OMOGOČANJE DOSTOPA DO KORPUSOV SLOVENSKIH SPLETNIH BESEDIL V LUČI PRAVNIH OMEJITEV

Tomaž ERJAVEC
Institut »Jožef Stefan«

Jaka ČIBEJ
Filozofska fakulteta, Univerza v Ljubljani

Darja FIŠER
Filozofska fakulteta, Univerza v Ljubljani

Erjavec, T., Čibej, J., Fišer, D. (2016): Omogočanje dostopa do korpusov slovenskih spletnih besedil v luči pravnih omejitev. Slovenščina 2.0, 4 (2): 189–219.
DOI: http://dx.doi.org/10.4312/slo2.0.2016.2.189-219.

Spletna besedila postajajo vse bolj relevanten vir informacij, korpuse tovrstnih besedil pa potrebujemo pri korpusnojezikoslovnih raziskavah in razvoju jezikovnih tehnologij za sodobno slovenščino. Čeprav so spletna besedila neposredno dostopna in je njihov zajem preprostejši od tiskanih, je izdelava takšnih korpusov še vedno zapletena, draga in zamudna. Ključno je, da poskrbimo, da se podobni podatki ne zbirajo večkrat, zato je nujno omogočiti njihovo čim večjo dostopnost čim širši raziskovalni skupnosti in zainteresirani javnosti. Tehničnih in prostorskih ovir za to sicer ni, vendar pri gradnji korpusa naletimo na številne omejitve v okviru zaščite avtorskih pravic, varstva osebnih podatkov in pogojev uporabe ponudnikov spletnih storitev. V prispevku predstavljamo pravno in dejansko stanje na teh področjih, opravimo pregled sorodnih tujih in domačih praks ter na primeru korpusa spletne slovenščine Janes predlagamo vrsto ukrepov, ki do največje možne mere omogočajo prosto in odprto razširjanje korpusov spletne slovenščine.

Ključne besede: spletna besedila, diseminacija korpusov, avtorske pravice, varstvo osebnih podatkov, prosti in odprti dostop

1 UVOD

Svetovni splet postaja vse pomembnejši vir informacij, saj vsebuje hitro rastočo količino besedil, spletna jezikovna produkcija pa že prehiteva tiskano in ponuja veliko količino podatkov, ki jih je mogoče uporabiti za različne raziskave s področja jezikoslovja, jezikovnih tehnologij, psiho- in sociolingvistike, besedilne analitike ipd. Zbiranje podatkov je v zadnjem desetletju močno olajšal tudi hiter tehnološki razvoj, ki je skorajda odpravil prostorske in druge logistične omejitve pri shranjevanju, arhiviranju in razširjanju raziskovalnih podatkov. Vendar poleg tehničnih faktorjev krajino digitalnih vsebin pomembno sooblikujejo tudi tržni dejavniki, zakonodaja in družbene norme (Lessig 1999), zato problematika zagotavljanja dostopa do izdelanih podatkovnih zbirk zgolj z odpravljanjem tehničnih omejitev še zdaleč ni rešena.

1.1 Pomen zagotavljanja prostega dostopa

Dostop do raziskovalnih podatkov je v zadnjih letih postal vroča tema tako v mednarodnem okviru kot pri nas (Kotar 2013, Štebe idr. 2013). Odprti dostop do raziskovalnih podatkov je denimo predviden v strategiji Evropa 2020 za trajnostno in vključujoče gospodarstvo (European Commission 2012), ki poudarja osrednjo vlogo znanja in inovacij pri spodbujanju rasti ter pomen hitrega razširjanja rezultatov raziskav (tako publikacij kot podatkovnih zbirk) čim širšemu krogu uporabnikov, saj to pospešuje znanstvena odkritja, spodbuja sodelovanje, povečuje transparentnost znanstvenega procesa, omogoča nove oblike raziskav in pospešuje prenos novih odkritij v evropska podjetja.

Tudi Evropska listina za raziskovalce – Kodeks ravnanja pri zaposlovanju raziskovalcev (Evropska komisija 2006: 13) v zvezi s širjenjem in izkoriščanjem rezultatov navajata, da morajo vsi raziskovalci zagotoviti, da bodo rezultate raziskav širili v druga raziskovalna okolja, rezultati pa naj bodo tržno izkoriščeni in/ali dostopni javnosti, kadarkoli se za to pojavi priložnost.

Načelo odprtega dostopa do podatkov je več kot smiselno, in sicer iz več razlogov. Prvič, pridobivanje podatkov za raziskave je kljub tehnološkemu

napredku še vedno drago in zamudno. Če so podatki uporabljeni le za namene projekta, v okviru katerega so bili zbrani in drugim raziskovalnim skupnostim niso na voljo, raziskovalci podobne ali celo enake podatke pogosto zbirajo in obdelujejo večkrat. Čas in finančna sredstva bi bilo v takšnih primerih mnogo bolj smiselno nameniti nadaljnjim raziskavam ali za gradnjo še neobstoječih podatkovnih zbirk. Drugič, brez dostopa do podatkov predhodnih raziskav ni mogoče ponoviti, preveriti ali nadgraditi, kar je osnova vsakega znanstvenega dela.

Tudi korpusnojezikoslovne raziskave pri tem niso izjema. Veliki, dobro metapodatkovno in jezikoslovno označeni korpusi so podlaga za sodobno slovaropisje, empirično jezikoslovje in razvoj jezikovnih tehnologij. Za zagotavljanje dostopnosti jeziko(slo)vnih podatkov za humanistične in družboslovne raziskave, s tem pa spodbujanje večkratne uporabe jezikovnih podatkov je bila ustanovljena evropska raziskovalna infrastruktura CLARIN (Common Language Resources Infrastructure), v Sloveniji in za slovenščino pa konzorcij CLARIN.SI (Erjavec idr. 2014). Infrastruktura CLARIN.SI je za slovenščino še posebej pomembna, saj se na ta način prvič pri nas omogoča trajno in stabilno deponiranje jezikovnih virov (kot so npr. korpusi) ter orodij za njihovo označevanje in raziskovanje. V tem okviru je bil storjen že velik korak z vzpostavitvijo repozitorija CLARIN.SI, ki omogoča hranjenje jezikovnih virob in je od začetka 2016 tudi certificiran repozitorij s strani DSA (Data Seal of Approval) in evropskega CLARIN-a.

1.2 Ovire pri zagotavljanju prostega dostopa

Kljub številnim iniciativam, ki podpirajo odprti dostop do podatkov, pa diseminacijo rezultatov raziskav ovirajo različne omejitve. Pri besedilnih korpusih so to predvsem avtorske pravice izvirnih besedil in vprašanja varovanja osebnih podatkov, zlasti v primerih, ko korpus vsebuje velike količine stvarnih besedil (npr. časopisnih člankov), zaradi česar ga je mogoče obravnavati kot agregirano bazo osebnih podatkov. Čeprav je namen korpusov

proučevanje jezika, ne pa ponovno izdajanje zaključenih besedil ali poizvedovanje po osebnih podatkih, so pogoji za diseminacijo besedil in podatkovnih baz zelo strogi, na kar opozarjajo tudi številni izdelovalci korpusov drugod po svetu (Spousta 2006, Baroni in dr. 2009, Beißwenger in dr. 2012a).

Položaj pri korpusih spletnih besedil je še nekoliko bolj občutljiv, saj problematika spletnih besedil še vedno ni eksplicitno predvidena v zakonodaji, zaradi česar številna vprašanja ostajajo nerazrešena. Trenutno tako v Sloveniji »Država s svojimi inštitucijami in zakonodajo na marsikaterem področju prej ovira dostop do znanja, kot ga omogoča«, zakoni in organi pri nas pa »dajejo prednost posameznikovi zasebnosti v škodo javne blaginje in ovirajo emancipiran vstop slovensko govorečih v globalno skupnost kulturno progresivnih nacij« (Hladnik 2016).

