

Милош В. УТВИЋ*
Универзитет у Београду
Филолошки факултет

Оригинални научни рад
Примљен: 2. 2. 2026.
Прихваћен: 15. 6. 2026.

РАЗВОЈ ЕЛЕКТРОНСКОГ КОРПУСА ДРАМСКИХ ТЕКСТОВА НА СРПСКОМ ЈЕЗИКУ И ЊЕГОВЕ ПРИМЕНЕ У ИЗУЧАВАЊУ СРПСКЕ КЊИЖЕВНОСТИ**

Током последње две деценије доступне су разне верзије једнојезичних и вишејезичних (махом паралелних) корпуса српског језика, специјализованих и општих. У раду се разматрају перспективе развоја специјализованог електронског корпуса драмских текстова на српском за чију су припрему почетни кораци углавном исти као и код осталих корпуса (утврђивање листе текстова, прикупљање и обрада грађе у форми дигиталног текста, као и одговарајућих библиографских метаподатака). Међутим, структурна анотација и аутоматска анализа текста захтевају додатни напор с обзиром на специфичан формат драмског текста и његову лексику разговорног језика. Рад се бави решавањем наведених проблема и могућим применама електронског корпуса у књижевним и лингвистичким истраживањима драмских текстова.

Кључне речи: корпус, драмски текст, српски језик, српска књижевност.

1. Увод

Први дијахронијски и пределектронски корпус српског језика аутора Ђорђа Костића (Костић 2001) настао је у периоду од 1957. до 1962. године, али се брзо суочио са обуставом финансирања, као и многи слични инострани пројекти у том тренутку, тако да је његова дигитализација започела тек деведесетих година двадесетог века (Костић 2012). Тек се осамдесетих година двадесетог века, у оквиру Групе за језичке технологије Универзитета у Београду, проистекле из Сталног семинара за математичку и рачунарску лингвистику на Математичком институту Српске академије наука и уметно-

* milos.utvic@fil.bg.ac.rs, <https://orcid.org/0000-0002-6909-9382>

** Ово истраживање подржало је Министарство науке, технолошког развоја и иновација Републике Србије (Уговор бр. 451-03-137/2025-03/ 200167).

сти, обнавља рад на изградњи корпуса српског, са фокусом на савремени и стандардни језик (Витас 2010: 258), што резултира фамилијом електронских корпуса СрпКор, доступном на вебу почев од 2002. године (Витас и др. 2024). СрпКор се брзо наметнуо као корисно средство за истраживање српског језика, а тиме и корпусна лингвистика као методологија, чиме се отворио пут развоју различитих врста корпуса српског: једнојезичних (само на српском) и вишејезичних (углавном паралелних, при чему је српски један од језика, најчешће преводни, односно циљни); корпуса који претендују да буду општи (СрпКор); и специјализованих за одређене домене. Поред Групе за језичке технологије, која је у међувремену прерасла у Друштво за језичке ресурсе и технологије (JePTex), и друге групе истраживача развијале су корпусе српског, на пример веб корпус *srWaC* (Љубешић и Клубичка 2016) центра ReLDI (скр. од Regional Linguistic Data Initiative, Регионална иницијатива о језичким подацима), као и Serbian Web Corpus PDRS 1.0 у оквиру пројекта *Јавни дискурс у Републици Србији* (Васершајт 2023). JePTex и ReLDI такође се баве образовањем лингвиста који се први пут сусрећу са рачунарском и корпусном лингвистиком, тако да свршени полазници таквих обука, самостално или у сарадњи са поменутим организацијама, развијају специфичне (пот)корпусе за потребе својих истраживачких радова.

У почетној фази истраживања описаног у овом раду корпуси СрпКор2013, СрпКор2021, *srWaC* и PDRS 1.0 били су предмет анализе са циљем да се установи колико драмских текстова садрже, пре свега оригинално насталих на српском језику. У ту сврху анализирана су имена аутора корпусних текстова тамо где су ти библиографски метаподаци били доступни (СрпКор2013), док су СрпКор2021 и веб корпуси *srWaC* и PDRS 1.0 анализирани само на основу конкорданци за упит "*prvi*" "*čin*" (којим се не прави разлика између великих и малих слова).

