Владан О. Павловић[1]
Универзитет у Нишу
Филозофски факултет
Департман за англистику

Прегледни рад
УДК 81`27:81`322
Примљен: 14. 4. 2019.
DOI: https://doi.org/10.46630/phm.11.2019.27

О ПРИМЕНИ ВЕЛИКИХ ЕЛЕКТРОНСКИХ ТЕКСТУАЛНИХ КОРПУСА У СОЦИОЛИНГВИСТИЧКИМ ИСТРАЖИВАЊИМА[2] [3]

Овај рад прегледног типа бави се великим електронским текстуалним корпусима (већим од 100 милиона речи) енглеског и српског језика, који су се развили током протеклих неколико година у склопу изразитог напретка у информационо-комуникационој сфери. Посебна пажња се посвећује могућностима њихове примене у области социолингвистике. Закључује се да они у датој области, и поред извесних ограничења, могу бити од значаја на више начина – у смислу изразито велике количине језичког материјала који садрже, какав раније у толиком обиму није постојао сакупљен на једном месту, у погледу напредних могућности претраге таквог материјала значајних из угла социолингвистике, као и у погледу добијања квантитативних података (укључујући и податке добијене применом статистичких тестова), који, уз одговарајућу квалитативну, теоријски утемељену интерпретацију, могу да понуде бољи увид у спрегу између језика и друштва.

Кључне речи: велики текстуални електронски корпуси, социолингвистика, културолошка лингвистика, научно-технолошка и информационо-комуникациона револуција, колострукциона анализа.

---

[1] vladan.pavlovic@filfak.ni.ac.rs

[2] Припремљено у оквиру пројекта *Традиција, модернизација и национални идентитет у Србији и на Балкану у процесу европских интеграција* (179074), којим руководи емеритус проф. др Љубиша Митровић, и који финансира Министарство просвете и науке Републике Србије, као и у оквиру пројекта *Иновације у настави и истраживања у области англистичке лингвистике и англоамеричке књижевности и културе,* који се изводи на Филозофском факултету Универзитета у Нишу (бр. 360/1-16-1-01).

[3] Рад је изложен на међународном научном скупу *Традиција, модернизација и идентитети 9: Улога универзитетске заједнице у унапређењу регионалног развоја и афирмације културе мира на Балкану*, и то у оквиру сесије под називом *Изазови научно-технолошке и информационо-комуникационе револуције у савремености*. Скуп је одржан 25. и 26. маја 2018. године на Филозофском факултету у Нишу, у организацији Центра за социолошка истраживања при датом факултету.

1. Увод

Електронски текстуални корпуси јесу групе текстова (од којих највеће садрже и више милијарди речи) који представљају аутентичну употребу језика од стране изворних говорника у разним комуникативним ситуацијама, односно апроксимацију лингвистичке компетенције изворних говорника, и који су лингвистички и технички обрађени тако да омогућују разне облике претраге, квалитативну и квантитативну анализу, тестирање постојећих знања (тзв. *corpus-based linguistics*, в. МЕКЕНЕРИ И ХАРДИ 2012: 5 *et passim*; ТОЊИНИ–БОНЕЛИ 2001: 65–81) и долажење до нових знања која би без увида које корпусни подаци нуде била тешко могућа (тзв. *corpus-driven linguistics*, в. МЕКЕНЕРИ И ХАРДИ 2012: 5; ТОЊИНИ–БОНЕЛИ 2001: 84–99).

При њиховом састављању, води се рачуна да они буду репрезентативни и балансирани у погледу заступљености говорног и писаног језика, и / или у погледу различитих језичких варијетета, попут жанрова (књижевног, журналистичког, академског, итд.), и / или регионалних варијетета (нпр. америчког, британског, аустралијског, канадског енглеског итд.) и / или временских периода (нпр. 1991. до 2000.) који су њима обухваћени (за детаље в. ДЕСАГИЈЕ 2017: 3–4, МЕКЕНЕРИ И ВИЛСОН 2001: 77–81; МЕКЕНЕРИ, ШАО и др. 2006: 13–20; ФРИЦИНАЛ И ХАРДИ 2014: 48–56; БАЈБЕР 1993).

У оквиру изразитог напретка у научно-технолошкој и информационо-комуникационој сфери, такви корпуси су и сами доживели изразит напредак у смислу обима, доступности, лингвистичке и техничке обраде (анотираности) и могућности претраге, и постају све значајнији како у лингвистичким истраживањима уопште, тако и у социолингвистичким истраживањима посебно, због чега и јесу предмет овог рада.

2. Класификација, развој и примери великих електронских текстуалних корпуса

Корпуси се могу класификовати на основу тога: 1) који број језика садрже, 2) који варијетет једног језика садрже, 3) да ли су општи или специјализовани, 4) да ли су статички или динамички, 5) да ли су синхронијски или дијахронијски, 6) који облик језика садрже – писани, говорни или оба, 6) величине, 7) извора корпусног материјала, 8) да ли садрже додатни мултимедијални материјал, као и на основу других сличних критеријума. Исти корпус се на основу поменутих критеријума може класификовати на различите начине (ДЕСАГИЈЕ 2017: 51–52; ХАНСТОН 2002:14–16). За алтернативну класификацију корпуса (и то пре свега општих корпуса), која укључује и велике текстуалне архиве као што су

LexisNexis, Google Books и Sketch Engine, читалац се упућује на ДЕЈВИС 2015: 11 *et passim*).

Први електронски корпуси почели су да се јављају 1960-их година, када је направљен Браунов корпус (Brown Corpus, који је објављен и у штампаном издању у КУЧЕРА И ФРАНСИС 1967), који садржи текстове на америчком енглеском и укупно један милион речи. Сличан корпус тада је направљен за британски енглески (Lancaster – Oslo/Bergen Corpus). Током 80-их година прошлог века електронски корпуси постали су доступнији, унапређене су и могућности њихове претраге, а самим тим и употребљивост у лингвистичкој (па и социолингвистичкој) анализи, како теоријској, тако и примењеној (нпр. при прављењу речника и граматика). Током 90-их година прошлог века направљен је Британски национални корпус, који има 100 милиона речи, који је једно време био највећи општи, синхронијски корпус једног језика (за шири историјски преглед развоја корпусне лингвистике, као и разних облика електронских корпуса в. БАЈБЕР И РЕПЕН, 2015: 2–4).

Посебан помак на пољу стварања корпуса, у смислу њихове величине, могућности претраге и доступности, учињен је током протеклих неколико година. Међу њима се, када је енглески језик у питању, посебно истиче група корпуса Универзитета Бригам Јанг (*Brigham Young University Suite of Corpora*), аутора професора Марка Дејвиса, доступних на линку: https://www.english-corpora.org/.

