Паралелизовани корпус за домен менаџмента: поступак припреме и могућности примене

УДК 81'322.2
УДК 81'373:005

САЖЕТАК: У раду је представљен паралелизовани енглеско-српски корпус за домен менаџмента као једна од већих колекција доступна у оквиру алата за претраживање паралелизованих текстова Библиша. Поред описа садржаја корпуса и критеријума за његов одабир, као и самог процеса компилације и припреме текстова за корпус, у раду су кроз примере анализе корпуса приказане и могућности његове примене у различитим врстама лингвистичких студија, првенствено у терминолошким, компаративним и контрастивним истраживањима језика струке, истраживањима специјализованог и академског дискурса, у процесу превођења, али и у настави енглеског језика струке.

КЉУЧНЕ РЕЧИ: паралелизовани корпус, доменски корпус, терминологија, терминолошка јединица, менаџмент

РАД ПРИМЉЕН: 29. децембар 2017.
РАД ПРИХВАЋЕН: 15. новембар 2018.

Јелена Анђелковић
plecasj@fon.bg.ac.rs
Факултет организационих наука

Даница Сеничић
danica.senicic@gmail.com

Ранка Станковић
ranka.stankovic@rgf.bg.ac.rs
Рударско-геолошки факултет

Универзитет у Београду
Србија

1. Увод

У ранијој лингвистичкој литератури под корпусом се подразумевала било која колекција текстова за коју се претпоставља да представља дати језик, дијалекат или други подсистем неког језика, а који се користи ради лингвистичке анализе (Pearson, 1998, 42). У новије време, са развојем рачунарске лингвистике и алата и система за обраду природних језика, мења се и дефиниција корпуса. Данас се

под корпусом подразумевају *машински читљиве колекције текстова* (McEnery and Wilson, 2001, 177), односно *колекције делова језичког текста у електронском облику, које су одабране на основу екстерних критеријума да представе, колико је то могуће, језик или језички варијетет као извор података за језичка истраживања* (Sinclair, 2005) Макенери, Ксао и Тоно (McEnery et al., 2006, 13) под корпусом подразумевају колекције *машински читљивих, аутентичних текстова* (укључујући и транскрипте писаног говора) који су *узорковани* тако да буду *репрезентативни* за одређени језик или језички варијетет. Постоје различите поделе и врсте језичких корпуса, а сам одабир врсте корпуса зависи од његове намене, то јест, врсте лингвистичке анализе која се спроводи. Док су корпуси општег језика (енгл. *general language corpora*) колекције писаног и / или говорног језика који представљају (или би требало да представе) језик као целину, специјализовани корпуси, корпуси језика за посебне намене или доменски корпуси (енгл. *specialized corpora / corpora for specific purposes / domain corpora*) јесу електронски доступне колекције текстова које представљају одређену специјализовану област комуникације, односно које су репрезентативне за посебне намене одређеног језика или посебан домен језичке употребе. Често су специјализовани корпуси састављени од текстова који припадају једном жанру (енгл. *genre-specific*), односно репрезентативни су само за одређену научну и стручну област, дисциплину или домен (енгл. *domain-specific или discipline-specific*).

Од посебног значаја за терминолошка, контрастивна и компаративна истраживања језика су паралелни и паралелизовани доменски корпуси (корпуси за посебне намене). Паралелни (енгл. *parallel*) су двојезични или вишејезични корпуси који садрже преводно еквивалентне текстове на два или више језика (Tognini-Bonelli, 2001, 6)(Pearson, 1998, 47). У неким случајевима, паралелни корпуси могу садржати текстове на само једном језику, тј. парове еквивалентних различитих превода на исти језик. Код паралелизованих корпуса (енгл. *aligned corpora*), са друге стране, паралелни (преводно еквивалентни) текстови су и паралелизовани, односно поравнати на нивоу пасуса, реченице или појединачних речи.

Основна карактеристика паралелизованих корпуса текстова за посебне намене јесте да су они примарни језички ресурси за вишејезичну обраду у оквиру рачунарске лингвистике. Ови корпуси омогућавају систематичну обраду велике количине терминолошких информација, аутоматску или полуаутоматску екстракцију термина и

њихових еквивалената на страном језику или језицима. Лексичко знање стечено коришћењем паралелизованих корпуса је стога неопходно за развој софтверских система и алата за аутоматску обраду природних језика (енгл. *natural language processing* – NLP), као што су системи за машинско превођење и генерисање билингвалних електронских терминолошких глосара, речника, лексикона и терминолошких база података (Véronis, 2013, 238).

Иако број и доступност паралелизованих колекција текстова који обухватају и српски језик нису на завидном нивоу, највише због чињенице да је често тешко пронаћи адекватне преводно еквивалентне текстове, у наставку поглавља представићемо неколико постојећих.

Први радови везани за паралелизоване текстове где је један од језика српски везују се за почетак деведесетих година, где се о конкорданцијама паралелизованих текстова говори у (Krstev C., 1994), а касније се паралелизује Платонова *Република* на 17 језика (Vitas, 1998b) и превод Орвеловог романа „1984" поравнат на седам језика(Vitas, 1998a).

Евротека[1] је двојезични, енглеско-српски, корпус правних текстова или делова правних текстова настао током процеса превођења правних тековина Европске уније на српски језик, за који је задужено Министарство за европске интеграције. Евротеку од 2009. године одржава и надограђује Сектор за комуникације овог министарства. За превођење правних текстова, те као и за креирање овог корпуса, користи се програм SDL Trados.[2], тј. његов алат Translator's Workbench

По броју језика који обухвата, издваја се и вишејезични паралелизовани корпус „MULTEXT-East "1984" annotated corpus 4.0" који је расположив на CLARIN.SI репозиторијуму, а развијен у оквиру пројекта MULTEXT-East – *Multilingual Text Tools and Corpora for Eastern and Central European Languages* (Erjavec and Ide, 1998). Корпус MULTEXT-East "1984" састоји се од енглеског оригинала романа „1984" аутора Џорџа Орвела и његовог превода на дванаест источно и централноевропских језика (бугарски, чешки, енглески, естонски, мађарски, македонски, персијски, пољски, румунски, српски, словачки и словеначки). Текстови су поравнати на нивоу реченице, а леме и морфосинтаксички описи су ручно валидирани (Erjavec et al., 2010). Паралелизовани корпус српског и енглеског језика srenWaC[3] (Ljubešić

[1] Евротека (на вебу)
[2] Trados (на вебу)
[3] srenWaC (на вебу)

et al., 2016) састоји се од електронских текстова различите тематике преузетих са домена .rs, а генерисан је аутоматски уз коришћење алата Spidextor.[4]

(Pazienza et al., 2005), као и други, описује различите приступе за екстракцију термина: 1) коришћење статистичких мера при одабиру добрих термина са листе кандидата, 2) идентификовање и препознавање израза користећи само лингвистичка и филтрирање специфичних синтактичких терминолошких образаца и 3) хибридне приступе који покушавају да користе ова два приступа заједно, узимајући у обзир синтаксичка својства и статистичке мере за препознавање термина. (Siddiqi and Sharan, 2015) наводи још два приступа: 4) коришћење метода машинског учења и 5) доменски специфичних ресурса знања (на пример онтологија) при екстракцији термина. У овом раду је коришћен хибридни приступ који комбинује препознавање синтаксичких образаца и статистичке мере, а који је описан у (Stanković et al., 2016a).