Површним увидом у списак аутора текстова корпуса СрпКор, у потрази за познатијим домаћим драмским писцима, као и анализом конкорданци, у верзији СрпКор2013 налазимо две драме Боре Станковића (*Коштана* и *Ташана*), две драме Љубомира Симовића (*Хасанагиница* и *Путујуће позориште Шопаловић*), једну драму Синише Ковачевића (*Зечји насип*), као и драме страних аутора (*Зојкин стан* Михаила Булгакова, *Мурлин Мурло* Николаја Кољаде, *Dealer's Choice* Патрика Марбера, *Подневна подела* Пола Клодела).¹ Анализом конкорданци у корпусу СрпКор2021 пронађене су драме *Коштана* и *Ташана* Боре Станковића, *Уговор* и *Светионик* Миодрага Ђукића, *Министарка воли кнедле* Леона Ковкеа, *Кактуси и руже* Ђорђа Лебовића, као и сатирични текстови Владимира Јокића у дневном листу *Данас* у форми кратког драмског текста (*Ричард Трећи: травестија*, *Хамлет: траве-*

¹ Не искључује се могућност да у анализираним корпусима има још драма које нису овом методологијом пронађене. На пример, у драмама Љубомира Симовића не користи се подела на чинове, већ на делове, тако да су драме тог аутора пронађене само на основу појаве његовог имена на списку аутора корпусних текстова, односно, конкорданце упита "*prvi*" "*čin*" (којим се не прави разлика између великих и малих слова) нису дале као резултат ниједну драму Љубомира Симовића.

стија, *Краљ Иби*, *Маратонци трче почасни круг*), као и драме које нису оригинално настале на српском (аутори: Софокле, Еурипид, Џорџ Бернард Шо, Албер Ками). Поједини домаћи драмски писци јесу присутни у поменутим корпусима (Милош Црњански, Александар Поповић, Бора Ћосић, Биљана Србљановић), али не са својим драмским текстовима.

Сличном анализом конкорданци у корпусу srWaC пронађене су драме *Наши очеви* и *Наши синови* Војислава М. Јовановића Марамба, *Камен Данијеле* Танасковић, *Прва награда* Александра Геца, *Краљевско дворско америчко злато* Марије Стојановић, *Мила Косте* Трифковића, *Острвска прича* Зорана Стефановића, са изузетком последње, све преузете са сајта Пројекта Растко.² Пронађене су и драме које нису оригинално настале на српском, већ су или преводи (на пример, *Закон* Володимира Виниченка) или пак текстови на хрватском.

Напоследку, на основу конкорданци издвојених из корпуса PDRS 1.0 пронађена је само једна драма на српском, *Дани леминга* Чедомира Јаничића, док су остали резултати махом били препричани драмски текстови, било као рекламе позоришних представа, било као резимеи за потребе ђака који не желе да читају лектуру.

Анализом претходно набројаних корпуса српског (СрпКор2013, СрпКор2021, srWaC, PDRS 1.0) намећу се два закључка.

1. У анализираним корпусима присутни су драмски текстови оригинално настали на српском, али у малом броју. Наиме, поређењем броја пронађених драмских текстова (9 текстова у СрпКор2013, око петнаестак текстова у СрпКор2021³, десетак текстова у srWaC и један драмски текст у PDRS 1.0) са укупним бројем текстова анализираних корпуса (5.038 текстова у СрпКор2013 са око 121 милион корпусних речи у збиру, 1.916 текстова у СрпКор2021 и укупно око 600 милиона корпусних речи, 1.353.238 текстова у srWaC са укупно око 477 милиона речи, 454.164 текстова у PDRS 1.0 са преко 600 милиона корпусних речи у збиру), произилази да је у сваком од тих корпуса процентуални удео драмских текстова у односу на укупан број корпусних текстова мањи од 1%.

