О *Корпусу студената англистике* (*КорСАнг*) и могућностима његове софтверске експлоатације

УДК 81'322.2

Мињa С. Радоња
minja.radonja@ff.ues.rs.ba

Срђан Р. Шућур
srdjan.sucur@ff.ues.rs.ba

Универзитет у Источном Сарајеву
Филозофски факултет Пале
Сарајево
Босна и Херцеговина

САЖЕТАК: Корпусна лингвистика у Босни и Херцеговини и Републици Србији не користи се довољно. Разлози за то су многи: недостатак језичких корпуса, бојазан од рачунарских метода, још увијек присутан традиционални приступ обраде података који је квалитативан или не сеже даље од дескриптивне статистике. Неријетко ћемо чути аргументе против рачунарског метода, као што је тај да велики број примјера који су резултат претраге може нарушити квалитет анализе и довести до извођења погрешних лингвистичких законитости. Ипак, развој технологије утицао је и на развој информатичке писмености у свим људским дјелатностима, укључујући и академску заједницу. У овом раду образложићемо на који начин је неколико катедри за англистику у Босни и Херцеговини и Србије у сарадњи са Друштвом за језичке технологије и ресурсе понудило једно рјешење ученичког корпуса, те описати како је текао процес прикупљања корпуса и какве опције претраге корпус нуди.

КЉУЧНЕ РЕЧИ: ученички корпус, корпусна лингвистика, паралелни корпуси, анотација корпуса

РАД ПРИМЉЕН: 16. август 2021.
РАД ПРИХВАЋЕН: 2. септембар 2021.

1. Увод

„Компјутерски ученички корпуси електронске су базе података формиране од аутентичне дискурсне продукције на страном/другом

језику по експлицитним критеријумима, тако да се могу користити за истраживање у оквиру усвајања другог језика и/или наставе страног језика. Базе података бележе се стандардизовано и хомогено, уз информације о пореклу података“ ((Granger 2002, 7), према (Марковић 2019, 13)). Корпус студената англистике (КорСАнг) ученички је корпус студената англистике србофоних говорника са 4 универзитета – Универзитет у Источном Сарајеву, Универзитет у Бањалуци, Универзитет у Београду и Универзитет у Новом Саду. Корпус је настао као производ два значајна национална пројекта: *Фразеолошка компетенција српских говорника енглеског кроз призму контрастивне анализе међујезика*, 19/6-020/961-46/18, који је реализован у периоду од 31. 12. 2018, и *Научни потенцијали анотираних ученичких корпуса у примијењеној лингвистици*, 19.032/961-135/19, који је реализован у периоду од 31. 12. 2019. Корпус КорСАнг је инспирисан првим ученичким корпусом енглеског језика србофоних говорника *ICLE-SE*, који је формиран у периоду од септембра 2015. до септембра 2016. године и састоји се од аргументативних састава студената англистике србофоних говорника писаним на енглеском језику. Досадашњи резултати истраживања на корпусу *ICLE-SE* представљени су у научним радовима (Марковић 2017, 2018, 2019, 2020; Radić-Bojanić 2019; Radonja 2019; Šućur 2019; Spajić and Suknović 2019; Tomović and Stefanović 2019). Више о корпусу *ICLE-SE* и начинима његове претраге видјети у (Шуђур 2020).

1.1 Структура корпуса

КорСАнг обухвата 327 текстова и укупно има око 140.000 речи. По саставу је трокомпонентан, тј. састоји се од три поткорпуса:

– Поткорпус превода студената англистике са енглеског језика (*КорПСАнг1*), који се састоји од превода књижевних (3 текста) и новинских чланака (3 текста) са енглеског на српски језик. Овај дио корпуса броји 127 задатака превода и 46.785 ријечи. То практично значи да се сваки превод може упарити са оригиналним текстом који је у овом поткорпусу на енглеском језику, а што се назива битекстом.[1].

1. „[Б]итекст представља текст и његов превод, односно преводе, представљене на такав начин да је између елемената њиховог логичког исказа успостављена експлицитна веза, на пример, на нивоу пасуса или реченица“ ((Vitas 2010, 273), према (Андоновски 2019, 17))

– Поткорпус превода студената англистике на енглески језик (*КорПСАнг2*), гдје су почетни текстови из жанра књижевности (1 текст) и новинског жанра (5 текстова). Коначан број превода овог поткорпуса је 130, а број ријечи 50.128. Као и у претходном поткорпусу, и *КорПСАнг2* је паралелни корпус.
– Поткорпус аргументативних састава[2] студената англистике писаних на српском језику као матерњем (*КорССАнг*) обухвата 70 састава и 40.012 ријечи.

Прве двије компоненте јесу паралелни ученички корпуси, док је трећа компонента осмишљена као референтни корпус корпусу *ICLE-SE*. „Под паралелним корпусима подразумевају се корпуси који садрже један или више оригиналних текстова и њихове преводе на један или више језика“ ((Tõny 2016, 11) према (Андоновски 2019, 16). О значају и примјени паралелних корпуса видјети детаљније (Андоновски 2019, 2021; Ристовић 2012, 2016)).