English	# words	language/dialect	time period
iWeb: The Intelligent Web-based Corpus **NEW!**	14 **billion**	US/CA/UK/IE/AU/NZ	2017
News on the Web (NOW)	1.5 **billion+**	20 countries / Web	2010-last month
Global Web-Based English (GloWbE)	1.9 **billion**	20 countries / Web	2012-13
Wikipedia Corpus	1.9 **billion**	English	2014
Hansard Corpus	1.6 **billion**	British (parliament)	1803-2005
Early English Books Online	755 million	British	1470s-1690s
Corpus of Contemporary American English (COCA)	**560** million	American	1990-2017
Corpus of Historical American English (COHA)	400 million	American	1810-2009
Corpus of US Supreme Court Opinions	130 million	American (law)	1790s-present
TIME Magazine Corpus	100 million	American	1923-2006
Corpus of American Soap Operas	100 million	American	2001-2012
British National Corpus (BYU-BNC)*	100 million	British	1980s-1993
Strathy Corpus (Canada)	50 million	Canadian	1970s-2000s
CORE Corpus	50 million	Web registers	2014

Слика 1 – Листа корпуса у породици корпуса Универзитета Бригам Јанг (извор: https://www.english-corpora.org/)

Као што се може видети, највећи корпус доступан на датом линку, а уједно и највећи корпус који је икада направљен, јесте *iWeb: The Intelligent Web-based Corpus* (https://www.english-corpora.org/iweb/), који садржи 14 милијарди речи у оквиру текстова објављених на интернету током 2017.

године у 6 земаља у којима се говори енглески језик – САД, Канади, Уједињеном Краљевству, Ирској, Аустралији и на Новом Зеланду, чији се регионални варијетети енглеског језика могу назвати варијететима „ужег круга“. Исечак радног окружења овог корпуса, доступног на линку: https://www.english-corpora.org/iweb/, дајемо ниже.

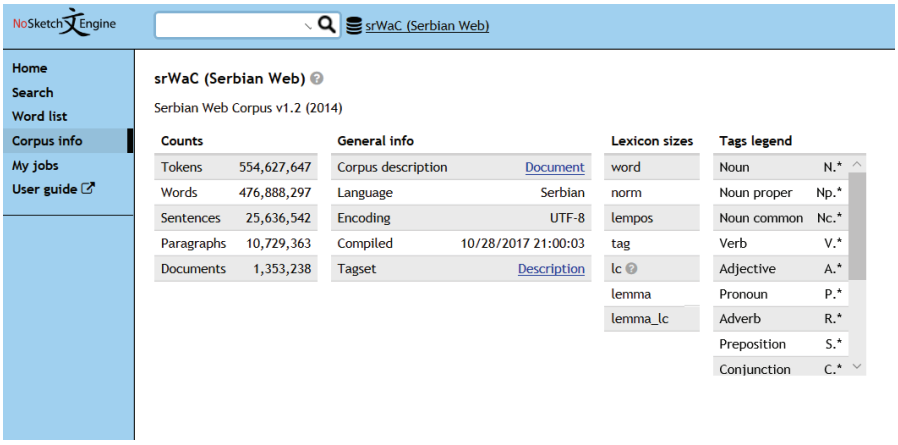
Још један корпус из дате породице корпуса – *News on the Web*, значајан је и због тога што је динамички. Наиме, он садржи текстове вести објављених на интернету почев од 2010. године у 20 земаља где се говори енглески језик, с тим што се он проширује тако што се у њега редовно аутоматски додају нови текстови са интернета на основу одређених критеријума. Отуда знак плус поред броја којим се на горњој слици означава тренутан апроксимативни број речи у датом корпусу, наиме – 1,5 милијарди.

Од осталих горенаведених корпуса у литератури која се односи на корпусна проучавања енглеског језика најчешће се наводе *Корпус савременог америчког енглеског* (*Corpus of Contemporary American English – COCA*), *Корпус историјског америчког енглеског* (*Corpus of Historical American English – COHA*) и *Корпус глобалног енглеског језика на интернету* (*Global Web-based English – GloWbE*). Поменути Британски национални корпус, који се такође налази на горњем списку и који је такође у честој употреби у лингвистичким истраживањима, оригинално не припада датој породици корпуса (створен је у Великој Британији пре него ли САД), али је горе наведен јер је у оквиру те породице корпуса добио функционалност коју раније није имао.

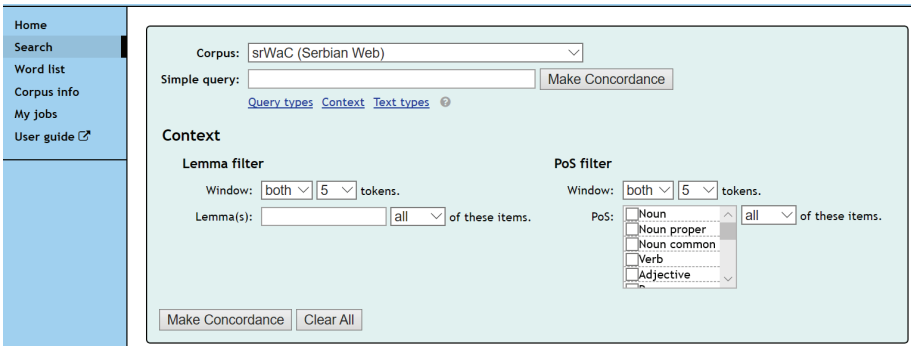
Слика 2 – Радно окружење корпуса *iWeb: The Intelligent Web-based Corpus* (извор: https://www.english-corpora.org/iweb/)

Када су у питању електронски корпуси српског језика, највећи је тзв. *SrWaC* (*Serbian Web-based Corpus*), тј. *Корпус српског језика заснован на језику на интернету*, који је настао аутоматским прикупљањем текстова са изабраних интернет презентација на домену .rs и њиховом одговарајућом лингвистичком и техничком обрадом (https://www.clarin.si/noske/run.cgi/corp_info?corpname=srwac), и који има око 477 милиона речи. Следе *Корпус твитова на српском језику* (https://www.clarin.si/noske/run.cgi/corp_info?corpname=tweet_sr), који има око 174 милона речи, и *Корпус савременог српског језика СрпКорп2013* на Математичком факултету Универзитета у Београду, који има око 122 милиона речи (http://www.korpus.matf.bg.ac.rs/prezentacija/korpus.html).

Технички детаљи и радно окружење у вези са првим наведеним корпусом дати су на сликама 3 и 4.



