Nadgradnja Gigafide: spletna besedila

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar, Vesna Mikolič

Abstract
This paper discusses the expansion of the Gigafida corpus, a Slovenian reference corpus, to include Internet content, i.e. webpages and user-generated content (tweets, blogs, forums and comments on news portals). The resources and tools available which are best suited to achieve this objective are discussed, and the Web crawling methodology used for this purpose is presented.

Keywords: reference corpus, Slovenian, dictionary, Internet content, Web crawling

Ključne besede: referenčni korpus, slovenščina, slovar, spletna besedila, pajkanje

1 UVOD

Leta 2012 smo poglavje z naslovom Spletna besedila v korpusu Gigafida (Logar Berginc et al. 2012: 45) začeli z ugotovitvijo, da postaja podajanje pisnega jezika v javni rabi vse manj domena tiska in vse bolj domena elektronskih medijev, ter ob tem zapisali podatek, da se je oktobra 2007 za internetne uporabnike v populaciji od 12 do 65 let v raziskavi RIS izreklo 66 % vprašanih. Novejši deleži so – pričakovano – še višji: po analizi Statističnega urada RS je imelo v prvem četrtletju leta 2014 dostop do interneta 97 % slovenskih gospodinjstev z otroki in 70 % gospodinjstev brez otrok, internet pa je v tem obdobju uporabljalo 72 % vseh oseb, starih od 16 do 74 let. Dodamo lahko še, da je

> 81 % teh oseb /.../ internet uporabljalo vsak dan ali skoraj vsak dan. Ti so ga v največjem odstotku (87 %) uporabljali za pošiljanje ali prejemanje e-pošte in za iskanje informacij o blagu ali storitvah. / 58 % oseb je v prvem četrtletju 2014 sodelovalo v družabnih spletnih omrežjih (v prvem četrtletju 2013 je bilo takih 53 %) (ibid.).

Pomemben podatek je tudi ta, da je 66 % uporabnikov do interneta dostopalo prek mobilnega telefona ali druge mobilne naprave (npr. bralnika). Internet je torej postal dostopen kjerkoli, in to ne le za branje, gledanje ter poslušanje, temveč tudi za pisanje in objavljanje besedil, slik, glasbe ipd. K široko dostopni javni besedi – nekoč omejeni na tisk, radio in televizijo – se tako lahko priglasi vsakdo, to pa v slovenščino v javni rabi prinaša nov segment: besedila, katerih jezikovna podoba kaže značilnosti, prej vezane predvsem na zasebne govorne položaje.

Oblikovalci aktualnih tujih referenčnih korpusov spletna besedila v korpuse vključujejo zelo različno. Pregled v Logar Berginc in Ljubešić (2013) je pokazal, da »skupna težnja, da bi se v referenčne korpuse vključevalo besedila z interneta in v kolikšnem obsegu bi to bilo, še ni jasno razvidna, če pa korpus že vsebuje ali bo vseboval besedila z interneta, se vanj v glavnem zajemajo besedila različnih žanrov« (ibid. 103). Tako imamo na eni strani npr. Oxford English Corpus, iz katerega nastajajo angleški slovarji Oxford, ki je skoraj v celoti sestavljen iz besedil s spleta, na drugi strani pa npr. Slovaški nacionalni korpus, na osnovi katerega se pravkar pripravlja Slovar sodobnega slovaškega jezika, ki spletnega dela sploh ne vsebuje (več gl. v Tabeli 1 v predhodnem poglavju).

Kot bo razvidno iz nadaljevanja poglavja, besedila s spletnih strani, komentarje pod prispevki na novičarskih portalih, blogovske zapise, tvite in forumska sporočila tu razumemo kot tvoren del javne pisne slovenščine, zaradi česar bi jih bilo treba zajeti tudi v korpus, ki bo podlaga za prihodnji referenčni slovar. Namreč: leksikografe bi morala zanimati leksika, ki jo v različnih okoliščinah

uporabljajo in ustvarjajo *vsi* govorci slovenščine, ne zgolj novinarji, prevajalci, pisatelji ipd. Tovrstno posebno pozornost zato danes terja prav (pol)javno pisno spletno komuniciranje, ki ga določajo okoliščine, kot so (ne)interaktivnost, (a)sinhronost, fizična (ne)prisotnost sogovornika in drugi situacijski dejavniki, katerih posledica je zelo interaktivna oblika komuniciranja z več prvinami spontanega govorjenega jezika ter z (za računalniško komuniciranje prilagojenimi) parajezikovnimi in prozodičnimi elementi (Crystal 2001). Naloga korpusa kot slovarskega vira torej mora biti zajem tudi te jezikovne realnosti, zato jo v poglavju osvetljujemo s štirih zornih kotov:

a) izhodiščnega stanja v Gigafidi (primerjalno s spletnim korpusom slovenščine slWaC$_2$),

b) raznovrstnosti spletnih besedilnih žanrov in njihove korpusne aktualnosti,

c) virov in orodij, ki so za tovrstno prihodnjo nadgradnjo Gigafide že na voljo (projekt JANES), ter

č) najustreznejše metodologije pajkanja, vključno s komentarjem možnosti gradnje spremljevalnega podkorpusa.

2 GIGAFIDA IN slWaC$_2$: STANJE, PRIMERJAVA, POVEZLJIVOST

Spletne strani, ki so bile vključene v Gigafido, in tehnologije za njihov zajem so natančneje opisane v že omenjenem poglavju knjige Logar Berginc et al. (2012: 45–67), zato naj spomnimo le, da je šlo pri vključevanju spletnih vsebin v Gigafido »v metodološkem smislu za prvi večji /dotedanji/ tak poskus pri nas, ki bi lahko oblikoval smernice za prihodnjo gradnjo referenčnih korpusov slovenščine ter nakazal nekatere zanimive (besedilnozvrstnoprimerjalne) jezikoslovne analize« (ibid.: 45). Gigafida tako vsebuje besedila 10 novičarskih portalov ter skupno 91 predstavitvenih spletnih strani podjetij (29) in ustanov (62). Pajkanje je bilo izvedeno v obdobju april 2010–april 2011, v korpus pa je prineslo več kot 185 milijonov besed, od kateri jih 63 % prihaja z novičarskih portalov (24ur.com, rtvslo.si, siol.net ipd.), 30 % s strani ustanov (gov.si, uni-lj.si, sazu.si, ijs.si itd.) ter 7 % s strani podjetij (eles.si, gorenje.si, kolosej.si itd.). Postopek za zajem besedil s spletnih strani je vseboval več korakov: izbor in pripravo programa za tri režime pajkanja (dnevno, mesečno ter enkratno), odstranjevanje spremnih in vnaprej pripravljenih besedil, detekcijo jezika ter na koncu še detekcijo dvojnikov in približnih dvojnikov. Izkazalo se je, da je za uresničitev na videz dokaj preproste naloge, tj. naloge vključitve spletnih besedil v referenčni korpus, potrebna dokaj kompleksna metodologija, ki pa smo jo – vključno z merili izbora spletnih strani in oceno pridobljenega – uspešno

preizkusili ter prilagodili slovenščini (več o najnovejših metodah pajkanja gl. v razdelku 5).