У наставку рада, у другом одјељку ћемо описати фазе припреме корпуса, од прикупљања и одабира текстова, преко обраде и паралелизације текстова, до израде метаподатака и формирања корпуса. Трећи одјељак приказује како се корпус може претраживати. Коначно, у закључку, сумирамо досадашње резултате и представљамо даље планове.

2. Припрема корпуса *КорСАнг*

У формирању корпуса укупно је учествовало 194 студената англистике. У периоду од 2016. до 2019. године формиран је *КорССАнг*, а од 2017. до 2019. године *КорПСАнг1* и *КорПСАнг2*. Од наведене три компоненте, најзахтјевније је било прикупити корпус састава на српском језику, због недовољне жеље студената да учествују у писању састава на матерњем језику, те је формирање овог дијела корпуса најдуже потрајало. Пројекат настанка паралелног *Корпуса студената англистике* састојао се од неколико етапа:

– припремна фаза,

2. „Аргументативни текст (расправа) је текст у којем аутор почиње од неке сумње или дилеме, а онда навођењем аргумената долази до разрешења. Аутор таквог текста тежи да доказима саговорника одн. читаоца убеди у исправност свог мишљења“ (CEO 2015, 79).

– прикупљање текстова,
– обрада текстова,
– паралелизација,
– вођење евиденције о метаподацима,
– публиковање текстова и метаподатака на платформу.

2.1 Припремна фаза

У припремној фази је дефинисано неколико кључних питања у вези са пројектом: главни и споредни циљеви пројекта; одређивање обима пројекта; прецизирање сарадње и расподјела задатака између чланова тима са Универзитета у Источном Сарајеву, Универзитета у Бањалуци, Универзитета у Београду и Универзитета у Новом Саду; прецизирање сарадње са члановима тима из JePTeха;[3] утврђивање временског оквира извршења појединачних етапа. За успостављање контаката са члановима тима била је задужена руководилац пројекта Јелена Марковић.

2.2 Прикупљање и одабир текстова

У сљедећој фази су чланови тима, заједно са руководиоцем пројекта, извршили одабир неколико текстова за превод са енглеског и на енглески језик и направили прелиминарни списак тема састава на српском језику. Приликом селекције текстова, водило се рачуна да текстови буду из књижевног и новинског жанра. Новински чланци покривају различите теме, као што су политика, економија, језик, здравље, друштвене мреже, шоу бизнис. У коначници је изабрано 6 текстова на енглеском, 6 текстова на српском и 24 теме састава на српском језику. Потом се приступило сакупљању превода и састава, што је био временски најзахтјевнији задатак који се састојао од организовања услова за превођење текстова и писање састава од стране студената. Чланови тима су организовали овај задатак на својим факултетима. Превођење је било временски ограничено, са изузетком од неколико радова, док је писање састава бар у трећини радова било неограничено. Уз задате писмене задатке, студенти су били дужни попунити пропратни документ профила учесника како би дали сагласност да се њихови преводи и есеји користе у сврхе истраживања, док су сами радови анонимни и евидентирани под шифром.

3. Друштво за језичке технологије и ресурсе – JePTex

2.3 Обрада и паралелизација текстова

Радни задаци који се тичу обраде текстова и паралелизације обављени су од стране сарадника на пројекту, под руководством Јелене Марковић и тима из JePTeха. Сарадници су прошли онлајн обуку за коришћење алата за паралелизацију како би се добили текстови у формату TMX[4] (Translation Memory eXchange) докумената. Упознати су са радом *Notepad++*[5] XML едитора, *Unitex/GramLab*,[6] Unitex модула за српски (Krstev 2008) и интегрисаног развојног окружења за паралелизоване корпусе *ACIDE* (Aligned Corpora Integrated Development Environment) (Utvić, Stanković, and Obradović 2008), који су омогућили адекватну обраду текста по прописаним правилима. У Notepad++ извршена је припрема изворног и преведеног текста, која подразумјева упаривање параграфа два документа, тако да сваком параграфу изворног текста одговара параграф превода. Тако припремљени фајлови даље су обрађени у *Unitex Visual IDE*, који омогућава сегментацију параграфа на реченице, сагласно језику на ком је писан текст. Након сегментације оригиналног и преведеног текста, фајлови су припремљени за паралелизацију, тј. „процес успостављања веза између одговарајућих варијанти јединица превођења, односно формирање сета јединица превођења“ (Андоновски 2019, 17). Паралелизација је омогућена коришћењем апликације *ACIDE* коју је развила Група за језичке технологије Универзитета у Београду (више о интегрисаном развојном окружењу за паралелизоване корпусе *ACIDE* може се пронаћи код (Utvić, Stanković, and Obradović 2008). Ова апликација омогућава аутоматску паралелизацију уз могућност провјере и корекција погрешака, што је члановима тима који су имали овај задатак помогло да успјешно припреме текстове за паралелни корпус. У текстовима који су припремани за паралелни корпус у оквиру овог

4. TMX је ISO стандард (ISO24616 2012) за складиштење такозваних преводилачких меморија (Translation Memories) и њихову размену између различитих софтверских преводилачких алата, као и између различитих фирми које се баве одржавањем преводилачких меморија (TMX 2005). Преводилачке меморије представљају збирке одредница у којима је текст изворног језика повезан са еквивалентним преводом текста циљног језика односно произведени TMX документ састављен је од добијених јединица превођења (Андоновски 2019, 22).