Слика 3. Технички детаљи везани за Корпус SrWaC (извор: https://www.clarin.si/noske/run.cgi/corp_info?corpname=srwac)



Слика 4. Радно окружење у корпусу SrWaC (извор: https://www.clarin.si/noske/run.cgi/first_form?corpname=srwac;align=)

Дејвис (2015: 30–31) процењује да ће сви тренутно постојећи општи корпуси, и то у светлу брзог напретка у области информатичких технологија, већ за пет до десет година бити застарели јер ће тада бити могуће имати приступ милијардама речи у оквиру разних текстова који су објављени на неки дан из такве блиске будућности, при чему ће тим речима аутоматски бити приписивани одговарајући метаподаци, тј. оне ће бити аутоматски анотиране за нпр. пол, узраст, приближну географску локацију / дијалекат и друге сличне демографске варијабле које се односе на говорника језика. Другим речима, биће могуће пратити коришћење речи, фраза или синтаксичких конструкција у реалном времену. То ће, наводи даље овај аутор, бити револуционарно бар са тренутне тачке гледишта. Управо због овакве експанзије великих електронских текстуалних корпуса (оних преко 100 милиона речи) у последњих неколико година, као и због њиховог очекиваног даљег брзог развоја, у овом раду превасходно се бавимо таквим пре него ли мањим корпусима или пак корпусима говорног језика (од којих ниједан није овако обиман), иако и они свакако имају своје место у социолингвистичкој, и уопште – лингвистичкој, анализи.

3. Социолингвистичка истраживања заснована на употреби електронских текстуалних корпуса

Социолингвистичка истраживања заснована на електронским корпусима релативно су ново поље истраживања у поређењу са етнографским методима, а њиховим зачетком могу се сматрати студије Бајбера, Финегана и сарадника настале током 1980-их (в. нпр. БАЈБЕР 1988; међу каснијим радовима ових аутора на тему корпусних истраживања у социолингвистици издвајају се БАЈБЕР, КОНРАД и др. 1998 и БАЈБЕР И РЕПЕН и др. 2010). У таквим студијама се наглашава да корпусна социолингвистика може да понуди увид у систематичност раслојавања језика, које може да се опише емпиријским и квантитативним методама, и даље интерпретира квалитативно у оквиру различитих теоријских приступа и њиховог аналитичког апарата. Додатно се скреће пажња (ФРИЦИНАЛ И ХАРДИ 2014: xv) да поље корпусне социолингвистике још увек није у потпуности укључено у шире поље социолингвистике, али да брзи развој електронских корпуса, квалитет анотираности језичког материјала доступног у њима (укључујући и анотирање значајно из социолингвистичког угла), као и пратећи развој напредних начина претраге таквих корпуса, чине да се поменуто превазилази (в. и ФРИЦИНАЛ И БРИСТОУ 2018: 6).

Нека од најзначајнијих новијих истраживања у области корпусне социолингвистике (уз примену текстуалних корпуса) јесу следећа: ГРИВ

2016, које се бави територијалним раслојавањем писаног америчког енглеског, КО 2016, које се бави анализом места прилога у десет варијетета писаног и говорног енглеског језика, и то на основу корпуса ICE (*International Coprus of English*), који садржи двадесетак поткорпуса, сваки са по око милион речи, као и ШМЕРЧАЊИ, ГРАФМИЛЕР и др. 2016, које се бави варијацијама у три синтаксичке конструкције у четири регионална варијетета енглеског. За проучавање регионалних варијетета користе се и Твитер и Фејсбук, као својеврсни корпуси, односно извори грађе за (социо)лингвистичка истраживања (в. нпр. АЈЗЕНШТАЈН, КОНОР и др. 2014 и ХУАНГ, ГУО и др. 2015). Текстуални електронски корпуси користе се и када су у фокусу разне демографске варијабле говорника (пол, род, етничка припадност, старост и сл.) и неретко су мањег обима, што је и примереније код истраживања ужих друштвених група и заједница говорника (в. нпр. КЛЕНСИ 2016; ТАЉАМОНТЕ 2016; ХЕЈ, ПЈЕРХАМБЕРТ и др. 2015). Напокон, текстуални електронски корпуси користе се и код језичког раслојавања везаног за регистар (журналистички, професионални, академски и сл.). У таква истраживања, поред осталог, спадају: ПИКЕРИНГ, ФРИЦИНАЛ и др. 2016; СИНКЛЕР 2000; РИЛЕМАН 2007. Нека од најновијих истраживања заснована управо на великим текстуалним електронским корпусима, пре свега у области регионалних дијалеката, укључујући и варијетет глобалног енглеског, јесу следећа: ДЕЈВИС 2018; БАЈБЕР, ЕГБЕРТ и др. 2018; БЕРБЕР САРДИЊА 2018 и БАЛАСУБРАМАНИЈАН 2018.

Може се приметити и да у оквиру појединих лингвистичких теорија и приступа, посебно оних који су супротстављени генеративној граматици и структурализму (где се превасходно води рачуна о идеализованом језичком знању, идеализованим говорницима и интроспекцији), тј. у теоријама и приступима који инсистирају на проучавању конкретне употребе језика, укључујући и употребу похрањену у оквиру електронских корпуса, такође долази до све веће социолингвистичке усмерености. То је случај са когнитивном лингвистиком у ужем смислу, односно когнитивном социолингвистиком у оквиру ње, о чему сведоче, поред осталих, следеће публикације: КРИСТИЈАНСЕН И ДЕРВЕН 2008 и ХЕРАРТС 2003. Сличан приступ може се приметити и у културолошкој лингвистици (*Cultural Linguistics*), као новом мултидисциплинарном пољу у оквиру кога се истражује веза између језика, са једне стране, и културом условљених концептуализација, са друге стране (в. нпр. ЈЕНСЕН 2017).

4. Додатне могућности примене великих електронских текстуалних корпуса у социолингвистичким истраживањима

Велики електронски текстуални корпуси могу пружити увид у податке значајне из социолигвистичког угла на начин на који то ранији корпуси – они мањег обима и ограниченијих начина претраге – нису могли.

На пример, у оквиру Корпуса глобалног енглеског језика на интернету (GloWbE), могуће је претраживати колокате именице *marriage* у облику придева у претпозицији (користећи претраживачку синтаксу *[j*] [marriage]* или *ADJ marriage**) у регионалним варијететима енглеског језика ужег круга (поменутом америчком, британском, канадском, ирском, аустралијском и новозеландском енглеском) и ширег круга (варијететима енглеског који се користе у Индији, Пакистану, Бангладешу, Сингапуру, Малезији, Филипинима, Хонг Конгу, и још у шест земаља). Као што се ниже може видети из тако добијених резултата, дата именица у варијететима ширег круга најчешће колоцира са придевима *permanent*, *customary*, *inter-caste*, *fixed-time*, *special*, *Hindu*, *Islamic*, *Indian*, док су у варијететима ужег круга ти колокати, поред осталог, *equal*, *same-gender*, *gay*, *anti-gay*, *open*, и слично, што свакако бар делом представља рефлексију одговарајућих (традиционалнијих односно либералнијих) културних и друштвених норми и интересовања у једној, односно другој групи друштава.

