Примена ТХМ алата за анализу корпуса просторних планова

УДК 811.163.41’322.2

Милена Милинковић
milenam@iaus.ac.rs
Институт за архитектуру и урбанизам Србије
Београд, Србија

САЖЕТАК: У раду су представљена квантитативна текстометријска истраживања *PPTXM* корпуса просторних планова у ТХМ окружењу. Захваљујући примени развијених метода и технологија у области корпусне лингвистике и алата *SrpNER* за обележавање именованих ентитета, указано је на различите лингвистичке и статистичке карактеристике текстуалног дела планских докумената. Коришћењем ТХМ алата током истраживања, добијени су резултати који су омогућили увид у учесталост појављивања различитих врста речи и именованих ентитета, како у целом корпусу, тако и њихову распоређеност и прогресију по партицијама овог корпуса.

КЉУЧНЕ РЕЧИ: ТХМ, текстометрија, именовани ентитети, дигитални корпуси, просторни планови..

РАД ПРИМЉЕН: 1. мај 2022.
РАД ПРИХВАЋЕН: 1. јун 2022.

1. Увод

Квантитативни приступ анализе текстова подразумева уочавање њихових статистичких карактеристика заснованих на унапред дефинисаним критеријумима класификације и квантификације (Bjekić et al. 2012). Класичан приступ квантитативне анализе текстова углавном је захтевао сарадњу већег броја истраживача, али и напоран и дуготрајан рад (Utvić 2013), што је изискивало значајне материјалне и временске ресурсе (Mehl and Gill 2010). Са појавом рачунара и

развојем савремених рачунарских технологија овакве анализе су умногоме олакшане. Без обзира да ли је у питању класичан или рачунарски подржан приступ, за било какву лингвистичку анализу неопходно је формирање адекватних и репрезентативних корпуса, како би истраживања вршена на њима могла да се примене и на друге текстове који задовољавају унапред постављене критеријуме. У лингвистици корпус се може описати као „колекција текстова за које се претпоставља да су репрезентативни за дати језик, дијалект или други подскуп језика, а који ће се користити за лингвистичку анализу“ (Francis 1975, 15). Оваква дефиниција може се применити на све корпусе, независно од тога који је носач информација у питању. Међутим, са развојем рачунарске лингвистике и формирањем електронских корпуса прихватљивија је дефиниција која под корпусом подразумева „колекцију аутентичних машински читљивих текстова који представљају репрезентативни узорак конкретног језика или језичког варијетета“ (McEnery, Xiao, and Tono 2006, 5). Електронски корпуси су подједнако као и пределектронски корпуси, намењени за језичка истраживања, али су за разлику од њих, прилагођени аутоматској или полуаутоматској обради и анализи.

Постоје различите поделе корпуса у зависности од обима, од броја језика који су у њега укључени, од намене, домена, од носача и других карактеристика (Popović and Vitas 2003). Једна од њих је подела на корпусе опште намене и специјализоване, тј. доменске корпусе. Корпус савременог српског језика – *CpnKop* представља корпус опште намене. Његов развој започет је 1981, а најновија верзија *CpnKop2021*, објављена је 2021. године са преко 600 милиона речи. У његовом саставу налазе се и различити доменски корпуси формирани током година. Доменски корпус кулинарства (Витас and Крстев 2016; Витас 2018; Vitas 2019, 2022), доменски корпус законских текстова Народне скупштине Републике Србије (Васиљевић 2014), доменски корпус из области библиотекарства (Тртовац 2017) и доменски корпус из области рударства (Обрадовић et al. 2017) само су неки од специјализованих корпуса који су током година развијали чланови Друштва за језичке ресурсе и технологије – JePTex.[1] Последњи доменски корпус формиран у оквиру Друштва, крајем 2021. године је доменски корпус просторног планирања. На подскупу овог корпуса вршена су истраживања алатима SrpNER и TXM која ће бити представљена у наредним поглављима.

1. Друштво за језичке ресурсе и технологије – JePTex

2. Алат TXM

TXM[2] је софтвер отвореног кода (Heiden 2010) чија је примена омогућена за три различита оперативна система: 64-битни Windows, Mac OS X и 64-битни Linux. Постоје две верзије овог софтвера. Десктоп верзија захтева инсталацију TXM софтвера после чега је кориснику омогућено да на сопственом рачунару увезе свој корпус и изврши жељене анализе, док за верзију TXM софтвера за веб инсталација није потребна већ корисник преко претраживача приступа различитим онлајн корпусима.

Да би текстови могли да се увезу у TXM окружење неопходно је да буду у одређеном формату и у адекватном кодном распореду. TXM алатом могуће је анализирати документа у TXT формату чистог текста или XML (*eXtensible Markup Language*) формату. Први формат мора да буде у препорученом UTF-8 кодном распореду, а други, поред наведеног кодног распореда мора да буде и у складу са TEI (*Text Encoding Initiative*)[3] упутствима. Степен сложености анализа, које је могуће спровести у TXM окружењу, зависи од нивоа репрезентације текстова. На текстовима у TXT формату могуће је спроводити само основне анализе будући да имају најнижи ниво репрезентације. За разлику од њих, текстови кодирани у XML-TEI формату могу имати мање или више богату репрезентацију, што директно утиче на комплексност анализа које је могуће применити на текстове.

TXM алат пружа корисницима најразличитије технике за статистичку обраду текстова, а његово радно окружење је изразито прегледно и лако за коришћење па је стога погодно за употребу како почетницима тако и стручним лицима која се баве статистичким истраживањима. С обзиром на то да је у TXM окружењу могућа хијерархијска организација текстуалних објеката на којима се истраживање спроводи, анализе је могуће вршити и на целом корпусу и на издвојеном поткорпусу или партицији јер су корпуси подељени на саставне јединице: текстуалне, структуралне и лексичке. Текстуалну чине сви текстови који се у њему налазе обележени неопходним метаподацима. Структуралне јединице корпуса су поглавља, пасуси, реченице и сл. Најнижи ниво хијерархијске организације чине лексичке јединице, тј. речи од којих се текст и састоји (Jaćimović 2019). Резултати задатих упита приказују се као конкорданце, табеле или графикони.