Pri omogočanju dostopa do korpusov je pomembna dimenzija tudi način njihove diseminacije. Za velik del raziskav zadostuje, da je korpus dostopen za iskanje in pregledovanje prek spletnega konkordančnika. Če je tak korpus brezplačno dostopen tretjim osebam, govorimo o *prostem dostopu* do korpusa. V prostem dostopu je na voljo že vrsta korpusov slovenščine (Erjavec 2013), a so besedila pri tem dostopna samo v iztržkih (kot omejeni kontekst iskanega izraza), posamezna besedila (ali celo celoten korpus) v korpusih pa je mogoče na (upravičeno) zahtevo izbrisati oz. dostop do njih delno ali popolnoma onemogočiti. Drug način dostopa je s prevzemom, kjer korpus kot podatkovno bazo v celoti prenesemo na lokalni računalnik. Če je tak korpus brezplačno dostopen tretjim osebam, govorimo o *odprtem dostopu*. Prevzem je nujen za razvoj jezikovnih tehnologij in za bolj poglobljene jezikoslovne raziskave, saj tu nismo več omejeni na funkcionalnost specifičnega konkordančnika, temveč je korpus mogoče analizirati z lastno programsko opremo na poljuben način. Je pa tu možnosti za potencialne zlorabe podatkov več, saj je na tak način mogoče reproducirati celotna besedila ali podrobno analizirati vedênje določene osebe. Poudariti je treba tudi, da v primeru pritožb lahko sporna besedila iz izvornega korpusa ali korpus sicer odstranimo, a nad že prevzetimi kopijami nimamo več

neposrednega nadzora.

V prispevku se osredotočamo predvsem na problematiko spletnih besedil in na pravne omejitve, ki jih je treba preseči, če želimo zagotoviti dostopnost takšnih korpusov za pregledovanje ali prevzem. Po razpravi o pravnih omejitvah pregledamo obstoječo prakso na tem področju, nato pa ponudimo predloge, kako jih (delno) preseči, kjer za primer uporabimo korpus spletne slovenščine, ki nastaja v sklopu projekta JANES (Fišer idr. 2016).

2 PRAVNE OMEJITVE

V tem razdelku predstavimo tri vidike pravnih omejitev pri zbiranju spletnih besedil, in sicer avtorske pravice, varovanje osebnih podatkov ter pogoje uporabe spletnih mest. Podroben pregled relevantne zakonodaje smo podali v Erjavec idr. (2015), tu zakone zgolj omenimo, predvsem pa razpravljamo o navzkrižju pravnih podlag z načeli o odprtem dostopu do raziskovalnih podatkov.

2.1 Avtorske pravice

V Sloveniji področje avtorskih pravic ureja Zakon o avtorskih in sorodnih pravicah (ZASP, Uradni list št. 21/95), ki določa, da je vsako pisno delo skupaj z vsemi sestavnimi deli avtorsko. Členi tega zakona npr. določajo, da lahko avtor uporabo svojega dela prepove in da avtorska pravica ugasne 70 let po smrti avtorja.

Čeprav so spletna besedila javno dostopna in zato včasih obravnavana kot javno dobro, so tudi tovrstna besedila zaščitena z avtorskim pravom (Hemming in Lassi 2002, King 2008, Vintar in Fišer 2009, Margaretha in Lüngen 2014), zaradi pogoste anonimnosti avtorjev na spletu in njihovega velikega števila pa je pridobivanje dovoljenj za uporabo nepraktično oz. z danimi sredstvi in v omejenem časovnem okviru trajanja projektov tipično neizvedljivo.

ZASP tudi določa, da ima avtor pravico skesanja, s katero lahko imetniku

prekliče materialne avtorske pravice in zahteva umik vsebine, če ima za to resne moralne razloge in če imetniku povrne s tem nastalo škodo.

2.2 Varovanje osebnih podatkov

Področje varovanja osebnih podatkov pri nas ureja Zakon o varstvu osebnih podatkov (ZVOP-1, Uradni list RS, št. 86/04). Zakon osebne podatke opredeli kot katerekoli podatke, ki se nanašajo na posameznika, ne glede na obliko, v kateri so izraženi. Med drugim zakon določa, da se osebni podatki lahko, ne glede na prvotni namen zbiranja, nadalje obdelujejo za zgodovinsko, statistično in znanstvenoraziskovalne namene, a obvezno v anonimizirani obliki, če posameznik ni predhodno podal pisne privolitve, da se podatki lahko obdelujejo neanonimizirano.

Z varstvom osebnih podatkov je povezana pravica do pozabe, s katero lahko posameznik zahteva izbris ali umik zastarele informacije, ki razvidno škoduje njegovemu položaju. Pravica do pozabe sicer ni izrecno predvidena v Zakonu o varstvu osebnih podatkov, a je postala svetovno znana zaradi tožbe španskega državljana proti podjetju Google, čigar spletni servis je ob iskanju njegovega imena prikazal povezavo na star časopisni članek, za katerega je tožnik menil, da danes ni več aktualen in obenem zelo negativno vpliva na njegovo sedanje življenje. Tožnik je s tožbo zmagal, odločitev Sodišča Evropske unije (Sodba Sodišča z dne 13. maja 2014 v zadevi C-131/12) pa je podjetju Google naložila, da mora na zahtevo uporabnikov vsebine deindeksirati.

Soroden primer v Sloveniji je odločba Informacijskega pooblaščenca Republike Slovenije št. 0612-63/2012/5, ki zahteva onemogočenje iskanja po »imenu in/ali priimku posameznika« v korpusu Nova beseda. Po naknadnem dogovoru med Pooblaščenko in ZRC SAZU je bil pogoj nato omiljen do te mere, da v korpusu Nova beseda sedaj ni mogoče iskati samo po sosledju dveh zaporednih osebnih lastnih imen, torej po imenu *in* priimku. Tudi do tega omiljenega ukrepa so jezikoslovci vseeno zelo kritični, saj je poleg želenega učinka onemogočil tudi npr. iskanje imen zgodovinskih ali namišljenih oseb, s čimer

se je okrnila uporabnost korpusa za celotno področje digitalne humanistike.

Pri tem ni zanemarljivo, da se je pritožba, ki je povzročila odločbo, nanašala na besedilo, ki je bilo objavljeno v tiskanem mediju pred razmahom interneta. Dandanes je pri spletnih besedilih v praksi težko vztrajati pri pravici do (popolne) pozabe, saj so vsebine kljub izbrisu iz indeksa iskalnikov ali z okrnjenjem funkcionalnosti konkordančnikov še vedno dostopne na spletu v izvorni obliki in tudi ob izbrisu s spletnega mesta ohranjene v arhivih spletnih vsebin, kot sta Wayback Machine in Spletni arhiv NUK.

ZVOP omejuje tudi iznos podatkov v tretje države (tj. v države izven Evropske unije ali Evropskega gospodarskega prostora), ki je dovoljen le v primeru, da tretja država zagotavlja ustrezno raven varstva osebnih podatkov. V praksi to pomeni, da bi bilo npr. informacije, ki so dostopne samo raziskovalcem, potrebno zamejiti na raziskovalce v teh državah, kar pa v znanstveni skupnosti, ki ji je inherentno povezovanje glede na raziskovalne interese, ne na narodnost, večinoma ni niti realno niti zaželeno.

2.3 Pogoji uporabe spletnih portalov

Pri spletnih besedilih je poleg zakonodaje o urejanju avtorskih pravic in varstva osebnih podatkov treba omeniti še pogoje uporabe, ki jih uporabniki spletnih platform najpogosteje sprejmejo ob ustvarjanju uporabniškega računa na spletnem portalu in z objavo vsebin. S pogoji uporabe lastniki spletnih mest (npr. družbenih omrežij, forumov, blogovskih strani ali novičarskih portalov, ki omogočajo komentiranje) objavo besedil na spletu najpogosteje omejijo z dodatnimi določili, npr. o tem, kako se naj uporabniki med sporazumevanjem vedejo (npr. sovražni govor, izzivanje ali objavljanje vsiljene pošte) ter ali in kako lahko tretje osebe objavljena besedila zajemajo in uporabljajo.