Паралелизоване и електронски доступне корпусе на енглеском и српском језику развија Група за језичке технологије на Математичком факултету[5] и Друштво за језичке ресурсе и технологије (JePTex)[6] чији су чланови углавном са Универзитета у Београду. Један од паралелизованих корпуса јесте енглеско-српски корпус *СрпЕнгКор*, који се састоји од текстова различитих жанрова (књижевност, новинарство, научни и стручни текстови из области права, медицине, образовања итд.) који су сегментирани и поравнати (у највећем броју случајева) на нивоу реченице.[7]

У оквиру Друштва Jertex је развијен је алат за претраживање паралелизованих колекција стручних текстова – Библиша.[8] Детаљан опис имплементације и начина употребе је представљен је у (Stanković et al., 2016b). Тренутно је путем Библише доступно неколико колекција стручних текстова, и то за научне и стручне области библиотекарства и информатике, рударства и геологије, стоматологије, архитектуре и урбанизма и других.

У следећим поглављима која следе описан је поступак компилације и обраде једне од већих колекција паралелизованих текстова доступних путем алата Библиша, колекције текстова за домен менаџмента.

[4] Spidextor (на вебу)
[5] Група за језичке технологије на Математичком факултету (на вебу)
[6] Друштво за језичке ресурсе и технологије (на вебу)
[7] СрпЕнгКор (на вебу)
[8] Библиша (на вебу)

Поред процеса израде, у раду су истакнуте и могућности примене ове колекције у различитим врстама лингвистичким истраживањима. Акценат није стављен на примарну намену паралелизованих доменских корпуса (у рачунарској лингвистици и менаџменту терминологије), већ на мање истражене могућности примене, и то пре свега у примењеној лингвистици и компаративном и контрастивном проучавању терминологије, језика струке и специјализованог дискурса.

2. Паралелизовани корпус за домен менаџмента

2.1 Садржај корпуса

Паралелизовани енглеско-српски специјализовани корпус за домен менаџмента састоји се од текстова који припадају жанру научног рада, а публиковани су у међународном научном и стручном часопису *Менаџмент: Часопис за теорију и праксу менаџмента (енгл. Management: Journal for theory and practice of management)*.[9] Реч је, дакле, о корпусу који је специфичан како за домен, односно научну и стручну област менаџмента (енгл. *domain-specific corpus*), тако и о корпусу који је специфичан за жанр научног рада (енгл. *genre-specific corpus*).

Корпус садржи 17 бројева часописа *Менаџмент* објављених између 2008. и 2012. године[10], са укупно 181 научним радом, око 30.000 реченица, и преко 600.000 речи по језику (прецизније, 611.651 реч на српском језику). Детаљнији приказ садржаја корпуса представљен је у табели 1.

Међународни научни и стручни часопис *Менаџмент: Часопис за теорију и праксу менаџмента* квартално издаје Факултет организационих наука Универзитета у Београду као водећа академска институција за ову област у Србији. Часопис има циљ да "омогући

[9] У мају 2017. године, часопис је променио назив из Management: Journal for Theory and Practice of Management у Management: Journal of Sustainable Business and Management Solutions in Emerging Economies.

[10] Часопис *Менаџмент* од 1996. (када је основан) до 2008. (закључно са бројем 46) излази само на српском језику. Од 2008. до 2013. године сваки број са идентичним садржајем се штампа и на енглеском и на српском језику, док се од 2013. године, почевши од броја 66, штампа искључиво на енглеском језику. Чињеница да су паралелни текстови доступни само за период између 2008. и 2012. године ограничила је и величину нашег корпуса на 17 бројева: од броја 47–48 до броја 65.

размену релевантних информација и комуникацију између научника, истраживача, менаџера и појединаца из различитих пословних области, који долазе са универзитета, института, компанија и јавних услужних предузећа”[11]. У време када су издати бројеви који су обухваћени нашим корпусом, часопис *Менаџмент* се налазио у категорији М51 по категоризацији Министарства просвете, науке и технолошког развоја Републике Србије. Сви радови на српском и енглеском језику, обухваћени корпусом, доступни су на сајту часописа.[12]

издање (број/година)	број радова (по језику)	број реченица (српски језик)
47-48/2008	12	2.187
49-50/2008	14	2.097
51/2009	9	1.503
52/2009	9	1.575
53/2009	10	1.194
54/2010	10	1.817
55/2010	10	1.750
56/2010	10	1.648
57/2010	10	1.502
58/2011	10	1.475
59/2011	10	1.501
60/2011	11	1.426
61/2011	14	2.301
62/2012	12	2.297
63/2012	10	1.815
64/2012	10	1.655
65/2012	10	1.583
$\sum$	181	29.326

Табела 1. Садржај корпуса

[11] Менаџмент: Часопис за теорију и праксу менаџмента (на вебу)

[12] Менаџмент, архива (на вебу). И поред јавне доступности, за њихово коришћење при изради паралелизованог корпуса добијена је дозвола тадашњег уредника часописа Менаџмент, проф. др Александра Марковића.

2.2 Предности и ограничења корпуса

Ауторство. Радови обухваћени корпусом дела су појединачних аутора, али и композитни текстови од више аутора. Аутори текстова су истраживачи у области менаџмента и представници академске заједнице, али и представници привреде из наше земље, региона и света. Увидом у метаподатке, уочили смо да су у 21 раду (11,.6%) од укупно 181 рада аутори (или аутор) искључиво страни држављани (ван простора бивше Југославије), а преосталих 160 радова (88,4%) написали су или домаћи аутори или аутори из региона, самостално или у коауторству.

На основу доступних метаподатака и информација са веб-сајта часописа није могуће тачно утврдити који језик су аутори користили за писање изворног текста, као ни да ли је текст преводио сам аутор или стручни преводилац; претпоставка је да су радови страних аутора изворно написани на енглеском језику, а затим преведени на српски, док су радови домаћих и аутора из региона највероватније изворно написани на српском језику. Иако овакав састав текстова у корпусу може значајно утицати на квалитет, односно прецизност и једнозначност терминологије која се у њима налази, он потенцијално може дати бољу слику о терминолошкој и другој језичкој варијацији условљеној прагматичким или социолингвистичким факторима. Из овог разлога није вршена селекција текстова на основу ауторства, односно на основу изворног језика текста.

