811.163.41:811.135.1]’322
https://doi.org/10.18485/cllc_romana60.2023.ch23

Саша Марјановић[1]
Универзитет у Београду
Филолошки факултет
https://orcid.org/0000-0002-6551-9538

Душица Терзић[2]
Универзитет у Београду
Филолошки факултет
https://orcid.org/0000-0001-6092-0819

ПАРАЛЕЛНИ КОРПУСИ С РУМУНСКИМ И СРПСКИМ ЈЕЗИКОМ: ПОСТОЈЕЋЕ СТАЊЕ И ДАЉЕ ПЕРСПЕКТИВЕ

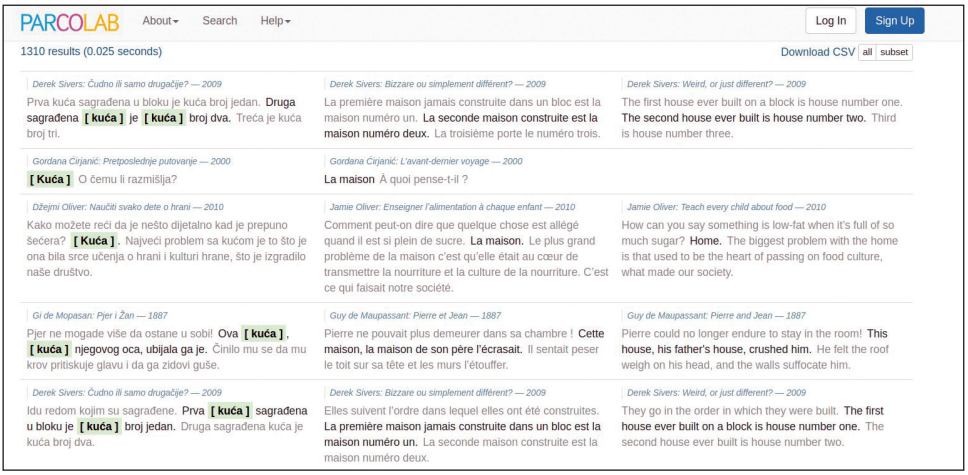
У раду се представљају постојећи електронски паралелни корпуси који садрже међусобно упарене текстове на румунскоме и српском језику. Забележени корпуси анализирају се према низу чинилаца: прво, с обзиром на то на који су начин и у којој мери доступни, те како им се може приступити; друго, имајући у виду изворни језик њихових текстова; треће, у погледу тога од којих су збирки текстова састављени, тога колико су ти текстови доменски разноврсни, каква је њихова уравнотеженост, као и у погледу тога који је обим појавница њима обухваћен; четврто, сагледава се јесу ли и на којим нивоима (лематском, морфосинтактичком, синтактичком и др.) појавнице у корпусима обележене; пето, показује се садрже ли забележени корпуси корисничко сучеље за претраживање и, ако садрже, које су могућности за претраживање текстова подржане. Циљ је рада да се представе предности и ограничења постојећих паралелних корпуса с румунским и српским језиком, како би се издвојиле смернице за разматрање у перспективи евентуалне израде нових румунско-српских обосмерних паралелних корпуса.

Кључне речи: паралелни корпус, румунски, српски, рачунарска обрада језика.

[1] sasa.marjanovic@fil.bg.ac.rs
[2] dusica.terzic@fil.bg.ac.rs

1. Увод

Терминолошки проблем који се тиче појма *паралелни* или *упоредни* корпус постоји откако се о вишејезичким корпусима говори у литератури (в. Xiao, 2010, стр. 159–160; уп. и Granger, 1996, стр. 37–38; Johansson, 2007, стр. 9). У овоме се раду паралелним корпусима називају претраживе електронске збирке текстова на одређеном језику који су упарени с преводима на једноме страном језику или више њих (Schmidt, Wörner, 2012, стр. xi). Ти изворни, али и преведени текстови најчешће се деле на мање сегменте (нпр. поглавља, пасусе, реченице, речи итд.), а исти се сегменти на разним језицима кроз процес аутоматског упаривања међусобно повезују помоћу разних рачунарских метода (нпр. Brown et al., 1991). Резултате аутоматског упаривања повезаних сегмената неопходно је потом ручно прегледати како би се осигурала тачност њихове упарености. Као пример овако дефинисаног паралелног корпуса, у приказу (1) налази се извод из резултата претраге речи *кућа* у вишејезичноме паралелном корпусу *ParCoLab* (Miletic et al., 2017).



Приказ 1. Извод из резултата претраге паралелног корпуса *ParCoLab* преко корпуснога сучеља.

Изворни се текстови још од открића камена у Розети пореде с одговарајућим преводима у сврхе лингвистичких истраживања, како у разним компаративним, тако и традуктолошким студијама (Johansson, 2007, стр. 4). Компаративна анализа као метод истраживања доживела је бурне критике средином прошлога века, па је брзо потиснута. Развој рачунара омогућио је, међутим, да се реше проблеми на које су критичари компаративне анализе дуго указивали. Израда *електронских* паралелних корпуса омогућила је

да се при истраживању ради на изузетно великом броју језичких података из разних врста текстова различитих аутора. Што су текстови бројнији и разноврснији, то је већи број и преводилаца чија ће билингвална интуиција играти улогу у истраживању. Самим тим, расту валидност и поузданост поређења (Johansson, 2007, стр. 5). Затим, предност паралелних корпуса лежи и у томе што се аутентичност података не може доводити у сумњу (Granger, 1996, стр. 38). Најзад, рачунарско доба омогућило је да се текстови у паралелним корпусима прикупљају систематски, као и то да се могу претраживати и анализирати помоћу рачунара (Johansson, 2007, стр. 5).

Употребна је вредност паралелних корпуса вишеструка. Први су електронски паралелни корпуси настајали како би омогућили развој стројног превођења (в. Brown et al., 1988), али је распон примена с временом порастао. Данас се они, у зависности од тога како су осмишљени и коме су намењени, могу користити за разноврсне лингвистичке компаративне и традуктолошке студије, затим у рачунарској обради језика, као и у низу језички оријентисаних струка, као што су превођење, лексикографија и глотодидактика. Нажалост, за језике као што су румунски и српски паралелни корпуси међусобно упарених текстова нису бројни. Опис постојећих паралелних корпуса који садрже упарене текстове на српскоме и румунском језику (§2) дајемо после овога уводног одељка (§1). Томе опису, у склопу трећег одељка (§3), следи збирни упоредни преглед свих корпуса по анализираним одликама. Потом се, у наредном одељку (§4), разматрају могућности даље израде румунско-српских обосмерних паралелних корпуса. Закључна се разматрања износе у последњем одељку (§5).

2. Опис постојећих корпуса

Досад, колико нам је познато, није било пројеката с циљем да се *директно* изради *посебан* двојезични једносмерни или двосмерни паралелни корпус који би укључио упарене текстове преведене с румунскога на српски језик, односно са српскога на румунски[3]. Текстови на румунскоме и српском језику, међутим, ипак улазе у састав десетине корпуса, али вишејезичних. Они су у тим корпусима, по правилу, аутоматски упарени с другим језиком, који се у конкретном случају сматра средишњим. Већином је тај средишњи језик енглески. Стога постојање румунско-српских упарених текстова, као и могућност њихове примене представља само додатну вредност тих корпуса, а ни у којем случају циљ сам по себи. Поред тога, постоје корпуси текстова на румунскоме који су упарени с хрватским и бошњачким преводом. Ти корпуси

[3] У даљем тексту поредак језика у предметноме језичком пару (*румунско-српски* и *српско-румунски*) варира искључиво из стилских разлога, па варијације у поретку не треба, уколико није посебно истакнуто, схватити као разлике у преводноме смеру.