2. What is TXM? (Шта је TXM?)
3. TEI (Text Encoding Initiative).

3. Корпус РРТХМ

За потребе овог истраживања формиран је корпус РРТХМ[4] који чине текстови просторних планова и представља део доменског корпуса из области просторног планирања. Планска документација чини специфичну врсту докумената карактеристичних за област просторног планирања. Постоје четири врсте просторних планова који се израђују на територији државе. То су Просторни план Републике Србије, Регионални просторни план, Просторни план јединица локалне самоуправе и Просторни план подручја посебне намене (Службени гласник РС 52/2021). Без обзира о ком плану је реч, после доношења одлуке о изради плана, основно је утврдити обухват подручја плана, урадити анализу тренутног стања, а потом у складу са условима и смерницама који су прецизирани у планским документима вишег реда, приступити његовој изради. Цео овај процес захтева сарадњу просторних планера и научника и стручњака из многих других области (архитектуре, урбанизма, демографије, екологије, социологије и др.). У почетним фазама израде плана припрема се елаборат за рани јавни увид како би се шира јавност упознала са сврхом и циљевима израде плана, са потенцијалним концепцијама и решењима развоја конкретног подручја и понуђеним планом заштите животне средине. Потом следе припремне активности које подразумевају прибављање подлога за израду графичког дела плана и прибављање услова и података неопходних за израду нацрта планског документа који садржи текстуални и графички део. Нацрт плана подлеже стручној контроли која траје 15 дана, а након поступања у складу са извештајем о обављеној стручној контроли план се излаже на јавни увид. Током ове фазе израде плана заинтересована физичка и правна лица подносе евентуалне примедбе на плански документ искључиво у писаном облику. По завршетку израде планови иду на усвајање у надлежне институције, а потом се текстуални део плана објављује у одговарајућем службеном гласилу (Службени гласник РС 32/2019).

Поменути корпус се састоји од шест просторних планова, четири плана подручја посебне намене, једног регионалног просторног плана и једног просторног плана јединице локалне самоуправе, а то су: *Регионални просторни план Златиборског и Моравичког управног округа*, *Просторни план подручја посебне намене Националног парка „Ђердап“*, *Просторни план подручја посебне намене слива*

4. Корпус просторних планова формиран за анализе у ТХМ окружењу.

акумулације „Грлиште“, *Просторни план подручја посебне намене слива акумулације „Ћелије“*, *Просторни план подручја посебне намене међународног водног пута Е-80 – Дунав (Паневропски коридор VII)* и *Просторни план општине Књажевац*. Једини просторни план који није укључен у овај корпус је *Просторни план Републике Србије*, међутим, имајући у виду да су наведеним плановима покривена различита подручја на територији Србије, као и да су њима покривене све разноликости географских карактеристика Србије, може се рећи да је наведени корпус репрезентативан.

4. Текстометрија алатом ТХМ

Текстометрија је методологија анализе текстуалних података чији су развој започели француски теоретичари Пјер Гиро (Pierre Guiraud), Шарл Милер (Charles Muller), Жан-Пол Бензекри (Jean-Paul Benzécri) и др., и своју примену је нашла у квантитативним лингвистичким истраживањима (MacMurray and Leenhardt 2011), али и у многим другим хуманистичким и друштвеним наукама (Jaćimović 2019). Текстометријском анализом, која је примењена на овај корпус, биће указано на различите лингвистичке и статистичке карактеристике текстуалног дела планских докумената. Од када је основана ова метода 80-их година двадесетог века, захваљујући богатој палети алата и софтвера који се користе најпре за обраду, а потом и анализу природних језика, као и низу међународних стандарда за структурирање података и текста, она представља много више од својеврсног бројања речи (Pincemin, Heiden, and Decorde 2020). Различити видови текстометријске анализе омогућени су ТХМ софтвером. Неки од примера текстометријске анализе у ТХМ окружењу спроведени су на једном од доменских корпуса који је развијен у Друштву за језичке ресурсе и технологије. У питању је SrpELTeC корпус[5] који је у том тренутку садржао 21 прозно дело на српском језику објављено у периоду 1840–1920. године (Trtovac, Milnović, and Krstev 2021). Мотив његове израде било је укључење текстова корпуса у вишејезичну збирку европских књижевних текстова (енгл. *European Literary Text Collection*). Добијени резултати су пружили увид у велику разноликост анализа коју је могуће вршити ТХМ алатима, истовремено указујући на специфичности употребе различитих врста речи у зависности од пола

5. Завршена колекција доступна на SrpELTeC

аутора и учесталости употребе појединих лема у текстовима SrpELTeC корпуса (Jaćimović 2019).

5. SrpNER и обележавање именованих ентитета

Један од многобројних задатака обраде природних језика је и препознавање именованих ентитета (*Named Entity Recognition* – NER). Ова врста обраде и анализе текста се односи на препознавање имена особа, организација, локација као и различитих нумеричких израза, укључујући проценте, новац, време и датуме. У скорије време технологија препознавања именованих ентитета примењује се и на екстраховање назива догађаја, производа, наслова књига и обележавање електронских адреса. Развој овог сегмента обраде природног језика траје већ скоро три деценије (Maurel, Friburger, and Eshkol-Taravella 2014), а за препознавање именованих ентитета у српском језику, у оквиру Друштва за језичке ресурсе и технологије, већ дуги низ година развијан је SrpNER систем. Начин рада и коришћења система за препознавање именованих ентитета у српском језику детаљно је описала проф. др Цветана Крстев (Krstev et al. 2014).