2. Иако су анализирани корпуси анотирани у одређеној мери на морфосинтаксичком нивоу (на пример, све корпусне речи обележене су бар врстом речи и лемом), као и на структурном нивоу (сегментација на реченице), драмски корпусни текстови нису додатно структурно анотирани на начин који би омогућио да се претрагом корпуса издвоје реплике појединачних ликова, дијаскалије и уопште да се претрага сузи на одређене структурне целине специфичне за драмски текст (чинови, сцене, дијалози).

У случају веб корпуса, односно корпуса чији су текстови највећим делом аутоматски преузимани са интернета, јасно је да су садржај корпуса, а тиме и присуство (или боље речено, одсуство) драмских текстова у корпусу, одређени избором адреса које су послужиле као извори. С друге стране,

² <https://www.rastko.rs>

³ Број није прецизан јер није јасно да ли и сатиричне новинске текстове у форми драмског текста треба третирати као „праве” драмске текстове.

може се закључити да у корпусима чији се само новински и административни текстови (на пример, усвојени закони) преузимају аутоматски, а текстови књижевноуметничког и научног и научно-популарног стила пажљиво бирају, постоји тенденција да се драмски текстови избегавају, и то није само случај са набројаним корпусима српског, већ и са корпусима енглеског језика направљеним са намером да буду референтни, попут Брауновог корпуса⁴ и Британског националног корпуса⁵ (БНК). Креатори Брауновог корпуса савсим су искључили драмске текстове јер им је намера била да направе корпус писаних текстова, а драмске текстове су третирали као „рекреирани говорни дискурс” (Франсис и Кучера 1979). БНК (4.049 текстова са приближно 100 милиона речи у збиру) садржи свега два текста сврстана у категорију драмског текста и Астон (2001: 74) наводи као разлог процене креатора корпуса у погледу удела фикције у продукцији и рецепцији текста у време планирања корпуса, као и процене удела драмских текстова у продукцији и рецепцији фикције.

Још један разлог избегавања драмских текстова у општим корпусима јесу значајни изазови у погледу њихове структурне анотације, које ћемо размотрити у одељку 3.

Астон (2001: 74) истиче и да је за извођење општих закључака о драмским текстовима (или о било којој другој категорији текста) неопходно креирати одговарајући специјализовани корпус који би садржао само текстове циљане категорије текста. У овом раду разматраћемо перспективе развоја специјализованог електронског корпуса драмских текстова на српском језику.

У одељку 4 рада се, у општим цртама, описују постојећи електронски корпуси драмских текстова и слични корпуси чија изградња може послужити као узор. Одељак **Error! Reference source not found.** осврће се на етапе у развоју (драмског) корпуса, посебно на анотацију специфичну за драмске текстове. Одељак **Error! Reference source not found.** разматра могућности претраге анотираног драмског корпуса помоћу алатке NoSketch Engine. Напослетку, одељак 5 даје преглед закључака и смернице будућег развоја.

2. Постојећи електронски корпуси драмских текстова

Ако изузмемо специјализоване корпуре које истраживачи креирају једнократно за потребе највише пар научних радова, при чему се резултати тих радова, у најбољем случају, могу делимично репродуковати, а сами корпуси углавном нису доступни у отвореном приступу, најзанимљивији електронски корпуси драмских текстова за потребе овог рада резултат су пројекта DraCor (Фишер и др. 2019). Већ сам назив пројекта сугерише да је његов циљ развој драмских корпуса (енгл. Drama Corpora), и то углавном

⁴ https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fbrown_1

⁵ <http://www.natcorp.ox.ac.uk/>

европских драмских текстова, при чему су најстарији извори корпусних текстова настали још током раздобља античке Грчке и старог Рима, а најмлађи у двадесетом веку, тако да више не подлежу ауторским правима јер су рокови заштите истекли. Реализовани драмски корпуси су електронски, доступни на отвореној платформи⁶ којом управљају Слободни универзитет у Берлину и Универзитет у Потсдаму, и могу се обрађивати како на локалном рачунару после директног преузимања, тако и кроз сучеље за програмирање апликација (енгл. Application Programming Interface, скр. API). DraCor за примерке својих корпуса користи термин „корпус који може да се програмира” (енгл. **programmable corpus**), у смислу да није само у питању корпус читљив за људе, већ и погодан за рачунарску обраду помоћу отворених стандарда и сучеља.

