to do with discouragement of students ' expressing personal opinion , which inevitably leads to suppression of leade

Слика 7. Напредна претрага помоћу трећег типа упита, за лему *courage*

Оваквим претрагама корпуса, примјеном разних доступних па-
раметара, могуће је вршити поређења, као и квалитативна и/или
квантитативна истраживања чији је предмет међујезик изворних го-
ворника једног или групе разних матерњих језика, те разни творбени
процеси и сл.

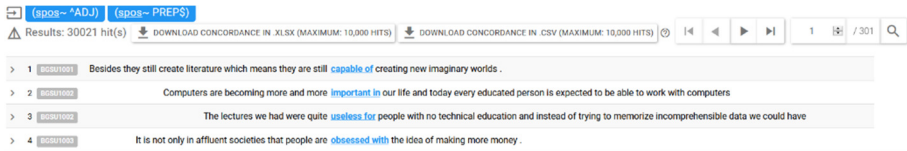
На крају, помоћу поједностављених етикета за врсте ријечи
вршимо претрагу претходно споменутих типова лексичких (*V+N*) и
граматичких колокација (*Adj+Prep*). Из целокупног корпуса издво-
јено је 66,619 примјера који садрже овај тип лексичких колокација
V+N.

[appos ~ "Vlex"] [appos ~ NS]	
Results: 66619 hit(s) DOWNLOAD CONCORDANCE IN XLSX (MAXIMUM: 10,000 HITS) DOWNLOAD CONCORDANCE IN CSV (MAXIMUM: 10,000 HITS)	
> 1	of a country is based on its enormous industrial corporations that are gradually replacing men with machines .
> 2	But I think that in our modern world which worships science and technology there is still a place for dreams and imagination .
> 3	Besides they still create literature which means they are still capable of creating new imaginary worlds .
> 4	People still use their imagination to paint pictures and to compose music .

Слика 8. Напредна претрага помоћу поједностављених етикета за упит
V+N

Но, важно је истаћи да је у резултатима претраге уочен и извје-
стан број колокација које садрже прошле партиципе у улози атри-
бута. Стога је претходно описана сложенија и свеобухватнија пре-
трага помоћу етикета за врсте ријечи (упитима попут *VVO+NN*,
VVD+NN и сл.) пожељна како би из резултата претраге аутоматски
били искључени такви примјери.

Претрагом типа граматичких колокација *Adj+Prep* добијамо 33,021 примјер.



Слика 9. Напредна претрага помоћу поједностављених етикета за упит *Adj+Prep*

И оваквим типом упита, као и претходно описаним, могуће је вршити разне облике поређења међујезика изворних говорника различитих матерњих језика, као и поређења с контролним корпусом који смо споменули, те другим, значајно обимнијим корпусима енглеског језика изворних говорника (попут *BNC – British National Corpus* и *COCA – Corpus of Contemporary American English*).

Приручник за *ICLEv3* (доступан у електронској форми на следећој веб-адреси: https://corpora.uclouvain.be/cecl/icle/ICLEv3_manual_100371-PUL-ICLEv3-V3_final.pdf), под називом *International Corpus of Learner English – Version 3*, чији су уредници Силвијана Грејнцер, Маите Дјупонт [Maïté Dupont], Фани Муније [Fanny Meunier], Хуберт Нетс [Hubert Naets] и Магали Пако [Magali Paquot] 2020. године издала је универзитетска штампарија Универзитета у Левену, Белгија (*Presses universitaires de Louvain: Louvain-La-Neuve, Belgium*).