Пре било каквих текстометријских анализа у корпусу PPTXM обележени су именовани ентитети алатом SrpNER. За потребе овог истраживања обележавани су геополитички појмови (`top.deogr`, `top.dr`, `top.geo`, `top.gr`, `top.hyd`, `top.reg`, `top.supreg`, `top.ul`), демоними (`demonym`) и називи организација (`org.com`, `org.gen`, `org.pol`, `org.rel`). Да би алатом TXM могла да се врше истраживања у корпусу, биле су неопходне извесне корекције које су подразумевале да се из етикета за обележавање именованих етикета, тачка замени доњом цртом, тј. да се етикете преименују тако што ће етикета `org.com` бити замењена етикетом `org_com`, етикета `top.gr` етикетом `top_gr` и тако редом. Једина етикета која је остала непромењена је етикета за обележавање демонима (табела 1). Сем наведеног било је потребно елиминисати угнежђене етикете именованих ентитета, које производи обележавање алатом SrpNER, а које нису подржане у TXM окружењу. Анотација организације *ЈКП „Водовод“ Зајечар* алатом SrpNER извршена је на следећи начин:

```
<org.com>
  <org.gen>JKP „Vodovod”</org.gen>
  <top.gr>Zaječar</top.gr>
</org.com>
```

У оквиру спољних етикета које се користе за обележавање комерцијалних организација `<org.com>` и `</org.com>` налазе се и угнежђене етикете за опште организације `<org.gen>` и `</org.gen>` којима је обележен *ЈКП „Водовод“* и етикете за градско насеље или насељена места `<top.gr>` и `</top.gr>` којима је обележен *Зајечар*. Уклањањем угнежђених етикета пун назив организације је обележен етикетом `<org.com>`.

ТХМ етикете	**објашњење етикета**	**Пример**
`demonym`	именице којима је означено становништво, етничке групе и придеви изведени из географских назива	*београдски*
`org_com`	комерцијалне организације	*ЈКП „Водовод“ Зајечар*
`org_gen`	опште организације (организације које нису на други начин разврстане)	*Институт за архитектуру и урбанизам Србије*
`org_pol`	политичке организације	
`org_rel`	религијски објекти (одн. организације)	*Црква Светог Марка*
`top_deogr`	део градског насеља или насељеног места	*Дедиње*
`top_dr`	држава	*Србија*
`top_geo`	географска обележја (низије, планине, брда, висоравни...)	*Авала*
`top_gr`	градско насеље или насељено место	*Београд*
`top_hyd`	Хидроними (реке, језера, извори...)	*Дунав*
`top_reg`	регион унутар државе	*Подунавље*
`top_supreg`	наддржавни регион	*Европа*
`top_ul`	улице или уопште градске локације	*Кнез Михаилова улица*

Табела 1. ТХМ етикете са објашњењима и примерима

6. Текстометријска обрада и анализа корпуса PPTXM у TXM окружењу

Први корак по увозу корпуса у ТХМ окружење подразумевао је аутоматску сегментацију, токенизацију, лематизацију и на крају етикетирање врстом речи (Jaćimović 2019). Ово је омогућено алатом

TreeTagger[6] интегрисаним у TXM софтвер који повезује сваку реч у тексту са лемом и придруженим морфосинтаксичким категоријама. Овако анотиран корпус је погодан за анализу претраживачем "Corpus Query Processor" (CQP) који уместо карактера као јединицу претраге подразумева корпусну реч што пружа могућност за различита фразеолошка истраживања.

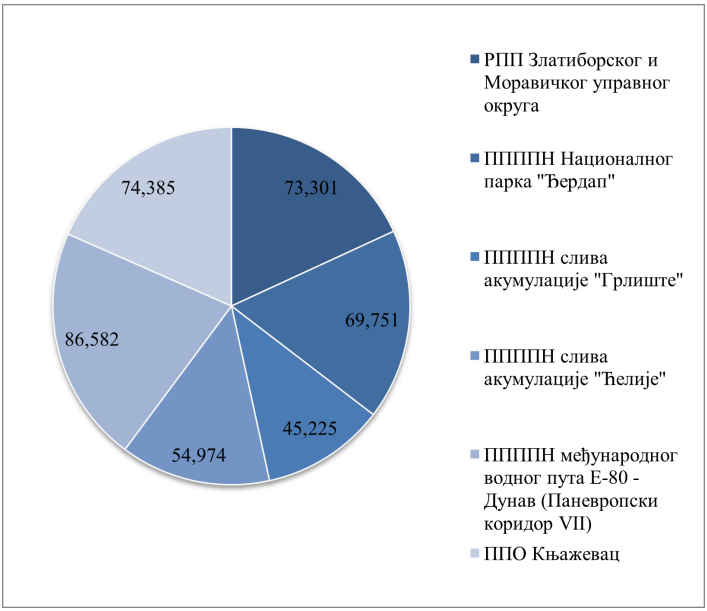
Број пасуса и реченица може се добити коришћењем CQP упита, који као резултат дају листе конкорданци. Упит за пасусе `/region[p]` даје као резултат 11.895 конкорданци, док се за упит `/region[s]` који се односи на реченице добијају 17.063 линије конкорданци.

Токенизацијом текстова утврђено је да у корпусу PPTXM постоји 487.213 токена (табела 2). Од тог броја токена појавних облика има 32.565, а различитих лема има 14.903. Када се од укупног броја токена одузму интерпункцијски знаци, долази се до податка да се корпус састоји од 404.218 речи. Број речи као ни број токена нису подједнако распоређени по партицијама овог корпуса (слика 1, табела 2).