такође чине предмет овога рада, будући да је језик у њима, као и језик у претходнима, само једна од варијанти полицентричног језика који се говори на подручју Босне и Херцеговине, Србије, Хрватске те Црне Горе. Сви се они у даљим редовима сматрају предметним корпусима, односно корпусима с предметним језицима. Постојећи корпуси представљају се сви у наредним тачкама овога одељка. У прегледу постојећих корпуса засебно се описује сваки забележени корпус. Методологија описа темељи се на низу одлика које су већ издвојене у претходној компаративно-контрастивној анализи неколико паралелних корпуса с паралелним корпусом *InterCorp* (Rosen, 2016, стр. 32–33; уп. §2.2). У томе су истраживању истакнуте следеће одлике анализираних корпуса: типови текстова према домену и број језика који је корпусом обухваћен, обим корпуса, обележеност појавница у текстовима, сегменти на којем су корпусни текстови упарени, удео ручног рада у изради корпуса, постојање корисничког сучеља, могућност слободног преузимања упарених текстова и укљученост метаподатака о текстовима, пре свега о изворном језику. Имајући у виду побројане одлике, забележени корпуси сагледавају се из више углова: прво, с обзиром на то на који су начин и у којој мери доступни, те како им се може приступити; друго, имајући у виду изворни језик њихових текстова; треће, у погледу тога од којих су збирки текстова састављени, тога колико су ти текстови доменски разноврсни, каква је њихова уравнотеженост, као и у погледу тога који је обим појавница њима обухваћен; четврто, сагледава се јесу ли и на којим нивоима (лематском, морфосинтактичком, синтактичком и др.) појавнице у корпусима обележене; пето и коначно, показује се садрже ли забележени корпуси корисничко сучеље за претраживање и, ако садрже, које су могућности за претраживање текстова подржане.

2.1. Први електронски корпус у којем су упарени румунски и српски језик настао је у оквиру пројекта *MULTEXT-East*, као и каснијих пројеката истраживачких група окупљених око истога руководиоца (Erjavec, 2012). Реч је о корпусу који укључује једино текст романа *1984* Џ. Орвела у енглеском изворнику и у преводу на румунски као романски језик, те на низу словенских језика (пољскоме, чешком, словачком, словеначком, српском, македонском, бугарском и рускоме), као и на персијскоме, естонском и мађарском језику. Корпус је веома малога обима. Састоји се од око сто хиљада појавница у енглескоме и свим осталим језицима. Текстови на свим језицима аутоматски су сегментирани до нивоа реченица, структурирани и упарени, а упареност је пажљиво ручно проверена. Појавнице су у текстовима лематизоване и обележене морфосинтактичким ознакама. Лематско и морфосинтактичко обележавање ручно је проверено како би тачност била у потпуности

засигурана. Корпус се налази у отвореном приступу и може се слободно преузимати у виду сирових датотека[4], али су датотеке с различитим језицима раздвојене. Не постоји, међутим, платформа путем које овај корпус могу претраживати просечно филолошки образовани корнисници. Премда се овај корпус малога обима може — теоријски гледано — примењивати у филолошким истраживањима, пројекти у склопу којих је настао тежили су развоју ресурса који би омогућили помак у рачунарској обради корпусних језика. Наиме, тежило се томе да се за све корпусне језике направе унифицирани и стандардизовани ресурси за морфосинтактичко обележавање, као и морфосинтактички лексикони у стројно читљивом формату. У вези с тим, важно је напоменути да су ови *MULTEXT-East* стандарди постали основом у рачунарском обележавању текстова управо на румунскоме и српском језику.

2.2. Други корпус од значаја за предметни пар језика представља вишејезични паралелни корпус *InterCorp*, који се од 2005. године развија на Карловом универзитету у Прагу (Чешка) (Čermák, Rosen, 2012; Rosen, 2016; Čermák, 2019). У изради корпуса, која и даље траје и која се спроводи кроз низ пројеката што су дуго подупирали Влада и Министарство просвете Чешке Републике, суделују како истраживачи у области рачунарске обраде језика, тако и факултетски наставници и студенти страних језика који су укључени у корпус. Циљ је овог амбициозног пројекта да се изради ресурс који би истраживачима и студентима тога факултета, а онда и шире, омогућио бројне компаративно-контрастивне примене у области филолошких наука. Од 2008. годишње се објави по једна нова верзија корпуса. У првој верзији корпус су чинили текстови упарени на двадесетак језика, али је број језика отад порастао: почетком наредне декаде у корпусу је било 30 језика (Čermák, Rosen, 2012), крајем декаде 40 (Čermák, 2019), док је најновијом верзијом из октобра 2023. обухваћено преко 60 језика. Кад су у питању предметни језици овога рада, румунски, те одвојено српски и хрватски обухваћени су откако је пројекат зачет, док је у најновијој верзији унет и бошњачки. Српски и хрватски текстови аутоматски су од верзије из 2016. године лематизовани и обележени морфосинтактичким ознакама, док за румунске текстове, који су према најновијим верзијама такође лематизовани и морфосинтактички обележени, нема података о години кад су анотације унете. Штавише, верзија из 2021. године садржала је за све ове језике и идентичне синтактичке ознаке према стандарду *Universal Dependencies*[5]. Што се приступа тиче, корпусни се текстови не могу преузимати, али се путем графички једноставног мрежног

[4] https://nl.ijs.si/ME. [Последњи је приступ свим повезницама у раду 15. 10. 2023].
[5] https://universaldependencies.org

корисничког сучеља *KonText*[6] могу претраживати према последњој, али и према претходним верзијама. Могуће је спроводити једноставне (према речи или леми), али и — зависно од тога како су текстови у доступним верзијама лематски, морфосинтактички и синтактички обележени — бројне сложене упите за претрагу корпуса, нарочито путем стандардизованог метајезика за корпусну претрагу (тзв. *CQL*). Као резултати упоредних претрага излазе увек најмање два језика, у зависности од броја одабраних језика.

Што се тиче састава корпуса, корпус се састоји од тзв. језгра и низа збирки. Језгро је главни део корпуса који у оквиру пројекта настаје интерно. Њега чине књижевна дела у изворнику објављена после 1945. како би језик корпуса био у највећој мери савремен. Средишњи је језик чешки, што значи да се сви текстови на другим језицима упарују с чешким. При томе, сви текстови нису упарени на свим језицима, пошто могућност упаривања језика увелико зависи од доступности текстова и степена ангажованости тима за поједине језике. Истовремено, чешки није увек језик изворника, већ изворници могу бити на било којем пројектом обухваћеном језику. Штавише, неки изворници нису ни укључени у корпус, иако су њихови преводи упарени на више језика, па тако и на чешкоме. Сви текстови који се припремају за језгро пролазе кроз процес дигитализације, којем следи фаза исправки грешака насталих у томе процесу. За то се ангажују сарадници на пројекту. Потом се текстови аутоматски упарују, а упаривање пре објављивања упарених текстова такође прегледају сарадници за поједине језике. Тиме се обезбеђује висок квалитет упарених сегмената, а последично и квалитет целога корпуса. Уз упарене текстове увек се детаљно похрањују њихови метаподаци. То при претраживању омогућава ограничавање упита према одговарајућим текстуалним параметрима. Кад је реч о обиму тектова за предметне језике, тренутно[7] је у језгру 45 текстова на румунском језику, од којих је 5 изворних; такође, језгро садржи 148 српских текстова, од којих је 13 изворно написано на српскоме језику, док хрватских текстова укупно има 324, од којих је 37 изворника. Међутим, румунско-српски поткорпус упарених текстова садржи у целини око 5,7 милиона појавница, од којих је само један текст изворно на српскоме (118 хиљада појавница), док изворних текстова на румунскоме нема. Даље, румунско-хрватски паралелни поткорпус садржи укупно око 4,8 милиона појавница, при чему је 112 хиљада појавница из једнога хрватског изворника, док румунских изворних текстова ни у њему нема.