Српска компонента треће верзије корпуса (*ICLE-SE*), на иницијативу проф. др Јелене Марковић, прикупљена је у периоду од септембра 2015. до јуна 2016. године. Тим који је радио на прикупљању српске компоненте, поред руководиоца проф. др. Јелене Марковић, чинила су три стручњака из области англистичке лингвистике: проф. др Ненад Томовић (Универзитет у Београду), проф. др Биљана Радић Бојанић (Универзитет у Новом Саду), проф. др Жељка Бабић (Универзитет у Бањој Луци), те три млада истраживача/сарадника: Миња Радоња (Универзитет у Источном Сарајеву), Јелена Матић (Универзитет у Београду) и Срђан Шућур (Универзитет у Источном Сарајеву). Српски поткорпус чини 202,621 ријеч, издвојена из 325 аргументативних састава на енглеском студената катедри за англистику на два универзитета у Републици Српској: Катедре за англистику Филозофског факултета у Источном Сарајеву и Катедре за англистику Филолошког факултета у Бањој Луци, као и студената катедри за англистику два универзитета у Србији: Катедре за англистику Филолошког

факултета Универзитета у Београду и Катедре за англистику Филолошког факултета Универзитета у Новом Саду (од чега су из Источног Сарајева – 75 састава, 48,932 ријечи, Бања Луке – 73 састава, 50,959 ријечи, Београда – 100 састава, 59,437 ријечи и Новог Сада – 77 састава, 43,293 ријечи).

Корпус *ICLEv3* и његова српска компонента *ICLE-SE* послужили су и служе као основа за научно-истраживачке пројекте, те за израду монографија, дисертација и научних радова⁵ из разних области англистичке лингвистике.

Корпус функционише на платформи *Corpora@*⁶, осмишљеној од стране Центра за обраду природног језика (фр. *Centre de Traitement Automatique du Langage – CENTAL*). Ова платформа омогућиће проширивање корпуса свим будућим корпусима који накнадно буду прикупљени у сарадњи са белгијским Центром за англистичку корпусну лингвистику. Како активности Центра трају пуних тридесет година,

⁵ Gries, S. & Wulff, S. (2020). Examining individual variation in learner production data: A few programmatic pointers for corpus-based analyses using the example of adverbial clause ordering. *Applied Psycholinguistics*, 1-21, 279–299. doi:<https://doi.org/10.1017/S014271642000048X>

Марковић, Ј. (2020). Концесивни конектори *though* и *however* у писању на енглеском језику код изворних и неизворних говорника. *Филолоџ – часопис за језик, књижевност и културу*, 21, 13–35. doi:<http://doi.org/10.21618/fil2021013m>

Radić-Bojanić, B. (2019). Neodređena zamenica ONE u pisanju kod neizvornih govornika engleskog jezika. *Godišnjak Filozofskog fakulteta u Novom Sadu*, 44 (2), 39–52. doi:<https://doi.org/10.19090/gff.2019.2.39-52>

Radonja, M. (2020). The use of interactive metadiscourse in Serbian students' writing in English. *Радови Филозофског факултета*, 21(2), 121–134. doi:<https://doi.org/10.7251/FIN1921121R>

Römer, U. et al. (2018). Verb-argument constructions in advanced English learner production: Insights from corpora and verbal fluency tasks. *Corpus Linguistics and Linguistic Theory*, 16(2), 303–331. doi:<https://doi.org/10.1515/cllt-2016-0055>

Schweinberger, M. (2020). How Learner Corpus Research can inform Language Learning and teaching – An analysis of adjective amplification among L1 and L2 English speakers. *Australian Review of Applied Linguistics*, 43 (2), 196–218. doi:<https://doi.org/10.1075/ara1.00032.sch>

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

⁶ <https://corpora.uclouvain.be/catalog/>

уз доступност све три верзије корпуса, отвара се мноштво могућности за вршење истраживања, нпр. међујезичке дијахроније⁷ и поређења с еволуцијом енглеског језика као изворног, те процеса попут американизације на плану лексике и сл.

Срђан Р. Шућур
Универзитет у Источном Сарајеву
Филозофски факултет
srdjan.sucur@ffuis.edu.ba

⁷ Gilquin, G. (2019). Diachronic learner corpus research: Examining learner language through the lens of time. *International Computer Archive of Modern and Medieval English Conference, ICAME 40*. <http://hdl.handle.net/2078.1/217739>