5. Notepad++

6. Unitex/Gramlab, Пакет за обраду корпуса применљив на различитим платформама.

пројекта, наилазили смо на примјере да у тексту превода недостају неки сегменти, или једном сегменту у оригиналном тексту одговара неколико сегмената у тексту превода, што се може јавити код превода. Сама обука је била изазов тиму лингвиста који се раније нису сусретали са оваквом врстом радних задатака, а стечене вјештине у току израде паралелног корпуса дале су нову перспективу на корпусну лингвистику. Као коначни резултат након обраде текста у *ACIDE*-у је *TMX* документ који се уноси у корпус.

2.4 Израда метаподатака и формирање корпуса

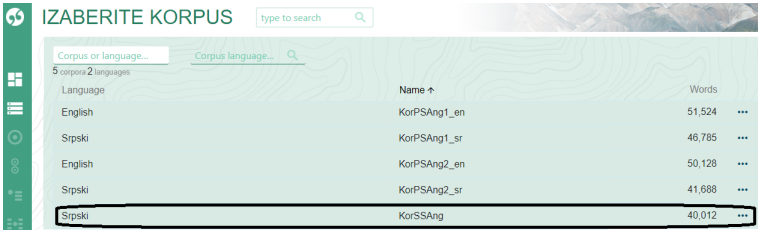
Паралелно са припремом текстова за корпус припремани су метаподаци који садрже информације о варијаблама као што су подаци о учеснику (узраст, пол, матерњи језик, подаци о образовању, године учења енглеског, познавању других страних језика, боравак на енглеском говорном подручју) и варијабле контекста језичке продукције, тј. подаци о томе у каквом су контексту ученици извршавали задатке (да ли је постојало временско ограничење за задатаке писања састава или превода, коришћење рјечника и приручника, да ли је задатак испитна обавеза или је писан ван испита, који је жанр, и сл.). Табеле са метаподацима рађене су посебно за преводе и саставе у засебним Excel документима. С обзиром да су радови који улазе у корпус анонимни, сваком документу је било потребно додијелити одговарајуће метаподатке заведене под истом шифром. Чланови тима из JePTeха припремљене документе су припремили за публиковање: додијелили метаподатке сваком документу, извршили анотацију врстом ријечи и лематизацију (детаљније о ресурсима који су омогућили ово видјети у (Stanković et al. 2020).

3. Претрага корпуса *КорСАнг*

Укратко смо описали како је текао процес прикупљања корпуса од припремне фазе све до његовог коначног електронског облика и изазови са којима смо се на том путу сусретали. *КорСАнг* садржи 136.925 ријечи: *КорПСАнг1* (46.785 ријечи, 127 превода на српски језик), *КорПСАнг2* (50.128 ријечи, 130 превода на енглески), *КорССАнг* (40.012 ријечи, 70 састава на српском језику). Један од већих изазова је био мотивисати студенте да пишу саставе на матерњем језику, што буди сумњу у опадање компетенције писања на матерњем језику (Šućur 2020, 144).

Напоменућемо и то да теме састава ограничавају могућност претраге. Од 24 понуђене теме, више од 40% састава је писано на четири теме: *Породица или посао: шта је важније*, *Како музика утиче на живот*, *Како замишљам доброг родитеља*, и *Од склапања брака до развода данас*. Стога очекујемо да корпус састава нуди богат вокабулар који припада семантичком пољу породичних односа, али сиромашнији када су у питању неке друге понуђене теме. У наставку ћемо, након уводних напомена о платформи *Sketch Engine*, најприје показати како је помоћу ње могуће вршити основну и напредну претрагу корпуса КорССАнг, а затим и претраживање паралелизованих корпуса *КорПСАнг1_ен* (чине га изворни текстови на енглеском) и *КорПСАнг1_ср* (чине га преводи на српски), односно *КорПСАнг2_ср* (чине га изворни текстови на српском) и *КорПСАнг2_ен* (чине га преводи на енглески).