С друге стране, поменуте збирке омогућују да се књижевна текстуална база из језгра *InterCorp*-а што је могуће више диверсификује. Оне представљају екстерно развијене ресурсе које се налазе у слободном

[6] https://www.korpus.cz/kontext/corpora/corplist

[7] https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze16

приступу путем портала пројекта *OPUS* (Tiedemann, 2012). Ови се екстерни ресурси интегришу с језгром и могуће их је претраживати истовремено с текстовима из језгра. Те збирке чине упарени вишејезични текстови преузети с новинских портала *Project Syndicate*[8] и *VoxEurop*[9], текстови правних тековина Европске уније упарени на 22 језика, тзв. *Acquis* (в. Steinbeger et al., 2014), записници Европског парламента на 21 језику, тзв. *Europarl* (в. Koehn, 2005), преводи Библије, као и паралелни корпус *OpenSubtitles2016* (Lison, Tiedemann, 2016, в. §2.5), захваљујући којем је број језика у последњој верзији увећан за 20. Премда у састав текстова на румунскоме језику улази и већина ових збирки, српски, хрватски и бошњачки текстови садрже само преводе Библије и филмске титлове, па се у перспективи претраживања упарених текстова на предметном пару језика могу добити резултати из последњих збирки текстова. Без обзира на то што су и текстови из збирки лематизовани и обележени као текстови из језгра, упареност није проверена и вероватноћа је грешака већа. Услед грешака у упареним сегментима, али и услед недостатака бројних метаподатака (нпр. податак о изворноме језику и сл.), ове збирке имају секундарни значај у односу на језгро.

2.3. Упарени текстови на српскоме и румунском језику део су и паралелног корпуса *SETimes* (Tyers, Alperen, 2010). Корпус је настао аутоматским прикупљањем новинских текстова с америчког мрежног портала *Southeast European Times*. Овај новински портал, који је деловао од 2002. до 2015. године, био је усмерен на вишејезичне вести на енглескоме с једне стране и балканским језицима с друге. Корпус стога укључује, поред текстова на предметним језицима, и текстове на енглескоме, као и на другим балканским језицима (албанскоме, бугарском, грчком, македонском, турском, али и на хрватском и бошњачком). При томе, сваки корпусни текст не садржи нужно преводе на свим корпусним језицима. Текстови су аутоматски сегментирани на реченице и аутоматски су упарени. Квалитет упарених сегмената за сваки пар језика проверен је ручно на узорку од хиљаду упарених реченица. Корпус је у потпуности у слободном приступу и могуће је датотеке с паровима језика у дорађеној верзији преузимати с одговарајућих страница[10]. То истовремено значи да не постоји платформа за претраживање овога корпуса за филолошке потребе. Корпус је, у ствари, првенствено састављен у рачунарске сврхе с циљем да се поспеши израда система за међусобно стројно превођење балканских језика с једнога на други, али и за обосмерно превођење балканских језика на енглески. У вези

[8] https://www.project-syndicate.org
[9] https://voxeurop.eu
[10] Нпр. http://nlp.ffzg.unizg.hr/resources/corpora/setimes

с тим, аутори наглашавају да изворни језик није познат, иако претпостављају да је већина текстова на балканским језицима преведена с енглеског језика. Кад је реч о броју појавница, највећи је румуско-српски поткорпус (са 8,9 милиона појавница), следи румунско-хрватски поткорпус (8,4 милиона појавница) и на крају румунско-бошњачки (5,7 милиона појавница). Будући да одређени број текстова, али не и сви, постоји упоредо у тзв. српској, хрватској и бошњачкој верзији, значај је овога корпуса (као и претходнога, уп. §2.2) у томе што се претраживањем текста на румунскоме језику могу добити три истојезичка превода.

2.4. Следећи корпус који садржи предметне језике чини вишејезични паралелни корпус *ParaSol* (von Waldenfels, 2011). У питању је резултат научноистраживачке сарадње Универзитета у Берну (Швајцарска) и Универзитета у Регенсбургу (Немачка). Корпус се темељи на старијој верзији овога ресурса, тзв. *RPC* (von Waldenfels, 2006). Та старија верзија садржала је међусобно упарене књижевне текстове на немачкоме језику, енглеском и на низу словенских језика. Она је проширена текстовима на већем броју језика, како германским, романским, балтичким, тако и другим европским језицима, али и на есперанту. Тиме укупан број језика премашује 30. Истовремено с проширењем, корпус је променио име. Суштина пројекта израде овога корпуса огледа се у изради ресурса који би најпре филолошки оријентисаним истраживачима и студентима омогућио компаративно-контрастивно поређење језика. Стога је приоритет у изради овога корпуса у томе да се укључе преводи једнога текста на што је могуће више језика. Сходно томе, број текстова на истоме језику у другоме је плану. По процесу дигитализације, сви се текстови аутоматски сегментирају на реченице, потом се врши аутоматско лематско обележавање и упаривање. Упареност се ручно не проверава, нити се грешке у упарењима посебно исправљају. Процес упаривања прати и аутоматско обележавање текстова ознакама за врсте речи. Поред румунскога, у корпус су посебно укључени српски и хрватски. Но, премда текстова на српскоме одн. хрватском има више, чак и изворника, с румунским су упарена само два романа: Булгаковљево дело *Мајстор и Маргарита*, чији је изворник руски, те роман Џ. К. Роулинг *Камен мудрости* из серијала о Харију Потеру, који је изворно написан на енглеском језику. То значи да су и српски односно хрватски и румунски у перспективи предметног језичког пара заступљени искључиво као преводни језици. У складу с тиме што је корпус намењен филолозима, доступан је за слободно претраживање путем сучеља на одговарајућој платформи[11]. Претраживање није интуитивно јер се упити за претраживање текстова формулишу

[11] https://www.parasolcorpus.org/Ursynow

искључиво путем метајезика за корпусну претрагу (тзв. *CQL*), а да то нигде на мрежним страницама није напоменуто. Коначно, истичемо да се корпусни текстови не могу преузимати.

2.5. Румунски и српски језик садржи и паралелни корпус *OpenSubtitles2016* (Lison, Tiedemann, 2016). Овај корпус, који се развијао десетак година и чија је последња верзија објављена 2016. године, састоји се од филмских и серијских титлова аутоматски упарених на преко 60 језика. Титлови су аутоматски преузети из базе репозиторијума титлова *OpenSubtitles*[12]. Репозиторијум је отвореног карактера, што значи да сваки корисник репозиторијума може слободно додавати нове преводе титлова за одговарајуће филмове и серије. Прикупљени титлови такође су аутоматски конвертовани, сегментирани до нивоа реченица и дорађени у погледу словних погрешака за све језике. Припремљеним документима додати су, опет аутоматски, метаподаци на основу података из базе филмова IMDB[13]. Временске одреднице које су саставни део титлова олакшале су и поспешиле аутоматско упаривање великог броја докумената на различитим језицима овога корпуса. Услед количине докумената који су укључени није, међутим, било сврсисходно ручно проверавати упареност сегмената, нити је било могуће исправљати грешке у њима. Но, будући да је предност овога корпуса у расположивости великог броја података и за језике с иначе недовољним бројем и квалитетом језичких ресурса, његова је примена велика. Прво, за неке од обрађених језичких парова ово је један од ретких корпуса и то корпус с не тако маленим обимом (уп. Rosen, 2016). То се јасно види на примеру предметних језика овога чланка: румунско-српски поткорпус садржи преко 24 хиљада упарених титлова, с 19,7 милиона реченица (око 276 милиона појавница); румунско-хрватски поткорпус броји 25 милиона упарених реченица из око 31 хиљаде титлова (око 345 милиона појавница); док румунско-бошњачки поткорпус чини око 12 хиљада титлова с 10,6 милиона реченица (око 148 милиона појавница). Овим бројкама *OpenSubtitles2016* истиче се као највећи паралелни корпус за предметни пар језика и, штавише, небројено по обиму појавница и количини докумената одскаче од осталих корпуса које приказујемо у овоме раду. Друго, *OpenSubtitles2016* даје приступ разноврсним говорним (односно псеудоговорним) језичким подацима (уп. Terzić et al., 2020), јер су прикупљени филмови различите тематске покривености. Ова велика количина података има значајну примену у даљем развоју система за стројно превођење између корпусних језика. Ваља, међутим, истаћи да у тој бази највероватније нема изворних