Прагматички фактори за одабир текстова. Приликом одабира текстова за паралелизовани корпус менаџмента, водили смо се првенствено прагматичким факторима: доступношћу адекватних преводно еквивалентних текстова за домен менаџмента у електронском облику, као и циљевима које желимо да постигнемо израдом и коришћењем оваквог корпуса, а то су лингвистичка и терминолошка истраживања у овој области и сродном научно-стручним областима. Оба ова фактора са собом доносе одређене предности и ограничења.

Величина корпуса. Величина представљеног корпуса (око 600.000 речи по језику) последица је доступности преведених научних радова из домена менаџмента на српском и енглеском језику. Иако се аутори који се баве корпусном лингвистиком не слажу у потпуности око оптималне величине корпуса (Roe, 1977; Fang, 1993; Gledhill, 2000), односно око идеалне величине специјализованог корпуса (Flowerdew, 2004, 18), сматрамо да је овде представљен корпус менаџмента адекватан

за многе, али не и за све врсте лингвистичких и терминолошких анализа у овом домену и сродним стручним доменима, због чега би за такве анализе дати корпус требало додатно проширити.

Жанр. Са фукционално-стилистичког гледишта, наш паралелизовани доменски корпус у потпуности припада једном текстуалном жанру језика струке, тачније жанру истраживачког научног рада. Припадност текстова једном жанру доприноси хомогености корпуса, будући да су приступ тематици и ниво специјалности уједначени, односно да регистарски, стилски и жанровски не постоје велике варијације између текстова. Иако је код општих корпуса препоручљива разноврсност жанрова, у терминолошким истраживањима и истраживањима језика струке сасвим су прихватљиви и веома често коришћени корпуси засновани на једном жанру. Имајући у виду да жанр истраживачког научног рада карактеришу једнообразност (одсуство разговорне и дијалекатске лексике), лексичка и терминолошка густина, као и информативност, логичност, целовитост и прецизност текста, сматрамо да је одабир овог жанра адекватан за терминолошку анализу, док је за одређене врсте лингвистичких истраживања потребно укључити и доступне текстове других жанрова.

Компилација и припрема корпуса. Процес припреме корпуса за анализу путем одговарајућих софтверских алата састојао се од неколико фаза: припрема и екстракција текста, поравнања текстова на нивоу пасуса и реченица, креирање докумената у форматима TEI/XML и TMX, снабдевање метаподацима и укључивање у базу.

2.3 Припрема и екстракција текста

Након одабира текстова уврштених у наш паралелизовани корпус за домен менаџмента и жанр научног рада, сви текстови су појединачно преузети у формату PDF са веб-странице часописа *Менаџмент*. Будући да је формат обичног текста (*.txt*) стандардан за софтвер намењен за обраду и анализу корпуса, све електронске текстове смо конвертовали у овај формат користећи програм Abby PDF Transformer. Ради лакше идентификације текстова, обележили смо их одговарајућим ознакама (на пример, називом Mng52_01-sr што се односи на датотеку која садржи први чланак часописа из броја 52 на српском језику). Проблеми на који смо повремено наилазили приликом припреме корпуса јесу да конверзија докумената из расположивог формата PDF у формат

обичног текста није препознала неке од карактера који се појављују у корпусу, пре свега дијакритичке ознаке у српским текстовима (сви текстови на српском језику писани су латиничним писмом), као и да се две колоне са текстом у формату PDF приликом конверзије у формат обичног текста често спајају у једну. Како би се умањиле грешке, поједине датотеке су конвертоване у формат програма Microsoft Word (*.doc*), затим кориговане и снимљене у формату обичног текста.

Након конвертовања текстова одстрањени су сви елементи нерелевантни за лингвистичку анализу, као што су табеле, графикони, формуле, литература, референце, садржај, заглавља, подножја и сл.

Овако припремљени текстови погодни су за лингвистичку анализу неанотираних (или „сирових") корпуса (енгл. *raw corpora*) на једном језику коришћењем јавно доступних и често бесплатних програма за анализу корпуса какви су, на пример, WordSmith[13] или AntConc.[14] За анализирање паралелних корпуса, међутим, потребно је извршити поравнање текста на неком од структурних нивоа (одељци, пасуси, реченице или речи), а за адекватно анализираање високо флективног језика какав је српски, неопходна је и морфосинтаксичка анотација овог дела корпуса.

Поравнање текстова на нивоу пасуса. После припреме појединачних текстова, парови одговарајућих изворних текстова (на српском језику) и циљних текстова (на енглеском језику) поравнати су на нивоу пасуса (нпр. Mng52_01-sr и Mng52_01-en). Овај процес обављен је у програму Notepad++, поређењем њихових садржаја и поравнањем пасуса, тако да се пасуси српског оригинала и пасуси енглеског превода, сваки у својој датотеци, морају налазити у истом реду. Приликом овог процеса, наишли смо на многобројне проблеме, као што су непревеђени, неадекватно преведени, изостављени или измештени пасуси. Ове проблеме решавали смо или проналаском делова пасуса који недостају у оригиналним документима у формату PDF, или простим одстрањивањем пасуса или делова пасуса за које не постоји еквивалент у другом језику. Основни разлог за овако захтеван и дуготрајан поступак јесте смањење тзв. буке/шума (енгл. *noise*) током корпусне анализе.

[13] WordSmith (на вебу)
[14] AntConc (на вебу)

Поравнање текстова на нивоу реченица и креирање XML документа. Трећи корак у припреми паралелизованог корпуса је креирање документа у формату XML (*eXtensible Markup Language*) поравнатих на нивоу реченица. Текстови у формату XML, поред основног текста могу да садрже и додатне интерпретативне лингвистичке податке, као што су информације о структури текста, информације о ауторима и верзијама текста, као и лингвистичку анотацију текста засновану на процесима токенизације, препознавање граница реченица, морфолошке анализе (укључујући лематизацију и анотацију врсте речи, енгл. *Part-of-Speech tagging* или *PoS tagging*), као и плитка синтаксичка анализа (енгл. *shallow parsing*).