Uporaba objav uporabnikov s strani tretjih oseb je v večini primerov določena zelo strogo. Čeprav uporabniki ob izdelavi računa in sprejetju pogojev uporabe materialne avtorske pravice nad svojimi pisnimi deli najpogosteje prenesejo na lastnika spletnega mesta, se politika lastnika pri posredovanju podatkov tretjim

osebam razlikuje od vira do vira, zato je treba vsak vir obravnavati posebej. Razmerja med ponudniki platform, zelo raznolikimi končnimi uporabniki (posamezniki, javne osebnosti, javne ustanove, nevladne organizacije, podjetja) in prav tako raznolikimi uporabniki podatkov (poleg raziskovalcev so to tudi številne javne institucije in podjetja, npr. za marketing) so zaradi njihovih različnih interesov kompleksna, zaradi vse večje vrednosti naraščujoče količine podatkov na portalih in zaskrbljenosti nad marginalizacijo zasebnosti končnih uporabnikov pa je pričakovati, da se bodo konflikti med njimi v prihodnje še poglabljali (Puschmann in Burgess 2014).

Poseben primer predstavljajo tviti, saj je te po eni strani zelo enostavno zbirati, ker že podjetje Twitter ponuja programski vmesnik (API), ki zajem tvitov omogoča in spodbuja. V Pogojih uporabe Twitter (2016a) določa, da lahko vsebine, ki jih objavijo uporabniki, vidijo tudi tretje osebe, razen če uporabnik to drugače določi v nastavitvah računa. Imetnik avtorskih pravic je še vedno uporabnik, a se z objavo vsebin na Twitterju strinja, da jih lahko Twitter (ali z njim povezane institucije) brezplačno uporablja, reproducira, preoblikuje in objavlja tudi v drugih medijih.

Vendar Twitter strogo omejuje diseminacijo tvitov tretjim osebam, kar raziskovalcem predstavlja resno oviro pri zagotavljanju transparentnosti in ponovljivosti eksperimentov nad korpusi tvitov. Celovit zajem podatkov otežuje netransparentno vključevanje tvitov na časovnico in vse večje omejitve dovoljenega zajema z API-ja, ki trenutno znaša okoli 2 % vseh tvitov na časovnici dnevno (Twitter 2016a), medtem ko je uporaba podatkov okrnjena zaradi strogih omejitev dovoljenega prikaza zajetih podatkov v neprogramatični tabelarični oz. PDF-obliki (Twitter 2016b). Diseminacijo zajetih podatkov močno ovira prepoved deljenja večjih količin zajetih podatkov brez eksplicitnega dovoljenja podjetja z izjemo ID-jev tvitov oz. uporabnikov (Twitter 2016c), težave pa povzroča tudi prepoved arhiviranja, obdelave, prikazovanja in deljenja izbrisanih tvitov oz. računov (Twitter 2016d). Dodatno oviro povzroča tudi pogosto spreminjanje pogojev, kar je zlasti problematično

za raziskave, ki so že v teku.

Na zanimivo kontradikcijo trenutne Twitterjeve politike je opozoril Beurskens (2014), ki kljub zgoraj omenjenim prepovedim za raziskovalce ameriški nacionalni knjižnici U. S. Library of Congress dovoljuje arhiviranje tvitov kot pomembno dokumentacijo sodobne kulture in zgodovine, in izpostavil nujnost posodabljanja zakonodaje za uporabo digitalnih podatkov v raziskovalne namene.

3 PREGLED PRAKS PRI RAZŠIRJANJU SPLETNIH KORPUSOV

Za dostopne korpuse spletnih vsebin se prakse pri upoštevanju pravnih omejitev med seboj zelo razlikujejo. V tem razdelku podamo primere korpusov, ki so bili zgrajeni namensko za konkretno raziskavo, obsežnih korpusov heterogenih spletnih besedil in specializiranih korpusov računalniško posredovane komunikacije.

3.1 Ad hoc korpusi spletnih vsebin

Predvsem pri gradnji namenskih (ad hoc) korpusov angleškega jezika, ki jih za konkretne raziskave večinoma zbirajo individualni raziskovalci ali manjše skupine, se pravnih vprašanj večinoma niti ne zavedajo oz. jih ne upoštevajo. Znana primera sta obsežna korpusa tvitov, ki so jih zbrali raziskovalci Univerze v Edinburgu (Petrović idr. 2010) s 100 milijoni tvitov in Stanforda (Yangon in Leskovec 2011) s pol milijarde tvitov, ki sta bila objavljena, nato pa umaknjena po pritožbi podjetja Twitter. Kljub temu na spletu še vedno najdemo korpuse tvitov, ki vsebujejo bistveno manjše količine besedil. Tako ima npr. korpus ročno označenih tvitov, zgrajen v okviru projekta TweetNLP na Univerzi Carnegie Mellon (Owoputi idr. 2013), samo okoli 2.500 tvitov, ki jih je možno neposredno prevzeti s spleta. Taki korpusi očitno niso dovolj znani ali veliki, da bi se Twitter zavedal njegovega obstoja oz. ga preganjal. Podobno velja za korpus slovenskih tvitov iz let 2007–2011 (Erjavec in Fišer 2013) z okoli 350.000 tviti (5 mio. besed), ki sicer ni dostopen za prevzem, je pa odprto

dostopen prek konkordančnika noSketchEngine na IJS že od leta 2011, pa do sedaj še nismo prejeli nobene pritožbe glede njegovega obstoja oziroma vsebine.

3.2 Splošni spletni korpusi

Stanje je še precej bolj sproščeno na področju korpusov, zbranih neposredno s spleta, saj bi se za razliko od korpusa tvitov, ki vsebujejo besedila enega samega ponudnika spletnih storitev, v primeru korpusov spletnih besedil morali nad vključenostjo posameznih besedil v pritožiti avtorji posameznih spletnih strani. Primer odprto (čeprav z omejitvijo na raziskovalno uporabo) dostopnih korpusov je zbirka korpusov CoW (Schäfer in Bildhauer 2012), ki zajema korpuse z več kot milijardo besed za vse večje evropske jezike. Podobno je s korpusi, ki jih ponuja repozitorij češke infrastrukture CLARIN, npr. korpus madžarskega spleta (Halacsy 2014), norveškega spleta (Guevara in Johannessen 2014) ter zbirka korpusov W2C (Majliš 2011) s spletnimi korpusi za prek 50 jezikov. Tudi podjetje Lexicom ponuja v sklopu svojega konkordančnika SketchEngine vrsto spletnih korpusov, kjer imajo korpusi serije TenTen po več kot deset milijard besed. Dostop do teh korpusov je sicer plačljiv in omejen na konkordančnik SketchEngine, kar pa ne spremeni dejstva, da podjetje nima dovoljenj za uporabo izvornih besedil.

Podobne primere najdemo tudi v Sloveniji. V sklopu konkordančnika noSketchEngine na IJS so dostopni veliki (od pol do prek dveh milijard besed) spletni korpusi slovenskega, hrvaškega, srbskega, bosanskega, japonskega, francoskega, italijanskega in nemškega jezika. Kljub temu da je večina teh korpusov prosto dostopnih že vrsto let, do sedaj še nismo prejeli nobene pritožbe glede njihove dostopnosti.

3.2 Specializirani korpusi računalniško posredovane komunikacije

Obstajajo pa tudi korpusi, ki so nastali kot rezultat namenskih raziskovalnih projektov, katerih cilj je izdelava skrbno načrtovanega in uravnoteženega ter

bogato označenega korpusa računalniško posredovane komunikacije, ki bo nato služil za najrazličnejše jezikoslovne in jezikovnotehnološke raziskave.

3.2.1 Korpusi s pridobljenimi pravicami

Nekateri raziskovalci so pri izboru besedil zelo previdni in v korpus vključujejo samo tista, za katera so pridobili eksplicitno privoljenje avtorjev. Rezultat so sicer pravno čisti korpusi, vendar za ceno obilice vloženega truda v pridobivanje dovoljenj, predvsem pa razmeroma majhni korpusi z največ nekaj sto tisoč besedami.

Nekateri izdelovalci korpusov soglasja avtorjev dobijo že pred samim zbiranjem besedil. Ta pristop sta npr. uporabila Spooren in van Charldorp (2014), ki sta v nizozemskem korpusu klepetalniških pogovorov ChatIG zbrala pogovore 188 srednješolcev, ki so pred tem podpisali soglasje za vključitev v korpus, prosto dostopen za znanstvene raziskave. Čeprav tak pristop odpravlja pravne omejitve objave korpusa, ni neproblematičen. Kot opozarjajo nekateri jezikoslovci (Corti idr. 2000, Beißwenger in Storrer 2009), lahko dejstvo, da so avtorji že vnaprej obveščeni o zbiranju podatkov, do določene mere vpliva na njihovo vedenje in s tem na avtentičnost zajete jezikovne rabe.