[12] https://www.opensubtitles.org
[13] https://www.imdb.com

титлова на српском језику (уп. Miletic et al., 2017; Terzić et al., 2020), а врло је вероватно да је таква ситуација и с титловима на румунскоме. То значи да су текстови на предметним језицима у овоме корпусу део преведених текстова. Ваља такође рећи да изворни језик није забележен, па тако није познат ни преводни смер прикупљених корпусних текстова. Коначно, упарени фајлови могу се слободно преузети преко портала пројекта *OPUS* (Tiedemann, 2012) и даље интегрисати у платформе за корпусно претраживање. Поред тога, овај је корпус такође већ интегрисан у друге платформе. Наиме, он је у лематски и морфосинтактички обележеној верзији доступан за отворено и слободно претраживање у оквиру платформе паралелног корпуса *InterCorp* (в. §2.2), али и путем комерцијалног система за претрагу корпуса *Sketch Engine* (Kilgarriff et al., 2004). Но, да би се приступило платформи овог последњег система, нужно је извршити претплату.

2.6. Наредни паралелни корпус, корпус катарскога Завода за рачунарство — првобитно назван *QCRI AMARA Corpus*, а потом *QCRI Educational Domain Corpus* (скраћено: *QED*) — такође је настао на интернетски расположивој грађи транскрипата говорних докумената на 225 језика (Abdelali et al., 2014). Ова је грађа аутоматски преузета с платформе *Amara*[14], која служи за титловање видео-записа и превођење њихових титлова на волонтерској основи (уп. §2.8, али и §2.5). Овде је, спрам претходнога корпуса, тј. корпуса филмских титлова, реч о транскриптима различитих снимљених образовних монолошких предавања и видео-снимака који се тичу различитих стручних домена људског деловања, као што су природне и друштвене науке, медицина, економија и сл. Прикупљени транскрипти аутоматски су обрађени, сегментирани до нивоа реченице и упарени. Упаривање су и овде олакшале временске одреднице из изворних датотека. На основу њих упарено је око 75% грађе. Преостала грађа, код које се временске одреднице међу језицима нису поклапале, упарена је поређењем реченичних дужина и упоређивањем кандидата за упаривање с двојезичним лексиконом аутоматски ексцерпираним из дела грађе чије су се временске ознаке поклапале. Румунско-српски паралелни поткорпус састоји се од 1.521 упареног транскрипта (око 6 милиона појавница); румунско-хрватски поткорпус чини 1.236 докумената (око 5,4 милиона појавница); док румунско-бошњачки броји 73 документа с око 0,3 милиона појавница. И овоме се корпусу слободно може приступити, отворен је за преузимање путем портала пројекта *OPUS* (Tiedemann, 2012) и то искључиво за потребе научноистраживачког рада. Његова превасходна функција била је да поспеши системе за стројно превођење. Он, сходно томе, није интегрисан

[14] https://amara.org

ни у једну платформу за претраживање паралелних корпуса, па се њиме просечно филолошки оријентисан корисник не може служити.

2.7. Поред корпуса из одељка §2.3 (в. горе) постоји још један паралелни корпус новинских текстова с упарењима на српскоме и румунском језику. Он се такође може преузимати с мрежног портала пројекта *OPUS* (Tiedemann, 2012; уп. Reimers, Gurevych, 2020) и није, колико нам је познато, интегрисан ни у једну пларформу за претраживање корпуса. Последња верзија која је у слободном приступу на поменутом порталу развијена је 2018. године. Корпус се састоји од 851 чланка који су преузети с новинског портала *Global Voices*[15], сегментирани и упарени у нивоу реченица на укупно 46 језика. При томе, сви чланци нису доступни за све језике обухваћене корпусом. Изворни чланци углавном представљају преузете новинске прилоге објављене на другим мрежним порталима, а преводе су израдили добровољци за потребе новинског портала *Global Voices*, при чему то по правилу нису, или барем углавном нису, професионални преводиоци. Изворни текстови написани су на разним језицима, па тако и на румунскоме и српском, али у корпусу нема података о изворном језику. Штавише, изворни текстови углавном се преводе с енглескога. По питању обима упарених текстова, румунско-српски поткорпус обухвата само 4 чланка с укупно 5,2 хиљаде појавница.

2.8. Вишејезични паралелни корпус *TED2020* (Reimers, Gurevych, 2020) састављен је од око 4.000 транскрипата ТЕД говора углавном на енглеском језику уз преводе на низ других језика, при чему је укупан број расположивих језика 108. Транскрипти који су ушли у састав овога корпуса чине део кратких мотивацијских и образовних говора који у просеку трају 20 минута. Говори покривају велику лепезу тематских домена. Они су ексцерпирани из текстуалне и видео-базе говора с одговарајућих мрежних страница које су у слободном приступу[16]. Корпус *TED2020* доступан је за преузимање путем портала пројекта *OPUS* (Tiedemann, 2012). Датотеке за румунско-српски језички пар садрже преко 220 хиљада упарених реченица из 2.179 говора (око 7,9 милиона појавница), за румунско-хрватски пар датотеке броје око 200 хиљада парова реченица (око 6,3 милиона појавница), док се за румунско-бошњачки пар датотеке састоје од око 10,7 хиљада упарених реченица (око 0,4 милиона појавница). Овај корпус настао је за потребе развоја система за стројно превођење, али с обзиром на велику лексичку разноврсност његових текстова може имати и друге, па тако и филолошке, примене. Корпус, међутим, не поседује сопствену платформу за претраживање, нити

[15] https://globalvoices.org
[16] https://www.ted.com

је, колико нам је познато, интегрисан у другу платформу за претраживање паралелних корпуса.

Поред овога корпуса, израђен је 2018. године још један паралелни корпус ТЕД говора под називом *MulTED* (Zeroual, Lakhouaja, 2022). *MulTED* укупно садржи 1.100 различитих говора на енглеском језику и такође обухвата преко стотину језика, с тим што сви говори нису доступни на свим језицима. Сви су говори аутоматски сегментирани и упарени на нивоу реченица. Поред тога, говори су аутоматски обележени универзалним ознакама за врсте речи. Спрам првога корпуса ТЕД говора, који је обимнији, предност овога другог огледа се у томе што су сви ТЕД говори у њему ручно разврстани у 11 тематских домена. Међутим, премда је у чланку и на пројекатским мрежним страницама (уп. Terzić et al., 2020) најављено да ће корпус ускоро бити доступан за преузимање, то и даље није могуће. Због тога није могуће ни известити о обиму упарених говора за предметне језике овога рада. У чланку се једино види да је укључено 1.003 говора у преводу на румунски (око 1,8 милиона појавница), 871 говор у српскоме преводу (око 1,4 милиона појавница) и 648 говора у хрватском преводу (око милион појавница), иако су, претпостављамо, укључени и бошњачки преводи.

2.9. На порталу пројекта *OPUS* (Tiedemann, 2012) индексиран је и доступан за преузимање и паралелни корпус преведених реченица на око 400 језика екстрахован из вишејезичке базе мрежних страница удружења *Tatoeba*[17] (уп. Reimers, Gurevych, 2020). Овај ресурс првенствено је настао у глотодидактичке сврхе како би корисници на мрежним страницима путем једноставног сучеља на лак начин могли претраживати кратке реченичне примере употребе речи у контексту на одговарајућем језику с преводом на други језик. При томе, преводи једне реченице на свим доступним језицима међусобно су повезани. Корисници, поред слободног претраживања, могу такође слободно уносити нове реченице на било којем језику, као и њихове преводе на жељени језик, а могу и кориговати и редиговати постојеће преведене реченице на било којем језику. Реченице које се преводе могу се изворно уносити на било којем језику, с тим што је изворни језик највећег броја реченица енглески. Корпус се може преузимати у виду упарених датотека за парове језика. Премда на страницама овог ресурса постоје упарене реченице на српскоме, хрватском и бошњачком с једне стране, и румунскоме с друге стране, на порталу пројекта *OPUS* могу се преузети једино датотеке за румунско-хрватски пар. Те упарене датотеке садрже, међутим, тек један пар реченица с укупно 8 појавница.