Пре поравнања текстова на нивоу реченице, неопходно је сегментирати их на том нивоу. Овај корак је аутоматски обављен помоћу система Unitex (Paumier, 2002) који се користи за креирање и претраживање корпуса. Реченице су сегментиране помоћу локалних граматика, форми за опис и препознавање лингвистичких феномена у тексту. Локалне граматике су имплементиране као коначни аутомати и трансдуктори којима корисник манипулише помоћу њиховог графичког приказа, тј. графова. Локалне граматике за препознавање краја реченице прилагођене су правопису српског језика и саставни су део лингвистичких ресурса за српски језик који се дистрибуирају са самим програмом Unitex. Резултат сегментације на реченице, излазни текст, садржи симбол {S} као граничник реченице, који се даље конвертује у одговарајуће етикете формата TEI/XML за означавање реченица (сегмената) у складу са Смерницама TEI P5.[15],незваничног стандарда за кодирање текста који се највише користи

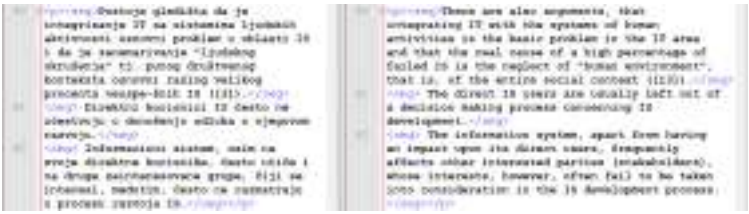
Обележавање текста (енгл. *markup*) на структурном нивоу пасуса и реченица олакшава процес упаривања изворних и циљних текстова.

У овом истраживању структурни нивои текста су обележени (Слика 1) етикетама (цео документ), <body> (садржина целог документа), <head> (наслови), <p> (пасуси) и <seg> (реченице).

Претраживање корпуса и генерисање конкорданција је подржано је морфолошком и семантичком експанзијом упита (Stanković et al., 2016b), тако да лематизација није била потребна.

Креирање TMX документа. Наредни корак у припреми корпуса јесте креирање документа у формату TMX (Savourel, 2004). Ради се XML

[15] TEI P5 (на вебу)



Слика 1. Пример припремљеног паралелног српско-енглеског текста у формату TEI/XML

.

спецификацији за размену података из преводилачких меморија (енгл. *Translation Memory eXchange*) и користи се често у алатима намењеним за превођење уз коришћење рачунара (енгл. computer-aided translation – CAT). Документи у формату TMX креирани су у програму ACIDE, интегрисаном окружењу за припрему паралелизованих корпуса које је развило Друштво за језичке ресурсе и технологије у Београду (Obradović et al., 2008, 563). ACIDE нуди графичко сучеље (енгл. *interface*) за поравнање и визуализацију поравнатих текстова, док се за само поравнање користе програмски пакети XAlign и Concordancier развијени у лабораторији LORIA.[16] у Француској (Bonhomme et al., 2001) Пример упарене реченице у TMX формату је приказан је на слици 2.

Слика 2. Пример преводилачке јединице са енглеском и одговарајућом српском реченицом у формату TMX

.

[16] LORIA (на вебу)

Напослетку, текстови у формату TMX укључени су у базу података креирану у оквиру платформе MongoDB.[17] Укључивањем у базу коришћењем алата Библиша[18] паралелни текстови су доступни за претраживање и даљу анализу.

Снабдевање корпуса метаподацима и укључивање у базу. Претходно припремљени TMX документи су укључени у Библишу (Stanković et al., 2016a) као седма колекција у оквиру базе паралелизованих текстова. Сама колекција подељена је на 17 потколекција које садрже текстове одговарајућих 17 бројева часописа *Менаџмент*. У оквиру сваке потколекције има између 9 и 12 докумената, односно радова. Свака потколекција (свеска часописа) и чланак имају свој јединствен идентификациони број. На пример, идентификациони број потколекције је 7.2011.59, што значи да је у питању потколекција (свеска часописа) седме колекције (часопис *Менаџмент*), да је свеска издата 2011. године и да је број свеске 59, док 7.2011.59.1 представља први чланак те свеске. Сваки од докумената у оквиру потколекција снабдевен је библиографским метаподацима (на енглеском и српском језику) који се односе на наслове, ауторе радова, њихове афилијације и контакте (имејл адресе), хипервезе ка чланцима у формату PDF, апстракте и кључне речи на оба језика, као и метаподатке од којих се састоји идентификациони број (редни број чланка, број часописа, година издања).

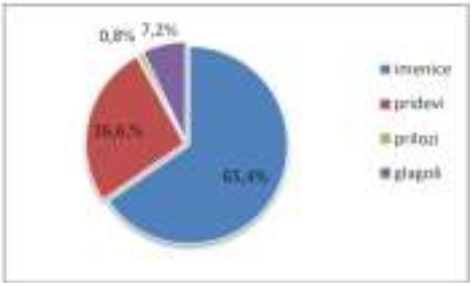
3. Анализа корпуса и могућности примене

Једна од важних предности које двојезични (у нашем случају паралелизовани) доменски корпус нуди у односу на једнојезичне јесте могућност анализе делова корпуса на сваком од два језика појединачно, али и могућност интерлингвалне компаративне и контрастивне анализе. У наредним редовима представићемо неке од основних могућности примене нашег доменског корпуса, прво путем екстракције једночланих и вишечланих терминолошких јединица у српском делу корпуса, а затим и екстракције парова преводних еквивалената на два језика.

[17] MongoDB (на вебу)
[18] Библиша (на вебу)

3.1 Екстракција кључних речи на српском језику

Након обављеног процеса компилације и обраде текста паралелизованог корпуса менаџмента, извршена је екстракција лексичких јединица на српском језику које се издвајају по критеријуму кључности (енгл. *keyness*), односно оних које се значајно фреквентније појављују у доменском корпусу него у референтном корпусу, тј. *Корпусу савременог српског језика СрпКор 2003*[19] (122 милиона речи), насталом на Математичком факултету Универзитета у Београду (Utvić, 2011, 36a-47a). Овај процес обављен је помоћу алата за развој и управљање лексичким ресурсима LeXimir који је развио JePTex, Друштво за језичке ресурсе и технологије (Stanković et al., 2011, 77-84). Анализом првих 500 кључних лексичких јединица издвојених из доменског корпуса установљено је да међу њима има 327 именица, 133 придева, 36 глагола и 4 прилога (Слика 3).



Слика 3. Дистрибуција врста речи у скупу од 500 изабраних кључних лексичких јединица

.

Детаљнијом анализом и филтрирањем добијене листе, издвојили смо кључне именице (Табела 2), придеве (Табела 3) и глаголе (Табела 4). Параметар кључност се рачуна као однос релативне фреквенције (изражене у милионитим деловима целине, тј. у ppm као јединицама мере) у доменском корпусу (менаџмента у овом случају) и одговарајуће релативне фреквенције у референтном корпусу опште намене, при чему су оба броја пре дељења увећана за један. Ознаке RFr и RFd

[19] СрпКор 2003 (на вебу)

редом представљају релативне фреквенције (у ppm) у референтном и доменском корпусу, док су AFr и Afd ознаке одговарајућих апсолутних фреквенција. Дакле, кључност рангира леме према односу фреквенција у доменском и референтном корпусу, а не само према фреквенцијама у доменском корпусу. Леме које се чешће појављују у доменском корпусу него у референтном, узимајући у обзир величине тих корпуса, биће при врху табеле и вероватно су оне термини домена.

