PRAVNA PODLAGA ZA ZAGOTAVLJANJE PROSTEGA DOSTOPA KORPUSOV SPLETNIH BESEDIL

Tomaž Erjavec
Institut »Jožef Stefan«, Ljubljana

Jaka Čibej, Darja Fišer
Filozofska fakulteta, Ljubljana

UDK 004.738.52:347.788:81'374

Korpusi spletnih besedil so uporabni pri izdelavi jezikovnih priročnikov, v korpusno-jezikoslovnih raziskavah in pri razvoju jezikovnih tehnologij. Izdelava takšnih korpusov je kljub neposredni dostopnosti besedil zapletena in draga, zato je zelo pomembno, da omogočimo njihovo čim večjo dostopnost čim širši raziskovalni skupnosti in splošni zainteresirani javnosti. Tehničnih ovir za to ni, obstajajo pa številne omejitve s področja zaščite avtorskih pravic, osebnih podatkov in pogojev uporabe ponudnikov spletnih storitev. V prispevku predstavljamo pravno in dejansko stanje na teh treh področjih v Sloveniji ter predlagamo ukrepe, ki do največje možne mere omogočajo prosto in odprto razširjanje korpusov spletne slovenščine.

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

Web corpora are useful in producing language reference materials, conducting research in corpus linguistics, and developing language technology applications. Despite the direct availability of web texts, building such corpora is complicated and expensive, which is why it is important to ensure their availability both to the academic community and the general public. But despite the lack of technical obstacles, a number of legal restrictions must be taken into account, e.g. copyright, personal data protection, and the terms of use of various web service providers. In this article, we provide an overview of the legal basis in this regard as well as the *de facto* state of things in Slovenia, and suggest a number of measures to enable free and open dissemination of corpora of internet Slovene.

web texts, dissemination of corpora, copyright, personal data protection, free and open access

1 Uvod

V zadnjih letih je odprti dostop do raziskovalnih podatkov postal vroča tema tako v mednarodnem okviru (EC 2012) kot tudi v Sloveniji (Kotar 2013; Štebe idr. 2013), saj je dozorelo spoznanje, da je pridobivanje podatkov za raziskave drago in zamudno, podatki pa so najpogosteje uporabljeni samo v objavi rezultatov raziskave in niso dostopni drugim raziskovalcem, zato se enaki ali podobni podatki pogosto zbirajo in obdelujejo večkrat. S tem se po nepotrebnem tratita javni denar in čas, ki bi ju lahko

namenili nadaljnjim raziskavam, obenem pa predhodnih raziskav ni mogoče ponoviti, preveriti ali nadgraditi, kar je osnova vsake znanosti.

Raziskave jezika z besedilnimi korpusi tu niso nobena izjema. Dostopnost velikih, dobro metapodatkovno in jezikoslovno označenih korpusov je osnova sodobnega slovaropisja, empiričnega jezikoslovja ter razvoja jezikovnih tehnologij. Pomembnosti večkratne uporabe jezikovnih podatkov se zaveda tudi Evropska unija, saj je bila za namene zagotavljanja dostopnosti podatkov za humanistične in družboslovne raziskave ustanovljena evropska raziskovalna infrastruktura CLARIN (Common Language Resources Infrastructure), v Sloveniji in za slovenščino pa konzorcij CLARIN.SI (Erjavec idr. 2014).

Kljub temu lahko za nadaljnje razširjanje raziskovalnih podatkov veljajo določene omejitve. Pri besedilnih korpusih so to avtorske pravice izvirnih besedil in vprašanje varovanja osebnih podatkov, saj korpusi pogosto vsebujejo velike količine stvarnih besedil (npr. časopisnih člankov) in jih zato lahko obravnavamo kot agregirane baze osebnih podatkov. Pogoji za diseminacijo besedil in podatkovnih baz so zelo strogi, čeprav je namen korpusov proučevanje jezika, ne ponovno izdajanje zaključenih besedil ali poizvedovanje po osebnih podatkih. Na težave s strogimi pogoji za dostopnost korpusov opozarjajo tudi številni izdelovalci korpusov drugod po svetu (Spousta 2006; Baroni idr. 2009; Beißwenger idr. 2012).

Danes se vedno več korpusov gradi neposredno iz besedil, dostopnih na spletu. Splet namreč vsebuje vse več relevantnih vsebin, pa tudi njihov zajem je tehnično mnogo enostavnejši od gradnje klasičnih korpusov tiskanih besedil. V prispevku bomo pregledali pravne omejitve diseminacije spletnih besedil in predlagali načine, kako jih (delno) preseči. Povod za to je korpus spletne slovenščine, ki ga gradimo v sklopu projekta Janes (Fišer idr. 2014).