[17] https://tatoeba.org

2.10. Најновији корпус који укључује предметне језике представља вишејезични паралелни корпус *RomCro* (Bikić-Carić i dr., 2023). Реч је о ресурсу који је настао у склопу факултетског романистичког пројекта на Филозофском факултету Свеучилишта у Загребу. Пројекат је започет 2019. године и трајао је три године. Мада је почетни пројекат завршен, назначена је могућност да ће се дорада и допуна ресурса наставити у оквиру новога пројекта. Циљ пројекта био је да се изради паралелни корпус који с једне стране укључује хрватски као средишни, а с друге стране пет романских језика, међу којима су шпански, француски, италијански, португалски и румунски. Сви текстови сегментирани су до нивоа реченице и међусобно су упарени. У састав корпуса ушла су искључиво књижевна дела како би језик у корпусу био општи, а не стручни. Тај критериј мотивисан је жељом да коначан корпус текстова има најширу могућу примену првенствено у контрастивним лингвистичким и традуктолошким истраживањима. Поред тога, стремило се ка томе да се упаре савремени текстови XXI века, али је на крају одређени број обрађених текстова у изворнику ипак настао у различитим декадама XX века. При одабиру текстова водило се и тиме да се за шпански и португалски поткорпус бирају текстови на европској варијанти тих језика, али је у крајњи резултат у португалски поткорпус укључено неколико преведених текстова на бразилској варијанти. Но, главни критериј био је доступност текстова, па је за сваки језик у ствари одабрано неколико романа који постоје у преводу на остале корпусне језике. Тако су у поткорпус с хрватским изворницима укључена три текста, а румунски поткорпус чине четири изворника. Но, 19 изворника на осталим романским језицима упарени су и с хрватским и румунским преводима, па укупан број изворних и преведених текстова који улазе у румунско-хрватски поткорпус достиже 20. Према броју речи, румунски поткорпус чини 3,1 милион појавница, а хрватски 2,8 милиона. При томе, изворни румунски текстови састоје се од 437 хиљада појавница, а хрватски од свега 175 хиљада, док су преостале појавнице део преведених текстова. Сви корпусни текстови прикупљени су у папирном или електронском формату. На њима је спроведен процес дигитализације како би се добио рачунарски читљиви документ. Овај процес подразумевао је ручни рад у одређеној мери, нарочито у погледу исправки грешака које су настале при дигитализовању папирних текстова. Дигитализовани документи аутоматски су сегментирани и упарени. Упарени текстови ручно су проверени. Проверавање је поверено студентима односних корпусних језика, након чега су пројекатски чланови извршили коначан преглед. С обзиром на то да сви учесници у овоме процесу нису имали компетенције у свим језицима, на истим текстовима радило је више учесника. Тиме је постигнут висок резултат у упаривању текстова. Упарени текстови унети су у систем за претрагу корпуса *Sketch Engine* (Kilgarriff

et al., 2004). При уносу у систем, текстови су аутоматски лематизовани и морфосинтактички обележени одговарајућим алатима који су интегрисани у систем. Корисници поменутог система, за који је приступ комерцијализован, корпус могу претраживати уз претходно одобрење администратора корпуса. Могућности претраге корпусних текстова веома су велике и корисничко сучеље прилично је разрађено. Истовремено, корпус је у слободном приступу на одговарајућој платформи доступан свим истраживачима за преузимање, али текстови ту нису ни лематизовани ни обележени. Штавише, како би се онемогућила повреда ауторских права дистрибуирањем текстова у целини, редослед реченица насумично је испреметан.

2.11. Као посебна скупина ресурса издвајају се корпуси и преводне меморије које су установе Европске уније ставиле на располагање широј јавности преко заједничког истраживачког центра. Језици у овим корпусима ограничени су на званичне језике Европске уније. За многе се језичке парове, пак, једино из ових корпуса могу извући паралелни ресурси. Иако је једна од мана објављених корпуса ограниченост текстова на правни и административни домен, њихова је предност у томе што садрже изразито велики број стручних и специфичних термина (Steinberger et al., 2014, стр. 687). За предметне језике значајна је, пре свега, преводна меморија Генералног директората Европске комисије за превођење *DGT-TM*. Она садржи текстове правних тековина Европске уније (фр. *acquis communautaire*). Грађа се састоји од свих правних регулатива усвојених у оквиру институција Европске уније преведених на 24 званична језика[18]. Првобитна верзија објављена је 2007. године када је упареност текстова ручно проверена. Све наредне верзије упарене су аутоматски узимајући енглески језик за централни. Будући да је Хрватска постала чланица Европске уније 2013. године, верзије на њеном службеном језику, који је такође предмет овога рада, објављене су након 2014. године. У изради превода учествују стручњаци с вишегодишњим искуством. При изради превода користе се савремене технологије. То увелико осигурава квалитет превода. Изворни језик, међутим, није познат мада се сматра да је већина текстова изворно написана на енглескоме језику. При преузимању паралелних текстова енглески се језик такође користи као средишни језик. Хрватски текстови садрже преко 42 милиона појавница, док румунски текстови броје преко 84 милиона[19]. Могуће је преузети базе у целости, као и одговарајући софтвер који омогућава издвајање само оних поткорпуса који су неопходни за упоредну претрагу. Корпуси су означени на лематском, морфосинтактичком и синтактичком нивоу. Они

[18]https://www.clarin.si/noske/run.cgi/corp_info?corpname=dgtud_hr&struct_attr_stats=1&subcorpora=1
[19] https://wt-public.emm4u.eu/Resources/DGT-TM_Statistics.pdf

су, дакле, прилагођени за употребу у рачунарској обради језика. С друге стране, платформа *NoSketchEngine* (Rychlý, 2007) пружа сучеље за претрагу преко 9.000 хрватских текстова из овога корпуса, који садрже 28 милиона појавница. Преко овога се сучеља могу паралелно претраживати румунско-хрватски текстови. Стога је овај корпус доступан не само корисницима који се баве рачунарском обрадом језика, већ и стручњацима из других области, па тако и просечно филолошки оријентисаним корисницима.

3. Упоредни преглед постојећих корпуса

Корпусе које смо засебно представили у претходном одељку у овоме приказујемо међусобно како бисмо добили општу слику стања паралелних корпуса за предметни пар језика. Одлике према којима су корпуси појединачно описани (уп. §2) сад збирно представљамо у таблици 1. Тиме омогућујемо упоредни преглед.

КОРПУС	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
Multext-East	књиж.	13	с	200х	л, м	+	+	не	+
InterCorp	књиж. +	40+	с, х	5,7м	л, м, (с)	+	+	+	не
SETimes	новин.	10	с, х, б	8,9м	--	не	не	не	+
Parasol	књиж.	30+	с, х	2р	л, м	+	не	+	не
OpenSubtitles2016	филм.	60+	с, х, б	276м	--	не	не	(не)	+
QED	наука	225	с, х, б	6м	--	не	не	не	+
Global Voices	новин.	46	с	5,2х	--	не	не	не	+
TED2020	говори	108	с, х, б	7,9м	--	не	не	не	+
MulTED	говори	100+	с, х, б	н	--	не	не	не	не
Tatoeba	општи	400+	с, х, б	н	--	не	(+)	+	+
RomCro	књиж.	6	х	5,9м	л, м	+	+	(не)	+
DGT-TM	правни	24	х	29,8м	л, м, с	не	+	(не)	+

Таблица 1. Упоредни приказ описаних корпуса према анализираним одликама: 1) типови текстова према домену; 2) укупан број језика; 3) присуство различитих стандарда: *с* – српски, *х* – хрватски, *б* – бошњачки; 4) обим румунско-*српскога* (односно, кад нема *српског*, румунско-хрватскога) поткорпуса: *х* – хиљаду, *м* – милион, *р* – роман, *н* – непознато; 5) обележеност појавница: *л* – лематски, *м* – морфосинтактички, *с* – синтактички; 6) метаподаци; 7) удео ручног рада; 8) постојање корисничког сучеља за претраживање; 9) могућност слободног преузимања.