Prav v času vključevanja spletnih besedil v Gigafido – leta 2011 – pa je nastal še en, v tem segmentu metodološko podoben korpus slovenščine: korpus slWaC (Ljubešić in Erjavec 2011), ki je bil leta 2014 nadgrajen v slWaC$_2$ (Erjavec in Ljubešić 2014). Korpus slWaC$_2$ vsebuje 1,2 milijardi pojavnic iz besedil, pridobljenih z dobrih 37.000 spletnih domen oz. 2,8 milijonov naslovov URL. Metodologija gradnje obeh različic korpusa slWaC je podrobneje predstavljena v obeh navedenih virih ter v Logar Berginc in Ljubešić (2013: 87–89).

Obstoj dveh obsežnih korpusov slovenščine je že spodbudil nekatere primerjave, ki so pokazale, kaj v enem in drugem korpusu *je*, omejeno (kolikor jih pač daje primerjava dveh entitet) pa tudi, kaj v obeh korpusih manjka. Primerjava po metodi frekvenčnega profila (Rayson in Garside 2000) med Gigafido ter korpusom slWaC$_2$ (Erjavec et al. 2015: 40) je tako med drugim pokazala, da je v slednjem več besedil, povezanih z računalništvom in spletom, ter uporabniških spletnih vsebin, medtem ko Gigafida vsebuje več besedil o športu, predvsem pa več za časopise značilnih besedil o notranji politiki, gospodarstvu in kaznivih dejanjih. Primerjava je pokazala tudi tehnične razlike, saj je bila Gigafida označena s programom Obeliks (Grčar, Krek in Dobrovoljc 2012), medtem ko smo za slWaC$_2$ uporabili program ToTaLe (Erjavec et al. 2005). Ker se orodji mestoma razlikujeta v označevanju, so se nekatere besede izpostavile kot zelo ključne za slWaC$_2$, ne pa za Gigafido, čeprav gre v resnici samo za razliko v obdelavi obeh korpusov. Tako je *le-ta* enkrat obravnavan kot ena pojavnica, drugič kot tri, *več* pa je npr. enkrat lematiziran kot »veliko«, drugič pa kar kot »več«. Pri kakršnemkoli povečevanju Gigafide bo zato treba pri izbiri orodij za jezikovno označevanje imeti v mislih smiselnost uporabe enotnega orodja za vse (prihodnje) slovenske korpuse (Erjavec et al. 2015).

Že objavljenim primerjalnim podatkom (Erjavec et al. 2015 ter Logar Berginc in Ljubešić 2013) tokrat dodajamo še podatke iz novejše primerjave med Gigafido in korpusom slWaC$_2$, narejene po metodi tematskega modeliranja (Blei et al. 2003; Sharoff 2010); s tem da bomo tu pri tematskih profilih obeh korpusov pozorni le na morebitne šibkosti Gigafide (Logar 2015). Tabeli 1 in 2 tako kažeta 20 za Gigafido in slWaC$_2$ najbolj značilnih tem.

Tabela 1: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po teži v korpusu Gigafida.

Tema	Teža*	Samostalniška lema
človek, moški, ženska, družina, življenje	4,835	otrok leto dan čas ženska življenje človek družina oče moški roka prijatelj glava žena mama mož sin starš hiša
šport	4,034	tekma mesto leto ekipa zmaga točka igra sezona igralec prvenstvo klub liga prvak trener minuta konec pokal krog reprezentanca
notranja politika	3,639	predsednik vlada država stranka svet minister leto zakon volitev predlog poslanec vprašanje komisija član odbor zbor seja politika ministrstvo
družba, RAZNO	3,631	človek življenje svet čas odnos način stvar država vprašanje družba primer beseda delo moč stran problem resnica leto občutek
prireditve v lokalnem prostoru	2,865	ura društvo leto prireditev dan sobota dom član vas mesto občina skupina nedelja šola srečanje gost obiskovalec dvorana delo
vojna, terorizem, kazniva dejanja	2,669	leto vojna država policija policist človek vojska dejanje orožje dan napad vojak žrtev sodišče oblast zapor kazen čas mesto
TV- in radijski program	2,622	film leto glasba oddaja tv poročilo skupina serija dan pesem festival čas koncert predstava program vloga gledališče del novica
promet	2,481	cesta pot dan nesreča leto ura voda voznik morje vozilo mesto meter promet letalo kilometer čas vožnja kraj avtomobil
gospodarstvo	2,47	leto odstotek država podjetje cena trg plača izdelek rast razvoj delo gospodarstvo proizvodnja delavec število področje strošek mesec sistem
izobraževanje	2,363	šola delo leto otrok program področje znanje študent projekt izobraževanje univerza fakulteta učenec razvoj starš organizacija učitelj center zavod
finance	2,292	milijon evro tolar leto banka družba podjetje odstotek delnica milijarda dolar denar vrednost cena prodaja delež dobiček trg sklad
lokalna (prostorska) politika	2,228	občina leto prostor gradnja objekt cesta projekt območje delo zemljišče mesto milijon stanovanje okolje podjetje načrt denar voda tolar
živali, narava, bivalno okolje	2,153	žival barva prostor vrsta pes voda hiša gozd del material les vrt tla drevo konj čas leto vrata oblika

Tema	Teža*	Samostalniška lema
pravo	2,145	zakon člen sodišče postopek pravica primer podatek organ odstavek dan oseba podlaga pogodba delo odločba sklad stranka določba zadeva
publikacije, kultura, umetnost	2,087	leto knjiga delo razstava stoletje cerkev mesto čas muzej svet ime zbirka umetnost avtor jezik zgodovina del slika beseda
avtomobilizem	2,021	m sit avtomobil motor km cena vozilo eur d e l model leto avto x n g r h
zdravje	1,942	bolezen zdravnik bolnik zdravilo telo človek zdravljenje leto koža težava dan zdravje primer rak bolnišnica bolečina kri celica čas
mediji	1,788	naslov stran številka medij novinar revija dan nagrada pošta časopis leto ime delo informacija oddaja članek televizija bralec vprašanje
informacijsko-komunikacijska tehnologija	1,491	računalnik sistem uporabnik podatek program slika stran naprava uporaba kartica telefon zaslon internet omrežje model oprema tehnologija možnost storitev
hrana	1,437	vino voda olje rastlina minuta meso sladkor g sol hrana zelenjava jed žlica okus sadje mleko krompir sok list

* »Teža« v drugem stolpcu pomeni razpršenost posamezne teme v korpusu.