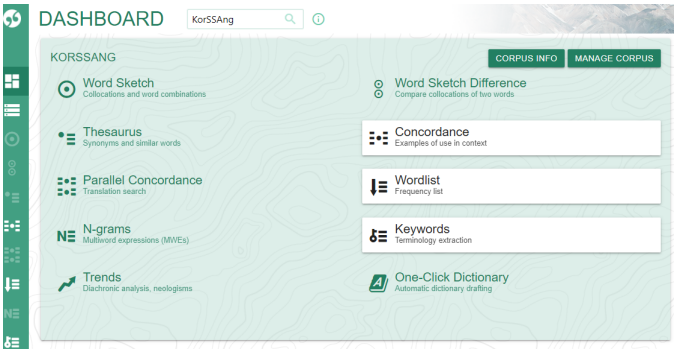
3.1 *Sketch Engine*



Слика 1. Одабир корпуса

Sketch Engine је онлајн платформа сачињена од низа робустних алата за претраживање електронских корпуса. Ову платформу је, у сарадњи са чешким програмером Павелом Рихлијем [*Pavel Rychlý*], осмислио британски лексикограф и корпусни лингвиста Адам Килгариф [*Adam Kilgarriff*], а њена комерцијална верзија доступна је од 2004. године (Kunilovskaya and Koviazina 2017, 503). Како истиче Kilgarriff et al. (2014, 15-16), Sketch Engine је потекао из свијета академских истраживања, гдје се данас користи у лингвистици и на катедрама за изучавање језика (у настави и истраживањима), затим у рачунарској лингвистици (енг. *Computational Linguistics*) и обради природног језика (енг. *Natural Language Processing*). Поред тога, користи се и у настави језика

(нарочито енглеског као страног), лексикографији, превођењу и настави превођења, затим анализи дискурса итд. (Kilgarriff 2014, 15-16).



Слика 2. Контролна табла

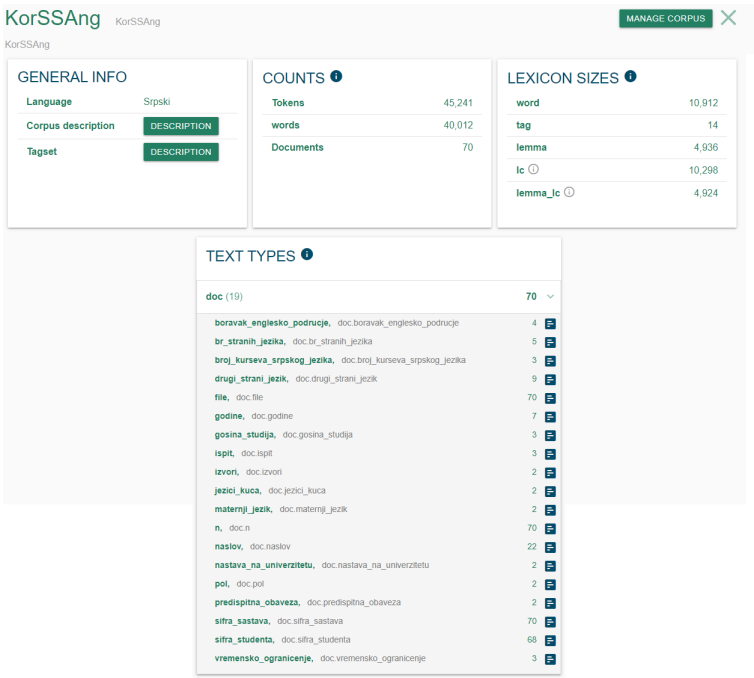
У наставку ћемо представити широко доступну некомерцијалну верзију ове платформе коју је прилагодио JePTex, која и поред ограниченог броја алата нуди мноштво могућности за претраживање и изучавање (паралелизованих) електронских корпуса ученичке продукције на матерњем и енглеском језику.

3.2 Пример претраге поткорпуса *КорССАнг*

У првом кораку (слика 1) бирамо један од пет доступних корпуса, *КорССАнг*, кога чине аргументативни састави на српском језику чији су аутори србофони ученици енглеског.

Тиме приступамо контролној табли (слика 2) која нуди неколико алата, и то: конкорданце (енг. *Concordance*), чијим одабиром је могуће вршити претрагу употребе неког термина у одговарајућем контексту, затим листе ријечи (енг. *Wordlist*), којим се термини претражују према учесталости, као и трећи алат – кључне ријечи (енг. *Keywords*) који може бити од користи у лексикографским истраживањима, при издвајању термина и сл.

У овом корисничком сучељу могуће је провјерити и метаподатке о одабраном корпусу (енг. *Corpus Info*) (слика 3).



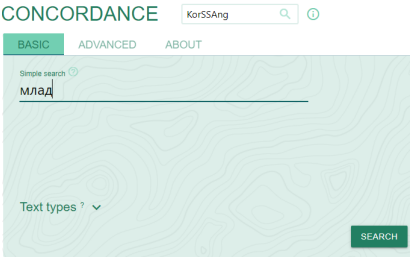
Слика 3. Метаподаци о корпусу *KorSSAng*

Овдје су доступни општи подаци о корпусу који садрже његов опис,[7] скуп ознака којима је етикетиран,[8] те број токена, ријечи, лема и докумената које га чине, а одабиром опције Manage corpus (горњи десни угао) могуће је модификовати корпус у смислу проширивања, нпр. додавањем елемената постојећем корпусу (енг. *Make bigger → Add texts*