Очигледно је да је вредност описаних паралелних корпуса с међусобно упареним текстовима на румунскоме и српском језику неупитна. Исто тако,

врло је добро што ти корпуси уопште постоје. Но, они истовремено имају низ ограничења. У наредном тексту међусобно упоређујемо предметне корпусе како би се јасно истакли њихове предности и недостаци.

Најпре, један од постојећих описаних корпуса (*MulTED*) уопште није доступан, иако је то од 2018. године најављено, или му се барем још увек не може приступити. С друге стране, већина је у слободном приступу, али у виду сирових фајлова који се могу преузети и претраживати у одговарајућим програмима или екстерним платформама за претраживање корпуса (нпр. *Multext-East*, *SETimes*, *OpenSubtitles2016*, *QED, Global Voices*, *TED2020*, *Tatoeba*, *RomCro* и *DGT-TM*). Незгодно је што корисник у том случају мора имати одговарајуће програмерске вештине за употребу корпуса. Стога просечно филолошки образован корисник од поменутих паралелних корпуса без корисничког сучеља нема веће користи. Једино корпуси *InterCorp*, *Parasol* и *Tatoeba* поседују платформу преко које се корпусна база бесплатно и у отвореном приступу може претраживати путем графички једноставнога корисничког сучеља. Но, док сучеље првога корпуса (*InterCorp*) омогућује низ различитих типова упита — од упита које су за просечног корисника најинтуитивнији, па до сложених упита који захтевају познавање имплементираног метајезика за задавање упита —, сучеље другога корпуса (*Parasol*) претражује се искључиво путем метајезика. То, међутим, нигде није видно истакнуто, па просечан корисник обичним уносом језичких израза у поље за претрагу добија невалидан одговор. Сучеље трећег корпуса (*Tatoeba*) залази у другу крајност: омогућени су интуитивни једноставни упити, али не и сложени. Ваља, међутим, нагласити, како би се предупредили погрешни закључци о немогућности директног претраживања већине корпуса, да су неки корпуси чак интегрисани у друге — већ постојеће — платформе за претраживање корпуса. Тако се *OpenSubtitles2016*, *RomCro* и *DGT-TM* могу путем сучеља платформе *Sketch Engine* претраживати једино уз претходну претплату, с тиме што је то за *RomCro* могуће уз додатно допуштење корпусног администратора да се корпус — на лични захтев — подели с претплаћеним корисником платформе *Sketch Engine*. Поред тога, *OpenSubtitles2016* може се претраживати путем сучеља чешког корпуса *InterCorp*, док се корпусу *DGT-TM* може приступити и путем сучеља платформе *NoSketch Engine*.

Сви ови корпуси настали су у оквиру пројеката чији је циљ био да се исти текстови упаре на већем броју језика, а не само на српскоме и румунском. Према броју језика, највећи су корпуси *Tatoeba* (преко 400 језика), *QED* (225 језика), *TED2020* и *MulTED* (оба с преко 100 језика), док су најмањи *RomCro* (6 језика), *SETimes* (10) и *Multext-East* (13). Преостали корпуси држе просек од 25–40 језика. Број језика, међутим, не корелира с обимом корпуса: *Tatoeba* садржи неупоредиво мање појавница него остали

корпуси с просечним бројем упарених језика. Мали је број појавница и у корпусима *Multext-East*, *Global Voices* и *Parasol*, пошто њихова база, у складу са замислима пројекта у оквиру којих су састављени, садржи мали број текстова на великом броју језика. Упадљиво су највећи *OpenSubtitles2016* (276 милиона појавница). Већина корпуса у просеку садржи око 6 милиона појавница (*InterCorp*, *SETimes*, *QED, TED2020* и *RomCro*). Кад је посреди корпус *DGT-TM*, који има релативно велик број појавница у румунскоме и хрватском поткорпусу, не може се са сигурношћу утврдити његов обим, јер без претходне екстракције поткорпуса није познато колико је текстова упарено између ова два језика. Прецизно одређивање обима свих корпуса омета и раздвајање српске, хрватске и бошњачке преводне верзије у већини корпуса (*InterCorp*, *SETimes*, *OpenSubtitles2016*, *QED, Global Voices* и *TED2020*). Наиме, текстови који су именовани као српски, хрватски и бошњачки немају сви верзију на свим трима стандардима. Кад би се квалитативно гледало који су, на пример, румунски текстови у чешкоме корпусу *InterCorp* упарени са српским и хрватским преводима и изворницима, могло би се лако доћи до податка да није реч о истима. С тим у вези, ако би се гледало колико је укупно различитих српских и хрватских докумената упарено с документима на румунскоме језику, добио би се већи обим него што је то сад описано (уп. §2.2). То важи за све поменуте корпусе у којима се налазе текстови на више стандарда. Такође, ваља нагласити да би без подробне анализе извора било погрешно тврдити да су одређени текстови приказани као бошњачки заиста и сви написани на бошњачком стандарду; наиме, текстови који су према корпусним спецификацијама наведени као бошњачки у корпусу *SETimes* не представљају новинске чланке на конкретном стандарду, већ чланке локализоване за подручје целе Босне и Херцеговине, у којој се службено употребљавају и српски и хрватски стандард.

Кад је реч о садржају корпусâ, видели смо у претходном одељку (§2) да су постојећи корпуси усмерени искључиво на један текстуални домен. Сходно томе, у највећем су броју описаних корпуса заступљени текстови књижевног домена (*Multext-East*, *InterCorp*, *Parasol* и *RomCro*). Остали корпуси укључују новинске текстове (*SETimes* и *Global Voices*), филмске и серијске титлове (*OpenSubtitles2016*), транскрибоване преводе низа мотивацијских говора на различите теме од друштвеног значаја (*TED2020* и *MulTED*), преводе транскрипата едукативних видео-записа и снимљених предавања из различитих научних области (*QED*), али и преведене и упарене текстове правних тековина Европске уније (*DGT-TM*). Но, с обзиром на то што база корпуса *InterCorp* садржи збирке екстерно упарених текстова (као што је збирка коју чини корпус *OpenSubtitles2016*), условно би се могло рећи да једино овај корпус није једнодоменски.

Корпусни су текстови у свим описаним ресурсима сегментирани, а сегменти су упарени до нивоа реченице. Упаривање је у свим случајевима извршено аутоматски, с тим што је, кад је реч о корпусима *Multext-East*, *InterCorp*, *Parasol* и *RomCro*, упареност ручно проверена и, по потреби, грешке исправљене. Документи који чине ове корпусе снабедевни су исцрпним метаподацима (нпр. подацима о изворном језику, аутору, језику с којега је текст преведен и сл.), па ови корпуси имају већи потенцијал за примену у областима као што је традуктологија, где су метаподаци од велике важности при изођењу закључака. Спрам претходних, корпус *DGT-TM* састављен је аутоматски, није било провере упарених сегмената, али су језици упарени с високим процентом тачности, будући да се правни текстови који улазе у састав овога корпуса преводе на тај начин да се може омогућити повезивање сегмената језичких верзија. Но, премда се већина текстова у овоме корпусу преводи с енглескога језика, не зна се ко су аутори и на којем су језику текстови изворно настали.