Tabela 2: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po teži v korpusu slWaC$_2$.

Tema	Teža	Samostalniška lema
človek, moški, ženska, družina, življenje	3,929	otrok dan čas leto človek ženska roka pes življenje stvar prijatelj moški glava mama ura družina starš svet konec
družba, RAZNO	3,266	človek življenje svet čas način odnos stvar družba otrok ljubezen beseda primer vprašanje resnica pot občutek ženska moč problem
notranja politika	2,626	vlada država predsednik stranka zakon svet leto predlog minister član poslanec komisija vprašanje zbor odbor politika skupina pravica mnenje
potovanja, turizem	2,524	pot mesto dan cesta ura leto čas vrh morje voda smer gora meter del dolina kraj gozd hotel stran
gospodarstvo, razvoj	2,36	podjetje področje razvoj sistem projekt delo leto trg storitev okolje država cilj organizacija program izdelek znanje rešitev sodelovanje tehnologija
finance	2,265	leto evro odstotek milijon podjetje banka država cena družba denar trg vrednost rast milijarda sredstvo delnica plača prodaja mesec

Tema	Teža	Samostalniška lema
šport	2,232	tekma ekipa mesto igra leto točka zmaga sezona igralec minuta prvenstvo klub liga konec tekmovanje prvak rezultat trener pokal
prireditve (film, glasba, gledališče)	2,139	film leto glasba skupina album pesem festival koncert skladba čas oder nastop predstava nagrada vloga dan zasedba oddaja zgodba
izobraževanje	2,072	šola otrok leto delo program študent učenec znanje starš ura izobraževanje fakulteta univerza čas študij delavnica področje učitelj dan
zdravje	2,059	telo bolezen koža zdravilo težava zdravljenje zdravnik dan leto bolnik bolečina človek zdravje celica primer čas kri otrok učinek
spletno nakupovanje	2,042	stran podatek uporabnik naslov storitev vsebina račun pošta cena ime nakup internet številka informacija izdelek naročilo ponudba dan paket
pravo	2,016	člen zakon sodišče postopek pravica odstavek oseba pogodba primer dan stranka podlaga sklad organ delo določba odločba podatek pogoj
lokalna (prostorska) politika	1,937	občina leto projekt društvo območje mesto delo prostor sredstvo objekt program član center gradnja zavod organizacija področje okolje ministrstvo
religija	1,887	leto cerkev človek vojna bog življenje dan mesto čas smrt vojska svet država oče maša ime beseda vera stoletje
publikacije, kultura, umetnost	1,826	leto knjiga delo jezik razstava avtor medij beseda fotografija nagrada zbirka revija del umetnost zgodba naslov čas svet dogodek
informacijsko-komunikacijska tehnologija	1,781	računalnik naprava sistem slika program telefon fotografija podatek uporabnik uporaba video zaslon stran aplikacija dokument model kamera različica oprema
avtomobilizem	1,697	vozilo avtomobil motor barva vožnja voznik model kolo avto del leto oblačilo znamka cesta hitrost obleka sedež oprema sistem
bivalno okolje	1,624	voda prostor energija hiša material sistem površina odpadek zrak objekt naprava del uporaba stanovanje temperatura okno okolje les plin
hrana	1,471	hrana voda olje rastlina vino mleko okus meso zelenjava vrsta jed sadje izdelek oseba količina dan kislina žival sladkor
svetovni splet	0,52	piškotek dan nastavitev seja mesto namen stran storitev uporaba informacija podatek oglaševanje klik gumb primer ura facebook možnost novica

V tabelah lahko prepoznamo tri teme, ki jih bo treba še posebej imeti v mislih pri izboru naslovov URL, s katerih bodo pri nadgradnji Gigafide pridobivana nova spletna besedila (o tvitih, forumskih sporočilih, komentarjih pod prispevki na novičarskih portalih in blogovskih zapisih gl. v nadaljevanju). Gre za teme: *potovanja, turizem*; *spletno nakupovanje* in nasploh *svetovni splet* (gl. zadnjo vrstico v Tabeli 2). Tema *religija* je edina, ki bi se jo dalo bolje vključiti v Gigafido že tudi s tiskanimi deli (je pa pri tem ključen odziv besedilodajalcev).

Samo po sebi se ob koncu primerjav v razmislek ponuja vprašanje morebitne kar neposredne vključitve spletnega korpusa slovenščine slWaC$_2$ v novo Gigafido. Z vidika bolj usmerjenega in kontroliranega, pa tudi časovno predvidljivega ter enakomernega zajema besedil z izrecnim namenom vključitve v referenčni korpus je na to vprašanje bolje odgovoriti negativno, ni pa treba, da bi bili tudi prihodnji nadgradnji obeh korpusov povsem ločeni. Nasprotno: kot je razvidno v razdelku 5, ju tesno povezuje metodologija izdelave, obstoj dveh korpusov sodobne slovenščine pa je koristen tudi v smislu medsebojnega dopolnjevanja ter izkaza medsebojnih razlik in pomanjkljivosti.

3 SPLETNA BESEDILNA ŽANRSKOST IN SLOVARSKI VIRI

Na spletu kot najvplivnejšem mediju 21. stoletja se srečujemo z različnimi komunikacijskimi okolji oz. področji in vsemi štirimi osnovnimi funkcijami besedil (Skubic 1995; Mikolič 2007): spoznavno, sporazumevalno, izvršilno ter umetnostnoizrazno. Na spletu se oblikujejo različne diskurzivne/govorne skupnosti, ki se v okviru določenega diskurza/govora odločajo za zanj značilne jezikovne izbire. Nekatere funkcije spletnih besedil so tako vezane na (bolj) neformalne govorne položaje in se udejanjajo v besedilih s številnimi nestandardnimi jezikovnimi prvinami, druga spletna besedila pa ustrezajo pojmu javnega komuniciranja v ožjem pomenu besede (Škiljan 1999) in sooblikujejo jezikovni standard. Jezikovna heterogenost spleta seveda močno vpliva na spremembe v jeziku in širjenje njegove leksike, zato je nujno ugotoviti, katera besedila morajo biti sestavni del korpusa, ki bo temeljni vir za sodobni slovar slovenskega jezika (ter hkrati, katerim se je mogoče (zaenkrat) odreči).