лема	кључност	RFr	RFd	AFr	Afd
индикатор	121,398	3,035	488,841	67	299
менаџмент	115,894	16,713	2051,824	369	1255
рачуноводство	115,656	3,306	497,015	73	304
бренд	107,933	1,857	307,365	41	188
портфолио	84,509	1,314	194,555	29	119
интернет	82,145	4,167	423,444	92	259
профитабилност	81,684	1,314	188,016	29	115
перформанса	81,194	4,892	477,396	108	292
подсистем	79,44	0,906	150,413	20	92
сертификација	69,345	1,042	140,603	23	86
конкурентност	67,896	4,529	374,397	100	229
евалуација	65,029	0,725	111,175	16	68
управљање	63,609	38,726	2525,95	855	1545
преференције	62,063	0,544	94,825	12	58
методологија	59,253	6,522	444,698	144	272

Табела 2. Кључне именице у доменском корпусу

Табеле 2, 3 и 4 показују да кључне терминолошке јединице не морају уједно бити и најфреквентније. Терминолошке јединице које су истовремено и изразито фреквентне и кључне за домен менаџмента истовремено су и најважније терминолошке јединице за ову област: *менаџмент* (кључност = 115,894, Afd = 1255) и *управљање* (кључност = 63,609, Afd = 1545), али и терминолошке јединице чија се синонимна употреба често доводи у питање (о овоме ће бити више речи у одељку 3.3). Са друге стране, на пример, глаголи попут *операционализовати* (кључност = 15,316, Afd = 13), *инкорпорирати* (кључност = 15,039, Afd

лема	кључност	RFr	RFd	AFr	Afd
пројектни	215,922	3,578	987,491	79	604
корпоративан	146,506	2,31	483,936	51	296
одржив	123,302	3,397	541,158	75	331
мотивациони	95,619	0,498	142,238	11	87
ефективан	85,99	2,491	299,19	55	183
рачуноводствени	84,553	2,763	317,174	61	194
екстерни	83,349	2,582	297,555	57	182
иновативан	83,031	1,178	179,841	26	110
управљачки	76,955	4,303	407,095	95	249
стратегијски	68,548	5,979	477,396	132	292
организациони	68,428	20,518	1471,427	453	900
проблемски	64,179	2,582	228,889	57	140
конкурентски	62,603	5,571	410,365	123	251
менаџерски	61,658	2,808	233,793	62	143

Табела 3. Кључни придеви у доменском корпусу

лема	кључност	RFr	RFd	AFr	Afd
фокусирати	47,821	3,397	209,27	75	128
имплементирати	40,691	2,038	122,619	45	75
генерисати	33,435	2,355	111,175	52	68
израчунавати	31,611	1,359	73,571	30	45
класификовати	20,779	2,038	62,127	45	38
базирати	19,884	10,644	230,524	235	141
рангирати	19,208	2,627	68,667	58	42
позиционирати	17,702	0,996	34,333	22	21
дефинисати	17,287	53,628	943,348	1184	577
формализовати	17,149	0,679	27,794	15	17
операционализовати	15,316	0,453	21,254	10	13
обухватати	15,087	36,778	568,952	812	348
креирати	15,048	14,494	232,159	320	142
инкорпорирати	15,039	1,132	31,063	25	19

Табела 4. Кључни глаголи у доменском корпусу

= 19) и *позиционирати* (кључност = 17,702, Afd = 21) имају релативно висок параметар кључности иако нису фреквентни у доменском корпусу.

3.2 Екстракција вишечланих термина на српском језику

Имајући у виду да су терминолошке јединице најчешће вишечлане (Krstev et al., 2015), а не једночлане, издвојена листа једночланих терминолошких јединица није довољна за терминолошку анализу. Стога је потребно укључити и вишечлане терминолошке јединице до којих се долази аутоматском екстракцијом помоћу синтаксичких графова развијених у оквиру програма Unitex. Коришћењем алата Лексимир екстрахују се кандидати према дефинисаним синтаксичким обрасцима (Stanković et al., 2016b)(Krstev et al., 2015), препознате лексичке јединице се лематизују, како би се објединила појављивања исте вишечлане леме. Узимајући у обзир вишезначност овакве лематизације, имплементиране су различите стратегије разрешавања вишезначности. Потом се за сваког терминолошког кандидата из генерисане листе, рачунају различите статистичке мере на основу којих се рангирају кандидати, након чега следи евалуација и одређивање термина који ће ући у терминолошки речник (Stanković et al., 2016b).

Увидом у двадесет најфреквентнијих вишечланих терминолошких јединица на српском језику у доменском корпусу за менаџмент (Табела 5) можемо закључити да су међу екстрахованим вишечланим терминима најфреквентније именичке синтагме са придевом као зависним чланом у препозицији (grf01, образац придев именица, у ознаци AXN, код кога се придев слаже у роду, броју и падежу са именицом), као у примерима *људски ресурси, информациони систем, електронско пословање, управљачко рачуноводство* итд.

Екстракција терминолошких јединица из српског дела нашег доменског корпуса и њихова детаљнија анализа може бити од великог значаја за проучавање семантичких, прагматичких и социолингвистичких аспеката терминолошких јединица менаџмента, затим контрастивних и компаративних проучавања стручне терминологије у домену менаџмента (посебно у односу на енглески језик), али може представљати и допринос терминолошкој језичкој политици, планирању, систематизацији и стандардизацији терминологије овог домена. Иако примери наведени у овом и претходном одељку указују првенствено на примену корпуса у оквиру терминолошких студија, доменски корпус за менаџмент је могуће

користити и у проучавању српског језика струке и специјализованог дискурса менаџмента, затим у анализи академског писања и анализи жанра научног рада на српском језику, али и другим врстама лингвистичких истраживања, што је у плану за будући истраживачи рад.