Назив просторног плана	**Обухват плана**	**Димензије партиција на основу броја токена**
Регионални просторни план Златиборског и Моравичког управног округа	9.184 km^2	90.724
Просторни план подручја посебне намене Националног парка „Ђердап“	1542 km^2	83.058
Просторни план подручја посебне намене слива акумулације „Грлиште“	400 km^2	53.907
Просторни план подручја посебне намене слива акумулације „Ћелије“	935 km^2	66.892
Просторни план подручја посебне намене међународног водног пута E-80 – Дунав (Паневропски коридор VII)	4.536 km^2	104.284
Просторни план општине Књажевац	1.202 km^2	88.348

Табела 2. Основне географске и језичке карактеристике корпуса PPTXM

6. TreeTagger - a part-of-speech tagger for many languages



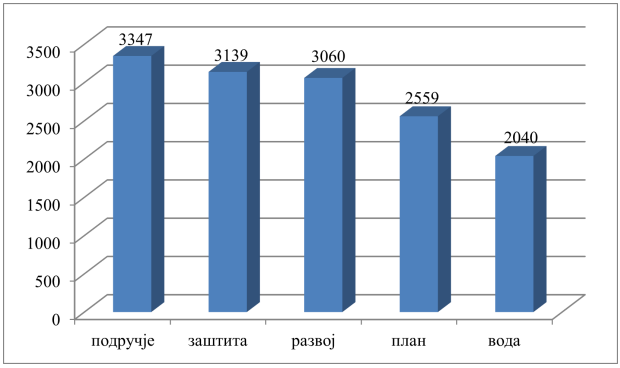
Слика 1. Екстрахован број речи по партицијама у корпусу РРТХМ.

Корпус РРТХМ, како је већ напоменуто, састоји се од шест планова (табела 2). Територијални обухват планова се знатно разликује од плана до плана, као и димензије текстова мерене бројем токена. Територије обухваћене плановима се крећу од 935 km^2 до 9.184 km^2, а димензије текстова од 53.907 до 104.284 токена. Међутим, уочљиво је да величина територије обухваћена планом није пропорцијална димензији планских текстова, што се може видети на примеру *РПП Златиборског и Моравичког управног округа* којим је покривено 9.184 km^2 територије и чији план садржи 90.724 токена и *ППППН међународног водног пута Е-80 – Дунав (Паневропски коридор VII)* чија територија је по површини готово дупло мања, а текст плана садржи око 14.000 токена више. Сличан пример је и са *ППППН Националног парка „Ђердап“* и *ППО Књажевац*. Први од наведених планова је по обухвату територије већи, а други садржи већи број токена.

Следећи корак у анализи корпуса представљао је утврђивање учесталости појављивања знакова интерпункције и различитих врста речи у текстовима. То је постигнуто употребом морфолошких етикета (енгл. *Part of Speech tags*) за српски језик. На основу добијених резултата

srpos	F	srpos	F
N (именица)	153.857	DET (детерминатори)	7.624
PUNCT (знак интерпункције)	82.995	ADV (прилог)	7.367
ADJ (придев)	78.039	AUX (помоћни глаголи)	6.991
ADP (предлог)	42.125	PART (речца)	5.917
CCONJ (напоредни везник)	33.368	X (остало)	3.878
PROPN (властита имена)	27.500	SCONJ (зависни везник)	2.261
VERB (глагол)	17.865	PRON (заменица)	893
NUM (број)	15.873	INTJ (узвик)	660

Табела 3. Листа учесталости свих заступљених вредности позиционог атрибута `srpos` у PPTXM корпусу



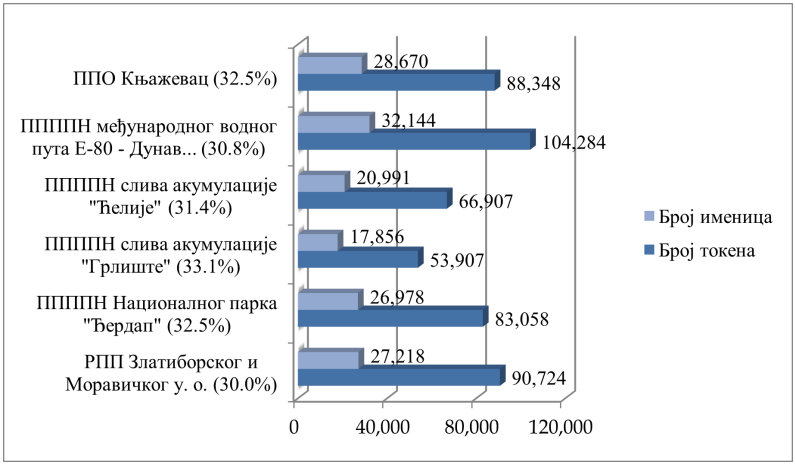
Слика 2. Приказ именица чија је фреквентност у корпусу већа од 2.000

(табела 3) може се видети је да је фреквентност именица највећа, за њима следи учесталост појављивања знакова интерпункције, а затим су наведене апсолутне фреквентности осталих врста речи. Поред укупног броја именица, упитом `[srpos="NOUN"]` утврђено је да постоји 110.587 појавних облика, као и 5.204 различите леме.