Opis spletne zvrstnosti in njenih ključnih dejavnikov je pravzaprav zelo nehvaležna naloga, tako zaradi obsežnosti oz. neobvladljivosti gradiva kot zaradi maloštevilnih raziskav spletnih žanrov in njihovih ciljnih javnosti (Crowston 2010: 17, 26). Kljub temu lahko na osnovi pregledane literature (Bishop 2009; Crowston 2010; Domingo in Heinonen 2008; Herring et al. 2004; Oblak et al. 2005) ugotovimo, da pri vseh avtorjih izstopata predvsem dve ključni merili, pomembni

tako za analizo spletne besedilne zvrstnosti kot za utemeljitev izbora spletnih besedil za korpus:

- avtorstvo oz. razmerje med sporočevalcem in naslovnikom (eden ali več avtorjev, isti, znan vir informacij ali več različnih, tudi anonimnih virov, formalno ali neformalno razmerje, ki zahteva, da je diskurz ne glede na število avtorjev in virov informacij bolj ali manj formalen in notranje konsistenten),
- funkcija ter z njo povezana notranja in zunanja zgradba oz. oblika besedila in večja ali manjša formalnost govora (tudi večkodnost, posodabljanje).

Z vidika avtorstva je osnovna delitev spletnih besedil na:

- klasične spletne strani (HTML), pri katerih je avtor en sam (npr. lastnik osebne spletne strani) ali je vir besedil isti in znan (npr. podjetje s svojo spletno stranjo) in pri katerih gre večinoma za enosmerno komuniciranje,
- žanre spletnih skupnosti (angl. *web-based community genres*; Bishop 2009), pri katerih besedila soustvarja več avtorjev in gre torej vedno za večsmerno komuniciranje,
- blogovske zapise kot vmesni žanr med eno- in dvosmerno obliko komuniciranja.

Za **klasične spletne strani** je v glavnem značilno enosmerno komuniciranje (najpogostejša izjema so tu sicer medijska spletna mesta, ki lahko vključujejo forumska sporočila, komentarje ali blogovske zapise bralcev). Vir besedil je pri spletnih straneh znan oz. lahko določljiv. Razmerje med sporočevalcem in naslovnikom je večinoma formalno, besedila nagovarjajo širšo javnost, zato je tudi diskurz v glavnem formalen. Med klasične spletne strani uvrščamo:

- spletne portale (tudi tipa Wikipedija, Wikivir, Wikiverza, Wikiknjige ipd.),
- medijska spletna mesta,
- komercialne in korporativne spletne strani,
- spletne strani vladnih in nevladnih organizacij ter lokalne samouprave.

Ena od oblik spletnih strani so tudi osebne spletne strani, ki pa so manj formalne in se lahko že približujejo obliki ter namenu blogovskega zapisa in žanrov spletnih skupnosti (npr. stranem na Facebooku).

Žanri spletnih skupnosti so h kolektivnemu delovanju usmerjena spletna mesta oz. interaktivne besedilne vrste računalniško posredovanega komuniciranja, pri katerem sodeluje več avtorjev, določajo pa ga prevladujoči akterji, komunikacijsko

okolje ali tema in notranja zgradba. Tudi jezikovne izbire tu določa narava interakcije med akterji, ti pa so zelo raznovrstni, pogosto tudi anonimni, zato je izraznost besedil zelo pestra, večinoma pa vključuje neformalne jezikovne prvine. Ker ti žanri vse bolj nadomeščajo govorno komuniciranje, gre pogosto za zapisan govorjeni jezik, pri nekaterih aplikacijah tudi za govorjena besedila. Med žanre spletnih skupnosti lahko uvrščamo besedila različnih spletnih orodij in družbenih omrežij, kot so:

- forumska sporočila (uporabniki razpravljajo na določeno temo),
- Twitter, Facebook, Myspace, LinkedIn (besedila, kot so: tviti, stanja oz. misli, komentarji stanj, fotografije, videoposnetki, hiperpovezave, skupine glede na interese, ustvarjanje dogodkov, povabila idr.),
- Instagram (objavljanje fotografij, povezave na Twitter ali Facebook),
- Ask.fm (uporabniki ustvarijo račun, ostali uporabniki jim postavljajo vprašanja, povezava na Twitter ali Facebook),
- Snapchat (mobilna aplikacija, prek katere uporabniki s svojimi prijatelji delijo misli, videoposnetke, fotografije itd., ki izginejo čez nekaj minut),
- Viber (mobilna aplikacija pametnega telefona, po kateri se komuniciranje odvija prek spleta, lahko je pisno ali govorno, vključuje pa imenik uporabnikov),
- spletne klepetalnice (različne kategorije – »sobe«, kjer se povezujejo uporabniki z enakimi interesi, namenjene so spoznavanju novih ljudi),
- komentarji novinarskih prispevkov, videoposnetkov itd. (razvijejo se v debato na določeno temo med ponavadi nepoznanimi uporabniki in delujejo na način foruma).

Blogovski zapisi so največkrat del publicistične besedilne vrste, namenjene širši javnosti in so pogosto v neposredni interakciji z njo. Avtor je ponavadi en sam in znan, saj se bralci največkrat prav zaradi avtorja odločajo za obisk bloga, kar veča njegovo branost in vpliv. Ker jih lahko pišejo profesionalni ali polprofesionalni pisci (novinarji in druge znane osebe iz javnega življenja, ki so bolj ali manj vešče javnega komuniciranja), pa tudi neprofesionalni avtorji, so jezikovne izbire odvisne od piščeve sporazumevalne zmožnosti, predvsem pa od vrste občinstva, ki ga želi pisec doseči. Po Domingo in Heinonen (2008) lahko ločimo naslednje vrste publicističnih blogovskih zapisov, ki se razlikujejo po profesionalnosti piscev in institucionaliziranosti okolja:

- državljanski blogovski zapisi (pišejo jih neprofesionalni pisci izven medijev),

- blogovski zapisi občinstev (pišejo jih neprofesionalni pisci v okviru medijev),
- novinarski blogovski zapisi (pišejo jih novinarji izven medijev),
- medijski blogovski zapisi (pišejo jih novinarji v okviru medijev).

Novinarskim blogovskim zapisom izven medijev so podobni blogi različnih oseb iz javnega življenja, ki so profesionalni ali polprofesionalni pisci (pisatelji, igralci, pevci, politiki ipd.). Enako lahko tudi besedila v četrti skupini, t. i. medijske blogovske zapise, pišejo tudi drugi profesionalni pisci, ne zgolj novinarji, npr. pisatelji, igralci, režiserji ipd.; tovrstni zapisi so pravzaprav redne rubrike z mnenji in komentarji (dokaj) stalnih avtorjev, ki ne izražajo nujno stališč medijev, v katerih gostujejo ali so pri njih zaposleni.