Graph	Pattern	лема	фрек.	бр. речи	рел. фрек.
grf01	AXN	људски ресурси	258	2	421.81
grf01	AXN	организациона наука	207	2	338.43
grf01	AXN	информациони систем	190	2	310.63
grf01	AXN	управни одбор	181	2	295.92
grf01	AXN	електронско пословање	162	2	264.86
grf01	AXN	пројектно финансирање	136	2	222.35
grf01	AXN	пројектни менаџмент	128	2	209.27
grf01	AXN	конкурентска предност	127	2	207.63
grf01	AXN	управљачко рачуноводство	122	2	199.46
grf01	AXN	пројектни менаџер	100	2	163.49
grf01	AXN	финансијски извештај	98	2	160.22
grf03	N2X	реализација пројекта	97	2	158.59
grf03	N2X	управљање ризиком	97	2	158.59
grf01	AXN	информациона технологија	95	2	155.32
grf01	AXN	каматна стопа	95	2	155.32
grf01	AXN	енергетска ефикасност	93	2	152.05
grf03	N2X	процес управљања	92	2	150.41
grf10	2XAXN	јавно-приватно партнерство	90	3	147.14
grf01	AXN	економска криза	86	2	140.6
grf01	AXN	финансијско средство	77	2	125.89

Табела 5. Најфреквентније вишечлане терминолошке јединице у доменском корпусу

3.3 Екстракција преводних еквивалената на енглеском језику

У претходна два одељка представљени су основни резултати анализе српског дела нашег доменског корпуса. Прави потенцијал овог паралелизованог корпуса, међутим, може се искористити

софтверским алатом Библиша. Комплетним текстовима који припадају представљеном корпусу могу приступити регистровани корисници путем веб-сајта http://jerteh.rs/biblisha/. Ограничено коришћење могуће је и без регистрације и овлашћења – тада је доступно првих девет реченица сваког текста приликом прелиставања докумената или 30 конкорданција при претраживању.

Уношењем упита – једночланих или вишечланих терминолошких јединица у претраживач Библише на једном од два језика – можемо доћи до парова конкорданција на енглеском и српском језику. Парови конкорданција не само да дају преводни еквивалент за задати упит, већ пружају и контекст употребе задате речи или израза на два језика. Уношењем енглеског термина *management* у претраживач Библише, на пример, добијамо све конкорданције ове терминолошке јединице на енглеском језику, као и паралелне реченице (преводе) на српски језик, из којих се могу издвојити преводи овог термина на српски (Табела 6), при чему су пронађени термини аутоматски истакнути плавом бојом. Треба поменути да Библиша, осим морфолошке експанзије упита, пружа могућност да се упит прошири и семантички коришћењем семантичке мреже WordNet и више терминолошких база. Уз то систем проналази еквивалнете у другом језику, чиме се омогућава издвајање паралелизованих реченица које: 1) проналазе еквиваленте у оба језика, 2) само у српском или 3) само у енглеском, што омогућава кориснику да користи систем за различите намене.

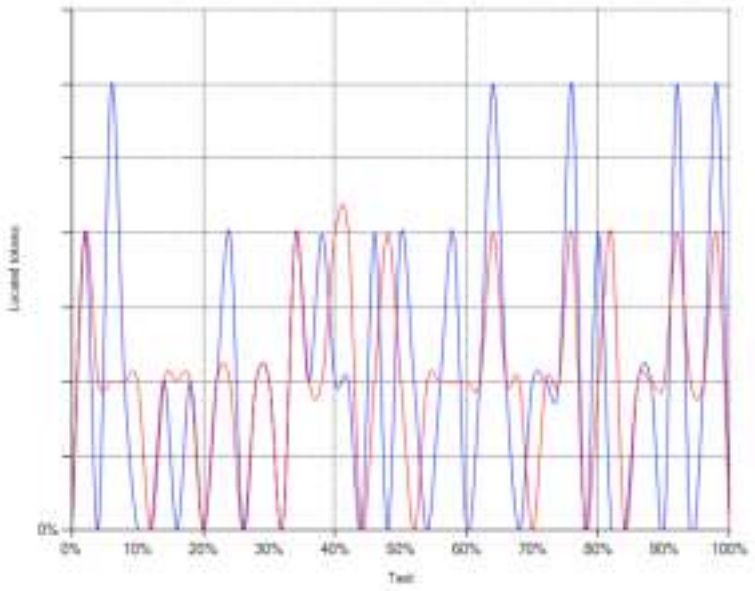
Табела 6 првенствено указује на различите преводне еквиваленте енглеског термина management на српски језик, и то као *менаџмент, управљање, руковођење, управа, управљачки, руководећи, менаџерски* итд., као и на примере у којима енглески термин није преведен него задржан у оригиналном облику. Уочени српски еквиваленти енглеског термина *management* указују како на његову полисемичност (*management* као процес управљања или као група људи која управља), тако и на синонимну употребу термина *менаџмент, управљање и руковођење* (као процеса), односно термина *менаџмент, управа и руководство* са (као тима људи). Осим тога, детаљнијом анализом могуће је утврдити и околности, односно контекст у коме се енгл. *management* преводи придевом, глаголом, или на неки други начин, али се тиме нећемо бавити у овом раду.

Детаљнијим упитом, односно претраживањем парова састављених од енглеског термина и сваког од његових преводних еквивалената на српском језику појединачно (нпр. енгл. *management* и срп. *управљање*,

енгл. *management* и срп. *менаџмент*, или енгл. *management* и срп. *руковођење*) можемо доћи до конкретних примера конкорданција у којима је овај енглески термин преведен на одређени начин. Слика 4 приказује дистрибуцију енглеског термина *management* и српског термина *менаџмент* кроз наш доменски корпус који је уређен дијахроно, од првих радова публикованих 2008 (лево) до последњих у корпусу публикованих 2012 (десно).

Метаподаци	Concordances (en)	Конкорданције (ср): 1260
Милићевић et al., 2009, vol. XIV:53, ID: 7.2009.53.1	The modern business prefers an integrative approach in the implementation of **management** tools in tracking ...	У савременом бизнису је пожељан интегративни приступ коришћењу **менаџерских** алата у праћењу ...
Милосављевић et al., 2012, vol. XVII:62, ID: 7.2012.62.5	As an integral function of global **management**, human resource **management** has a task to explore and define, ... what it is that makes the organization unique on the market.	**Менаџмент** људских ресурса, као интегрална функција глобалног **менаџмента**, има задатак да истражује и дефинише, ... шта је то што организацију чини јединственом на тржишту.
Ђурић Д., 2010, vol. XIV:55, ID: 7.2010.55.2	n40 Essentially, financial accounting is part of **management** accounting and reporting.	н40 Суштински, финансијско рачуноводство представља део **управљачког** рачуноводства и извештавања.
Mason R., 2012, No. 63, ID: 7.2012.63.1	The executives in our studies, even those that recognize the opportunities of this still emerging environment, still have widely divergent views on the most effective **management** models for realizing these opportunities.	Директори обухваћени нашом студијом, чак и они који су свесни могућности које даје ово ново окружење које је још у настајању, још увек се разликују у ставовима о томе који су најефективнији модели за реализацију ових могућности.
Пинтерић У., 2008, vol. XIII:49/50, ID: 7.2008.49-50.7	n78 Slovenian scientists wrote about introducing new public **management** elements into the work at all levels of Slovenian public administration ...	н78 Словеначки аутори писали су о увођењу елемената нове јавне **управе** у пословање на свим нивоима словеначке јавне управе ...
Станић С., 2008, vol. XIII:49/50, ID: 7.2008.49-50.6	n75 The internal factors include: the amount of media budget, the competence of **management** and administrative structure within the media department of the company or the hired marketing agency.	н75 Интерни фактори обухватају: величину медијског буџета, способности **руководеће** и административне структуре у оквиру медијског одељења компаније или ангажоване маркетинг агенције.
Домазет et al., 2009, vol. XIV:51, ID: 7.2009.51.4	n160 The Customer Relationship **Management** combines the business strategy and technology aiming to identify, attract and retain long-term relations with customers ...	н160 Customer Relationship Management комбинује пословну стратегију и технологију са циљем да идентификује, привуче и одржи дугорочне односе сакупцима ...
Вулић et al., 2012, No. 63, ID: 7.2012.63.7	The **management** is making organizational improvements in the country.	**Руководство** се организационо усавршава у земљи.
Hitka et al., 2009, vol. XIV:51, ID: 7.2009.51.8	n11 Managers from the area of manpower **management** have to deal with ... the problem ...	н11 Менаџери који **управљају** људском радном снагом морају да нађу прави одговор на питање ...
Панић С., 2012, No. 63, ID: 7.2012.63.9	To facilitate a two-way communication between the **management** and the employees, the company implemented three modes of communication...	Да би олакшала двосмерну комуникацију између **управе** и запослених, компанија је увела 3 начина комуникације...
Барјактаровић et al., 2011, vol. XVI:61, ID: 7.2011.61.1	A very important constituent of the overall bank **management** process is the implementation of the corporate governance principles.	Врло битан елемент целокупног процеса **управљања** банком јесте и приме-на принципа корпоративног управљања.