2 Načini dostopa do korpusnih podatkov

Korpuse lahko ločimo po načinu diseminacije. V prispevku predstavljamo prosti in odprti dostop, ki sta po našem mnenju za slovenski kontekst najbolj zanimiva, saj omogočata najširši dostop do izdelanih korpusov.

Prosti dostop pomeni, da je korpus mogoče brezplačno pregledovati na spletu preko konkordančnika. 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) pa je mogoče na (upravičeno) zahtevo izbrisati oz. do njih delno ali popolnoma onemogočiti dostop.

Pri *odprtem dostopu* je korpus dovoljeno prevzeti, tj. kot podatkovno bazo v celoti prenesti na lokalni računalnik. Tak pristop je nujen za razvoj jezikovnih tehnologij in za bolj poglobljene jezikoslovne raziskave, saj pri tem nismo več omejeni na specifičen konkordančnik, temveč je korpus mogoče analizirati z lastno programsko opremo. Pri tem je možnosti za zlorabe več, saj je na tak način mogoče zajeti in ponatisniti celotna besedila ali podrobno analizirati vedênje določene osebe. Sporna

besedila lahko iz izvornega korpusa sicer izbrišemo ali korpus celo zaklenemo, a nad že prevzetimi kopijami nimamo več nadzora.

Pravna definicija odprtega dostopa ni preprosta in zanj je bila izdelana že vrsta licenc, predvsem za odprto programsko kodo (npr. GNU General Public License – GPL). Za besedila in druga digitalna avtorska dela so najbolj razširjene licence Creative Commons,[1] ki pod različnimi pogoji omogočajo prevzem in nadaljnjo diseminacijo del. Z licenco Creative Commons omogočimo najširšo distribucijo in uporab(lja)nost izdelanih korpusov, a imamo obenem najmanj nadzora nad nadaljnjo usodo vsebovanih besedil.

Z zavedanjem, da vseh jezikovnih podatkov ni mogoče dati v prosto uporabo, deluje pri infrastrukturi CLARIN delovna skupina za pravna vprašanja distribucije virov,[2] ki je izdelala licence med ponudniki jezikovnih virov v repozitorije CLARIN in serijo pogojev uporabe, ki jim mora zadostiti besedilojemalec: od povsem odprte uporabe do dovoljenja za individualno delo za točno določen namen.

3 Pravne omejitve

3.1 Avtorske pravice

V Sloveniji področje avtorskih pravic ureja Zakon o avtorski in sorodnih pravicah (Uradni list št. 21/95), ki določa, da je vsako pisno delo skupaj z vsemi sestavnimi deli avtorsko, avtorska pravica pa ugasne 70 let po smrti avtorja. Korpusi spadajo tudi med baze podatkov, z uvrstitvijo del v bazo pa ne smejo biti prizadete pravice avtorjev. Avtorji imajo tudi pravico do skesanja: lahko zahtevajo umik vsebine, če ta resno moralno ali gmotno vpliva na njihov položaj. To pravico upošteva npr. Twitter, pri katerem lahko uporabniki prekličejo svoje tvite, ki so nato izbrisani iz baze.

Pri (zlasti referenčnih) korpusih večinoma tiskanih besedil se vprašanje avtorskih pravic ponavadi ureja s pisnimi sporazumi med izdelovalci korpusa in imetniki avtorskih pravic (Logar Berginc idr. 2012; Kupietz, Lüngren 2014). Pri spletnih besedilih je pridobivanje dovoljenj zaradi pogoste anonimnosti avtorjev na spletu in njihovega velikega števila zelo nepraktično in realno izvedljivo le pri gradnji manjših korpusov in z zbiranjem podatkov v krajšem časovnem obdobju (Glaznieks, Stemle 2014; Spooren, van Charldorp 2014).

3.2 Varstvo osebnih podatkov

Področje varovanja osebnih podatkov pri nas ureja Zakon o varstvu osebnih podatkov (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 tudi, da se lahko podatki ne glede na prvotni namen zbiranja nadalje obdelujejo za znanstvenoraziskovalne namene, a samo v anonimizirani obliki.

1 http://creativecommons.org, http://creativecommons.si
2 https://www.clarin.eu/governance/legal-issues-committee

Iznos podatkov v tretje države (izven EU oziroma EGP) je dovoljen le v primeru, da tretja država zagotavlja ustrezno raven varstva osebnih podatkov.

Z varstvom osebnih podatkov je povezana t. i. pravica do pozabe, ko zastarela informacija posamezniku očitno škoduje. Ta pravica 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 podjetju Google nalaga, da mora na zahtevo uporabnikov vsebine deindeksirati.