Z vidika *funkcije* besedila razvrščamo v skupine širših besedilnih zvrsti in ožjih besedilnih vrst, in sicer glede na naslednje skupne lastnosti: namen oz. vplivanjsko vlogo, naslovnika, referenco, zunanjo in notranjo zgradbo besedila ter z njimi povezano večjo ali manjšo notranjo konsistentnost in formalnost diskurza (prim. Mikolič 2013; Nidorfer Šiškovič 2013). Po teh lastnostih smo naredili preliminarno klasifikacijo besedil v spletnem okolju, pri čemer smo se naslonili na Crowstona (2010), ki povzema ključne tipologije spletnih besedil glede na namen in obliko. Tako lahko poskusimo besedilno zvrstnost na spletu opisati v okviru spodnjih skupin besedilnih vrst (povzeto po Mikolič in Rolih 2015), ki vse, razen pogovarjalnih z večjim obsegom vsebinskih in jezikovnih elementov zasebnega komuniciranja, svoj namen dosegajo v javnem komuniciranju s ciljnimi javnostmi ali pa širšo javnostjo, zato se v večini primerov v njih oblikuje formalni govor in standardni jezik:

- *Pogovarjalne in vsaj v delu zasebne besedilne vrste*: elektronska pošta, spletne klepetalnice, tviti in druga besedila družbenih omrežij (npr. Twitter, Facebook) ter forumska sporočila. Pripis »zasebnosti« pri teh besedilih s povezovalno-izrazno komunikacijsko vplivanjsko vlogo sloni na zanje značilnem večjem obsegu prvin zasebnega jezika oz. vsebinskih elementov zasebnih komunikacijskih sfer (Škiljan 1999), sociolektov in idiolektov (Skubic 2004).
- *Predstavitvene in promocijske besedilne vrste*: osebne spletne strani, spletne strani podjetij, ustanov, organizacij, društev ipd. Gre za besedila, namenjena širši javnosti, pri katerih se prepletata predstavitvena in usmerjevalna vplivanjska vloga. Osebne spletne strani imajo poleg predstavitvenega tudi samopromocijski namen, istočasno pa nastajajo z namenom vzpostavljanja družbenih mrež, saj želi avtor obiskovalcem svoje spletne strani sporočiti svoja stališča, poglede, poročati o svojem delu in s tem

vzpostaviti tesnejše medsebojne vezi. Predstavitveni namen spletnih strani različnih podjetij in organizacij se pogosto tesno prepleta s komercialnimi nameni.

- *Oglasne in komercialne besedilne vrste*: oglasna sporočila, zbirke povezav (angl. *link collections*), spletne trgovine, spletni portali za trženje in prodajo. Usmerjevalna vplivanjska vloga teh besedil se kaže v njihovem vplivanju na nakupno ravnanje naslovnikov.
- *Poročevalske in širšepublicistične besedilne vrste*: novinarska besedila različnih žanrov, spletne izdaje tiskanih medijev, prispevki, povezani z življenjskimi stili (kuharski recepti, nasveti v obliki t. i. tutorialov, vodiči za zdravo telo itd.).
- *Programerske besedilne vrste*: tehnični podatki/pomoč/podpora, poročila o težavah (angl. *problem reports*), pogosta vprašanja ali FAQ (angl. *frequently asked questions*). Tu gre za (poljudno)strokovna besedila, namenjena širši javnosti, ki jih na spletno stran postavijo upravljalci oz. programerji spletnih strani.
- *Akademske besedilne vrste* (dostopne npr. prek Googlovega Učenjaka): znanstvena in strokovna besedila s spoznavno in predstavitveno vplivanjsko vlogo, namenjena akademski ter strokovni javnosti.
- *Uradne in uradovalne besedilne vrste*: zapisniki sej državnih organov, zakonodajne strani, strani borze, pravilniki itd.; e-uprava, e-prijave ipd. Namen teh besedil z izvršilno vplivanjsko vlogo je seznaniti širšo javnost s ključnimi postopki, pravili in zakoni v državi ter hkrati omogočiti uradovanje z upravnimi organi prek spletnih obrazcev.
- *Literarne in polliterarne besedilne vrste*: umetnostna besedila, za katera je značilna skladnost z jezikovnim standardom z namernim odstopanjem od njega in individualizacijo jezikovnega stila. Najpogostejši polliterarni spletni vrsti sta blogovski zapis in spletni dnevnik.

Nedvomno je večina naštetih spletnih besedilnih vrst – čeprav še premalo raziskanih tako v slovenskem kot mednarodnem merilu – tudi v slovenščini zelo živih v smislu njihovega uresničevanja, branosti in vpliva tako v okviru različnih vrst družbenega komuniciranja kot v okviru razvoja jezika. Zaradi hitrega spreminjanja spletnih orodij lahko nekatere vrste hitro zastarajo (ta hip so v zatonu npr. spletne klepetalnice), namesto njih pa se pojavijo nove, lahko s podobnim ali povsem drugačnim namenom ter samosvojo jezikovno podobo. Prav zato je treba pri pripravi slovarskih opisov ne le *upoštevati* spletno jezikovno stvarnost, temveč ji tudi sproti *slediti*.

Ob predpostavki, da se bo iz nadaljnjih analiz spletnega gradiva zgornji nabor spletnih besedilnih vrst potrdil kot relevanten tudi za slovenski jezik, se zdi pri nadgradnji Gigafide utemeljeno upoštevati tisti del besedilnih vrst, ki imajo znanega avtorja oz. vir, svoj namen dosegajo v javnem komuniciranju in zato bolj neposredno sooblikujejo jezikovni standard. Kot taka se kažejo predvsem besedila vseh skupin klasičnih spletnih strani in tistih osebnih spletnih strani, ki so široko brane, dalje blogovski zapisi profesionalnih in polprofesionalnih piscev, tviti in Facebookove strani oseb ter ustanov, ki imajo večji vpliv na splošno jezikovno rabo (število sledilcev, medijski odzivi). Z vidika funkcije gre za del pogovarjalnih besedilnih vrst ter večji del predstavitvenih in promocijskih besedil, dalje poročevalske in širšepublicistične besedilne vrste, uradne in uradovalne ter literarne in polliterarne besedilne vrste – vse, kot že rečeno, ob pogoju večje vplivnosti in široke branosti ter na način, da bo iz taksonomskih kategorij korpusa vidno, za katere žanre gre.

4 NESTANDARDNO ZAPISANA SPLETNA BESEDILA PROJEKTA JANES

Kot smo zapisali že zgoraj, je za jezik pisnega spletnega komuniciranja značilna pogosta raba nestandardnih jezikovnih oblik, kot so nestandardni zapis besed in specifične krajšave. Zaradi tega je jezikoslovna analiza in posledično tudi avtomatska obdelava tovrstnih vsebin otežena (Sproat et al. 2001), prizadevanja za premostitev teh ovir pa so trenutno ena bolj vročih tem na področju računalniškega jezikoslovja. Leta 2014 se je tudi za sodobno slovenščino začelo delo na tem področju, in sicer v okviru temeljnega raziskovalnega projekta JANES (Jezikoslovna analiza nestandardne slovenščine), ki ima za enega od ciljev zgraditi vire in razviti orodja ter vzpostaviti metodologijo za proučevanje jezika spletnega komuniciranja (Fišer et al. 2014).