SEC 1 (India, Sri Lanka, Pakistan,...): 644,753,594 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	PERMANENT MARRIAGE	319	5	0.5	0.0	122.7
2	CUSTOMARY MARRIAGES	64	2	0.1	0.0	61.5
3	INTER-CASTE MARRIAGE	28	1	0.0	0.0	53.8
4	HINDU MARRIAGES	27	1	0.0	0.0	51.9
5	CUSTOMARY MARRIAGE	154	9	0.2	0.0	32.9
6	AFRICAN MARRIAGE	39	3	0.1	0.0	25.0
7	HINDU MARRIAGE	165	13	0.3	0.0	24.4
8	FIXED-TIME MARRIAGE	150	0	0.2	0.0	23.3
9	SPECIAL MARRIAGE	82	7	0.1	0.0	22.5
10	TEMPORARY MARRIAGE	586	55	0.9	0.0	20.5
11	BLESSED MARRIAGE	26	4	0.0	0.0	12.5
12	INDIAN MARRIAGES	32	5	0.0	0.0	12.3
13	BLISSFUL MARRIAGE	50	8	0.1	0.0	12.0
14	ISLAMIC MARRIAGE	141	27	0.2	0.0	10.0
15	EARLY MARRIAGES	184	44	0.3	0.0	8.0
16	ISLAMIC MARRIAGES	24	6	0.0	0.0	7.7

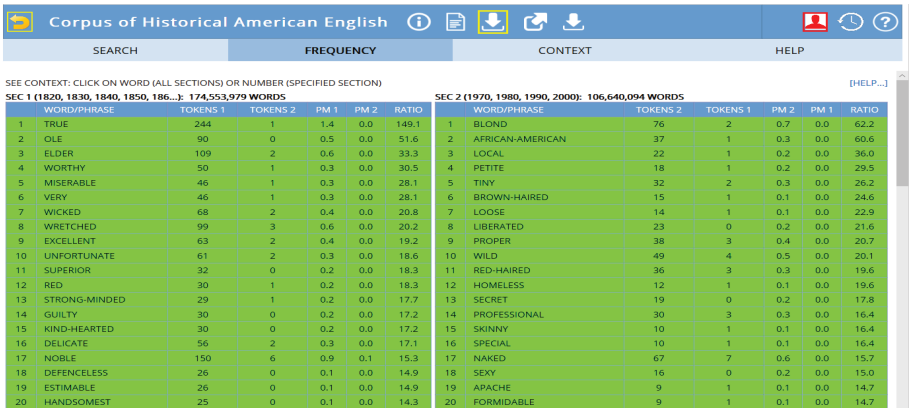
SEC 2 (United States, Canada, Grea...): 1,239,817,686 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	SACRAMENTAL MARRIAGE	51	1	0.0	0.0	26.5
2	EQUAL MARRIAGE	562	14	0.5	0.0	20.9
3	SAME-GENDER MARRIAGE	21	1	0.0	0.0	10.9
4	ANTI-GAY MARRIAGE	145	7	0.1	0.0	10.8
5	CATHOLIC MARRIAGES	50	3	0.0	0.0	8.7
6	CELESTIAL MARRIAGE	32	2	0.0	0.0	8.3
7	BIBLICAL MARRIAGE	27	2	0.0	0.0	7.0
8	GAY MARRIAGE	10740	809	8.7	1.3	6.9
9	HETEROSEXUAL MARRIAGES	103	8	0.1	0.0	6.7
10	OPEN MARRIAGES	24	2	0.0	0.0	6.2
11	HOMOSEXUAL MARRIAGE	457	43	0.4	0.1	5.5
12	SAME-GENDER MARRIAGES	21	2	0.0	0.0	5.5
13	NATIONAL MARRIAGE	51	5	0.0	0.0	5.3
14	HETEROSEXUAL MARRIAGE	343	34	0.3	0.1	5.2
15	SAME-SEX MARRIAGE	5415	550	4.4	0.9	5.1
16	HEALTHY MARRIAGES	29	3	0.0	0.0	5.0

Слика 5. Колокати именице *marriage* у регионалним варијететима енглеског језика ширег круга (лево) и ужег круга (десно); извор: *Корпус глобалног енглеског језика на интернету* (*GloWbE*)

Поред наведеног, дати корпуси могу се користити и за претрагу придева као колоката именица у различитим временским периодима. На пример, поређење придева као колоката именице *women* који се могу јавити на првој, другој, трећој и четвртој позицији пре и после ове именице у периоду од 1820. до 1910. године, са једне стране, и од 1980. до 2000. године, са друге

стране, у оквиру Корпуса историјског америчког енглеског језика (COCA) показује да су међу најчешћим колокатима ове именице у првом наведеном периоду били, поред осталог, придеви који се односе на године старости жене и пратећу мудрост и животно искуство (*ole* (=*old*), *elder*), придеви који се односе на личне врлине (*kind-hearted*, *noble*, *worthy*, као и *true*, који се тада употребљавао као данашњи придев *honest*), потом придеви који описују физичку слабост жене (*delicate*, *defenceless*), као и придеви негативног значења (*wicked*, *wretched*, *unfortunate*). У овој групи придева ретки су они који се односе на физички изглед жене у смислу лепоте и привлачности (такав је придев у суперлативу *handsomest*, који се данас превасходно односи на мушкарце). Обрнуто, у другом периоду налазимо значајно више придева који се односе на физички изглед жене (*blond*, *petit*, *tiny*, *skinny*, *brown-haired*, *red-haired*, *sexy*), потом на придев *African-American*, који се односи на етничко порекло и политички је коректан термин за референцирање на припаднике дате етничке заједнице у САД, као и придеве попут *wild*, *loose* и *liberated*, који се могу односити на женску сексуалност и / или општу еманципацију.


Corpus of Historical American English

SEARCH | FREQUENCY | CONTEXT | HELP

SEE CONTEXT: CLICK ON WORD (ALL SECTIONS) OR NUMBER (SPECIFIED SECTION) [HELP...]