V Sloveniji je podoben in zelo relevanten primer odločba Urada informacijske pooblaščenke, ki je zahtevala onemogočenje iskanja po sosledju lastnih imen oseb v korpusu Nova beseda.[3] To je med jezikoslovci povzročilo ogorčenje, saj je ta ukrep onemogočil tudi iskanje imen namišljenih ali zgodovinskih 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, pa tudi ob izbrisu s spleta so lahko ohranjene v arhivih spletnih vsebin, kot sta Wayback Machine[4] in Spletni arhiv NUK.[5]

3.3 Pogoji uporabe ponudnikov spletnih storitev

Lastniki spletnih mest objavo besedil na spletu pogosto omejijo s pogoji uporabe, ki določajo tudi, ali in kako lahko tretje osebe objavljena besedila zajemajo in uporabljajo. Ti pogoji so večinoma zelo strogi. Forum avtomobilizem.com npr. v svojem pravnem pouku navaja, da je kakršnokoli reproduciranje, javno objavljanje, spreminjanje ipd. kateregakoli dela njihove spletne strani brez pisnega dovoljenja podjetja Domenca Labs, d. o. o. (lastnika portala), prepovedano in lahko tudi sodno preganjano.

4 Odpravljanje ovir dostopa

V tem razdelku obravnavamo načine, kako lahko korpuse kljub pravnim preprekam napravimo prosto oziroma odprto dostopne. Ker idealne rešitve za vse identificirane ovire ni, predstavljamo nabor metod, ki v večji ali manjši meri posegajo v besedilo in tako obratnosorazmerno obidejo pravne ovire. Najprimernejša rešitev oz. kombinacija rešitev je odvisna od vira besedil in od stopnje dostopnosti korpusa.

3 http://mailman.ijs.si/pipermail/slovlit/2012/004192.html
4 http://web.archive.org
5 http://arhiv.nuk.uni-lj.si

4.1 Odpravljanje težav z avtorskimi pravicami

Dovoljenje za uporabo

Uporabniki družbenih medijev s sprejetjem pogojev uporabe najpogosteje na ponudnike prenesejo materialne avtorske pravice za svoja besedila, zato bi bilo od večjih ponudnikov mogoče pridobiti dovoljenje za uporabo s sklenitvijo sporazuma (Kupietz, Lüngren 2014). Pri tem je lastnike avtorskih pravic spletnih vsebin treba ozaveščati, zakaj je gradnja tovrstnih korpusov pomembna, za kakšne namene bodo besedila uporabljena in na kakšen način bodo diseminirana.

Vzorčenje

Iz celotne zbirke besedil lahko za korpus vzamemo le naključno izbrane zamejene enote, npr. odstavke ali stavke (Östling, Wirén 2013). Ker s tem uporabimo zgolj tolikšen delež besedil, kot je brez kršenja avtorskih pravic dovoljen z zakonodajo, se je vzorčenje uveljavilo kot klasičen način za omogočanje odprtega dostopa do korpusov. Vzorčenje odstavkov so npr. uporabili pri izgradnji uravnoteženega (ccKRES) in referenčnega korpusa (ccGigaFida) slovenskega jezika (Logar Berginc idr. 2012), ki sta odprto dostopna pod licenco Creative Commons. Vzorčenje bistveno zmanjša velikost končnega korpusa, razbije pa tudi koherenco besedila, ki zato ni več primerno za pomensko analizo, analizo diskurza ipd.

Premešanje

Metoda, ki je podobna vzorčenju, a ne zmanjšuje količine podatkov, je *premešanje*, ki le spremeni izvorni vrstni red segmentov. Pristop s premešanimi stavki sta za odprto distribucijo velikih korpusov spletnih besedil več evropskih jezikov uporabila Schäfer in Bildhauer (2012). V najstrožjem scenariju lahko za segmente vzamemo kar različne n-terčke[6] besed, vsakemu pa dodamo še njegovo frekvenco v korpusu oz. število besedil. Takemu seznamu težko še rečemo korpus, je pa kljub temu koristen za analizo jezika in je s pravnega vidika lahko povsem odprt.

4.2 Odpravljanje težav z osebnimi podatki

Anonimizacija

V metapodatkih in znotraj besedil so z vidika varovanja osebnih podatkov glavni potencialni problem osebna lastna imena, pa tudi e-poštni naslovi in v manjši meri naslovi URL. V občutljivih besedilnih zvrsteh (npr. zdravstvene kartoteke) se problem osebnih imen večinoma rešuje z anonimizacijo (Corti idr. 2000; Teutsch idr. 2009; Petrović idr. 2010). Pri korpusu Janes je situacija nekoliko drugačna, saj so zajeta besedila javno dostopna na spletu. Iz korpusa torej o neki osebi ne bomo izvedeli nič več, kot bi z brskanjem po spletu, zato ni jasno, ali je imena res treba anonimizirati.