3. Етапе у развоју и структурна анотација драмског корпуса

Припрема произвољног електронског корпуса састоји се од следећих етапа (Утвић 2014а):

- предобрада (претходна обрада) текстова корпуса:
 - прикупљање, дигитализација и класификација текстова за корпус;
 - конверзија текстова у одговарајући формат електронског текста;
 - лингвистичка обрада и анотација електронских текстова за корпус;
- индексирање и компресија текстова корпуса.

На почетку развоја корпуса, док се још тестирају различите методологије, односно технологије, привлачно је кренути од опортунистичког корпуса, то јест корпуса састављеног од релевантних текстова који су већ на располагању у форми електронског текста. У том смислу најпре су размотрени следећи извори постојећих електронских драмских текстова:

- серијске публикације *Савремена српска драма* и часопис *Драма* на веб сајту Удружења драмских писаца Србије (УДПС 2025);
- антологије српске драме на веб сајту Интернационалних театарских сусрета Брчко дистрикта БиХ (Сусрети 2025);
- драме дигиталне едиције *Антологија српске књижевности* (АСК 2025) некадашњег Учитељског факултета Универзитета у Београду (сада Факултет за образовање учитеља и васпитача).

Поред наведених извора, за потребе развоја паралелног двојезичног корпуса драмских текстова може се искористити и архива на веб сајту часописа *Мостови*, чије свеске, поред стручних чланака, садрже и српске преводне одабраних страних драмских текстова (Мостови 2025). При томе, формати текстова свих поменутих извора нису исти и захтевају различит степен кон-

⁶ <https://dracor.org>

верзије у формат текста погодан за креирање корпуса. АСК (2025) садржи драме у формату .docx (Microsoft Word) са ћириличним текстом кодираним Уникодом, тако да је погодан за обраду, посебно аутоматску структурну анотацију на нивоу пасуса. Драмски текстови осталих извора доступни су у формату PDF и форматирали тако да један пасус није једна физичка линија (знак за крај реда није само на крају пасуса), већ се свака логичка линија пасуса (која се види на екрану) завршава знаком за нови ред, те је потребно спајање линија пасуса у једну физичку линију како би се омогућила аутоматска анотација пасуса. Неки драмски текстови поменутих извора доступни су на ћирилици, неки на латиници. Уз све наведено, драме са сајта УДПС (2025), нажалост, не користе Уникод већ застареле YUSCII фонтове, напуштене још почетком века, што има за последицу да је екстраховани текст на енглеској латиници (јер је тако откуцан), а велике и мале варијанте слова којих нема у енглеској латиници (ш, ж, ђ, ч, ђ, љ, њ), представљене су интерпункцијским знацима попут ~, @, [,], { , }, | , \ , ` , ^, односно словима Q, q, W и w. Према томе, у оквиру предобrade неопходно је уједначити писмо изворних текстова конверзијом у циљно писмо, а у случају YUSCII текстова и конвертовати текстове у кодну шему UTF-8. CorpusPreprocessor (Утвић 2014а: 272), алатка развијена за потребе предобrade корпусних текстова на српском, омогућава аутоматску конверзију писма и YUSCII текстова у жељено циљно писмо и UTF-8, али у случају YUSCII текстова, поготово текстова код којих се истовремено појављују и ћирилица и латиница (на пример, римски бројеви и стране речи на латиници), неопходне су накнадне ручне исправке, посебно се поједини карактери користе и у свом изворном значењу (интерпункција) и као YUSCII решења за приказ слова којих нема у енглеској латиници.