7. *KorSSang*

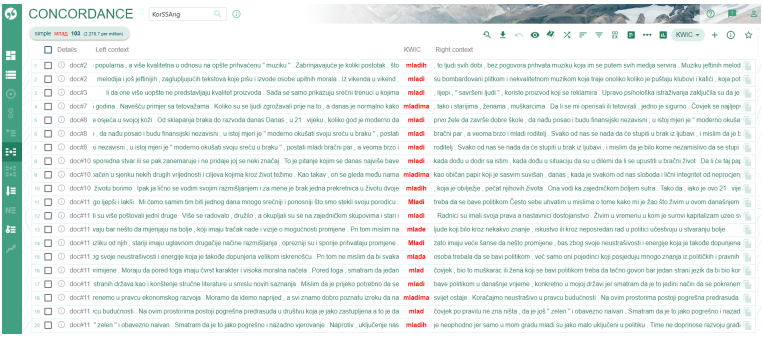
8. 1. N (именица, енг. *Noun*), 2. A (придјев, енг. *Adjective*), 3. V (глагол, енг. *Verb*), 4. PRO (замјеница, енг. *Pronoun*), 5. NUM (број, енг. *Number*), 6. PREP (приједлог, енг. *Preposition*), 7. CONJ (везник, енг. *Conjunction*), 8. INT (узвик, енг. *Interjection*), 9. PAR (партикула, енг. *Particle*), 10. ADV (прилог, енг. *Adverb*), 11. PREF (префикс, енг. *Prefix*), 12. ABB (скраћеница, енг. *Abbreviation*), 13. RN (римски број, енг. *Roman numeral*), 14. PUNCT (интерпункцијски знак, енг. *Punctuation*), 15. SENT (ознака за крај реченице, енг. *Sentence end marker*), 16. ? (ријечи које нису српског поријекла, или суфикси у сложеницама, енг. *Non-Serbian words or suffixes in compounds*).

to corpus), или формирањем и измјеном поткорпуса (енг. *Subcorpora* → *Manage subcorpora*) и сл.



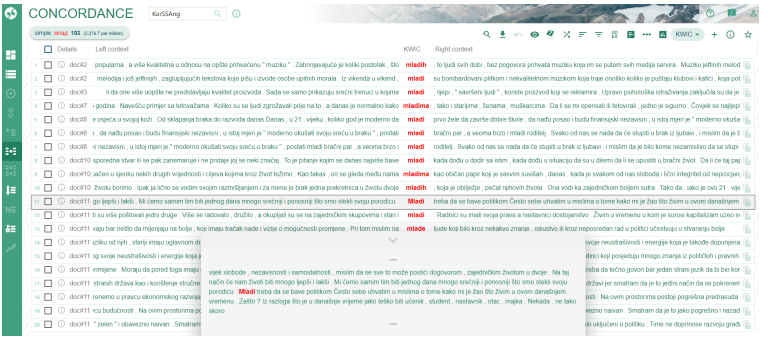
Слика 4. Основна претрага леме „млад“

Даље, одабиром менија за претрагу конкорданци, у основном типу претраге (енг. *Basic*), користећи ћирилично писмо, вршимо упит за лему „млад“ (слика 4), пошто су аутори састава припадници млађе популације, па се на датом корпусу може очекивати значајна дистрибуција овог термина.



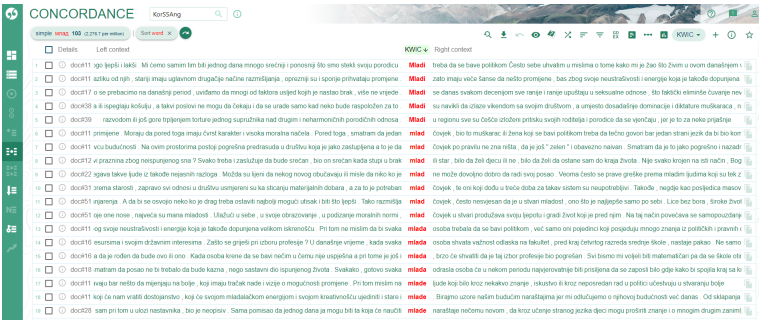
Слика 5. Резултати основне претраге за лему „млад“

Претрагом добијамо 103 резултата које чине облици придјева „млад“, као и облици поименичених придјева „млади“, поређаних редослиједом којим се јављају у нумерисаним документима који чине поткорпус КорССАнг (слика 5).



Слика 6. Шири контекст употребе одабране леме „млад“

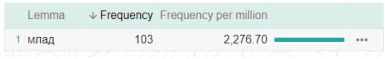
Одабиром неког од издвојених резултата добија се шири контекст у коме се дата лема јавља (слика 6).



Слика 7. KWIC леме „млад“.