Temu so se uspešno izognili sodelavci projekta DiDi (Glaznieks in Stemle 2014), ki so avtentična besedila pridobili tako, da so uporabnike družbenega omrežja Facebook z Južne Tirolske s spletno anketo povabili k sodelovanju, ko so jim ti odobrili dostop, pa so s pomočjo Facebook aplikacije avtomatsko zajeli njihova že objavljena javna sporočila na zidu in (če so se uporabniki s tem strinjali) zasebna sporočila v nabiralniku. Podobno pot so ubrali sodelavci projekta What'sUp v Švici (Dürscheid 2015), ki so uporabnike prosili, da jim po elektronski pošti posredujejo svoje pretekle pogovore z aplikacije WhatsApp. Če so bila med poslanimi pogovori prisotna tudi sporočila uporabnikov, za katerih dovoljenja niso dobili, ta niso bila vključena v korpus.

Tudi pri spletnih besedilih so imetniki materialnih avtorskih pravic pogosto organizacije (tiskovne agencije, televizijske hiše, založbe ipd.), zato se pri

gradnji korpusov avtorske pravice lahko upošteva s sklepanjem sporazumov s temi organizacijami. Gre za pogodbe, ki določajo, katera besedila (in v kakšnem obsegu) so lahko vključena v korpus. Postopek je zamuden in naporen, zlasti v primeru velikega števila virov oz. besedilodajalcev. Pristop sklepanja sporazumov z organizacijami imetnicami avtorskih pravic je bil uporabljen že večkrat (priporoča ga tudi Olohan (2004)), npr. pri izgradnji korpusa sodobne arabščine (Al-Sulaiti in Atwell 2003) in referenčnega nemškega korpusa DeReKo (Kupietz in Lüngren 2014).

3.2.2 Korpusi brez pridobljenih pravic

Posebej v primerih, ko izdelovalci korpusa nimajo privoljenj avtorjev, se pogosto poslužujejo različnih prijemov, kako narediti korpus manj sporen za kršitve avtorskih pravic oz. zasebnosti. Ena pogostejših strategij je omejevanje na avtorje oz. uporabnike, ki so javne osebnosti (politiki ali zvezdniki), zato je njihove objave na družbenih omrežjih mogoče razumeti kot del njihovega javnega delovanja. Tako je bil na primer zgrajen in objavljen korpus tvitov francoskih politikov (Longhi et al. 2014). Zanimiv primer prizadevanja odprtega dostopa podatkov za raziskovalne namene je korpus tvitov Stephana Danna (2010), ki je svojo celotno časovnico tvitov objavil pod licenco CC BY-SA, kar bi lahko postalo zelo koristno, če bi mu sledilo večje število ostalih uporabnikov, kar je v tujini že razmeroma pogosta praksa pri blogih.

Med najbolj razširjenimi tehničnimi rešitvami za korpuse brez zbranih privoljenj avtorjev pa je omogočanje umika vsebin na zahtevo, kar je uporabno pri korpusih, ki s spleta zajemajo besedila, ki so že javno dostopna. Besedila so vključena v korpus brez pridobivanja izrecnega dovoljenja, avtor pa lahko od sestavljalcev korpusa zahteva, da se njegova vsebina umakne (Baroni idr. 2009). Ta pristop je primeren za korpuse, ki so dostopni prek konkordančnikov, manj pri korpusih za prevzem, saj nad že prenesenimi različicami nimamo več neposrednega nadzora in iz njih zadevne vsebine ne moremo odstraniti.

Posebnost spletnih korpusov je, da je s seznamom spletnih naslovov in programsko opremo za zajem njihovih besedil korpus mogoče v veliki meri ponovno zgraditi. Ta pristop k distribuciji korpusa je ubrala skupina WaCkY (Baroni idr. 2009), ki je izdelala odprta orodja za zajem korpusov s spleta in z njimi zgradila velike korpuse (1–2 milijardi besed) večjih evropskih jezikov ter jih naredila prosto dostopne. Korpusi niso neposredno na voljo v odprtem dostopu, a je WaCkY objavil seznam naslovov URL, katerih besedila so vključena v korpuse. S tem seznamom in njihovimi orodji je na tak način mogoče znova zgraditi (skoraj) enak korpus, pri čemer avtorske pravice z njegovo redistribucijo niso kršene. Podoben princip se tipično uporablja tudi za redistribucijo korpusov tvitov, kjer korpus vsebuje samo identifikacijske številke tvitov, API podjetja Twitter pa omogoča prevzem samih besedil za tvite, ki medtem niso bili izbrisani. Kljub vsemu pa ima ta pristop tudi težave: za gradnjo sta potrebni odlična internetna povezava, namestiti in razumeti je treba tudi programsko opremo, korpus pa nato vsakič znova zbirati počasi, da ne preobremenimo posameznih strežnikov.

Klasičen način za izogib problemu z avtorskimi pravicami je tudi vzorčenje, kjer iz besedil za korpus vzamemo samo redke naključno izbrane segmente, tj. jezikoslovno zamejene enote, kot so posamezni stavki (Östling in Wirén 2013). Seveda z vzorčenjem bistveno zmanjšamo velikost končnega korpusa. Metoda, ki je podobna vzorčenju, a brez pomanjkljivosti zmanjševanja količine podatkov, je premešanje, kjer le spremenimo izvorni vrstni red segmentov. Pristop s premešanimi stavki sta za distribucijo že omenjenih spletnih korpusov CoW uporabila Schäfer in Bildhauer (2012). Z obema metodama razbijemo koherenco besedila, s čimer sicer preprečimo reprodukcijo celotnih besedil, vendar takšna besedila niso več primerna za pomensko analizo, analizo diskurza ipd.

V najstrožjem scenariju lahko za segmente vzamemo kar različne n-terčke besed, vsakemu pa dodamo še njegovo frekvenco v korpusu. Takemu seznamu sicer težko še rečemo korpus, je pa kljub temu koristen za analizo jezika in nima

nobenih pravnih zadržkov glede popolne odprtosti.

3.2.3 Zaščita osebnih podatkov

V principu tudi osebni podatki, ki so zbrani v korpusu, ne bi smeli biti javno dostopni, in bi jih bilo torej treba bodisi izbrisati bodisi anonimizirati (Medlock 2006, Teutsch idr. 2009, Petrović idr. 2010, Lee in Woods 2012). Osebni podatki, v primeru spletnih korpusov predvsem (uporabniška) imena, so lahko del metapodatkov (uporabniško ime avtorja) ali samega besedila (uporabniška ali lastna imena v besedilu), pri čemer se identifikacija slednjih tipično izvrši avtomatsko z uporabo programov za označevanje imenskih entitet. Uporabniška imena ali identificirana lastna imena v besedilih se lahko izbrišejo ali nadomestijo s šifro. Izbris imen v besedilu je za uporabnike korpusa škodljiv, saj močno vpliva na koherenco besedil, kar onemogoča besediloslovne ali diskurzivne analize, zato se pogosto uporabi samo delen izbris, kjer lastna imena nadomestimo z naključnimi ali enoznačnimi šiframi (psevdonimizacija), kjer slednje omogoča opazovanje objav po (sicer neznanemu) avtorju. S tem ohranimo funkcionalnost korpusa in obenem zaščitimo osebne podatke.

4 PRIMER KORPUSA JANES

Korpus Janes (Fišer idr. 2016) je obsežen (prek 150 mil. besed) korpus uporabniško ustvarjenih slovenskih spletnih vsebin, ki ga sestavlja pet podkorpusov: tviti, blogi, forumi, komentarji na spletne novice in pogovorne strani na Wikipediji. Izdelava korpusa naj bi bil končana do srede 2017, ko naj bi ga tudi objavili. Pri tem se pojavi vprašanje, kako lahko to storimo v največji možni meri in se hkrati čim bolje zavarujemo pred ugovori ponudnikov vsebin, avtorjev in oseb, omenjenih v korpusu.