CorpusPreprocessor користи се и током припреме текста за лингвистичку анотацију, односно конверзију текста у вертикал(изова)ни формат (Утвић 2014а: 129), као и за саму анотацију, односно као графичко окружење за примену софтверске алатке TreeTagger (Шмид 1997; Шмид 1999). Међу доступним параметарским датотекама (језичким моделима) за српски, који су наменски направљени за TreeTagger, изабран је SrpKor4Tagging-TreeTagger (Станковић и Шкорић 2021), дериват система морфолошких електронских речника српског језика (у даљем тексту: СМР), чији су аутори Цветана Крстев и Душко Витас (Крстев 2008). Модел SrpKor4Tagging-TreeTagger омогућава аутоматску анотацију токена корпуса⁷ придруживањем информација о највероватнијој врсти речи и леми, укључујући и неке морфолошке категорије за поткласе именица, придева и глагола: граматички род, као и степене поређења за придеве (Станковић и др. 2020). При томе, модел користи два скупа ознака (енгл. tagset), Универзални скуп ознака врста речи (енгл. Universal Part-of-Speech tagset, скр. UPoS) са 17 категорија (Petrov et al 2012) и дериват СМР-а који је више усклађен са савременим граматицама српског.

⁷ Полазећи од неалфанумеричких карактера као сепаратора, корпусне речи дефинишемо као непразне ниске карактера између два узастопна сепаратора, док су токени заједнички назив за сепараторе који нису белине (тј. нису размак, табулатор, знак за нови ред) и корпусне речи (Утвић 2014б: 244).

Смернице Иницијативе за обележавање текста (енгл. Text Encoding Initiative Guidelines, скр. TEI Guidelines, у даљем тексту: Смернице TEI), актуелна верзија TEI P5 (2025), послужиле су као основни извор приликом разматрања питања структурне анотације драмског текста, посебно поглавље 7, које је посвећено анотацији драмских текстова. Изабран је минимални скуп TEI P5 елемената којим се описују лица, сценски говор и дидаскалије (Слика 1). Значења појединачних елемената су :

- елементом `div` и његовим атрибутом `type` описују се веће структурне целине, попут чинова (`type='act'`) и сцена (`type='scene'`), док се елементом `head` описују наслови тих целина;
- елемент `castList` описује скуп ликова у драми. Ликови могу бити подељени у именоване групе (на пример ликови у кући и ликови у Сабирном центру) описане елементом `castGroup`. Име лика обележава се елементом `role`, а његов опис елементом `roleDesc`. Сваки лик у документу мора имати придружен јединствен идентификатор (вредност атрибута `xml:id`);
- сценски говор описује се елементом `sp`, који садржи ознаку говорника (`speaker`) и текст који изговара у форми пасуса (`p`). Говорник се идентификује вредношћу атрибута `who` елемента `sp`, која мора бити референца (#) на неки од идентификатора ликова дефинисаних у оквиру елемента `castList` (то јест, вредност атрибута `xml:id` неког елемента `role`);
- дидаскалије се обележавају елементом `stage` који садржи пасусе (`p`).

Током тестирања различитих извора, алатки за преобраду текста и примене структурне анотације у складу са Смерницама TEI у сврху разраде етапа процеса формирања драмског корпуса на српском, припремљено је укупно 9 драма (Табела 1).

р. б.	Назив драме	Аутор
1.	Београдска трилогија	Биљана Србљановић
2.	Буре барута	Дејан Дуковски
3.	Сабирни центар	Душан Ковачевић
4.	Балкан Експрес	Гордан Мухић
5.	Један	Небојша Ромчевић
6.	Хадерсфилд	Угљеша Шајтинац
7.	Сироти мали хрчки	Гордан Мухић
8.	Балкански шпијун	Душан Ковачевић
9.	Радован Трећи	Душан Ковачевић

Табела 1: Списак драма припремљених у складу са Смерницама TEI (драме 1–6 припремио докторанд Лазар Савковић, 7–8 Милош Утвић, а 9. драму студенти основних студија Катедре за библиотекарство и информатику).