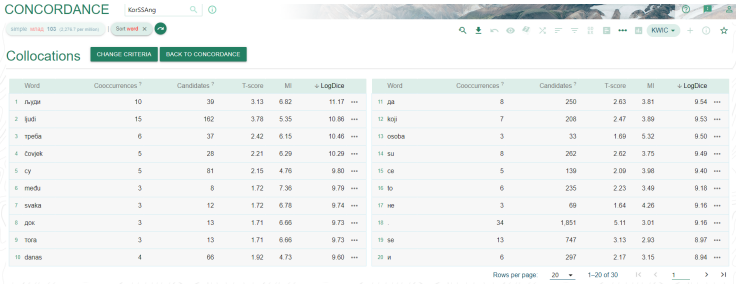
Добијене резултате претраге могуће је поређати према лијевом и десном контексту, или према кључним ријечима у контексту (енг. *Key Word in Context - KWIC*), који је и најуобичајенији формат за приказивање низова конкорданци (слика 7), те их је могуће сачувати у неколико формата, и то: *TXT*, *CSV*, *XLS*, *XML*, а тренутни приказ могуће је сачувати и као *PDF*.

Платформа омогућава провјеру учесталости (слика 8) и дистрибуције тражене леме у читавом корпусу, као и термина са којима се најчешће јавља у колокацији (слика 9) на растојању од 1-5 мјеста са лијеве и



Слика 8. Учесталост леме „млад“ у корпусу *КорССАнг*

десне стране тражене леме (у овом случају је то именица људи,[9] у 25 примјера).



Слика 9. Термини са којима се лема „млад“ најчешће јавља у колокацији у корпусу *КорССАнг*

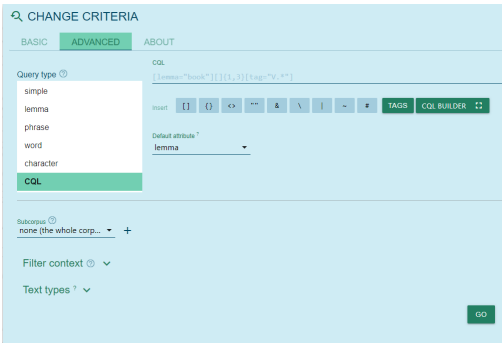
У наставку вршимо напредну претрагу[10] помоћу *CQL* упита (енг. *Context Query Language*), а приступамо јој преко менија приказаног на слици 4, одабиром опције *Advanced* (слика 10).

За потребе овог приказа користимо алат *CQL Builder*, интегрисан на платформи *Sketch Engine* (слика 11), чија сврха је генерисање упита специфичне синтаксе за напредно претраживање корпуса, помоћу скупа ознака из фусноте 8. У нашем случају, тражићемо глаголе који се јављају на растојању од 3 мјеста десно од леме „млад“.

Претрага по овом упиту даје 94 резултата (слика 12), али је неопходно мануелно издвојити јединствене примјере, како би се искључили они који

9. Поређења ради, претрага синтагме „млади људи“ (енг. *young people*) у упоредивом корпусу *LOCNESS* (тј. његовој америчкој компоненти, коју чини приближно 150 хиљада ријечи) даје 12 резултата.

10. Поред претраге *CQL* упитима, напредну претрагу могуће је вршити и помоћу „једноставног“ облика ријечи (енг. *simple*), гдје се претражују сви облици те ријечи који не садрже велико (почетно) слово, затим претрага лема, фраза, ријечи (гдје је свеједно садрже ли велика слова или не), те претрага посебних карактера, попут интерпункцијских знака, бројева, итд.



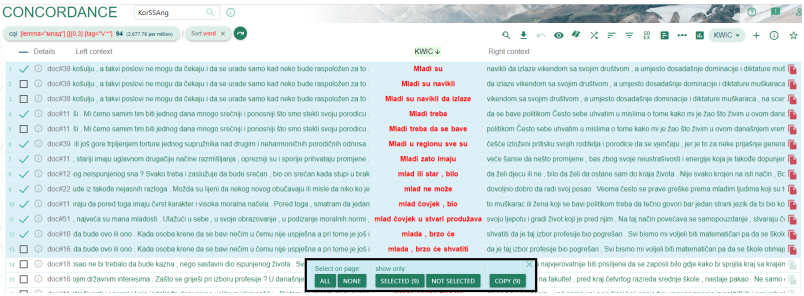
Слика 10. Одабир типа напредне претраге

Слика 11. Креирање CQL упита `[lemma="млад"] [ ]0,3[tag="V.*"]`

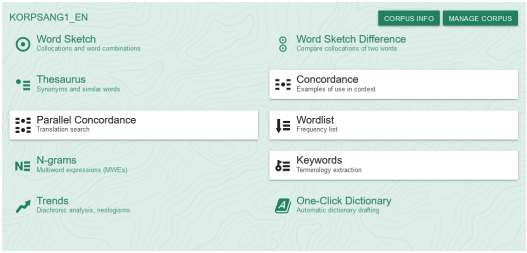
се, због одабира специфичног растојања од 3 мјеста на коме се глагол налази, понављају неколико пута.

У наставу ћемо представити претрагу паралелизованих корпуса *КорПСАнг1_ен* и *КорПСАнг1_ср*, односно *КорПСАнг2_ен* и *КорПСАнг2_ср*. Одабиру ових корпуса такође приступамо преко менија на слици 1.