У вези с метаподацима, ваља истаћи и квалитет превода текстова који су ушли у сваки корпус. Будући да су у део корпуса (*Multext-East*, *InterCorp*, *SETimes*, *Parasol*, *RomCro* и *DGT-TM*) ушли како ауторизовани текстови, тако и потписани (односно, у случају последњег поменутог корпуса, подробно из правничке перспективе проверени) преводи тих текстова, квалитет је превода у њима на задовољавајућем нивоу. С друге стране, текстове у другим описаним корпусима (*QED, Global Voices*, *TED2020* и *MulTED*) првенствено преводе аматери и волонтери, па овакви корпуси могу имати мању вредност због потенцијално нижег квалитета. Даље, у корпусу *Tatoeba* преводе на одговарајући језик обезбеђују волонтери, не само изворни говорници датог језика, већ и лица којима је дотични језик нематерњи и који по струци нису нужно преводиоци. На тај начин крајње је упитан квалитет превода у томе корпусу, а последично и квалитет текста у целоме корпусу. Коначно, ауторство превода уопште није могуће утврдити у највећем описаном корпусу за предметни пар језика (*OpenSubtitles2016*). Може се чак претпоставити да је део преведених текстова у њему настао применом стројнога превођења с енглескога језика. То потенцијално може значити да су неки упарени текстови на српскоме и румунском језику у оба случаја производи стројног превођења.

Што се тиче даље рачунарске обраде текстова који су ушли у корпусе, може се констатовати да су појавнице у тек неколико испитаних корпуса (*Multext-East*, *InterCorp*, *Parasol*, *RomCro* и *DGT-TM*) лематски и морфосинтактички обележене. Обележеност појавница може се сматрати додатном вредношћу ових корпуса јер увелико побољшавају могућности претраживања корпусних текстова када су интегрисани у одговарајућу

платформу с корисничким сучељем. Додатно, треба истаћи да су само појавнице у корпусу *DGT-TM* и појавнице у једној од старијих верзија корпуса *InterCorp*, која се и даље може претраживати, такође обележене синтактичким ознакама за односе према другим појавницама у реченици.

4. Могућности развоја предметних корпуса

Нема сумње да могуће примене (§1) паралелних корпуса који укључују међусобно упарене текстове на румунскоме и српском језику (§2–3) представљају један од темељних основа за даљи развој румунистике у Републици Србији, али и околним истојезичким земљама. Исто вреди за србистику у Румунији и Молдавији. Јасно је и то да је обучавање студената у употреби овога типа корпуса стожер који би омогућио помак у сврсисходнијем оспособљавању стручног кадра у области лингвистике, преводилаштва, глотодидактике и других примењених додирних филолошких области. С обзиром на такав значај паралелних корпуса, од изузетне је билатералне језичке (али и мултилатералне државне) важности покренути пројект израде паралелног корпуса с текстовима на румунскоме и српском језику који би био у отвореном приступу и који би се слободно и бесплатно могао претраживати путем платформе с интуитивним корисничким сучељем, као што је платформа паралелног корпуса *ParCoLab* (Miletic et al., 2017; в. §1). Истини за вољу, први су кораци у томе смислу већ постигнути у оквиру корпуса *InterCorp*, пошто су у њега, поред текстова из тзв. књижевног језгра, укључене и разноврсне збирке, од којих за српски с румунским постоји велика збирка филмских титлова *OpenSubtitles2016* и збирка с библијским текстом, али не и остале доступне збирке и постојећи корпуси. Постојећи корпуси у слободном приступу могли би се сви одмах интегрисати у поменуте платформе (уз претходно решавање питања права њихове дистрибуције), као и униформно обележити одговарајућим ознакама за румунски односно српски језик на више нивоа не би ли се увећале могућности претраживања корпусних текстова. Тако би се добио вишедоменски обележени паралелни корпус с предметним паром језика.

Но, имајући у виду да већину тренутно упарених текстова у постојећим корпусима чине преведени текстови, а да је количина изворно писаних текстова на румунскоме и српском језику у тим корпусима занемарива, пожељно би било да се у нови (или само допуњени) корпус унесу текстови директно преведени с румунскога на српски и са српскога на румунски језик. Додатно, збирку тих директно преведених текстова ваљало би издвојити, попут текстуалног језгра у чешкоме корпусу *InterCorp* (в. §2.2). Стварање таквог поузданог језгра текстова упарених на румунскоме и српском погодовало би примени корпуса у областима у којима је важно утврдити

који је језик изворни, а који циљни, као и то који је преводни смер међу дотичним језицима. Како би се овај циљ испунио, неколико је решења. У првом реду, у такав би се корпус могла укључити релативно бројна дела српске књижевности с преводом на румунски језик, као и дела изворне румунске књижевности у српскоме преводу (уп., нпр., Алмажан, 2005). Такође, могуће би било у корпус унети преведена дела румунске мањинске књижевности настале на територији Републике Србије. У другоме реду, с обзиром на то да се у Републици Србији у општинама с бројном румунском националном мањином основношколска и средњошколска настава одвија на румунскоме језику, на располагању су бројни скорији уџбеници различитих издавача, али и уџбеници из југословенског периода, за низ стручних предмета. Будући да су ови уџбеници изворно писани на српскоме, а преведени на румунски језик, те будући да је превод, пошто је ауторизован и званично одобрен, високог квалитета, увећали би се домени који би ушли у састав такавог новог, односно допуњеног корпуса. Такође, у нови би се корпус могли унети двојезични административни списи и подзаконска акта која су на снази у општинама Републике Србије у којима је румунски један од мањинских језика. Штавише, томе би се могли прикључити преводи транскрипата српских и југословенских филмова на румунски, будући да су већ израђени транскрипти мањег броја тих филмова на српском језику (уп. Terzić et al., 2020). На тај би се начин добио поуздан вишедоменски корпус с изворним текстовима на једном од корпусних језика. Поред тога, будући да се на јавном сервису Радио-телевизије Војводине свакодневно емитује дневник на румунскоме језику, при чему део вести представљају вести преведене са српскога језика, и ова збирка текстова, која непрекидно расте, представља сигурну основу за развој паралелног корпуса за предметни пар језика. Треба закључно рећи, премда, као што смо овде истакли, могућности за реализацију оваквог пројекта постоје, да је ово подухват који очекује нове нараштаје румуниста у Србији и околним истојезичким земљама, као и србиста у румунофонским срединама.

5. Закључна разматрања

У раду је описано 11 паралелних корпуса с текстовима на српскоме и румунском језику. Пошло се од чињенице да су сви описани корпуси део вишејезичних ресурса, код којих, уз само два изузетка (али тек у незнатној мери), текстови нису изворно написани на румунскоме и српском језику, већ су преводи текстова с трећега језика. Стога смо констатовали да су у описаним корпусима међусобно упарени већином текстови на предметним језицима као преводнима. Констатовали смо такође да су корпуси неједнаког обима, да има крајње малих, али и веома — релативно говорећи — великих корпуса. Поред

тога, показали смо да су корпуси састављени од текстова из једнога домена, најчешће књижевног, али да постоје и корпуси с текстовима из новинског, филмског и правног домена, као и с преведеном интернетском текстуалном грађом различитих мотивацијских и образовних говора и предавања. Даље, констатовали смо да само неколико корпуса поседује платформу на којој се преко корисничког сучеља интерно лематизовани и морфосинтактички (па и синтактички) обележени текстови могу претраживати, као и то да је већину текстова (врло често аутоматски прикупљене и упарене, али необележене) могуће у сировом формату слободно преузимати и интегрисати у друге екстерне платформе за претраживање корпуса. То је, показали смо, добра страна постојећих ресурса, јер уз сва садржајна ограничења — пре свега, једнодоменски извор текстова, неуравнотежен обим појавница, упитан ниво квалитета преведених текстова, неодговарајућа упареност и непостојање текстуалних метаподатака — могу у виду одскочне даске послужити у даљем развоју првенствено директно и обосмерно преведених текстова за румунско-српски језички пар. Пошто смо истакли да такви директно преведени текстови не изостају, те да их има у више текстуалних домена, наредни је корак у овој области покретање активности којим би се такав један пројекат од билатералног (али и мултилатералног) значаја успешно остварио.