Истим упитом установљено је и које именице су најучесталије у целом корпусу, а пет најфреквентнијих је приказано на слици 2 на којој се види да је учесталост именица *подручје*, *заштита* и *развој* у корпусу већа од 3.000 и да се оне чешће користе у плановима

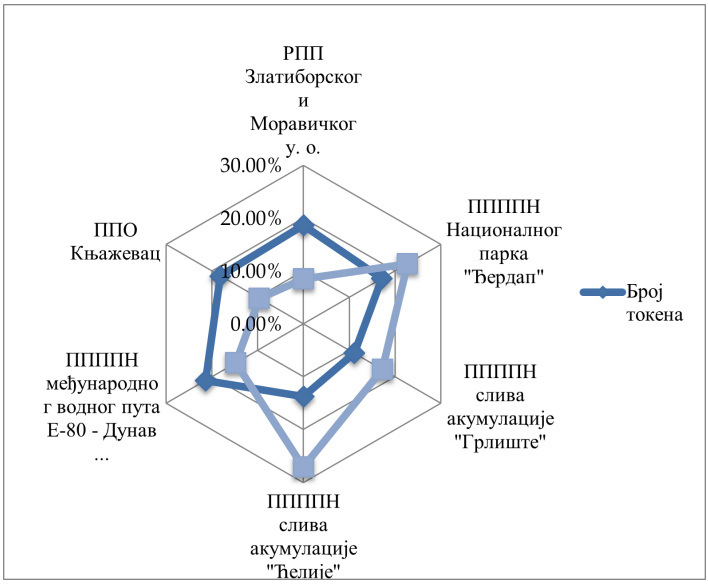
Као што је већ речено у оквиру сваке партиције постоји одређен број токена. На исти начин могуће је утврдити колико је нека од врста речи заступљена у сваком од текстова PPTXM корпуса. Ако се посматра однос броја именица и токена по партицијама корпуса, може се закључити да је тај однос поремећен само када су у питању *РПП*

Златиборског и Моравичког управног округа и *ППО Књажевац*. По броју токена *РПП Златиборског и Моравичког управног округа* налази се на другом, а по броју именица на трећем месту, док је у случају *ППО Књажевац* ситуација супротна. Остали планови су задржали пропорционалан однос заступљености токена и именица (слика 3).



Слика 3. Однос заступљености токена и именица по партицијама корпуса РРТХМ.

Ситуација је нешто другачија када је у питању однос токена и заменица у партицијама РРТХМ корпуса. Укупан број токена, као што је већ утврђено, у целом корпусу је 487.213, а укупан број заменица добијених упитом `[srpos ="PRON"]` је 893. Заступљеност токена по партицијама није пропорцијална заступљености заменица. На слици 4 је приказан однос заступљености токена и заменица у сваком од шест текстова просторних планова. Највећа учесталост употребе заменица је у *ПППППН слива акумулације „Ћелије“*, мада је број токена у овом плану већи једино од броја токена у *ПППППН слива акумулације „Грлиште“*. Са друге стране *ПППППН међународног водног пута Е-80 – Дунав (Паневропски коридор VII)*, као највећа партиција рачуната на основу броја токена, по процентуалној заступљености заменица налази се тек на четвртом месту. *ППО Књажевац* и *ПППППН Националног парка „Ђердап“*, по броју токена заузимају треће, односно четврто место са

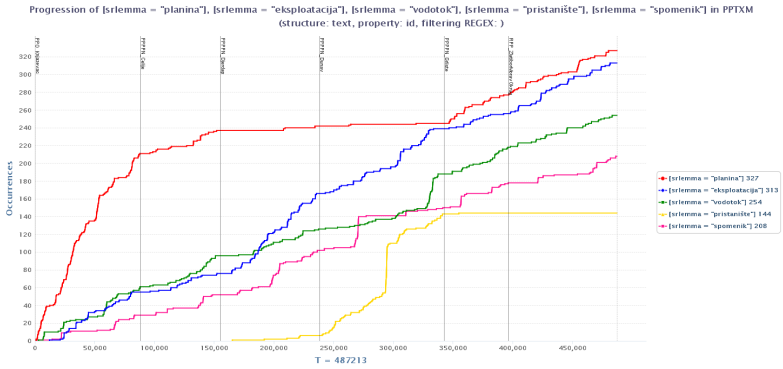


Слика 4. Однос заступљености токена и заменица по партицијама корпуса РРТХМ.

разликом од свега 5.000 токена. Међутим, у *ППППН Националног парка „Ђердап“* заступљеност заменица је готово два и по пута већа него у *ППО Књажевац*.

Досадашње анализе представљају основни облик статистичке обраде текстова, а алати у ТХМ окружењу омогућавају много сложеније анализе различитих облика учесталости и заступљености у текстовима. Оно што је карактеристично за ТХМ окружење је чињеница да је у корпусу поред укупног појављивања одређене врсте речи и фреквентности појединачних облика омогућено пратити и прогресију и кумулативну фреквенцију различитих врста речи кроз саставне делове корпуса и кроз цео корпус.

На графикону прогресије, приказаном на слици 5, дати су примери за пет именица: *планина*, *експлоатација*, *водоток*, *пристаниште*, *споменик*. Црвеном линијом је приказана кумулативна фреквентност именице *планина*. Приметно је да се она најчешће јавља у *ППО Књажевац*, док је њено појављивање у *ППППН Националног парка „Ђердап“* (5) и *ППППН међународног водног пута Е-80 – Дунав (Паневропски коридор VII)* (3) на нивоу статистичке грешке. У



Слика 5. Приказ прогресије пет именица у текстовима PPTXM корпуса; партиције редом одговарају *ППО Књажевац*, *ПППП Ћелије*, *ПППП Ђердап*, *ПППП Дунав*, *ППП Грлиште*, *РПП Златиборски и моравички округ*.

преостала три план, именица *планина*, најређе је употребљавана у *ПППП слива акумулације „Ћелије“*, нешто чешће у *ПППП слива акумулације „Грлиште“*, а најчешће у *РПП Златиборског и Моравичког управног округа*. Оваква учесталост именице *планина* по партицијама PPTXM корпуса је и очекивана зато што и подручје *ППО Књажевац*, као и *РПП Златиборског и Моравичког управног округа* имају изражен брдско-планински рељеф.