SEC 1 (1820, 1830, 1840, 1850, 186...): 174,553,979 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	TRUE	244	1	1.4	0.0	149.1
2	OLE	90	0	0.5	0.0	51.6
3	ELDER	109	2	0.6	0.0	33.3
4	WORTHY	50	1	0.3	0.0	30.5
5	MISERABLE	46	1	0.3	0.0	28.1
6	VERY	46	1	0.3	0.0	28.1
7	WICKED	68	2	0.4	0.0	20.8
8	WRETCHED	99	3	0.6	0.0	20.2
9	EXCELLENT	63	2	0.4	0.0	19.2
10	UNFORTUNATE	61	2	0.3	0.0	18.6
11	SUPERIOR	32	0	0.2	0.0	18.3
12	RED	30	1	0.2	0.0	18.3
13	STRONG-MINDED	29	1	0.2	0.0	17.7
14	GUILTY	30	0	0.2	0.0	17.2
15	KIND-HEARTED	30	0	0.2	0.0	17.2
16	DELICATE	56	2	0.3	0.0	17.1
17	NOBLE	150	6	0.9	0.1	15.3
18	DEFENCELESS	26	0	0.1	0.0	14.9
19	ESTIMABLE	26	0	0.1	0.0	14.9
20	HANDSOMEST	25	0	0.1	0.0	14.3

SEC 2 (1970, 1980, 1990, 2000): 106,640,094 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	BLOND	76	2	0.7	0.0	62.2
2	AFRICAN-AMERICAN	37	1	0.3	0.0	60.6
3	LOCAL	22	1	0.2	0.0	36.0
4	PETITE	18	1	0.2	0.0	29.5
5	TINY	32	2	0.3	0.0	26.2
6	BROWN-HAIRED	15	1	0.1	0.0	24.6
7	LOOSE	14	1	0.1	0.0	22.9
8	LIBERATED	23	0	0.2	0.0	21.6
9	PROPER	38	3	0.4	0.0	20.7
10	WILD	49	4	0.5	0.0	20.1
11	RED-HAIRED	36	3	0.3	0.0	19.6
12	HOMELESS	12	1	0.1	0.0	19.6
13	SECRET	19	0	0.2	0.0	17.8
14	PROFESSIONAL	30	3	0.3	0.0	16.4
15	SKINNY	10	1	0.1	0.0	16.4
16	SPECIAL	10	1	0.1	0.0	16.4
17	NAKED	67	7	0.6	0.0	15.7
18	SEXY	16	0	0.2	0.0	15.0
19	APACHE	9	1	0.1	0.0	14.7
20	FORMIDABLE	9	1	0.1	0.0	14.7

Слика 6. Колокати именице *women* у периоду од 1820. до 1910. године (лево) и од 1980. до 2000. године (десно); извор: *Корпус историјског америчког енглеског језика* (*COCA*)

Велики текстуални корпуси могу дати увид и у промену значења одређених речи кроз време. Тако поређење именских колоката речи *gay* у паровима типа *gay* + *именица* у периоду од 1820. до 1910. године, са једне стране, и у периоду од 1970. до 2000. године, са друге стране (в. нижу слику) такође у *Корпусу историјског америчког енглеског*, као и додатна претрага ширег контекста у коме се овакви колокати јављају, показује промену значена дате речи од *весео*, *-ла*, *-о* (*gay world*, *gay company*, *gay colours*, *gay throng*, *gay spirits*, *gay voices*) у значење које она данас има (*gay men*, *gay rights*, *gay marriage*, итд).

SEC 1 (1820, 1830, 1840, 1850, 186...): 174,553,979 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	GAY WORLD	95	3	0.5	0.0	19.3
2	GAY COMPANY	67	3	0.4	0.0	13.6
3	GAY LAUGH	54	2	0.3	0.0	16.5
4	GAY LIFE	54	17	0.3	0.2	1.9
5	GAY COLORS	51	6	0.3	0.1	5.2
6	GAY PARTY	49	1	0.3	0.0	29.9
7	GAY SCENE	46	3	0.3	0.0	9.4
8	GAY VOICE	39	0	0.2	0.0	22.3
9	GAY THRONG	37	0	0.2	0.0	21.2
10	GAY SMILE	35	1	0.2	0.0	21.4
11	GAY FLOWERS	35	0	0.2	0.0	20.1
12	GAY ATTIRE	33	1	0.2	0.0	20.2
13	GAY SPIRITS	31	0	0.2	0.0	17.8
14	GAY LAUGHTER	30	2	0.2	0.0	9.2
15	GAY SOCIETY	29	2	0.2	0.0	8.9
16	GAY TONE	26	2	0.1	0.0	7.9
17	GAY TIME	25	2	0.1	0.0	7.6
18	GAY SEASON	24	0	0.1	0.0	13.7
19	GAY CAVALIER	22	0	0.1	0.0	12.6
20	GAY VOICES	21	1	0.1	0.0	12.8

SEC 2 (1970, 1980, 1990, 2000): 106,640,094 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	GAY MEN	186	3	1.7	0.0	101.5
2	GAY RIGHTS	104	0	1.0	0.0	97.5
3	GAY MARRIAGE	82	0	0.8	0.0	76.9
4	GAY COMMUNITY	67	0	0.6	0.0	62.8
5	GAY PEOPLE	61	19	0.6	0.1	5.3
6	GAY MAN	45	7	0.4	0.0	10.5
7	GAY BAR	33	1	0.3	0.0	54.0
8	GAY LIBERATION	24	0	0.2	0.0	22.5
9	GAY BARS	24	0	0.2	0.0	22.5
10	GAY ACTIVISTS	24	0	0.2	0.0	22.5
11	GAY PRIDE	21	0	0.2	0.0	19.7
12	GAY ACTIVIST	20	0	0.2	0.0	18.8
13	GAY COUPLES	18	1	0.2	0.0	29.5
14	GAY LIFE	17	54	0.2	0.3	0.5
15	GAY GAMES	14	0	0.1	0.0	13.1
16	GAY FRIENDS	12	10	0.1	0.1	2.0
17	GAY MARRIAGES	10	0	0.1	0.0	9.4
18	GAY SEX	10	0	0.1	0.0	9.4
19	GAY PERSON	10	2	0.1	0.0	8.2
20	GAY YOUTH	10	15	0.1	0.1	1.1

Слика 7. Колокати речу *gay* у претпозицији у периоду од 1820. до 1910. године (лево) и од 1970. до 2000. године (десно); извор: *Корпус историјског америчког енглеског језика* (*COCA*)

Додатно, дати тип корпуса може помоћи и да се прати фреквенција употребе појединих речи или групе синонима кроз време. То могу бити следеће речи / синтагме којима се означавају припадници једне од етничких скупина у САД: *negro*, *blacks*, *people* / *persons of colour*, и поменути придев *African-American*. С тим у вези овде ћемо кратко размотрити дијахронијску фреквенцију две од датих речи на основу *Корпуса историјског америчког енглеског језика*. Наиме, добијени подаци показују углавном значајан пад у употреби придева *negro* након 1860. године, а нарочито слабу употребу након 1970-их и надаље, када је готово изопштен из употребе, вероватно на основу примедби које су на употребу овог придева имале вође покрета за еманципацију дате етничке скупине америчког друштва, док управо током 1970-их долази до значајног раста фреквенције јављања придева *blacks*, чија употреба у потоњим деценијама такође бележи пад.