Podatke v korpusu Janes lahko razdelimo na tri skupine. Najpomembnejša so (različno strukturirana) besedila v petih podkorpusih, ki so dodatno (ročno in avtomatsko) jezikoslovno označena. Vsak »odstavek« besedil je razdeljen na stavke (povedi) in tokeniziran v besede in ločila. Besede so označene z

normalizirano (standardizirano) obliko, lemo in oblikoskladenjsko oznako. Druga skupina podatkov so podatki o besedilih (metapodatki), ki so, še posebej pri tvitih, zelo bogati: poleg uporabniškega imena itd. tudi spol avtorja, sentiment besedila in njegova jezikovna »standardnost«. Tretja skupina so manjši (od 1.000 do 100.000 besed) vzorci iz podkorpusov Janes, kjer so bile bodisi avtomatsko pripisane oznake ročno popravljene (Čibej idr. 2016), ali pa nove oznake ročno pripisane za namene konkretne raziskave, npr. strategij krajšanja (Goli idr. 2016) in napačnega postavljanja vejice v tvitih (Popič idr. 2016). Takšni vzorci so nepogrešljivi za natančne korpusnojezikovne raziskave in za razvoj programov jezikovnih tehnologij, kjer lahko služijo kot učne in testne množice. Pogoji dostopa pri teh ročno označenih korpusih so lahko bolj odprti kot za celoten korpus, saj po definiciji vsebujejo samo majhne vzorce, torej »citate« iz celotnega korpusa.

4.1 Pridobivanje dovoljenj lastnikov materialnih avtorskih pravic

Materialne avtorske pravice za večino vrst spletnih virov, ki so vključeni v korpus Janes (forumi, blogi, komentarji na spletne novice), pripadajo slovenskim podjetjem, ki so lastnik spletnih portalov, s katerih so bila besedila zajeta. Ker se pogoji uporabe od vira do vira razlikujejo, je treba k vsakemu pristopiti posebej. Podobno kot pri zagotavljanju pravic do uporabe tiskanih besedil, pri katerih se sporazume sklene z založbami, se bomo tudi tu trudili skleniti sporazume, ki bi določali, kako bodo vsebine uporabljene. Nekatera uredništva to že predvidevajo, saj so na njihovih spletnih mestih navedeni kontaktni podatki, kamor se je treba obrniti v primeru uporabe uporabniških vsebin.

Kot nadomestilo za dovoljenje za diseminacijo korpusa bi lahko podjetjem tudi ponudili celovit in polno označen del korpusa z njihovimi besedili, ki ga lahko nato uporabijo za izboljšanje svoje platforme. V primeru, da ne bomo mogli pridobiti privoljenja lastnikov, pa bomo njihova besedila umaknili iz podkorpusov za javno objavo.

V korpusu Janes sta zajeta tudi vira, kjer lastnik spletne platforme ni slovensko podjetje. Eden od njiju vsebuje pogovorne in uporabniške strani z Wikipedije, pri čemer je Wikipedija od vseh navedenih virov z vidika avtorskih pravic najmanj problematična, saj je vse gradivo (z nekaj manjšimi izjemami, kot so npr. uradni logotipi Wikipedije) na voljo pod licenco Creative Commons Attribution-ShareAlike (CC BY-SA 3.0), ki omogoča kopiranje, spreminjanje in diseminiranje gradiva pod enakimi pogoji (odprti dostop).

Drugi primer je podkorpus tvitov, kjer podjetje Twitter pogosto spreminja pogoje dostopa, vendar, kot že omenjeno, v splošnem velja, da manjši korpusi tvitov v praksi niso problematični, je pa zaradi možnosti izbrisa tvitov ali računa težko odprto diseminirati velike korpuse, kot je to podkorpus tvitov Janes, še posebej za prevzem. Lahko pa, kot je bilo že omenjeno in bo natančneje tudi obrazloženo v nadaljevanju, ponudimo programsko rešitev, s pomočjo katere lahko uporabnik korpus, skupaj z jezikoslovnimi oznakami, sam ponovno ustvari.

4.2 Licence in preverjanje identitete

Pri korpusu Janes nameravamo ločiti med dostopom poljubnega (neregistriranega) uporabnika, ki bo korpus uporabil za poljuben namen in dostopom registriranih raziskovalcev za raziskovalne namene oz. poučevanje.

Splošni dostop ima vrsto prednosti, saj lahko do korpusa dostopamo brez registracije, omogoča pa tudi komercialno uporabo, poljubno redistribucijo (nadgrajenih) korpusov, s čemer spodbujamo informatizacijo in proučevanje slovenskega jezika. Vendar moramo ravno zaradi tega biti bolj previdni, kaj korpus vsebuje in kako je zaščiten pred zlorabo oz. pritožbo. Zato bodo korpusi, namenjeni splošni uporabi, imeli več omejitev, ki jih opisujemo v nadaljevanju.

Pri različici korpusov za uporabo v akademske namene bo omejitev znotraj korpusa manj, saj naj bi bil uporabnik raziskovalec in se bo obvezal uporabljati korpus samo za raziskovalne namene oz. poučevanje in, da ga ne bodo redistribuirali. Pri taki uporabi je pomembno, da poznamo identiteto

uporabnika, saj je sicer kakršnokoli strinjanje s pogoji uporabe brezpredmetno. Do pred kratkim proces identifikacije legitimnih uporabnikov ni bil enostaven in v večini primerov omejen na dejstvo, da ima nek uporabnik e-poštni naslov akademske institucije, nato pa mu je bilo potrebno dodeliti uporabniški račun za vsako platformo (kot npr. konkordančnik ali repozitorij) posebej.

Na tem področju se je v zadnjem času zgodil velik premik z uvajanjem preverjanja identitete skozi AAI (Authentication & Authorization Infrastructure), ki ga v Evropi uporabljajo raziskovalne in izobraževalne institucije, združene v mrežo EduGain (naslednik EduRoam). V Sloveniji je registracija prek AAI omogočena prek matične institucije raziskovalca ali pedagoga že prek 100 ustanovam, tudi preko slovenskega Arnesa. Pri takšnem dostopu matična institucija preveri fizično identiteto uporabnika in mu na svoji AAI platformi (ponudnik identitete) dodeli uporabniško ime in geslo, medtem kot ponudnik storitve prepusti postopek registracije izbranemu ponudniku identitete. Na temelju registracije AAI delujeta tako repozitorij CLARIN.SI kot tudi LINDAT veja konkordančnika Kontext, ki ga bomo v prihodnje ponudili kot storitev CLARIN.SI s korpusi, ki so trenutno dostopni prek noSketchEngine na IJS (Erjavec 2013).[1]

Delovna skupina za pravna vprašanja CLARIN je za urejanje razmerja med repozitorijem in besedilojemalcem izdelala več licenc in seznam najbolj razširjenih, od odprtih (t. i. licence PUB), prek takih, kjer identificirani besedilojemalec lahko neposredno prevzame jezikovni vir, vendar samo za namene raziskav (ACA), ga pa tipično ne sme redistribuirati, do takih, kjer se mora dodatno digitalno podpisati pod pogoje uporabe in morebiti vnesti še dodatne osebne podatke (RES), s čemer so lahko tudi podvrženi še dodatnim omejitvam. Repozitorij CLARIN.SI omogoča tudi nastavitve, pri katerih je vnašalec vira v repozitorij obveščen o vsaki zahtevi po prevzemu, dodatno pa lahko zahtevamo, da se tega vsakemu besedilojemalcu posebej eksplicitno

[1] Dostopen v beta različici na https://www.clarin.si/kontext/.

omogoči.

4.3 Dostop prek konkordančnika

Dostop do korpusa prek konkordančnika je v principu manj problematičen, saj lahko uporabnik naenkrat vidi samo iztržke iz celotnih besedil, čeprav to velja bolj za klasična besedila kot za spletna (npr. tviti ali spletni komentarji), kjer so lahko posamezna besedila zelo kratka, tako besedila kot metapodatki njihovih avtorjev pa tudi lahko najdljivi na spletu.

Korpus Janes bomo instalirali na že omenjeni konkordančnik Kontext, kjer so bolj napredne funkcije dostopne šele po registraciji AAI, omogoča pa tudi omejitev največje dolžine prikazanega konteksta iskanega izraza.