Именица *пристаниште* (жута линија) у целом корпусу појављује се 144 пута, међутим, како се може видети на графикону, ова именица није употребљавана у *ППО Књажевац*, *ПППП слива акумулације „Ћелије“* и *РПП Златиборског и Моравичког управног округа*, једва да је споменута у *ПППП слива акумулације „Грлиште“*, незнатно више у *ПППП Националног парка „Ђердап“*, а једини просторни план у ком је присуство ове именице изражено је *ПППП међународног водног пута Е-80 – Дунав (Паневропски коридор VII)*. Имајући у виду да је овим просторним планом обухваћено цело подручје Подунавља у Републици Србији, тј. читав ток Дунава у држави, потпуно је разумљиво да се именица пристаниште користи у свим фазама представљања Плана.

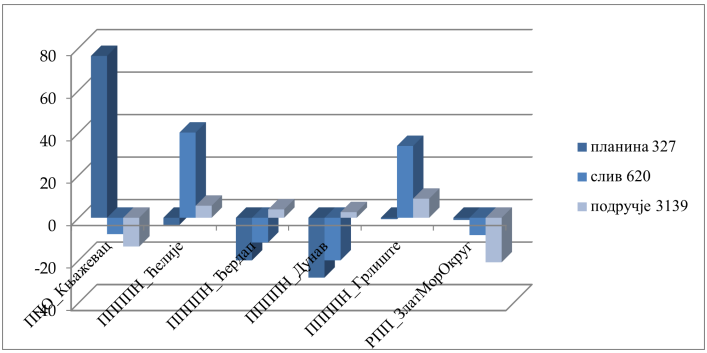
Именица *водоток* (зелена линија) је заступљена у свих шест планова и мада је, нешто чешће њено појављивање у *ППО Књажевац* и *ПППП међународног водног пута Е-80 – Дунав (Паневропски коридор VII)*, употреба ове именице није занемарљива ни у преостала четири плана.

Будући да се цела територија Републике Србије одликује изузетним хидропотенцијалом и развијеним речним системом, оваква учесталост у свих шест текстова корпуса се могла и претпоставити.

Распрострањеност природног и културног наслеђа на територији државе Србије може се потврдити и учесталошћу појављивања именице *споменик* (ружичаста линија) на нивоу целог корпуса, али и његових саставних делова. Именица *експлоатација* (плава линија) је још једна од оних именица чије присуство је приметно у свим текстовима РР-ТХМ корпуса. Наравно, на то је утицала чињеница да Србија поседује извесне резерве различитих врста минералних сировина и да је њихова експлоатација заступљена у свим делова Републике.

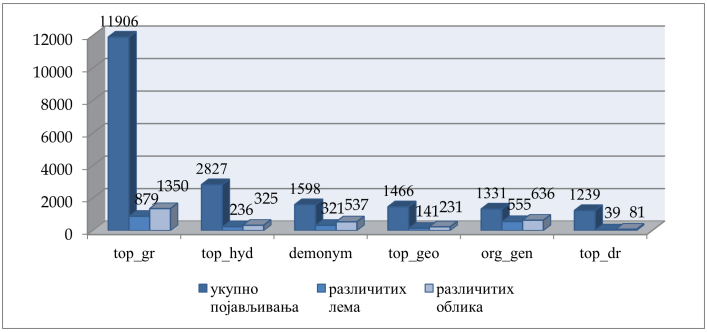
ТХМ алат омогућава још једну значајну врсту анализе која пружа могућност праћења дистрибуције учесталости појављивања одређених речи по партицијама. На слици 6 приказане су *планина*, *слив* и *подручје*. Свака од њих има различит степен заступљености у целом корпусу. Именица *подручје* је са 3.139 појављивања најзаступљенија именица на нивоу читавог корпуса, док су именица *слив* са 620 и *планина* са 327 појављивања знатно мање коришћене у планским документима. Резултати, након обраде у ТХМ окружењу, показују изразито високу фреквентност именице *планина* у *ППО Књажевац*, значајно изнад просека у односу на њену употребу на нивоу целог корпуса, док је њена фреквентност испод просека у преосталих пет просторних планова, а посебно у *ПППРН међународног водног пута Е-80 – Дунав (Паневропски коридор VII)*. Коришћење именице *слив* најизраженије је у *ПППРН слива акумулације „Ћелије“*, а позитиван скор има и у *ПППРН слива акумулације „Грлиште“*. У сва четири преостала плана коришћење ове именице је испод просека, а као и код именице *планина*, то је најизраженије у *ПППРН међународног водног пута Е-80 – Дунав (Паневропски коридор VII)*. Именица *подручје* негативну заступљеност, у односу на цео корпус, има у *РПП Златиборског и Моравичког управног округа* и *ППО Књажевац*. У осталим партицијама заступљеност ове именице је изнад просека у односу на цео текст.

Имајући у виду чињеницу да је у ТХМ платформу увезен корпус који је претходно обележен етикетама именованих ентитета, па је у њему, поред утврђивања фреквенције различитих врста речи, омогућена и анализа којом се утврђује и степен заступљености именованих ентитета, коришћењем опције `Index` и постављањем различитих упита као што су `/region[top_gr]`, `/region[top_hyd]`, `/region[demonym]` и др. Слика 7 приказује шест најфреквентнијих именованих ентитета



Слика 6. Специфичност употребе именица у РРТХМ корпусу.

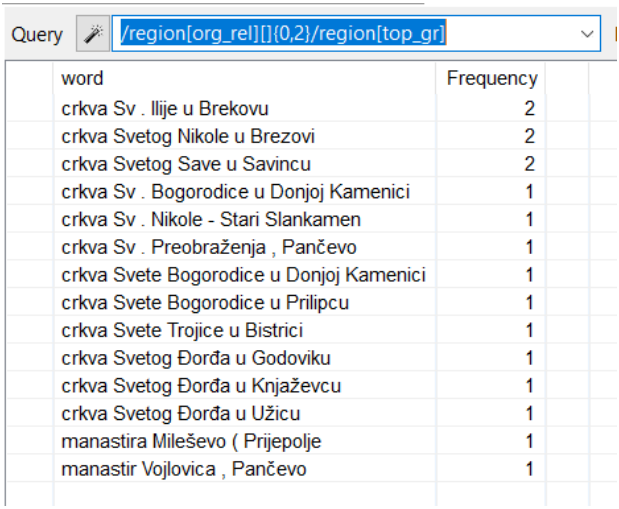
у РРТХМ корпуса. Поред укупне фреквентности приказан је и број различитих лема и различитих облика за сваки именовани ентитет. Убедљиво најзаступљенији тип именованог ентитета је *top_gr*, којим су обележена градска насеља и друга насељена места.