4.1 Korpus spletne slovenščine JANES

V trenutno različico korpusa JANES smo vključili štiri zvrsti uporabniških spletnih vsebin, in sicer tvite, forumska sporočila, komentarje pod prispevki na novičarskih portalih in blogovske zapise. Tviti se zajemajo z namenskim orodjem TweetCat (Ljubešić et al. 2014), pri čemer zajem poteka sproti, trenutno že več kot dve leti. Za zajem forumskih sporočil in novičarskih portalov, s katerih smo pridobili komentarje uporabnikov, smo izbrali po nekaj virov, ki so v slovenskem spletnem prostoru najbolj priljubljeni, ponujajo največ jezikovne produkcije in/ali predstavljajo pomemben del slovenskega spletnega prostora. Ker se spletna mesta po sestavi med seboj razlikujejo, smo uporabili ciljno pajkanje (več gl. v

razdelku 5.2) in za vsak vir posebej napisali ekstraktor besedila. Za gradnjo podkorpusa blogovskih zapisov smo uporabili kar deduplicirano različico splošnega korpusa slovenskega spleta slWaC$_2$, iz katerega smo zajeli vsa besedila, pri katerih se v domeni pojavi niz »blog«. Rešitev je začasna, saj za razliko od podkorpusov forumov in komentarjev zanje zaenkrat še nismo izdelali ciljnega ekstraktorja, tako da nimamo ohranjene notranje strukture blogovskih zapisov, npr. razdelitve na glavno besedilo in komentarje nanj.

Vsi našteti podkorpusi so bili združeni v korpus JANES, v katerem so poenoteni in s tem tudi poenostavljeni metapodatki posameznih besedil. Podkorpusi in korpus JANES so zapisani v formatu XML, ki omogoča strukturiranje korpusa, zapis metapodatkov in konsistenten zapis znakov po standardu Unicode. Korpus smo tudi jezikoslovno označili. Prvi korak označevanja sta bili tokenizacija in stavčna segmentacija, za kar smo uporabili standardno knjižnico mlToken za slovenski jezik, ki je del programa ToTaLe (Erjavec et al. 2005). V naslednjem koraku smo besedne pojavnice normalizirali z metodo, ki temelji na statističnem strojnem prevajanju črk, naučena pa je bila na 1.000 ključnih besedah iz korpusa tvitov glede na korpus Kres (Logar Berginc et al. 2012: 77–97) in na njihovih ročno normaliziranih oblikah (Ljubešić et al. 2014). Z orodji za standardno slovenščino programa ToTaLe smo nato normalizirane besede še oblikoskladenjsko označili in lematizirali (več o jezikovnem označevanju gl. v Erjavec et al. 2015a).

Korpus JANES, ki je, kot je razvidno iz zgornjega, omejen na javne spletne uporabniške vsebine, trenutno vsebuje dobrih 161 milijonov pojavnic. Največji delež v korpusu predstavljajo tviti (38 %), sledijo jim forumska sporočila (29 %), blogovski zapisi (24 %) in komentarji pod prispevki na novičarskih portalih (9 %). Že trenutni korpus je tako uporaben za leksikografske namene, saj je vsebinsko komplementaren z Gigafido, je zadosti velik in tudi razmeroma raznovrsten. Seveda pa bi bilo koristno še povečati število in raznovrstnost virov, predvsem pri forumskih sporočilih in komentarjih pod prispevki na novičarskih portalih.

4.2 Nestandardni jezik v korpusu JANES

Korpus JANES sicer vsebuje uporabniške spletne vsebine, nikakor pa niso vsa besedila v njem napisana v nestandardni slovenščini. Ročni pregled manjšega števila naključno izbranih tvitov je celo pokazal (Ljubešić et al. 2015), da je takšnih besedil v korpusu manj kot tretjina, česar sprva nismo pričakovali. A izkazalo se je, da družbena omrežja za obveščanje javnosti in promocijo uporabljajo tudi številna podjetja ter ustanove, kot so časopisne in druge medijske hiše, javne ustanove itd., ki tipično generirajo več besedil kot zasebni uporabniki ter v svojem uradnem komuniciranju uporabljajo standardno slovenščino.

Za proučevanje nestandardne slovenščine bi bila zato koristna ločitev spletnih (in drugih) besedil, ki so zapisana v nestandardnem jeziku, od vseh ostalih. Za ta namen smo razvili metodo za merjenje stopnje nestandardnosti (Ljubešić et al. 2015), pri kateri smo v prvem koraku analizirali večje število besedil in ugotovili, da je treba ločiti tehnično od jezikoslovne nestandardnosti. Prva pomeni npr. to, da pisec piše vse besede z malo začetnico ali ne uporablja ločil, druga pa npr. to, da besede zapisuje pogovorno, uporablja sleng ali drugo nestandardno leksiko. Za ti dve razsežnosti smo oblikovali navodila za označevalce, ki so nato ročno označili manjše vzorce iz podkorpusov JANES z oceno njegove tehnične oz. jezikoslovne nestandardnosti od 1 (zelo standardno) do 3 (zelo nestandardno).

Definirali smo okoli trideset značilk, za katere smo predpostavili, da bi lahko služile kot delne mere ene ali druge vrste nestandardnosti. Značilke so bodisi na nivoju znakov (npr. razmerje števila ločil glede na število vseh znakov v besedilu), na nivoju nizov (npr. razmerje besed z veliko začetnico do vseh besed) ali pa na nivoju leksike, pri čemer smo pozorni predvsem na razmerje leksike opazovanega besedila do besed v oblikoskladenjskem leksikonu Sloleks (Dobrovoljc et al. 2015), npr. razmerje kratkih besed, ki so v besedilu, a jih ni v Sloleksu. S temi značilkami smo na učni množici naučili regresor, ki lahko novim besedilom pripisuje obe meri nestandardnosti. S tem ko smo vsa besedila v korpusu JANES opremili z metapodatkom o stopnji tako tehnične kot jezikoslovne nestandardnosti, smo dobili možnost, da se pri korpusnih študijah osredotočimo samo na tisti del korpusa, ki je napisan v želenem tipu in stopnji (ne)standardnosti, leksikografu pa ta podatek lahko služi kot dobra redakcijska orientacija, saj lahko npr. za proučevanje nestandardnih zapisov besed iz korpusa izdvoji samo besedila, označena kot zelo jezikovno nestandardna.

5 METODOLOGIJA PAJKANJA

Pajkanje je proces, pri katerem z avtomatskimi metodami s spleta zajemamo dokumente bodisi za izdelavo indeksov spletnih iskalnikov, pridobivanje drugih informacij s spleta ali pa za izdelavo korpusov. Če je pri prvem najpomembnejši visok priklic, smo pri drugem bolj osredotočeni na pridobivanje jezikovnih vsebin, saj je bolje izgubiti dele zajetih dokumentov kot pa izdelati zelo šumen korpus, ki bi poleg zveznega besedila vseboval tudi elemente, kot so glave in noge spletnih strani, navigacijo itd.