Табела 6. Конкорданције енглеског термина management и паралелне конкорданције на српском језику



Слика 4. Дистрибуција српског термина менаџмент и енглеског *management* кроз корпус

.

Дат је и преглед преводних еквивалената термина *management* на српски језик, заједно са бројем конкорданција у којима се дати одређени преводни еквиваленти појављују (Табела 7), стога можемо закључити да одговор на упит садржи различите флективне облике термина на српском језику, а не само облик номинатива (лему) у ком је задат упит.

Могућности примене паралелизованог доменског корпуса за менаџмент које су илустроване изнетим примерима могу бити од користи стручним преводиоцима, истраживачима у области компаративне и контрастивне лингвистике и терминологије, наставницима и студентима енглеског језика струке, као и другим профилима корисника.

Прво, претраживање доменског паралелизованог корпуса за менаџмент путем алата Библиша може помоћи стручним преводиоцима да у процесу превођења разреше терминолошке и друге језичке недоумице и пронађу одговарајући преводни еквивалент у контексту језичке употребе, посебно имајући у виду недостатак адекватних

расположивих терминографских и лексикографских ресурса на српском језику.

енгл.*management*		пример		
Преводни еквивалент	Број конкорданција	извор	енглески	српски
управљање	1431	Варјактаровић et al., 2011, vol. XVI:61, ID: 7.2011.61.1	A very important constituent of the overall bank **management** process is the implementation of the corporate governance principles.	Врло битан елемент целокупног процеса **управљања** банком јесте и примена принципа корпоративног **управљања**.
менаџмент	1089	Митрић et al., 2012, No. 65, ID: 7.2012.65.5	The fields of her scientific and professional interests are related to Accounting and Finance.	Њени главни истраживачки и наставни интереси везани су за област рачуноводства и финансијског **менаџмента**.
руковођење	14	Michalski G., 2008, vol. XIII:49/50, ID: 7.2008.49-50.12	**n95** Operating cycle management should also contribute to realization of this fundamental aim.	**н95** Постизању овог основног циља треба да допринесе и руковођење пословним циклусом.
управа	26	Savoiu et al., 2008, vol. XIII:49/50, ID: 7.2008.49-50.1	**n16** 55 Development of Slovenian selfgovernment in the new public **management** perspective	**н16** 55 Развој локалне самоуправе у Словенији у светлу Нове јавне управе
управљачки	83	Петровић С., 2009, vol. XIV:51, ID: 7.2009.51.6	**n120** • Organizations are too complicated to be understood by means of one **management** model. . .	**н120** • Организације су превише компликоване да би могле бити схваћене коришћењем једног **управљачког** модела. . .
руководећи	2	Петковић М., 2009, vol. XIV:51, ID: 7.2009.51.1	**n11** . . . that are presented by the number and the density of communications among organizational parts, **management** positions or members of a team.	н11 које се представљају бројем и густином комуникација између делова организације, **руководећих** позиција или чланова једног тима.
менаџерски	33	Петковић et al., 2012, No. 64, ID: 7.2012.64.7	. . . organizational design is a **management** lever (tool) used to achieve a balance between effectiveness and efficiency. . .	...организациони дизајн **менаџерска** полуга (алат), којом се балансира између ефективности и ефикасности. . .

Табела 7. Преводни еквиваленти енглеског термина management са примерима из корпуса

Друго, овај корпус је такође користан ресурс у компаративним и контрастивним проучавањима српског и енглеског језика струке, нпр. у контрастирању карактеристика академског писања на два језика, али и проучавању терминолошке (синонимијске и друге) варијације, недоследности и празнина које се неминовно јављају у српском језику као пасивном примаоцу научно-технолошког трансфера и трансфера

знања из развијених (махом англофоних) земаља.
Треће, за разлику од једнојезичних корпуса који се већ дуже време наилазе на примењујуну у учењу страних језика и креирању наставних материјала, педагошка употреба паралелизованих корпуса представља релативну новину о којој говоре, на пример, Danielsson and Mahlberg (2003), Granger (1998), а за српски језик и Ristović (2012).

Наш паралелизовани доменски корпус пружа могућности директне и индиректне примене у настави. Индиректно, енглески део овог корпуса може се користити као полазна основа за креирање наставних материјала и тестова, али и наставних планова и програма за стручни и академски енглески језик у области менаџмента и организације (нпр. за студенте менаџмента или за специјализоване курсеве енглеског језика), и то ослањањем на листе фреквентних и кључних речи и израза ради креирања тзв. „лексичких силабуса”(McEnery and Xiao, 2011). Директна експлоатација корпуса односи се на учење вођено подацима (енгл. *data-driven learning*), процес у оквиру кога ученици и студенти самостално користе корпус (Römer, 2011). Другим речима, студенти енглеског језика струке за област менаџмента, уз надзор наставника и адекватну обуку за коришћење корпуса и Библише, могли би самостално да користе корпус ради откривања одређениха граматичких, лексичких, дискурзивних и других правила и карактеристика, како би извршили анализу грешака у академском писању (најчешће насталих услед интерференције матерњег, тј. српског језика), али и у процесу превођења стручних текстова са једног језика на други.