Слика 7. Најзаступљенији именовани ентитети у РРТХМ корпусу.

Упитом `/region[org_rel]` утврђено је да се у целом корпусу појављују 33 различита облика именованих ентитета *org_rel* којим су обележени религијски објекти и да је за њих евидентирано 29 различитих лема са укупно 56 појављивања. Најфреквентнији су *Манастир Милешева* (6), *Црква Св. Петра* (4), *Црква Св. Николе*, *Св. Тројице*, *Св. Ђорђа* и *Манастир Рача* са три појављивања, затим осам религијски објеката који се у корпусу појављују два пута и преостали

именовани ентитети *org_rel* који се јављају само једном. Међутим, у TXM окружењу омогућено је постављање и сложенијих упита. Тако се упитом `/region[org_rel][]0,2/region[top_gr]` утврђује фреквентност именованих ентитета *org_rel* после чијег назива следи и именовани ентитет *top_gr*. Као резултат се добија потпуно другачији редослед у броју појављивања религијских објеката, као што је и приказано на слици 8. У овом случају, као најфреквентније су излистане *Црква Св. Илије у Брекову*, *Светог Николе* у Брезови и *Светог Саве* у Савинцу.


Query /region[org_rel][]{0,2}/region[top_gr]

word	Frequency
crkva Sv . Ilije u Brekovu	2
crkva Svetog Nikole u Brezovi	2
crkva Svetog Save u Savincu	2
crkva Sv . Bogorodice u Donjoj Kamenici	1
crkva Sv . Nikole - Stari Slankamen	1
crkva Sv . Preobraženja , Pančevo	1
crkva Svete Bogorodice u Donjoj Kamenici	1
crkva Svete Bogorodice u Prilipcu	1
crkva Svete Trojice u Bistrici	1
crkva Svetog Đorđa u Godoviku	1
crkva Svetog Đorđa u Knjaževcu	1
crkva Svetog Đorđa u Užicu	1
manastira Mileševo (Prijepolje	1
manastir Vojlovica , Pančevo	1

Слика 8. Учесталост појављивања религијских објеката на основу упита `/region[org_rel][]0,2/region[top_gr]`.

7. Закључак

У раду су приказане разноврсне лингвистичке и статистичке анализе, коришћењем TXM алата, на текстовима просторних планова у којима су претходно обележени именовани ентитети коришћењем SrpNER-а, система за препознавање именованих ентитета у српском језику. Указано је на заједничке карактеристике текстова планске документације, као и на извесне специфичности појединачних просторних планова, који су

као део доменског корпуса просторног планирања уврштени у српски корпус *СрпКор2021*, који тренутно садржи више од 600 милиона речи.

Просторни планови, који представљају главни фокус овог рада, чине специфичну врсту докумената карактеристичних за област просторног планирања. Како би била омогућена детаљнија истраживања ових текстова, креиран је корпус РРТХМ који обухвата шест докумената којима су обрађени различити региони Републике Србије: Златиборски и Моравички управни округ, сливови акумулација „Грлиште“ и „Ћелије“, Национални парк Ђердап, Међународни пловни пут Е-80 – Дунав (Паневропски коридор VII) и општина Књажевац.

Анализом је обухваћено преко 400.000 речи, чија се заступљеност креће од 11,18% у *ППППН слива акумулације „Грлиште“* до 21,42% у *ППППН међународног водног пута Е-80 – Дунав (Паневропски коридор VII)*. Корпус садржи више од пола милиона токена, од којих трећину чине именице.

Треба напоменути да, сем приказаних анализа, ТХМ окружење пружа и друге могућности и да њихова употреба, као и употреба представљених текстометријских приступа зависе од унапред постављених захтева као и од потреба и склоности истраживача. Оно што је неоспорно је да ТХМ платформа омогућава приказ особености корпуса у целини, као и његових саставних делова. Из тог разлога је примена ових метода веома значајна за разнолике квантитативне анализе и утврђивање карактеристичних обележја текстова. Неки од представљених резултата, као што су заступљеност конкретних именица у целом корпусу и учесталост њиховог појављивања у појединачним плановима, могли су се и претпоставити на основу подручја која су планом обухваћена. Међутим, за добијене резултате о заступљености заменица по партицијама корпуса не може се дати неко логичко објашњење само на основу сазнања о обухвату планова, већ је за њихово разумевање неопходно спровести додатна истраживања и анализе.

Захвалност

Ово истраживање је вршено за потребе израде докторске дисертације пријављене на Филолошком факултету Универзитета у Београду под називом „Развој библиотечких и језичких ресурса за организовање и проналажење информација о просторном планирању“. Алати коришћени током истраживања развијани су (SrpNER) или су прилагођени за

различита лингвистичка истраживања за српски језик (TXM) у оквиру Друштва за језичке ресурсе и технологије. Посебну захвалност ауторка дугује професоркама др Цветани Крстев и др Ранки Станковић без чијег личног ангажовања, истраживања представљена овим радом не би била могућа.

Литература

Bjekić, Jovana, Ljiljana Lazarević, Milica Erić, Elena Stojimirović, and Teodora Đokić. 2012. "Razvoj srpske verzije rečnika za automatsku analizu teksta (LIWCser)." *Psihološka istraživanja* XV (1): 85–110. https://doi.org/10.5937/PsIstra1201085B.