Glavno varovalko nam bo predstavljala možnost, da se lahko zahteva izbris določenega dela (stavka, odstavka ali besedila) iz korpusa, torej rešitev, ki je podobna tisti, ki jo uporablja Google. Vsak strukturni element v korpusu bo opremljen z identifikatorjem, uporabnik pa bo naprošen, da poda razlog za zahtevo po izbrisu ter identifikator strukturnega elementa, ki naj se izbriše. Ta rešitev ima prednost, da se lahko iz korpusa sporne dele odstrani, preden pride do bolj resnih pritožb (npr. odločbe informacijske pooblaščenke ali tožbe), slabost pa, da je za vsako pritožbo potrebno iz korpusa element odstraniti in korpus ponovno indeksirati za konkordančnik. Ker se korpus nato razlikuje od prejšnje različice, ima to za posledico tudi slabšo ponovljivost raziskav. Skrbništvo nad umikom na zahtevo za Kontext bo prevzel CLARIN.SI.

4.4 Prevzem

Prevzem korpusa Janes oz. njegovih podkorpusov bo omogočen prek repozitorija CLARIN.SI. Urejanje dostopa za prevzem je kompleksnejše kot prek konkordančnika, saj tu uporabnik prevzame celoten korpus, kar nato omogoča poljubne analize in uporabe. Takšen dostop najbolj koristi računalniškim jezikoslovcem oz. razvijalcem jezikovnih tehnologij, saj imajo potrebna znanja, da lahko pridobljene baze obdelajo. Slednja imajo tudi

nekateri humanisti in družboslovci, ki lahko prevzete podatkovne zbirke uporabijo za svoje raziskave, so pa v slovenskem prostoru zaenkrat manj razširjena med jezikoslovci, ki zato v veliki meri ostajajo omejeni na uporabo korpusov prek konkordančnika.

Za čim boljšo dostopnost in uporabnost prevzetih korpusov morajo biti ti dobro dokumentirani in zapisani v katerem od standardnih formatov, vendar je diskusija teh vprašanj izven okvira pričujočega prispevka (glej Beißwenger idr. 2012b). Pri pravnih vprašanjih pa sta ključni predobdelava korpusov in licenca, pod katero je korpus objavljen.

V repozitorij CLARIN.SI bomo deponirali dve vrsti korpusov Janes, in sicer avtomatsko jezikoslovne označene (pod)korpuse projekta skupaj z metapodatki ter ročno označena vzorčena besedila, ki bodo na voljo pod dvema vrstama dostopa oz. licenc: za splošno ter za akademsko uporabo.

Korpus	Dov.	AAI	Licenca	MetaA	TekstA	Vzorec	Prem.	Zajem
Slo. viri	da	ne	CC	da	da	da	da?	ne
		da	CLARIN	da	da	ne?	ne	ne
Tviti	ne	ne	CC	ne	ne	ne	ne	da
		da	CLARIN	delno	delno	ne	ne	ne
Wiki	ne	ne	CC	ne	ne	ne	ne	ne
Ročni	ne	ne	CC	da	ne	da	da	ne

Slika 1. Parametri predvidene distribucije korpusov Janes prek CLARIN.SI.

Slika 1 poda parametre za predvideno distribucijo korpusov projekta Janes. V prvem stolpcu je ime (pod)korpusa, pri čemer ločimo podkorpuse, ki so zajeti s slovenskih spletnih platform (blogi, komentarji na novice, forumi) od tvitov, komentarjev na Wikipediji ter ročno označenih korpusov. Drugi stolpec določa, ali bomo lastnika platforme prosili za dovoljenje za redistribucijo korpusov – to bomo storili samo za slovenske lastnike, saj Twitter jasno določa pogoje uporabe, težko si pa tudi predstavljamo, da bi za nas delali izjemo. Za Wikipedijo dovoljenja ne potrebujemo, ročno označeni korpusi pa so dovolj majhni in tudi že vzorčeni, tako da predvidevamo, da dovoljenja ne

potrebujemo.

Tretji in četrti stolpec razdelita dostop do korpusov glede na to, ali se mora uporabnik registrirati in pod kakšno licenco dobi dostop do korpusa. Za splošno uporabo predvidevamo zelo permisivno licenco *Creative Commons Priznanje avtorstva – Deljenje pod enakimi pogoji* (CC BY-SA 4.0), saj bi radi spodbujali čim širšo uporabo korpusov, ob tem pa vsaj moralno spodbudili uporabnike, da svoje izsledke delijo z drugimi pod istimi pogoji, kot so pripisani korpusu. Po drugi strani bomo za akademske uporabnike uporabili eno od licenc CLARIN, predvidoma Academic use + Attribution + Non-commercial + No redistribution, po kateri lahko besedilojemalci korpus uporabijo samo za raziskave in poučevanje, ob tem pa korpusa ne smejo redistribuirati. Slednji pogoj je potreben, da se prepreči nekontrolirano nadaljnje razširjanje korpusov, saj bi se tako hitro izgubili pogoji uporabe kot tudi odgovornost raziskovalcev, ki so korpus prevzeli.

Zadnjih pet stolpcev poda, kako bomo obdelali korpus pred distribucijo. Stolpec »MetaA« določa, ali bomo anonimizirali metapodatke korpusnih besedil, predvsem uporabniško ime – druge metapodatke (npr. spol uporabnika, čas objave) bomo pustili v korpusu, saj ne identificirajo uporabnika, lahko pa koristijo raziskovalcem.

Stolpec »TekstA«, določa, ali bomo opravili anonimizacijo lastnih imen v besedilih. Ker imamo v Sloveniji že precedenčno odločbo informacijske pooblaščenke, ki prepoveduje samo sosledja imena in priimka, bi lahko pri vsakem takem sosledju ohranili samo priimek, izpust pa ustrezno označili. Tako bo npr. namesto »*kje je dr. Bozo Predalic, ko ga clovek rabi.*« v korpusu imeli »*kje je dr. ✖ Predalic, ko ga clovek rabi.*« Prednost tega pristopa je, da ne pokvari koherence besedila in v besedilo ne uvaja izmišljenih šifer.

Poseben primer anonimizacije so tviti, kjer bo anonimizacija samo delna. Uporabniška imena (ki se začenjajo s @) bomo anonimizirali z naključnimi šiframi, tako v metapodatkih (avtor) kot v besedilu. S tem preprečimo

enostavno identifikacijo piscev, obenem pa ohranimo njihov »korpus«, število naslovnikov v tvitu pa tudi pokaže na širši namen pisanja.

Stolpca Vzorec in Premešano določata, ali bomo korpus vzorčili in premešali njegove (vzorčene) dele. Ročno označeni korpusi so že tako ali tako vzorčeni, pri ostalih pa to predvidevamo samo pri podkorpusih slovenskih ponudnikov in še tam pogojno, kar bo odvisno od vrste dogovora, ki ga bomo dosegli s ponudniki platform.

Zadnji stolpec »Zajem« določa, ali bo za korpus potreben ponoven zajem besedil. To predvidevamo samo pri korpusu tvitov, kjer korpus – vsaj javno dostopen – ne bo vseboval besedil tvitov, pač pa samo njihove identifikatorje ter programsko opremo, s katero je mogoče tvite prevzeti od podjetja Twitter. Ta klasična rešitev ima pomanjkljivost, da ne upošteva jezikoslovnih oznak na besedah, saj leme in še posebej normalizirane oblike besed v veliki meri prenesejo tudi samo besedilo. To bomo rešili tako, da bomo v korpusu shranili samo razliko normalizirane oblike oz. leme do izvorne besede, priložena programska oprema pa bo omogočila rekonstrukcijo izvornih oznak.

6 ZAKLJUČEK

Transparentnost in ponovljivost eksperimentov sta temeljni načeli znanstvenega dela, raziskovalci pa so zavezani tudi k načelu širjenja in izkoriščanja podatkov. To pomeni, da bi morali vsi raziskovalci zagotoviti, da bodo podatki in rezultati, pridobljeni v raziskavah, v čim večji meri dostopni tako znanstveni skupnost kot splošni javnosti. Vendar, kot smo obravnavali v pričujočem prispevku, je to načelo v primeru korpusov spletnih besedil marsikje v navzkrižju z zakonodajo.