Иако наведени типови резултата заиста могу бити значајни како за анализу у смислу језичке рефлексије одређених друштвених и културних тема и интересовања, па и друштвено и културно условљених норми и образаца понашања, тако и у смислу могућности претраге које код ранијих корпуса – пре појаве породице корпуса Универзитета Бригам Јанг, нису били могући, треба нагласити да се сви они ослањају на дескриптивно-статистичке податке (и њихове нормализоване верзије).

Прави значај великих текстуалних корпуса, могло би се тврдити, лежи у количини језичког материјала на које је могуће применити не само дескриптивну статистику него и релевантне статистичке тестове,

и на тај начин такође обрађивати одговарајуће социолингвистичке теме (поред других тема из области лингвистике). У нижим редовима ћемо се тако осврнути на један облик статистичке лингвистичке анализе и говорити о могућностима њене примене на једну конструкцију из енглеског језика и њен еквивалент из српског језика, са нагласком на социолингвистички аспекат такве анализе, а такође уз помоћ материјала који нуде велики електронски текстуални корпуси. Том приликом, имајући у виду да је ово рад прегледног типа, а с обзиром и на просторна ограничења, овде нећемо изводити цело истраживање већ ћемо само изнети његов могућ нацрт.

Статистички метод, тачније породицу статистичких метода, коју бисмо овде у најкраћем представили је *колострукциона анализа*, којом се анализирају односи између речи и граматичких структура у којима се те речи јављају, најчешће у оквиру теорије познате под називом конструкциона граматика (в. ГОЛДБЕРГ 1995; ХОФМАН И ТРАУЗДЕЈЛ 2013; ХИЛПЕРТ 2014б). Њоме се утврђује које се речи или парови речи јављају у статистички значајној мери у различитим деловима једне граматичке конструкције или више њих (отуда термин *колострукција*, као сливеница настала повезивањем термина *конструкција* и *колокација*). Овај метод развили су С. Т. Грис и А. Стефанович, а неки од најзначајнијих радова из дате области јесу следећи: ГРИС 2013; ГРИС И СТЕФАНОВИЧ 2004а; 2004б; СТЕФАНОВИЧ И ГРИС 2003; 2005, *inter alia*.

Постоје три варијанте колострукционе анализе – анализа простих колексема (*simple collexeme analysis*), анализа дистинктивних колексема (*distinctive collexeme analysis*) и анализа коварирајућих колексема (*covarying collexeme analysis*). Првом се истражује које се лексичке јединице јављају у статистички значајној мери на једном месту у једној граматичкој конструкцији (нпр. у конструкцији *keep on V-ing*, као у примерима *keep on singing*, *keep on reading* итд.). Другом се, поред осталог, истражује да ли постоји статистички значајна разлика у употреби лексема у једној конструкцији у различитим варијантама употребе те конструкције (нпр. у оквиру различитих језичких регионалних и / или социјалних варијетета). Трећом се истражује да ли има статистички значајне разлике у коварирању парова лексема које се јављају на два различита места у истој конструкцији (нпр. у конструкцији *be ADJECTIVE to VERB*, као у примерима *She is tough to deal with*, *He is difficult to please*). Потребни подаци који се добијају на основу корпусне грађе (при чему су управо велики текстуални корпуси од великог значаја управо због обима) за обављање све три статистичке анализе, и то применом Фишеровог теста, а типично у оквиру софтверског пакета R (https://www.r-project.org/, https://www.rstudio.com/) детаљно су представљени у раду СТЕФАНОВИЧ 2013 (в. и

ХИЛПЕРТ 2014а). Применом тог теста добијају се вредности *p*. Као што је познато, што су оне мање, тј. ниже од вредности 0,05, која се типично узима за граничну вредност статистичке значајности, то се може сматрати да је већа повезаност одређених лексема и одређене конструкције.

Размотримо сада конструкцију *ADJECTIVE / NOUN enough to VERB* у енглеском језику и еквивалентну конструкцију *превише ПРИДЕВ / ИМЕНИЦА да ГЛАГОЛ* у српском језику. Оне су, као и бројне друге конструкције, значајне поред осталог и зато што могу садржати лексеме и генерално – износити садржај, који може бити значајан из социолингвистичког угла (и сродног угла горе поменуте културолошке лингвистике), као у следећим примерима: *He was not man enough to face the truth back then*, *She was woman enough to own up to her error*, *Био је превише мушко / мушкарац да дозволи да му други виде сузу у оку*, *Да ли си довољно мушко / мушкарац да прихватиш изазов?*, *Није био довољно стар да бисмо му понудили такву руководећу позицију*, итд.

С наведеним у вези, а уз помоћ датог статистичког метода, било би корисно истражити које се лексеме (и из којих семантичких поља) јављају у статистички значајној мери у примерима када је једна варијабла већ дата – нпр. *MAN enough to VERB* и *довољно МУШКО / МУШКАРАЦ да ГЛАГОЛ* (упор. ЈЕНСЕН 2017). То би представљало пример примене анализе простих колексема. Друго, могло би се истражити да ли постоје парови колексема (*ADJECTIVE/NOUN–VERB* односно *ПРИДЕВ/ИМЕНИЦА–ГЛАГОЛ*) који се у наведеним конструкцијама јављају у статистички значајној мери. То би представљало пример примене анализе дистинктивних колексема. Треће, могло би се установити које јединице *ADJECTIVE/NOUN*, односно *ПРИДЕВ/ИМЕНИЦА*, као и који парови јединица *ADJECTIVE/NOUN–VERB* и *ПРИДЕВ/ИМЕНИЦА–ГЛАГОЛ* јесу дистинктивни за енглески односно српски језик. То би представљало пример примене анализе коваријајућих колексема. Значај таквих анализа из социолигвистичког угла огледао би се, поред осталог, у утврђивању конкретних глагола са којима се поједини придеви и именице у овој конструкцији комбинују, што би се могло сматрати бар делимичном рефлексијом културом условљених концептуализација рода, старости и сличних варијабли у сваком од два језика, као и у поређењу таквих концептуализација у дата два језика и сходно томе – у две културе.