4. Индексирање и претрага анотираног драмског корпуса

Алатка NoSketch Engine подржава индексирање и претрагу корпуса са уграђеном структурном и лингвистичком анотацијом у формату вертикалног текста (Утвић 2014а: 129–130), као и у формату XML. Корпус индексиран помоћу NoSketch Engine третира се као табела чији сваки ред одговара или појединачној позицији у корпусу (токену) или XML-етикети, прва (једина обавезна) колона садржи конкретан токен или XML-етикету, док остале (необавезне) колоне садрже одређени тип информација придружен токену (на пример, врсту речи и лему).

```
<head>Sabirni centar</head>
...
<castList>
<castGroup>
<head>LIČNOSTI U KUĆI:</head>
<castItem><role xml:id="profesor">МИHAJLO PAVLOVIĆ, profesor u penziji.</role>
<roleDesc>Blagodaran, tih čovek. Pri kraju života pokušao da osnuje Zavičajni
muzej.</roleDesc></castItem>
<castItem><role xml:id="angelina">TETKA ANGELINA, proživela vek u kući
profesorovoj.</role> <roleDesc>Podigla Ivana i Sonju.</roleDesc></castItem>
<castItem><role xml:id="petar">PETAR, bivši učitelj.</role>
<roleDesc>Napustio prosvetiteljsko zvanje i prešao u „slobodne
arheologe“.</roleDesc></castItem> ... </castGroup>
<div type="act"><head>I ČIN</head>
<div type="scene">
<head>I. UMIRANJE PROFESORA MIHAJLA PAVLOVIĆA</head>
<stage><p>Tokom dvadesetak godina napornog rada i strpljivog izučavanja iskopina „daleke
provincije“ negdašnje Rimske Imperije, profesor Mihajlo Pavlović ispunio je sve delove svoje
kuće raznovrsnim i bogatim arheološkim predmetima [...] Sa ulaznih vrata, iz hodnika, oglasi
se zvono. Ono prizva iz susedne sobe tetka Angelinu... sićušnu, bogobojažljivu staricu u crnini.
Žena, davno zašla u sedmu deceniju života, [...] izađe u hodnik. Tamo, u polumraku, pozdravi
Petra... sredovečnog, koštunjavog čoveka u sivom, dva broja većem i širem odelu. Petar u
naručju nosi platneni zavežljaj. Uznemireno se osvrće po dnevnoj sobi. Tetka ga ponudi da
sedne, ali on odbi.</p></stage>
<sp who="#petar"><speaker>PETAR:</speaker><p> Ne, ne, hvala. Kako je
profesor?</p></sp>
<sp who="#angelina"><speaker>TETKA ANGELINA:</speaker><p> Loše, jako
loše.</p></sp>
```

Слика 1: Одломак из драме *Сабирни центар* Душана Ковачевића (кодиран у формату TEI P5 XML).

Атрибути XML-етикета којима се означава почетак или крај корпусног текста садрже информације о самом тексту (на пример, библиографске метаподатке о аутору, наслову, години издања, издавачу итд.). Колоне корпуса индексираног алатком NoSketch Engine још се називају **позициони атрибути**, док се за различита имена XML-елемената у корпусу користи назив **структурни атрибути**.

Упити за претрагу корпуса задају се на корпусном упитном језику CQL⁸ (SketchEngineDoc 2019), који филтрира корпус (то јест редове табеле) општом дозвољених вредности позиционих и структурних атрибута помоћу регуларних израза, као и употребом логичких (буловских) операција.

За потребе овог рада направљен је мали корпус састављен од драма наведених у Табела 1, конвертованих из формата TEI P5 XML у формат вертикалног текста. Делимична морфолошка анотација у верзији корпуса за тестирање реализована је позиционим атрибутима `pos` (ознака врсте речи) и `lemma` (лема), који садрже одговарајуће информације о токenu. Табела 2 илуструје CQL-упите који користе позициони атрибут `lemma` и структурни атрибут `sp`.

Табела 2: Примери претраге драме *Сабирни центар* засноване на њеној структурној анотацији.

Циљ претраге	Упит	Значење упита
Наћи све реплике које изговара учитељ Петар	<code><sp who="#petar"/></code>	Сви елементи <code>sp</code> чији је идентификатор говорника <i>petar</i> .
Наћи појављивања леме <i>континенталац</i> у оквиру реплике коју изговара Јанко Савски	<code>[lemma= „kontinentalac”] within <sp who="#janko"/></code>	Сваки токен чија је лема <i>kontinentalac</i> и који се налази у оквиру елемента <code>sp</code> чији је идентификатор говорника <i>janko</i> .