Најприје бирамо корпус *КорПСАнг1_ен* и приступамо контролној табли (слика 13), која се, самим тиме што је овај корпус паралелизован са корпусом *КорПСАнг1_ср*, од контролне табле на слици 2 разликује по томе што је доступан и алат за претраживање паралелизованих конкорданци (енг. *Parallel Concordance*).



Слика 12. Резултати претраге по CQL упиту `[lemma="млад"] [ ]0,3[tag="V.*"]`



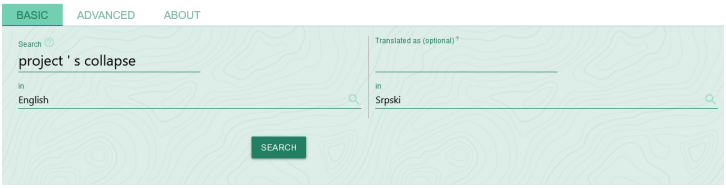
Слика 13. Контролна табла за паралелизовани корпус

Затим бирамо алат за претраживање паралелизованих конкорданци и вршимо основну претрагу именичке синтагме *project's collapse* (која је дио веће релативне клаузе – *which then led to the project's collapse*),[11] присутне у једном од изворних текстова на енглеском језику, како бисмо провјерили на који начин је преведена на српски језик (слика 14).

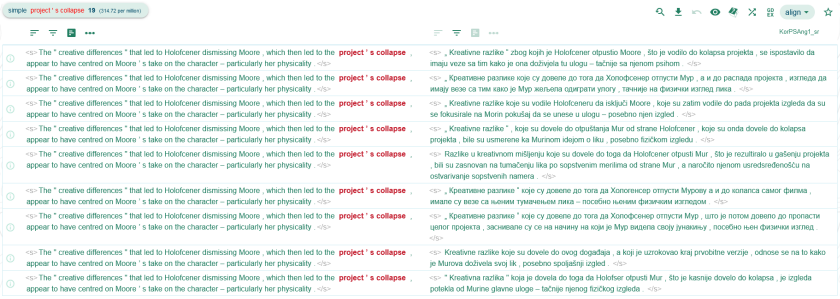
Претрага синтагме *project's collapse* у паралелизованом корпусу даје 19 резултата (слика 15).

У детаљном прегледу добијених резултата може се видјети да је тражена синтагма на енглески језик преведена на неколико различитих начина, претежно именичким синтагмама: *колапс пројекта* (4x), *пропаст (целог) пројекта* (4x), *распад пројекта* (3x), *пад пројекта*, *гашење пројекта*, *колапс (самог) филма*, *колапс*, *крај првобитне верзије*, *крах пројекта*, *пропадање пројекта*, те се, у једном случају, у преводу

11. Напредну претрагу ове синтагме могуће је извршити помоћу сљедећег CQL упита: `[lemma="project"][word="'"][word="s"][lemma="collapse"]`.



Слика 14. Основна претрага синтагме *project's collapse*



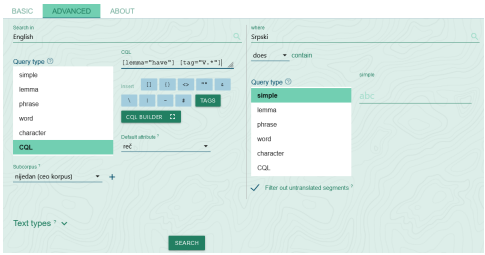
Слика 15. Резултати основне претраге синтагме *project's collapse*

налази у оквиру клаузе, *које су довеле до тога [...] да цео пројекат пропадне*.

Како сложени глаголски облици неријетко представљају изазов у учењу страних језика, у наставку вршимо напредну претрагу перфекатских облика, и то помоћу *CQL* упита[12] `[lemma="have"] [tag="V.*"]` (слика 16).

Претрага по овом упиту даје 432 резултата (слика 17), међу којима су облици прошлог перфекта, садашњег перфекта, перфекатског инфинитива и партиципа.

12. За потребе овог *CQL* упита користимо ознаке којима су етикетиране компоненте корпуса на енглеском, Penn Treebank Tagset.

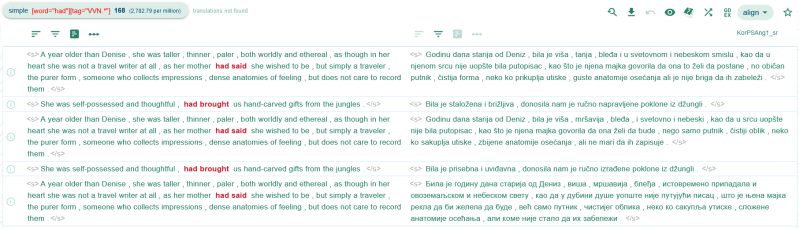


Слика 16. Напредна ретрага перфекатских облика CQL упитом `[lemma="have"] [tag="V.*"]`