Obstajata dva osnovna pristopa k pajkanju jezikoslovno zanimivih podatkov. Prvi, *generični* pristop, za vse zajete dokumente uporablja isti postopek obdelave in ima to prednost, da je enostavnejši za implementacijo ter pokriva zelo raznovrstne tipe dokumentov, ima pa tudi vrsto slabosti. Zajeti podatki so bolj šumni

in slabše strukturirani, saj npr. ne razlikujejo naslovov in podnaslovov od samega besedila, ravno tako pa ne vsebujejo potencialno koristnih metapodatkov o besedilu, kot so npr. datum in čas objave, ime avtorja ter razvrstitve dokumentov v vsebinske kategorije. Drugi pristop je *ciljni*. Pri njem je implementacija pajkanja prilagojena posameznemu izvoru dokumentov. Prednosti ciljnega pristopa so manjša količina šuma, boljša struktura zajetih besedil in večja količina metapodatkov, slabost pa je ta, da je treba program za zajem prilagoditi vsakemu izvoru posebej, kar je večinoma časovno zahteven proces.

Generični pristop se uporablja pri izdelavi velikih zbirk besedil, ki temeljijo bodisi na neki vrhnji spletni domeni (kot npr. ».si«) oz. posameznem jeziku (dober primer tega pristopa je korpus slWaC). Ciljni pristop je primernejši za izdelavo manjših zbirk besedil, pri katerih so zaradi specifik raziskav še posebej pomembni struktura in metapodatki besedil (primer tega pristopa je v prejšnjem razdelku predstavljeni korpus JANES).

Pajkanje se tipično začne z začetnim naborom spletnih dokumentov, nato pa pajek s pomočjo povezav v dokumentih zbira nove dokumente. Vprašanje, ki se tu postavlja, je, kako omejiti nabor zajetih dokumentov, da ne bodo bodisi v napačnem jeziku ali napačne zvrsti glede na namene pajkanja. Razlikujemo dva osnovna pristopa k določanju dokumentov za pajkanje. Prvi temelji na *omejitvah naslovov URL*, npr. na vrhnji domeni ».si« ali na »med.over.net«, drugi pa na *seznamu ključnih besed*, ki definirajo ciljno področje diskurza, kot je npr. okolje, turizem, kulinarika itd. Za slednje je tipično, da se za zajem dejanskih naslovov URL za pajkanje uporablja katerega od programskih vmesnikov (API) spletnih iskalnikov. Večina pajkanja za splošne korpuse spletnih besedil se izvaja s pomočjo omejitev naslovov URL (na ta način je pajkanje potekalo tudi pri obstoječi Gigafidi), za specializirane korpuse pa je primernejša izbira s pomočjo seznama ključnih besed, pri čemer sta bolj znani orodji za ta pristop BootCaT (Baroni in Bernardini 2004) in WebBootCaT (Baroni et al. 2005).

Na spletu najdemo dokumente v več različnih formatih. Najpogostejši so dokumenti HTML, ki so za zajem v korpus problematični zato, ker je velik del njihove vsebine lahko namenjen izgledu, dostikrat pa vsebujejo tudi ponavljajoče se dele, ki za besedilni korpus predstavljajo šum. Drugi format dokumentov, ki tudi vsebujejo veliko jezikoslovno zanimivih podatkov, vendar se jih bistveno redkeje zajema in obdeluje, pa so dokumenti v formatu PDF. Problem pri zajemu besedil iz takih dokumentov je, da je format PDF namenjen tiskanju, zato ni kodiran kot niz znakov, temveč kot nabor znakov s svojim položajem na posamezni strani. V nadaljevanju se zato osredotočamo predvsem na opis obdelave dokumentov HTML (za dokumente PDF bi bilo namreč treba prilagoditi proces luščenja vsebine dokumentov).

Posebej je treba omeniti še spletne platforme, na katerih, vsaj izvorno, besedila niso postavljena na splet kot dokumenti HTML (ali PDF), pač pa so prejemnikom poslana kot posamezna sporočila, podobno kot SMS-sporočila ali elektronska pošta. Tu je daleč najbolj znana platforma Twitter, sistem, ki omogoča pošiljanje krajših sporočil sledilcem. Twitter ponuja tudi programske vtičnike API, ki omogočajo zajemanje tvitov posameznikov ali tem. Kot je bilo prikazano zgoraj, smo v korpus JANES vključili tvite, ki smo jih zbrali z orodjem TweetCat (Ljubešić et al. 2014), ki je bilo namensko izdelano za gradnjo korpusov tvitov manjših jezikov. To orodje s pomočjo začetnega seznama specifično slovenskih besed identificira uporabnike, ki tvitajo pretežno v slovenščini, nato pa prek prijateljev in sledilcev postopoma širi nabor uporabnikov ter zbira njihove tvite skupaj z metapodatki.

5.1 Postopki pri generičnem pajkanju

Kot rečeno, se generično pajkanje uporablja predvsem takrat, kadar je cilj pridobiti velike količine besedil (z več kot milijardo pojavnic) ali pa so človeški viri za pridobivanje podatkov omejeni.

Osnovnih korakov, ki se izvajajo pri generičnem pajkanju jezikoslovno relevantnih podatkov in jih uporablja tudi sistem, s katerim smo zgradili korpus slWaC, je več. Prvi korak je *izdelava seznama spletnih strani*, ki so izhodišče za pajkanje. V primeru jezikov z manjšim številom govorcev, kakršna je slovenščina, je to večinoma nekaj bolj znanih spletnih strani v izbranem jeziku. Drugi korak je *pajkanje*, ki se, tehnično gledano, tipično izvaja v večnitnem načinu in s pregledovanjem povezav v širino, pri čemer se seznam strani, ki jih je treba zajeti, sproti dopolnjuje z identifikacijo povezav z že zbranih spletnih strani. Ko je dokument zajet, je treba najprej ugotoviti, kateri *kodni nabor znakov* uporablja. Ta podatek naj bi bil sicer zapisan v metapodatkih dokumenta HTML, vendar v praksi dostikrat manjka ali ni pravilen glede na dejansko kodiranje dokumenta. Ugotavljanje kodnega nabora zato večinoma poteka na podlagi primerjave distribucije bajtov v besedilu dokumenta z distribucijo vnaprej pripravljenih dokumentov z znanimi kodiranji.