5. Закључак и будући рад

У раду је изложен приступ изградњи и претрази корпуса драмских текстова комбинавањем лингвистичке и структурне анотације. Идентификовано је више извора драмских текстова, од којих по обиму и значају највише обећава УДПС (2025), односно часопис *Савремена српска драма*, чија ће обрада, посебно аутоматска структурна анотација у складу са Смерницама TEI P5, бити предмет посебног рада. Почетна верзија корпуса показује да се анализом може обухватити и специфична структура драмског текста. Будући

⁸ У енгл. оригиналу Corpus Query Language, скр. CQL.

рад обухватиће статистичку анализу корпуса, као и анализу релација између ликова у драмском тексту.

Извори

- АСК 2025:** Дигитална едиција *Антологија српске књижевности*. Универзитет у Београду – Учитељски факултет. <<http://www.antologijasrpskeknjivosti.rs/>>. 30.09.2025.
- Васершајт 2023:** Wasserscheidt, Philipp. Serbian Web Corpus PDRS 1.0. *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042. <<http://hdl.handle.net/11356/1752>>. 30.09.2025.
- Љубешић и Клубичка 2016:** Ljubešić, Nikola and Klubička, Filip. Serbian web corpus srWaC 1.1. *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042. <hdl.handle.net/11356/1063>. 18.09.2025.
- Мостови 2025:** Архива часописа Мостови. Удружење књижевних преводилаца Србије. <<https://www.ukps.rs/mostovi-arhiva/>>. 30.09.2025.
- SketchEngineDoc 2019:** Documentation for Expert Users. Sketch Engine Documentation <<https://www.sketchengine.eu/documentation/>>. 30.09.2025.
- SrpELTeC:** The Serbian Part of the ELTeC Collection. Друштво за језичке ресурсе и технологије (JePTeX). <<https://noske.jerteh.rs/#dashboard?corpname=srpELTeC>>. 30.09.2025.
- СрпКор 2013:** Витас, Душко и Утвић, Милош. Корпус савременог српског језика (СрпКор), верзија СрпКор2013. Група за језичке технологије Универзитета у Београду, <<https://noske.jerteh.rs/#dashboard?corpname=SrpKor2013>>. 30.09.2025.
- СрпКор 2021:** Корпус савременог српског језика (СрпКор), верзија СрпКор2021. Друштво за језичке ресурсе и технологије (JePTeX). <<https://noske.jerteh.rs/#dashboard?corpname=SrpKor2021>>. 30.09.2025.
- Станковић и Шкорић 2021:** Stanković, Ranka and Škorić, Mihailo, SrpKor4Tagging-TreeTagger. Version 1.0.0 (automatically assigned). <<https://doi.org/10.57771/bvkk-jv85>>. 30.09.2025.
- Сусрети 2025:** Интернационални театарски сусрети Брчко дистрикта БиХ. <<https://susreti.co.ba>>. 30.09.2025.
- TEI 2025:** TEI Consortium (eds.), TEI P5: Guidelines for Electronic Text Encoding and Interchange, Version 4.10.2. Last updated on 4th September 2025, revision bcfa98f42. <<http://www.tei-c.org/P5/>>. 30.09.2025.
- УДПС 2025:** Удружење драмских писаца Србије. <<https://www.drama.org.rs>>. 30.09.2025.