4. Закључак

Посебан значај корпуса представљен у претходним поглављима лежи у чињеници да пре његове израде није постојао ниједан други електронски доступан и обележен корпус српског језика из области менаџмента нити сродних дисциплина (економија, маркетинг, организација и друге). Осим тога, електронски корпус може се непрестано надограђивати новим материјалом, чиме он остаје релевантан и савремен. Паралелизовани доменски корпуси радова представљају примарни ресурс за екстракцију терминолошких јединица и израду секундарних терминолошких ресурса – двојезичких терминолошких речника и терминолошких база и њихове сталне надоградње новим терминима. Осим тога, овакви корпуси су корисни и у области статистичког и неуронског машинског превођења. Као

ресурс за превођење, паралелизовани корпус који је поравнат на нивоу реченице може се користити за креирање преводилачких меморија и као подршка процесу превођења. Иако је у раду приказана екстракција само на српском језику, овакав ресурс је могуће користити и за двојезичну екстракцију терминологије. Осим наведених примена паралелизованог корпуса за домен менаџмента, треба нагласити многобројне могућности његове примене у другим врстама лингвистичких истраживања, пре свега у оквиру терминолошких, компаративних и констрастивних истраживања језика, студија превођења, учења и подучавања енглеског као страног језика струке, али и у анализи дискурса менаџмента, семантичким, прагматичким и социолингвистичким студијама.

Литература

Bonhomme, P, TMH Nguyen and S O'ROURKE. "XAlign: l'aligneur de Langue & Dialogue, 2001", ,

Danielsson, P and M Mahlberg, "There is more to knowing a language than knowing its words: Using parallel texts in the bilingual classroom". *English for Specific Purposes World. Online Journal for Teachers* Vol. 3, no. 6 (2003)

Erjavec, Tomaž and Nancy Ide. "The MULTEXT-East Corpus". У *Proceedings of the First International Conference on Language Resources and Evaluation*, 971–74. Citeseer, 1998,

Erjavec, Tomaž, Ana-Maria Barbu, Ivan Derzhanski, Ludmila Dimitrova, Radovan Garabík et al.. "MULTEXT-East" 1984" annotated corpus 4.0", (2010)

Fang, Cheng-yu, "Building a corpus of the English of computer science". *English Language Corpora: Design, Analysis and Exploitation. Amsterdam and Atlanta, GA: Rodopi* (1993): 73–8

Flowerdew, Lynne. "The argument for using English specialized corpora to understand academic and professional language". *Discourse in the professions: Perspectives from corpus linguistics* (2004): 11–33

Gledhill, Chris. "The discourse function of collocation in research article introductions". *English for Specific Purposes* Vol. 19, no. 2 (2000): 115–135

Granger, Sylviane. "The computer learner corpus: A testbed for electronic EFL tools". *Linguistic databases* (1998): 175–88

Krstev, Cvetana, Ranka Stankovic, Ivan Obradovic and Biljana Lazic. “Terminology Acquisition and Description Using Lexical Resources and Local Grammars.”. У *TIA*, 81–89. 2015

Krstev C., Vitas D. “Konkordancije paralelizovanih tekstova”. *Zbornik radova XXXVIII konferencije ETRAN, Niš, juni 1994*, 229–230. 1994

Ljubešić, Nikola, Miquel Esplà-Gomis, Sergio Ortiz Rojas, Filip Klubička and Antonio Toral. “Serbian-English parallel corpus srenWaC 1.0”, 2016

McEnery, Anthony M and Anita Wilson. *Corpus linguistics: an introduction.*. Edinburgh University Press, 2001

McEnery, Tony and Richard Xiao. “What corpora can offer in language teaching and learning”. *Handbook of research in second language teaching and learning* Vol. 2 (2011): 364–380

McEnery, Tony, Richard Xiao and Yukio Tono. *Corpus-based language studies: An advanced resource book*. Taylor & Francis, 2006

Obradović, I, R Stanković and M Utvić. “An integrated environment for development of parallel corpora”. *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen* (2008): 563–578

Paumier, Sébastien. “Manuel d’utilisation du logiciel Unitex”. *Université de Marne-la-Vallée*, 2002

Pazienza, Maria Teresa, Marco Pennacchiotti and Fabio Massimo Zanzotto. “Terminology extraction: an analysis of linguistic and statistical approaches”. У *Knowledge mining*, 255–279. Springer, 2005,

Pearson, Jennifer. *Terms in context*, Vol. 1. John Benjamins Publishing, 1998

Ristović, Zoran. “From corpus to classroom: The use of aligned corpora in English language teaching”. *Infoteka* Vol. 13, no. 2 (2012): 52–66

Roe, Peter Joseph. “The Notion of Difficulty in Scientific Text.”. Докторска дисертација, University of Birmingham, 1977

Römer, Ute. “Corpus research applications in second language teaching”. *Annual review of applied linguistics* Vol. 31 (2011): 205–225

Savourel, Y. “TMX 1.4 b Specification, The Localisation Industry Standards Association (LISA)”, 2004

Siddiqi, Sifatullah and Aditi Sharan. “Keyword and keyphrase extraction techniques: a literature review”. *International Journal of Computer Applications* Vol. 109, no. 2 (2015)

Sinclair, John. “Corpus and Text: Basic Principles. Wynne, M.(Ed.) Developing Linguistic Corpora: A Guide to Good Practice: 1-16”, , 2005

Stanković, Ranka, Ivan Obradović, Cvetana Krstev and Duško Vitas. “Production of morphological dictionaries of multi-word units using a multipurpose tool”. У *Proceedings of the Computational Linguistics-Applications Conference*, 77–84. 2011

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Biljana Lazić and Aleksandra Trtovac. "Rule-based automatic multi-word term extraction and lemmatization". У *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC*, 507–514. 2016a

Stanković, Ranka, Cvetana Krstev, Duško Vitas, Nikola Vulović and Olivera Kitanović. "Keyword-based search on bilingual digital libraries". У *Semanitic Keyword-based Search on Structured Data Sources*, 112–123. Springer, 2016b

Tognini-Bonelli, Elena. *Corpus linguistics at work*, Vol. 6. John Benjamins Publishing, 2001

Utvić, Miloš. "Annotating the corpus of contemporary Serbian". У *Proceedings of the INFOtheca '12 Conference*, 2011

Véronis, Jean. *Parallel Text Processing: Alignment and use of translation corpora* Vol. 13. Springer Science & Business Media, 2013

Vitas, Krstev C., D. "Electronic edition of Serbian translation of Orwell's 1884 aligned with 7 languages by Duško Vitas, Cvetana Krstev". (1998a)

Vitas, Nenadić G. Krstev C., D. "Electronic edition of Serbian translation of Plato's Republic aligned with 17 languages by Duško Vitas, Goran Nenadić, Cvetana Krstev". (1998b)