Pri generičnem pajkanju ne obstaja vnaprej predvidljiva predloga videza dokumenta, zato je treba za *zajem vsebine* uporabiti splošne programe, kot so jusText (Pomikálek 2011) in Boilerpipe (Kohlschütter et al. 2010). Ta korak, ki zaradi svoje generičnosti dokument strukturira največ do obsega odstavka, ne zajame metapodatkov o besedilu in tipično tudi ne odstrani vsega nebesedilnega šuma iz dokumenta. Na osnovi zajete besedilne vsebine dokumenta je nato treba *identificirati jezik* dokumenta. Splet je namreč večjezično okolje, zato je ta korak nujen pri izdelavi korpusa. Orodje, ki daje pri tem dobre rezultate, je program langid.py (Lui in Baldwin 2012), napisan v programskem jeziku Python. Zadnji

korak je *odstranjevanje (približnih) dvojnikov*, saj se enaka ali skoraj enaka besedilna vsebina pogosto pojavlja na več različnih naslovih URL. Postopki odstranjevanja (približnih) dvojnikov večinoma temeljijo na računanju preseka n-gramov besed med dvema dokumentoma. Tipična hevristika je, da če se 7-grami dveh dokumentov prekrivajo v več kot polovici primerov, lahko enega od njiju odstranimo kot približnega dvojnika.

Opisanih šest korakov se večinoma izvaja ločeno, zaradi česar je postopek pajkanja daleč od optimalnega. Izjema je SpiderLing (Suchomel in Pomikálek 2012), ki združuje korake od pajkanja do identifikacije jezika v integriran postopek, pri katerem posamezni koraki medsebojno komunicirajo s ciljem optimizacije količine prevzetih podatkov in velikosti končnega korpusa.

5.2 Postopki pri ciljnem pajkanju

Ciljno pajkanje se uporablja, kadar se pajka manjša količina podatkov oz. kadar je človeških virov za izvedbo postopkov dovolj. Takšno pajkanje ima tri osnovne korake. Specializirani korpusi se najpogosteje gradijo na osnovi določene vsebine, ne na osnovi spletnih domen. Prvi korak je zato *identificiranje spletnih domen* oz. njihovih delov, za katere se predvideva, da so bogati z želenimi vsebinami. Pri tem je treba upoštevati tudi tehnično-pravne omejitve posameznih izvorov, npr. ali spletno mesto prepoveduje pajkanje (datoteka robots.txt), ali izvor ponuja programski vmesnik API za zajem podatkov (npr. Twitter) in ali morda celo omogoča prevzem celotne baze besedil (npr. Wikipedija). To dvoje izrazito olajša zajem, medtem ko uporaba tehnologij, kot so pošiljanje POST, AJAX ipd., zelo oteži izdelavo ekstraktorjev. Naslednji korak je *pajkanje*, ki večinoma zajame vse oz. čim več dokumentov z izbranih domen. Najbolj zahtevno in zamudno je pisanje *ekstraktorjev*, v katerih mora programer opisati shemo HTML-dokumentov za vsak posamezen vir, pri čemer je lahko struktura dokumentov zelo kompleksna, npr. pri zajemu časopisnih prispevkov, pri katerih bi hoteli zajeti tudi zaporedje komentarjev vsakega prispevka.

5.3 Spremljevalni korpusi

Splet je izrazito naklonjen gradnji spremljevalnih korpusov, saj se njegova vsebina nenehno spreminja in dopolnjuje, pri čemer je potem, ko je sistem pajkanja postavljen, ponovno zbiranje podatkov ter beleženje razlik oz. povsem novih dokumentov enostavno. To velja tako za generično kot za ciljno pajkanje, pri čemer je generični pristop robustnejši, saj lahko posamezni izvori spremenijo obliko svojih strani, s čimer stari ciljni ekstraktorji prenehajo pravilno delovati.

Najboljši pri sprotnem zajemanju spleta so spletni iskalniki, predvsem Google, pa tudi lokalni, kot je Najdi.si, saj ti nepretrgoma in intenzivno pregledujejo splet ter na njem iščejo nova besedila. Težko si sicer zamislimo, da bi za jezikoslovne potrebe lahko uporabljali tako intenzivno pajkanje, nam pa lahko služi vsaj kot zgornja meja možnega. Od posameznega projekta je odvisno, kako se bodo tako glede na svoje potrebe kot tudi zmožnosti sodelujoči raziskovalci odločili glede pogostosti ponovnega pajkanja izbranih vsebin. Za leksikografske namene bi bil spremljevalni korpus, ki bi nastajal sproti vse od prve faze projekta dalje, gotovo dragocen za zaznavanje večjih in nenadnih leksikalnih sprememb, ki jih povzročajo dogodki ter pojavi, o katerih v nekem obdobju bolj intenzivno poročajo mediji in posledično vzbujajo tudi večje zanimanje govorcev – potencialnih uporabnikov slovarja. Potem ko je prva različica slovarja narejena in že na voljo, slovar pa bi hoteli sproti vsebinsko posodabljati, pa postane izdelava spremljevalnega korpusa in metod za ugotavljanje nove leksike, pomenskih sprememb ali sprememb značilnega besedilnega okolja še veliko bolj pomembna, pravzaprav ključna.

6 SKLEP

V sodobnem jezikoslovju so paradigme, ki na rabo nestandardnih jezikovnih različic v internetnem pisnem komuniciranju gledajo kot na odraz nepopolnosti ali osiromašenosti sporazumevalnih zmožnosti, preživete, saj analize jezikovne rabe na internetu ugotavljajo sposobnost uporabnikov, da se prilagodijo računalniškemu mediju oz. da zmožnosti medija izrabijo za zadovoljevanje svojih komunikacijskih potreb, da si prizadevajo skrajšati in poenostaviti pisanje, predvsem pa da pisanje približajo svoji identiteti ter govoru (Herring 2001).

Internetno komuniciranje danes tvorno dopolnjuje in spreminja podobo javne pisne slovenščine do mere, ki je sodobni slovar gradivno ne more več zaobiti. V prispevku smo skušali prikazati, kako je mogoče spletni del Gigafide nadgraditi tako v obsegovnem kot v tematskem in žanrskem smislu, ter opozorili, da ga je treba v korpus umestiti na razviden način (tj. z bolj razdelanimi taksonomskimi kategorijami). Del spletnih žanrov je zapisan v nestandardni slovenščini, kar v korpusno jezikoslovje prinaša še dodaten jezikovnotehnološki izziv: premostitev ovir za njihovo avtomatsko obdelavo, vendar pa viri in orodja, s katerimi si je mogoče pri tem pomagati, v slovenskem prostoru že nastajajo, preizkušene pa so bile tudi že različne metode pajkanja. Namen na predlagani način nadgrajene Gigafide je torej zajeti javnosti namenjeno pisno produkcijo spletne slovenščine v širokem smislu; izbor in interpretacija podatkov iz takega korpusa za potrebe slovarja pa bosta nato prepuščena odločitvam akterjev naslednje faze tega procesa – leksikografom.