ЛИТЕРАТУРА

- Астон 2001:** Guy Aston, Text Categories and Corpus Users: A Response to David Lee (Commentary), *Language Learning & Technology*, 5/3, 73–76. <<https://doi.org/10.64152/10125/44566>>. 30.09.2025.
- Витас и др. 2024:** Duško Vitas, Ranka Stanković and Cvetana Krstev, The Many Faces of SrpKor, in: Jasmina Moskovljević Popović and Ranka Stanković (eds), *International Conference South Slavic Languages in the Digital Environment JuDig: Book of Abstracts*, Belgrade: University of Belgrade – Faculty of Philology.
- Костић 2001:** Ђорђе Костић, *Квантитативни опис структуре српског језика: Корпус српског језика*, Београд: Институт за експерименталну фонетику и патологију говора и Лабораторија за експерименталну психологију Филозофског факултета Универзитета у Београду.
- Костић 2012:** Aleksandar Kostić, Elektronski korpus srpskog jezika od XII do XVIII veka, u: *Nove tehnologije i standardi: Digitalizacija nacionalne baštine 2012. Jednanaesta nacionalna konferencija sa međunarodnim učešćem*, Beograd: Matematički fakultet Univerziteta, 12–13.
- Крстев 2008:** Cvetana Krstev, *Processing of Serbian – Automata, Texts and Electronic Dictionaries*, Belgrade: Faculty of Philology, University of Belgrade.
- Петров 2012:** S. Petrov, D. Das, and R. McDonald, A Universal Part-of-Speech Tagset, in: Nicoletta Calzolari et al. (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association (ELRA).
- Рихли 2007:** Pavel Rychlý, Manatee/Bonito – A Modular Corpus Manager, in: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, Brno : Masaryk University, 65–70.
- Станковић и др. 2020:** Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić, Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian, in: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 3954–3962, Marseille, France: European Language Resources Association.
- Тртовац и др. 2022:** Aleksandra Trtovac, Vasilije Milnović and Cvetana Krstev, The Serbian Part of the ELTeC – from the Empty List to the 100 Novels Collection, *Infotheca – Journal for Digital Humanities*, 21/2, 7–25. <<https://doi.org/10.18485/infotheca.2021.21.2.1>>. 30.09.2025.
- Утвић 2014а:** Miloš Utvić, *Izgradnja referentnog korpusa savremenog srpskog jezika*, Doktorska disertacija, Beograd: Univerzitet u Beogradu, Filološki fakultet, <<https://fedorabg.bg.ac.rs/fedora/get/o:10061/bdef:Content/download>>. 30.09.2025.
- Утвић 2014б:** Милош В. Утвић, Листе учестаности Корпуса савременог српског језика, *Научни састанак слависта у Вукове дане*, 43/3, 241–262.
- Фишер и др. 2019:** F. Fischer, I. Börner, M. Göbel, A. Hechtl, C. Kittel, C. Milling and P. Trilcke, Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama, in: *Proceedings of Digital*

Humanities (DH2019): “Complexities”, Utrecht University, <<https://doi.org/10.5281/zenodo.4284002>>. 30.09.2025.

Франсис и Кучера 1979: W. N. Francis and H. Kucera, *Brown Corpus Manual*. Providence, Rhode Island, US: Department of Linguistics, Brown University. <<https://listings.lib.msu.edu/public-corpora/cd421/manuals/brown/INDEX.HTM>>. 30.09.2025.

Шмид 1997: Helmut Schmid, Probabilistic Part-of-Speech Tagging Using Decision Trees, in: D. B. Jones *et al.* (eds.) *New Methods in Language Processing*, Routledge, 154–164.

Шмид 1999: Helmut Schmid, Improvements in Part-of-Speech Tagging with an Application to German, in: S. Armstrong *et al.* (eds.) *Natural Language Processing Using Very Large Corpora*, Dordrecht: Springer, 13–25.

Miloš V. Utvić

A DEVELOPMENT OF AN ELECTRONIC CORPUS OF DRAMA TEXTS
IN SERBIAN AND ITS APPLICATIONS IN THE STUDY OF SERBIAN
LITERATURE

Summary

Various versions of monolingual and multilingual (mostly parallel) corpora of the Serbian language, both specialized and general, have been available over the past two decades. This paper discusses the prospects for the development of a specialized electronic corpus of drama texts in the Serbian language, the initial steps for the preparation of which are largely the same as for other corpora (determining the list of texts, collecting and processing material in the form of digital text, as well as appropriate bibliographic metadata). However, structural annotation and automatic text analysis require additional effort given the specific format of the drama text and its colloquial vocabulary. The paper deals with solving the above problems, as well as possible applications of the electronic corpus in literary and linguistic research of drama texts.

Keywords: corpus, drama, Serbian language, Serbian literature.