Novi načini širjenja podatkov so v digitalni dobi pokazali potrebo po reformi intelektualne lastnine, saj neažurna zakonodaja ovira prost pretok idej in spodbujanje ustvarjalnosti (Olson 2013: 93–94, Močnik idr. 2008, Hladnik 2016). Korpusnojezikoslovna skupnost bi si morala organizirano prizadevati za

spremembe zakonodaje, ki bi bolj izrecno opredelile pogoje in načine dela s korpusi. Pri teh pobudah bi bila ključna aktivna vloga konzorcija CLARIN.SI, ARRS in zakonodajnih organov, kot tudi pravna pomoč in podpora v raziskovalnih organizacijah izvajalkah projektov zbiranja korpusov. Nenazadnje bi bilo potrebno ali vsaj zelo koristno v projektih že vnaprej predvideti sredstva za reševanje pravnih ovir pri distribuciji korpusov in pripravi korpusov za distribucijo.

Pomembno bi bilo tudi ozaveščati in spodbujati izdelovalce korpusov, da svoje korpuse objavijo. Še vse prevečkrat se nek korpus zgradi samo za namene določene raziskave, nato pa se za njim izgubi vsaka sled ali v najboljšem primeru ostane za nove raziskave na voljo samo raziskovalcem oz. instituciji, v okviru katere je bil zgrajen. To je seveda najbolj udobno za izdelovalce, saj je v korpus, ki naj bi se redistribuiral, potrebno vložiti več dela, lahko pa jim celo nudi prednosti, saj imajo nato monopol nad zbranimi besedili, ki ga lahko izkoriščajo v novih projektih. Zato bi bilo potrebno nagrajevati objavljanje korpusov tudi skozi mehanizem točkovanja objav v COBISS oz. SICRIS. Podatkovne baze in korpusi sicer že imajo svojo alinejo v objavah, vendar z malo točkami – izjema je Arhiv družboslovnih podatkov ADP, kjer se objave individualno evalvirajo in, v primeru objav podatkov, ki so ocenjeni kot mednarodno relevantni, tudi ocenijo s primernim številom točk. Tudi za korpuse bi se bilo v okviru CLARIN.SI koristno potruditi za podoben sistem točkovanja. Nenazadnje pa je objavljanje korpusov v interesu izdelovalcev tudi zato, ker bodo, ob zadostni raziskovalni etiki uporabnikov, njihovi korpusi citirani v objavah raziskav, ki so te korpuse uporabile.

Tudi javnost bi bilo treba obveščati o rezultatih dela in o namenih, za katere se podatki zbirajo, ter poudarjati družbene koristi korpusnojezikoslovnih raziskav, npr. da bodo podatki lahko podlaga za slovarje in jezikovne tehnologije, ki bodo slovenščini nudile boljšo podporo v digitalni dobi. V primeru spletnih korpusov je treba o tem ozaveščati tudi lastnike spletnih portalov, jim transparentno predstaviti projekte in rezultate, jih vključiti v

proces in se jim zahvaliti za sodelovanje ter na tak način vzpostaviti primere dobrih praks, ki bodo privedli do potencialnih sprememb zakonodaje.

Dokler zakonodaja ne bo prilagojena digitalnemu mediju, pa morajo zbiralci korpusov krmariti med željo po čim bolj liberalnem objavljanju korpusov v čim bolj neokrnjeni obliki na eni strani in neživljenjsko zakonodajo, ki preprečuje skoraj kakršnokoli redistribucijo, na drugi. Kot je bilo ilustrirano v prispevku, so nevarnosti kršenja zakonodaje dostikrat bolj namišljene kot dejanske, saj je osnova za preganjanje kršitve pritožba nad vsebovanostjo konkretnega besedila ali besedil v korpusu. Glede na to, da raziskovalna dejavnost ponudnikom spletnih storitev ne povzroča poslovne škode, je (uspešno) sodno preganjanje raziskovalcev razmeroma malo verjetno. V skladu z raziskovalnimi etičnimi standardi pa ostaja pomembna ustrezna zaščita zasebnosti oseb, katerih imena so vključena v korpus.

Če se izdelovalci korpusov po najboljših močeh potrudijo, da je korpus primerno predobdelan, ponujen pod primernimi licencami in omogočena zahteva za izbris posameznih besedil, je velika verjetnost, da bodo korpus lahko s pridom uporabljali drugi raziskovalci ali celo podjetja, ob tem pa to nikomur ne bo povzročilo škode. V prispevku smo navedli vrsto ukrepov, kako se tovrstnim težavam izogniti, v prihodnje pa jih bomo implementirali pri gradnji dostopnih različic korpusa Janes.

ZAHVALA

Raziskava, opisana v prispevku, je bila opravljena v okviru nacionalnega temeljnega projekta »Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine« (J6-6842, 2014–2017), ki ga financira ARRS.

LITERATURA

Al-Sulaiti, L.; Atwell, E. (2004): Designing and developing a corpus of contemporary Arabic. Zbornik šeste konference TALC.

Baroni, M.; Bernardini, S.; Ferraresi, A.; Zanchetta, E. (2009): The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation 43/3. 209–226.

Beißwenger, M.; Ermakova, M.; Geyken, A.; Lemnitzer, L.; Storrer, A. (2012b): DeRiK: A German Reference Corpus of Computer-Mediated Communication. Zbornik konference Digital Humanities 2012. Alliance of Digital Humanities Organizations (ADHO).

Beißwenger, M.; Ermakova, M.; Geyken, A.; Lemnitzer, L.; Storrer, A. (2012b): A TEI Schema for the Representation of Computer-mediated Communication. V: Journal of the Text Encoding Initiative, Issue 3.

Beißwenger, M.; Storrer, A. (2008): Corpora of computer-mediated communication. V: A. Lu¨deling and M. Kytö (ur.). Corpus linguistics: An international handbook. Vol. 1, 292–309. Berlin and New York: Walter de Gruyter.

Beurskens, M. (2014): Legal Questions of Twitter Research V: V: Weller, K.; Bruns, A.; Burgess, J.; Mahrt, M.; Puschmann, C.: Twitter and Society. Peter Lang.

Beurskens, M. (2014): Legal Questions of Twitter Research. V: Weller, K.; Bruns, A.; Burgess, J.; Mahrt, M.; Puschmann, C.: Twitter and Society. Peter Lang.

Corti, L.; Day, A.; Backhouse, G. (2000): Confidentiality and Informed Consent: Issues for Consideration in the Preservation of and Provision of Access to Qualitative Data Archives. Forum Qualitative Sozialforschung/Forum: Qualitative Social Research 1/3. http://www.qualitative-research.net/index.php/fqs/article/view/1024/2207

Čibej, J.; Arhar Holdt, Š.; Erjavec, T.; Fišer, D. (2016): Razvoj učne množice za izboljšano označevanje spletnih besedil. Zbornik konference Jezikovne tehnologije in digitalna humanistika.

Čibej, J.; Fišer, D.; Erjavec, T.; Arhar Holdt, Š. (2016): Razvoj učne množice za izboljšano označevanje spletnih besedil. JTDH 2016.

Dann, S. (2010): Twitter content classification First Monday, Volume 15, Number 12 http://firstmonday.org/ojs/index.php/fm/article/view/2745/2681

Dürscheid, C. (2015): Interaktionsräume ohne Grenzen? Texte in den neuen Medien. V: Dalmas, Martine idr. (ur.): Texte im Spannungsfeld von medialen Spielräumen und Normorientierung. Pisaner Fachtagung 2014 zu interkulturellen Perspektiven der internationalen Germanistik. München: Iudicum, 74–88.

Erjavec, T. (2013): Korpusi in konkordančniki na strežniku nl.ijs.si. Slovenščina 2.0, ISSN 2335-2736, letn. 1, št. 1, str. 24-49. http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_03.pdf.

Erjavec, T.; Čibej, J.; Fišer, D. (2015): Pravna podlaga za zagotavljanje prostega dostopa korpusov spletnih besedil. Smolej, M. (ur.). OBDOBJA 34: Slovnica in slovar – aktualni jezikovni opis. Ljubljana: Znanstvena založba Filozofske fakultete, 193–199.

Erjavec, T.; Javorše., J.; Krek, S. (2014): Raziskovalna infrastruktura CLARIN.SI. Zbornik Devete konference Jezikovne tehnologije. Ljubljana: Institut »Jožef Stefan«. 19–24.