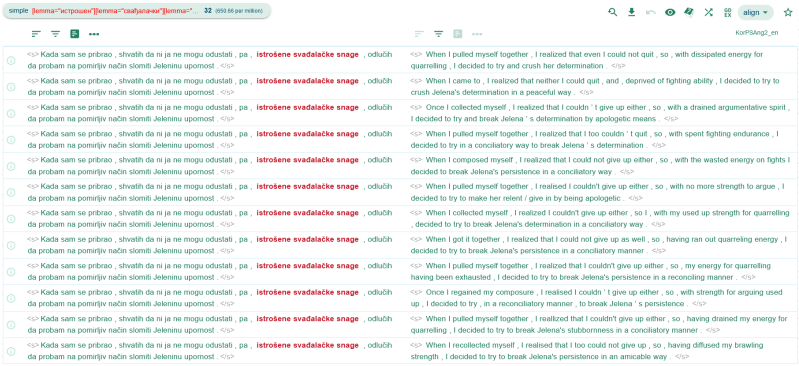
Слика 17. Резултати претраге по CQL упиту `[lemma="have"] [tag="V.*"]`

Претрагу је могуће и додатно сузити помоћу низа финијих ознака,[13] те, уколико нпр. желимо да је ограничимо на облике прошлог перфекта, можемо то учинити *CQL* упитом `[word="had"][tag="VVN.*"]`, гдје је *VVN* ознака за прошли партицип главног глагола. Претрага по овом упиту даје 168 резултата (слика 18).

13. *VB* → основни облик глагола *BE* (*be*), *VBD* → прошли облик глагола *BE* (*was*, *were*), *VBG* → герунд/садашњи партицип глагола *BE* (*being*), *VBN* → прошли партицип глагола *BE* (*been*), *VBP* → садашњи облик глагола *BE*, изузев трећег лица једнине (*am*, *are*), *VBZ* → треће лице презента једнине глагола *BE* (*is*), *VH* → основни облик глагола *HAVE* (*have*), *VHD* → прошли облик глагола *HAVE* (*had*), *VHG* → герунд/садашњи партицип глагола *HAVE* (*having*), *VHN* → прошли партицип глагола *HAVE* (*had*), *VHP* → садашњи облик глагола *HAVE*, изузев трећег лица једнине (*has*), *VHZ* → треће лице презента једнине глагола *HAVE* (*has*), *VV* → основни облик глагола, *VVD* → прошли облик глагола, *VVG* → герунд/садашњи партицип глагола, *VVN* → прошли партицип глагола, *VVP* → садашњи облик глагола, изузев трећег лица једнине, *VVZ* → треће лице презента једнине глагола.



Слика 18. Резултати претраге облика прошлог перфекта по CQL упиту `[word="have"] [tag="VVN.*"]`



Слика 19. Резултати основне претраге синтагме *istrošena svađalačka snaga*

На крају, преко менија на слици 1 бирамо корпус *КорПСАнг2_ср* и помоћу *CQL* упита `[lemma="истрошен"] [lemma="свађалачки"][lemma="снага"]` вршимо напредну претрагу синтагме *istrošene svađalačke snage*, да бисмо провјерили како је преведена на енглески језик (слика 19).

Претрага по овом упиту даје 32 резултата, од којих наводимо неколико, различите сложености: *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*, итд.

4. Закључак

У овом раду смо укратко описали кораке настанка Корпуса студената англистике (*КорСАнг*), и представили могућности његове претраге. Познато је да је недостатак нарочито ученичких корпуса хендикеп за истраживаче који се баве примјењеном лингвистиком, те се надамо да ће наши напори да тај проблем пренебрегнемо резултирати већој популарности корпусне лингвистике као истраживачког метода на овим просторима нарочито из области англистичке лингвистике. Досадашњи резултати коришћења корпуса *КорСАнг* су представљени у радовима: (Šućur 2020; Spajić and Suknović 2019; Tomović and Stefanović 2019; Марковић and Станковић, у штампи), а у изради је и једна докторска дисертација чији је дио корпуса базиран на *КорСАнгу*. Научни потенцијал овог корпуса може се очекивати у мноштву научних и стручних радова, монографија и научно-истраживачких пројеката. Додаћемо и то да је у плану израда друге верзије корпуса КорСАнг, са проширеним бројем састава и превода, и са могућношћу интегрисања нових софтверских алата који ће пружити лакшу претрагу и угоднију употребу корпуса корисницима.

Захвалност

Овај рад је заснован на истраживањима која су спроведена у оквиру два национална пројекта: *Научни потенцијали анотираних ученичких корпуса у примијењеној лингвистици*, 19.032/961-135/19, и *Фразеолошка компетенција српских говорника енглеског кроз призму контрастивне анализе међујезика*, 19/6-020/961-46/18. Пројекти су суфинансирани од стране *Министарства за научнотехнолошки развој, високо образовање и информационо друштво*, Бања Лука. Захваљујемо *Друштву за језичке ресурсе и технологије – JePTex*, а нарочито проф. Ранки Станковић, на сарадњи на пројекту 19.032/961-135/19. Посебно захваљујемо координатору пројекта, проф. Јелени Марковић, што нам је пружила прилику да учествујемо у поменутим пројектима, као и за сталну подршку у виду охрабрења, савјета, мотивације и стручности.

Литература

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 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.