ЛИТЕРАТУРА

Алмажан, А. (2005). *Међусобни преводи дела српских и румунских књижевника (1994–2004): писмени рад за стручни испит из библиотечке делатности*. [б. и.].

Abdelali, F. A., Guzman, F., Sajjad, H. и Vogel, S. (2014). The AMARA Corpus: Building parallel language resources for the educational domain. У N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk и S. Piperidis (Ур.), *The Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)* (стр. 1856–1862). European Language Resources Association (ELRA).

Bikić-Carić, G., Mikelenić, B. и Bezlaj, M. (2023). Construcción del RomCro, un corpus paralelo multilingüe. *Procesamiento de Lenguaje Natural*, *70*, 99–110.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R. и Roossin, P. (1988). A Statistical Approach to Language Translation. У *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*

(стр. 71–76). John von Neumann Society for Computing Sciences.

Brown, P. F., Lai, J. C. и Mercer, R. L. (1991). Aligning sentences in parallel corpora. У *ACL '91: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkley (стр. 169–176). Association for Computational Linguistics.

Čermák, F. и Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics, 17* (3), 411–427.

Čermák, P. (2019). InterCorp: A parallel corpus of 40 languages. У I. Doval и M. Teresa Sánchez Nieto (Ур.), *Parallel Corpora for Contrastive and Translation Studies: New resources and applications* (стр. 93–101). John Benjamins Publishing Company.

Erjavec, T. (2012). MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation, 46* (1), 131–142.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. У K. Aijmer, B. Altenberg и M. Johansson (Ур.), *Language in contrast. Symposium on Text-based Cross-linguistic Studies*, *Lund, Sweden, March 1994* (стр. 38–51). Lund University Press.

Johansson, S. (2007). *Seeing through multilingual corpora. On the use of corpora in contrastive studies*. John Benjamins.

Kilgarriff, A., Rychlý, P., Smrž, P. и Tugwell, D. (2004). The Sketch Engine. У *Proceedings of the 11th EURALEX International Congress* (стр. 105–116).

Koehn P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. У *Proceedings of Machine Translation Summit X: Papers* (стр. 79–86).

Lison, P., и Tiedemann, J. (2016). *OpenSubtitles2016:* Extracting Large Parallel Corpora from Movie and TV Subtitles. У N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk и S. Piperidis (Ур.), *Proceedings of the 10th International Conference on Language Resources and Evaluation LREC 2016* (стр. 923–929). European Language Resources Association (ELRA).

Miletic, A., Stosic, D. и Marjanović, S. (2017). *ParCoLab:* A Parallel Corpus for Serbian, French, and English. У K. Ekštein и V. Matoušek (Ур.), *Text, Speech, and Dialogue* (стр. 156–164). Springer International Publishing.

Reimers, N. и Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. У B. Webber, T. Cohn, Y. He и Y. Liu (Ур.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (стр. 4512–4525). Association for Computational Linguistics.

Rosen, A. (2016). InterCorp – a look behind the façade of a parallel corpus. У E. Gruszczyńska и A. Leńko-Szymańska (Ур.), *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora* (стр. 21–40). Instytut Lingwistyki Stosowanej.

Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. У P. Sojka и A. Horák (Ур.), *1st Workshop on Recent Advances in Slavonic Natural Language Processing* (стр. 65–70). Masaryk University.

Schmidt, T. и Wörner, K. (Ур.). (2012). *Multilingual corpora and multilingual corpus analysis (Vol. 14)*. John Benjamins Publishing.

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M. и Gilbro, S. (2014). An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation Journal (LRE), 48* (4), 679–707.

Terzić, D., Marjanović, S., Stosic, D. и Miletic, A. (2020). Diversification of Serbian-French-English-Spanish Parallel Corpus ParCoLab with Spoken Language Data. У P. Sojka, I. Kopeček, K. Pala и A. Horák (Ур.), *Text, Speech, and Dialogue 12284* (стр. 61–70). Springer.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. У N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk и S. Piperidis (Ур.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (стр. 2214–2218). European Language Resources Association (ELRA).

Tyers, F. M. и Alperen, M. S. (2010). SETimes: A parallel corpus of Balkan languages. У S. Piperidis, M. Slavcheva и C. Vertan (Ур.), *Proceedings of the MultiLR Workshop at the Language Resources and Evaluation Conference, LREC2010* (стр. 49–53).

von Waldenfels, R. (2006). Compiling a parallel corpus of Slavic languages: Text strategies, tools, and the question of lemmatization in alignment. У B. Brehmer, V. Ždanova и R. Zimny (Ур.), *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9* (стр. 123–138). Verlag Otto Sagner.

von Waldenfels, R. (2011). Recent Development in ParaSol: Breadth for Depth and XSLT based web concordancing with CWB. У D. Majchrakova и R. Garabik (Ур.), *Natural Language Processing, Multilinguality, Proceedings of Slovko 2011* (стр. 156–162). Tribun.

Xiao, R. (2010). Corpus creation. У N. Indurkhya и F. J. Damerau (Ур.), *Handbook of natural language processing, vol. 2* (стр. 147–165). CRC Press.

Zeroual, I. и Lakhouaja, A. (2022). MulTed: A multilingual aligned and tagged parallel corpus. *Applied Computing and Informatics*, *18* (1/2), 61–73.

Saša Marjanović, Dušica Terzić

PARALLEL CORPORA WITH ROMANIAN AND SERBIAN LANGUAGES: CURRENT STATUS AND FUTURE PROSPECTS

Summary

The paper introduces existing electronic parallel corpora containing texts aligned between the Romanian and Serbian languages. These corpora are evaluated based on various criteria. Firstly, their availability and accessibility are discussed. Secondly, the source language of the texts is considered. Thirdly, an analysis is conducted regarding the composition of text collections, their diversity across domains, the balance between them, and their coverage of word tokens. Fourthly, the extent of token annotation in the corpora is investigated at different linguistic levels, including lemmas, morphosyntactic features, and syntactic roles. Fifthly, it is determined whether the analyzed corpora offer a user interface for text queries, and if so, what search capabilities are supported. The aim of this paper is to present both the strengths and limitations of existing Romanian-Serbian parallel corpora while outlining future considerations for the development of new bidirectional parallel corpora between Romanian and Serbian.

Keywords: parallel corpus, Romanian, Serbian, natural language processing.

Saša Marjanović, Dušica Terzić

CORPUSURI PARALELE CONȚINÂND LIMBILE ROMÂNĂ ȘI SÂRBĂ: STADIUL ACTUAL ȘI PERSPECTIVE VIITOARE

Rezumat

Articolul prezintă corpusurile electronice paralele existente care conțin texte aliniate în limbile română și sârbă. Aceste corpusuri sunt evaluate în funcție de diverse criterii. În primul rând, sunt discutate disponibilitatea și accesibilitatea lor. În al doilea rând, se ia în considerare limba sursă a textelor. În al treilea rând, se analizează compoziția colecțiilor de texte, cât de variate sunt în funcție de domenii, care este echilibrul dintre ele și numărul de cuvinte incluse. În al patrulea rând, se investighează gradul de adnotare la diferite niveluri (leme, trăsături morfosintactice, roluri sintactice) a cuvintelor din corpusuri. În al cincilea rând, se stabilește dacă corpusurile analizate oferă sau nu o interfață de căutare și, în caz afirmativ, ce capacități de căutare prezintă. Scopul acestei lucrări este de a prezenta atât punctele forte, cât și limitele corpusurilor paralele româno-sârbe existente, evidențiind în același timp câteva direcții de luat în considerare în vederea dezvoltării de noi corpusuri paralele bidirecționale între română și sârbă.

Cuvinte-cheie: corpus paralel, limba română, limba sârbă, procesarea limbajului natural