5. Закључна разматрања

Употреба великих текстуалних корпуса у социолингвистичким истраживањима свакако има и својих ограничења. Један од разлога је тај што су социолингвисти нередко заинтересованији за говорни језик, који је у

текстуалним корпусима ређе присутан (нпр. у облику транскрибованих дискурса) (в. нпр. ЊУМАН 2008, који рад је цео посвећен неопходности формирања говорних корпуса за потребе социолингвистичких истраживања). Друго, социолингвистика, поред осталог, настоји да испита употребу језика у конкретном интерперсоналном и друштвеном контексту, који неретко није јасно доступан у великим електронским текстуалним корпусима (нпр. такви корпуси неретко не пружају довољно података о социо-економском статусу, етничкој припадности, полу, старости, нивоу образовања, степену религиозности и сличним демографским варијаблама које се односе говорнике језика чија је језичка продукција ушла у корпусну грађу). Треће, текстуални корпуси типично не садрже податке о интонацији говорника (ако је у питању говорни језик, као облик језика који нпр. чини 20% грађе у Корпусу савременог америчког енглеског језика). Самим тим практично је немогуће на основу датог извора грађе проучавати одговарајуће социофонетске обрасце. И четврто, социолингвистику неретко занимају сасвим специфични језички варијетети или језичке ситуације, као што је језичка пракса одређене уске заједнице или друштвене групе, у ком смислу углавном општи корпуси који типично садрже стандардне варијетете не морају бити од значаја (в. Kendall 2011: 368).

Ипак, на основу горенаведеног може се закључити да велики текстуални корпуси могу бити од значаја у социолингвистичким истраживањима на више начина: прво, у погледу изразито велике количине језичког материјала који садрже а какав раније у толиком обиму није постојао сакупљен на једном месту, друго – у погледу могућности претраге таквог материјала које претходни корпуси (средње и мале величине) нису поседовали, и треће – у погледу добијања квантитативних података (укључујући и податке добијене применом статистичких тестова) који, уз одговарајућу квалитативну, теоријски утемељену интерпретацију, могу да понуде објективан увид у тенденције и законитости присутне у корпусном материјалу, као аутентичном језичком материјалу, и тако омогуће бољи увид у спрегу између језика и друштва у коме се тај језик употребљава.

Цитирана литература

АЈЗЕНШТАЈН, КОНОР и др. 2014: Eisenstein, Jacob and Brendan O'Connor, Noah A. Smith, Eric P. Xing. "Diffusion of lexical change in social media". *PLoS ONE*, 9(11) (2014): p. 23–45.

БАЈБЕР 1988: Biber, Douglas. *Variation across speech and writing*. Cambridge, UK: Cambridge University Press, 1988.

БАЈБЕР 1993: Biber, Douglas. "Representativeness in corpus design." *Literary and Linguistic Computing* 8(4) (1993): p. 243–257.

БАЈБЕР, КОНРАД и др. 1998: Biber, Douglas and Susan Conrad, Randy Reppen. *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press, 1998.

БАЈБЕР И РЕПЕН и др. 2010: Biber, Douglas and Randy Reppen, Eric Friginal "Research in corpus linguistics". In: Kaplan, Robert B. (Ed.), *The Oxford handbook of applied linguistics* (2nd ed.), Oxford: Oxford University Press, 2010, p. 548–570.

БАЈБЕР И РЕПЕН 2015: Biber, Douglas and Randy Reppen (eds.) *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 2015.

БАЈБЕР, ЕГБЕРТ и др. 2018: Biber, Douglas and Jesse Egbert, Meixiu Zhang. "Using Corpus- Based Analysis to Study Register and Dialect Variation on the Searchable Web". In: Friginal Eric (Ed.), *Corpus-Based Sociolinguistics*, London: Routledge. 2018, p. 83–111.

БЕРБЕР САРДИЊА 2018: Berber Sardinha, Tony. "Variation in Global English: A Collocation-Based Analysis". In: Friginal, Eric (Ed.), *Corpus-Based Sociolinguistics*, London: Routledge, 2018, p. 112–135.

БАЛАСУБРАМАНИЈАН 2018: Balasubramanian, Chandrika "Indian English: A Pedagogical Model (Even) in India?" In: Friginal, Eric (Ed.), *Corpus-Based Sociolinguistics*. London: Routledge, 2018, p. 136-153.

ГОЛДБЕРГ 1995: Goldberg, Adele E. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press, 1995.

ГРИС 2013: Gries, Stefan Thomas. "Data in Construction Grammar." In: Hoffman, Thomas and Graeme Trousdale (Eds.), *The Oxford Handbook of Construction Grammar*, Oxford: OUP, 2013, p. 93–108.

ГРИС И СТЕФАНОВИЧ 2004a: Gries, Stefan Thomas and Anatol Stefanowitsch. "Extending collostructional analysis: A corpus-based perspective on 'alternations' " *International Journal of Corpus Linguistics* 9(1) (2004a): 97–129.

ГРИС И СТЕФАНОВИЧ 2004б: Gries, Stefan Thomas and Anatol Stefanowitsch. "Co-varying collexemes in the *into*-causative." In: Achard, Michel and Suzanne Kemmer. *Language, culture, and mind*, Stanford, CA: CSLI Publications, 2004b, 225–36.

ГРИВ 2016: Grieve, Jack. *Regional variation in written American English*. Cambridge, UK: Cambridge University Press, 2016.

ДЕЈВИС 2015: Davies, Mark. "Corpora: An Introduction". In: Biber, Douglas and Randy Reppen (Eds.), *Cambridge Handbook of English Corpus Linguistics*, Cambridge: Cambridge University Press, 2015, p. 11–31.

ДЕЈВИС 2018: Davies, Mark. "Using Large Online Corpora to Examine Lexical, Semantic, and Cultural Variation in Different Dialects and Time Periods". In: Friginal, Eric (Ed.), *Corpus-Based Sociolinguistics*. London: Routledge, 2018, p. 19–82.

ДЕСАГИЈЕ 2017: Desagulier, Guillaume. *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*. Cham: Springer International Publishing AG, 2017.

ЈЕНСЕН 2017: Jensen, Kim Ebensgaard. "Corpora and Cultural Cognition: How Corpus-linguistic Methodology Can Contribute to Cultural Linguistics". In: Sharifian, Farzad (Ed.), *Advances in Cultural Linguistics*, Singapore: Springer, 2017, p. 477–505.