Evropska komisija (2006): Evropska listina za raziskovalce. Kodeks ravnanja pri zaposlovanju raziskovalcev. http://ec.europa.eu/euraxess/pdf/brochure_rights/kina21620b7c_si.pdf

Evropska komisija (2012): Towards better access to scientific information: Boosting the benefits of public investments in research.

Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. https://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf

Fišer, D., Erjavec, T., Ljubešić, N. (2016): JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. Slovenščina 2.0, 4 (2): 67–100.

Glaznieks, A.; Stemle, E. (2014): Challenges of building a CMC corpus for analyzing writer's style by age: The DiDi project. Journal for Language Technology and Computational Linguistics 29/2. 31–57.

Goli, T.; Osrajnik, E.; Fišer, D. (2016): Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter. Zbornik konference Jezikovne tehnologije in digitalna humanistika.

Guevara, E.; Johannessen, J. (2014): NoWaC (Norwegian Web as Corpus), LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11372/LRT-343.

Halacsy, P. (2014): Hungarian Web Corpus, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11372/LRT-348.

Hemming, C.; Lassi, M. (2002): Copyright and the web as corpus. http://hemming.se/gslt/copyrightHemmingLassi.pdf

Hladnik, M. (2016): Nova pisarija. WikiKnjige. https://sl.wikibooks.org/wiki/Nova_pisarija

King, B. (2009): Building and analysing corpora of computer-mediated communication. Contemporary corpus linguistics, 301-320.

Kotar, M. (2013): Odprti dostop v Evropski uniji in v Sloveniji. Knjižničarske

novice 23/10. http://www.nuk.uni-lj.si/knjiznicarskenovice/v2/podrobnostClanek.aspx?id=778

Kupietz, M.; Lüngen, H. (2014): Recent Developments in DeReKo. Language Resources and Evaluation 43/3. 209–226.

Lee, C.; Woods, K. (2012): Automated Redaction of Private and Personal Data in Collections: Toward Responsible Stewardship of Digital Heritage. The Memory of the World in the Digital age: Digitization and Preservation, 2012. Vancouver, BC.

Lessig, L. (1999): Code and other laws of cyberspace. New York, NY: Basic Books.

Longhi, J.; Marinica, C.; Borzic, B.; Alkhouli, A. (2014): Polititweets : corpus de tweets provenant de comptes politiques influents 1. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang.fr: Nancy. http://hdl.handle.net/11403/comere/cmr-polititweets/cmr-polititweets-tei-v1

Majliš, M. (2011): W2C – Web to Corpus – Corpora, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11858/00-097C-0000-0022-6133-9.

Margaretha, E.; Lüngen, H. (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. JLCL, 29(2), 59-82.

Medlock, B. (2006): An introduction to NLP-based textual anonymisation. Zbornik pete mednarodne konference Language Resources and Evaluation (LREC).

Močnik, M.; Bogataj Jančič, M.; Kovačič, M.; Milohnić, A. (2008): Upravljanje avtorskih in sorodnih pravic v digitalnem okolju. Končno poročilo raziskovalnega projekta.

http://www.uil-sipo.si/fileadmin/upload_folder/prispevki-mnenja/Raziskava_Upravljanje-ASP_2008.pdf

Olohan, M. (2004): Introducing corpora in translation studies. Routledge.

Olson, K. (2013): Intellectual Property. V: Stewart, Daxton (ur.). Social Media and the Law: A Guidebook for Communication Students and Professionals. New York: Routledge, 75-98.

Olutobi, O.; O'Connor, B.; Dyer, C.; Gimpel, K.; Schneider, N.; Smith, N. (2013): Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In Proceedings of NAACL 2013. http://www.cs.cmu.edu/~ark/TweetNLP/#pos

Östling, R.; Wirén, M. (2013): Compounding in a Swedish Blog Corpus. Computer mediated discourse across language. Stockholm: Stockholm University. 45–63.

Owoputi, O.; O'Connor, B.; Dyer, C.; Gimpel, K.; Schneider, N.; Smith, N. A. (2013): Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Petrovič, S.; Osborne, M.; Lavrenko; V. (2010): The Edinburgh Twitter Corpus. Zbornik konference NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media. Los Angeles: Association for Computational Linguistics. 25–26.

Popič, D.; Fišer, D.; Zupan, K.; Logar, P. (2016): Raba vejice v uporabniških spletnih vsebinah. Zbornik konference Jezikovne tehnologije in digitalna humanistika.

Puschmann, C.; Burgess, J. (2014): The Politics of Twitter Data V: Weller, K.; Bruns, A.; Burgess, J.; Mahrt, M.; Puschmann, C.: Twitter and Society. Peter Lang.

Riccardi, G. (ur.) (2015): *Sensei project D8.4 – Second Ethical Issues Report*.

http://www.sensei-conversation.eu/wp-content/uploads/2016/02/D8.4SecondEthicalIssuesReport_v5.3_updated_final.pdf

Schäfer, R.; Bildhauer, F. (2012): Building Large Corpora from the Web Using a New Efficient Tool Chain. Zbornik konference Eighth International Conference on Language Resources and Evaluation (LREC'12).

Sodba Sodišča z dne 13. maja 2014 v zadevi C-131/12. http://curia.europa.eu/juris/document/document.jsf?text=&docid=152065&amppageIndex=0&doclang=sl&mode=lst&dir=&occ=first&part=1&cid=276332

Spooren, W.; van Charldorp, T. (2014): Challenges and experiences in collecting a chat corpus. Journal for Language Technology and Computational Linguistics 29/2. 1–15.

Spousta, M. (2006): Web as a Corpus. Zbornik konference WDS'06. Praga: Matfyzpress. 179–184.

Štebe, J; Bezjak, S.; Lužar, S. (2013): Odprti podatki: načrt za vzpostavitev sistema odprtega dostopa do raziskovalnih podatkov v Sloveniji. Ljubljana: FDV.

Teutsch, P.; Piat, F.; Reffay, C. (2009): Anonymizing and sharing corpora of online training courses. Zbornik konference Interaction Analysis and Visualization for Asynchronous Communication, Workshop CSCL'2009. International Society of the Learning Sciences. 1–6.

Twitter (2016a). Terms of service. http://twitter.com/tos

Twitter (2016b): Developer Display Requirements https://dev.twitter.com/overview/terms/agreement-and-policy

Twitter (2016c): Developer Rules of the Road https://dev.twitter.com/overview/terms/agreement-and-policy

Twitter (2016d): Privacy Policy https://twitter.com/privacy

Vintar, Š.; Fišer, D. (2009): Gradnja in analiza korpusov za prevodoslovne raziskave. V: Kocijančič-Pokorn, Nike (ur.). Sodobne metode v prevodoslovnem raziskovanju, (Zbirka Prevodoslovje in uporabno jezikoslovje). Ljubljana: Znanstvena založba Filozofske fakultete, 2009, str. 80-109.

Legal framework of textual data processing for Machine Translation and Language Technology research and development activities/Open Data and Web crawling Case Studies. Wiki Books. https://en.wikibooks.org/wiki/Legal_framework_of_textual_data_processing_for_Machine_Translation_and_Language_Technology_research_and_development_activities/Open_Data_and_Web_crawling_Case_Studies

Yang, J.; Leskovec, J. (2011): Temporal Variation in Online Media. ACM International Conference on Web Search and Data Mining (WSDM '11). http://snap.stanford.edu/data/twitter7.html

OVERCOMING LEGAL LIMITATIONS IN DISSEMINATING SLOVENE WEB CORPORA

Web texts are becoming increasingly relevant sources of information, with web corpora useful for corpus linguistic studies and development of language technologies. Even though web texts are directly accessable, which substantially simplifies the collection procedure compilation of web corpora is still complex, time consuming and expensive. It is crucial that similar endeavours are not repeated, which is why it is necessary to make the created corpora easily and widely accessible both to researchers and a wider audience. While this is logistically and technically a straightforward procedure, legal constraints, such as copyright, privacy and terms of use severely hinder the dissemination of web corpora. This paper discusses legal conditions and actual practice in this area, gives an overview of current practices and proposes a range of mitigation measures on the example of the Janes corpus of Slovene user-generated content in order to ensure free and open dissemination of Slovene web corpora.

Keywords: web texts, corpus dissemination, copyright, privacy, free and open access