Francis, W Nelson. 1975. "Problems of Assembling, Describing, and Computerizing Corpora." *Research Techniques and Prospects. Papers in Southwest English,(1),* 15–38.

Heiden, Serge. 2010. "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation,* 389–398. Tohoku University, Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University. https://aclanthology.org/Y10-1044.

Jaćimović, Jelena. 2019. "Textometric methods and the TXM platform for corpus analysis and visual presentation." *Infotheca - Journal for Digital Humanities* 19 (1): 30–54. https://doi.org/10.18485/infotheca.2019.19.1.2.

Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. "A system for named entity recognition based on local grammars." *Journal of Logic and Computation* 24 (2): 473–489. https://doi.org/10.1093/logcom/exs079.

MacMurray, Erin, and Marguerite. Leenhardt. 2011. "Textometry and Information Discovery: A New Approach to Mining Textual Data on the Web." In *Proceedings of the 2011 International Conferece on Artifical Intelligence, ICAI 2011,* edited by H.R. Arabnia, D. Fuente, E.B. Kozerenko, and J.O. Olivas, II:605–611. Las Vegas: Worldcomp'11.

Maurel, Denis, Nathalie Friburger, and Iris Eshkol-Taravella. 2014. "Enrichment of Renaissance Texts with Proper Names." *Infotheca* 15 (1): 29a–41a. https://infoteka.bg.ac.rs/index.php/en/archives/2014/1/infoteka-15-1-2014-30-41.

McEnery, Tony, Richar Xiao, and Yukio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge. https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/CBLS.htm.

Mehl, Matthias R., and Alastair J. Gill. 2010. "Automatic text analysis." In *Advanced methods for conducting online behavioral research,* edited by S. D. Gosling and J. A. Johnson, 109–127. Washington: American Psychological Association. https://doi.org/10.1037/12076-008.

Pincemin, Bénédicte, Serge Heiden, and Matthieu Decorde. 2020. "Textometry on Audiovisual Corpora Experiments with TXM software." In *JADT 2020 : 15es Journées internationales d'Analyse statistique des Données Textuelles.* http://lexicometrica.univ-paris3.fr/jadt/JADT2020/jadt2020_pdf/PINCEMIN_HEIDEN_DECORDE_JADT2020.pdf.

Popović, Ljubomir, and Duško Vitas. 2003. "Konspekt za izgradnju referentnog korpusa srpskog standardnog jezika." In *Naučni sastanak slavista u Vukove dane*. Beograd, Novi Sad: MSC.

Trtovac, Aleksandra, Vasilije Milnović, and Cvetana Krstev. 2021. "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.

Utvić, Miloš. 2013. "Izgradnja referentnog korpusa savremenog srpskog jezika." PhD diss., Univerzitet u Beogradu, Filološki fakulte. https://nardus.mpn.gov.rs/handle/123456789/4091.

Vitas, Duško. 2019. "Food as Text." *Infotheca - Journal For Digital Humanities* 19 (2): 139–161. https://doi.org/10.18485/infotheca.2019.19.2.7.

Vitas, Duško. 2022. "From Onions to Champagne – Food and Drink in the SrpELTeC Corpus." *Infotheca - Journal for Digital Humanities* 21 (2): 88–118. https://doi.org/10.18485/infotheca.2021.21.2.5.

Васиљевић, Небојша М. 2014. "Аутоматска обрада правних текстова на српском језику." PhD diss., Универзитет у Београду, Филолошки факултет. https://nardus.mpn.gov.rs/handle/123456789/4091.

Витас, Душко, and Цветана Крстев. 2016. “Оглед из гастрономатике [Experiments in gastronomatics].” In *Теме језикословне у србистици кроз дијахронију и синхронију [Linguistic topics in Serbian through diachrony and synchrony],* edited by Јасмина Дражић, Исидора Бјелаковић, and Дејан Средојевић, 1–10. Novi Sad: Филозофски факултет.

Витас, Душко М. 2018. “Храна из нежељене поште : (анатомија језика брзе хране [Spam Food: (Anatomy of Fast Food Language)]).” In *Српски језик и његови ресурси: теорија, опис и примене Научни састанак слависта у Вукове дане,* edited by Божо Ћорић and Александар Милановић, 21–35. Београд: Међународни славистички центар, Филолошки факултет, Универзитет у Београду. https://doi.org/10.18485/msc.2018.47.3.ch2.

Обрадовић, Иван, Александра Томашевић, Ранка Станковић, and Лазић Биљана. 2017. “Увођење доменских и семантичких маркера за област рударства у српске електронске речнике.” In *Научни састанак слависта у Вукове дане - Српски језик и његови ресурси: теорија, опис и примене,* edited by Рајна Драгићевић and Александар Милановић, 147–158. Београд: Међународни славистички центар на Филолошком факултету. https://doi.org/10.18485/msc.2017.46.3.ch10.

Службени гласник РС. 52/2021. *Закон о планирању и изградњи.* Бр, 72/2009, 81/2009 - испр., 64/2010 - одлука УС, 24/2011, 121/2012, 42/2013 - одлука УС, 50/2013 - одлука УС, 98/2013 - одлука УС, 132/2014, 145/2014, 83/2018, 31/2019, 37/2019 - др. закон, 9/2020 и 52/2021), 52/2021.

Службени гласник РС. 32/2019. *Правилнику о садржини, начину и поступку израде докумената просторног и урбанистичког планирања.* Бр. 32/2019), 32/2019.

Тртовац, Александра. 2017. “Аутоматизација библиотека у Србији - историјски преглед.” *Библиотекар,* no. 2, 101–114. https://bds.rs/wp-content/uploads/2018/01/bibliotekar-2-2017-7.pdf.