КЛЕНСИ 2016: Clancy, Brian "Hurry up baby son all the boys is finished their breakfast: Examining the use of vocatives as markers in Irish traveler and settled family discourse". In: Amador-Moreno, Carolina P. and Kevin McCafferty, Elaine Vaughan, (Eds.), *Pragmatic markers in Irish English*, Amsterdam: John Benjamins Publishing Company, 2016, p. 229–247.

КРИСТИЈАНСЕН И ДЕРВЕН 2008: Kristiansen, Gitte and René Dirven (Eds.). *Cognitive sociolinguistics: Language variation, cultural models, social systems*. Berlin and New York: Mouton de Gruyter, 2008.

КЕНДАЛ 2011: Kendall, Tyler. "Corpora from a Sociolinguistic Perspective". *Revista Brasiliera de Linguística Aplicada* 11(2) (2011): 361–389.

КО 2016: Ko, Edwin (2016) *A corpus- based study of variation and change in adverb placement across world Englishes*. Unpublished PhD Thesis. Georgetown: University of Georgetown. <https://repository.library.georgetown.edu/bitstream/handle/10822/1040757/Ko_georgetown_0076M_13327.pdf?sequence=1&isAllowed=y> 20. 03. 2018.

КУЧЕРА И ФРАНСИС 1967: Kučera, Henry and Nelson W. Francis. *Computational analysis of Present-Day American English*. Providence, RI: Brown University Press, 1967.

МЕКЕНЕРИ И ХАРДИ 2012: McEnery, Tony and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press, 2012

МЕКЕНЕРИ И ВИЛСОН 2001: McEnery, Tony and Andrew Wilson. *Corpus Linguistics*. 2. ed. Edinburgh: Edinburgh University Press, 2001.

МЕКЕНЕРИ, ШАО и др. 2006: McEnery, Tony and Richard Xiao, Yukio Tono. *Corpus-based Language Studies: An Advanced Resource Book*. New York / London: Routledge, 2006.

ЊУМАН 2008: Newman, John. "Spoken corpora: Rationale and application". *Taiwan Journal of Linguistics* 6(2) (2008): p. 27–58.

ПИКЕРИНГ, ФРИЦИНАЛ и др. 2016: Pickering, Lucy and Eric Friginal, Shelley Staples (Eds.), *Talking at work: Corpus- based explorations of workplace discourse*. London: Palgrave Macmillan, 2016.

РИЛЕМАН 2007: Rühlemann, Cristoph. *Conversation in context: A corpus-driven approach*. London: Continuum, 2007.

СИНКЛЕР 2000: Sinclair, John. "Lexical grammar", *Naujoji Metodologija* 24 (2000): p. 194–224.

СТЕФАНОВИЧ И ГРИС 2003: Stefanowitsch, Anatol and Stefan Thomas Gries. “Collostructions: Investigating the interaction between words and constructions.” *International Journal of Corpus Linguistics* 8(2) (2003): p. 209–243.

СТЕФАНОВИЧ 2013: Stefanowitsch, Anatol. “Collostructional Analysis.” In: Hoffman, Thomas and Graeme Trousdale (Eds.), *The Oxford Handbook of Construction Grammar*, Oxford: OUP, 2013, p. 290–306.

ТАЉАМОНТЕ 2016: Tagliamonte, Sali A. “So sick or so cool? The language of youth on the internet”. *Language in Society* 45(1) (2016): p. 1–32.

ТОЊИНИ–БОНЕЛИ 2001: Tognini-Bonelli, Elena. *Corpus Linguistics at Work*. Amsterdam: John Benjamins, 2001.

ФРИЦИНАЛ И БРИСТОУ 2018: Friginal, Eric and Mackenzie Bristow “Corpus Approaches to Sociolinguistics”. In: Friginal, Eric (Ed.), *Corpus-Based Sociolinguistics*, London: Routledge, 2018, p. 1–16.

ФРИЦИНАЛ И ХАРДИ 2014: Friginal, E. and Jack A. Hardy. *Corpus-based sociolinguistics: A guide for students*. New York: Routledge, 2014.

ХАНСТОН 2002: Hunston, Susan. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press, 2002.

ХЕЈ, ПЈЕРХАМБЕРТ и др. 2015: Hay, Jennifer B. and Janet B. Pierrehumbert, Abby J. Walker, Patrick LaShell “Tracking word frequency effects through 130 years of sound change“. *Cognition*, 139 (2015), 83–91.

ХЕРАРТС 2003: Geeraerts, Dirk. “Cultural models of linguistic standardization”. In: Dirven, René and Roslyn M. Frank, Martin Pütz (eds.), C*ognitive Models in Language and Thought. Ideology, Metaphors and Meanings*, Berlin: Mouton de Gruyter, 2003, p. 25–68.

ХИЛПЕРТ 2014a: Hilpert, Martin. “Collostructional analysis: Measuring associations between constructions and lexical elements.” In: Glynn, Dylan and Justyna A. Robinson (Eds.), *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*, Amsterdam and Philadelphia: John Benjamins Publishing Company, 2014a, p. 391–404.

ХИЛПЕРТ 2014б: Hilpert, Martin. *Construction Grammar and its Application to English*. Edinburgh: Edinburgh University Press. 2014b.

ХОФМАН И ТРАУЗДЕЈЛ 2013: Hoffman, Thomas and Graeme Trousdale (Eds.), *The Oxford Handbook of Construction Grammar*. New York and Oxford: Oxford University Press, 2013.

ХУАНГ, ГУО и др. 2015: Huang, Yuan and Diansheng Guo, Alice Kasakoff and Jack Grieve. “Understanding U.S. regional linguistic variation with Twitter data analysis.” *Computers, Environment and Urban Systems*, 59 (2015): p. 244–255.

ШМЕРЧАЊИ, ГРАФМИЛЕР и др. 2016: Szmrecsanyi, Benedikt and Jason Grafmiller, Benedikt Heller, Melanie Röthlisberger. “Around the world in three alternations: Modeling syntactic variation in varieties of English”. *English World Wide* 37 (2016): p. 109–137.

Vladan O. Pavlović

ON THE USE OF MASSIVE ELECTRONIC TEXTUAL CORPORA IN SOCIOLINGUISTIC RESEARCH

The paper focuses on massive textual electronic corpora (those with more than 100 million words) of English and Serbian in sociolinguistic research. It concludes that such corpora, which have been developed in the recent years, can, despite their limitations, be prolifically used in the area in a number of ways that were not possible with previous more limited corpora, especially when it comes to advanced methods of their search and the quantitative data they can offer (including those obtained by means of statistical tests), as a valuable input for theoretical interpretation of such data and for gaining a better insight into the relationship between language and society.

Key words and phrases: massive textual electronic corpora, sociolinguistics, cultural linguistics, IT revolution, collostructional analysis.