Identifikacija imen oseb je razmeroma enostavna s seznami lastnih imen in priimkov Statističnega urada Republike Slovenije.[7] V slovenskem kontekstu bi lahko

6 N-terčki so zaporedja besed (ali drugih enot), ki so dolgi *n* enot, pri čemer je *n* večinoma v razponu od 3 do 7.

7 https://www.stat.si/ImenaRojstva/sl

pri anonimizaciji implementirali tudi že omenjeno precedenčno odločbo informacijske pooblaščenke, tako da bi pri vsakem sosledju lastnih imen ohranili samo prvo ime, izpust pa ustrezno označili (npr. »Dr. Janez <gap/> je dejal …«). Prednost tega pristopa je, da ne pokvari koherence besedila in v besedilo ne uvaja izmišljenih šifer.

5 Zaključek

Po najstrožji interpretaciji zakonskih določil je kakršnokoli kopiranje (spletnih) besedil brez predhodnega pisnega dovoljenja vseh nosilcev avtorskih pravic nedovoljeno, a je treba za vsak primer posebej pretehtati na eni strani potencialno škodo, ki bi jo nosilci pravic imeli s kopiranjem in omogočanjem dostopa do njihovih besedil, na drugi pa koristi, ki jih omogočanje dostopa do jezikoslovno bogato označenih korpusov prinaša za znanost, predvsem za slovenistiko in računalniško jezikoslovje. Problematika avtorskih pravic in varstva osebnih podatkov je v okviru korpusov še vedno brez celovitih in sistemskih rešitev, zato bi si morala znanstvena skupnost organizirano prizadevati za spremembe zakonodaje, ki bi bolj izrecno opredelile pogoje in načine dela s korpusi.

V prispevku smo nakazali probleme pri zbiranju, uporabi in diseminaciji korpusov spletnih besedil ter definirali več načinov, kako se takim problemom izogniti z zagotovitvijo dovoljenja za uporabo, vzorčenjem, premešanjem in anonimizacijo. V nadaljevanju raziskav bi ustreznost teh možnosti preverili s konzultacijo z Uradom informacijske pooblaščenke in s področnimi pravniki ter z implementacijo nekaterih predlaganih načinov v praksi.

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

BARONI, Marco, BERNARDINI, Silvia, FERRARESI, Adriano, ZANCHETTA, Eros, 2009: The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43/3. 209–226.

BEIßWENGER, Michael, ERMAKOVA, Maria, GEYKEN, Alexander, LEMNITZER, Lothar, STORRER, Angelika, 2012: DeRiK: A German Reference Corpus of Computer-Mediated Communication. *Zbornik konference Digital Humanities 2012*. Alliance of Digital Humanities Organizations (ADHO).

CORTI, Louise, DAY, Annette, BACKHOUSE, Gill, 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 [On-line Journal]* 1/3. http://www.qualitative-research.net/index.php/fqs/article/view/1024/2207

ERJAVEC, Tomaž, 2013: Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0*. 1. Škofja Loka: Trojina. 24–49. http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_03.pdf

ERJAVEC, Tomaž, JAVORŠEK, Jan Jona, KREK, Simon, 2014: Raziskovalna infrastruktura CLARIN.SI. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Devete konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 19–24.

EUROPEAN COMMISSION, 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, Darja, ERJAVEC, Tomaž, ZWITTER VITEZ, Ana, LJUBEŠIĆ, Nikola, 2014: JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Devete konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«.

GLAZNIEKS, Aivars, STEMLE, Egon W., 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.

KOTAR, Mojca, 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, Marc, LÜNGEN, Harald, 2014: Recent Developments in DeReKo. *Language Resources and Evaluation* 43/3. 209–226.

LOGAR BERGINC, Nataša, GRČAR, Miha, BRAKUS, Marko, ERJAVEC, Tomaž, ARHAR HOLDT, Špela, KREK, Simon, 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina/FDV.

ÖSTLING, Robert, WIRÉN, Mats, 2013: Compounding in a Swedish Blog Corpus. *Computer mediated discourse across language*. Stockholm: Stockholm University. 45–63.

PETROVIĆ, Saša, OSBORNE, Miles, LAVRENKO, Victor, 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.

SCHÄFER, Roland, BILDHAUER, Felix, 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, Wilbert, van CHARLDORP, Tessa, 2014: Challenges and experiences in collecting a chat corpus. *Journal for Language Technology and Computational Linguistics* 29/2. 1–15.

SPOUSTA, Miroslav, 2006: Web as a Corpus. *Zbornik konference WDS'06*. Praga: Matfyzpress. 179–184.

ŠTEBE, Janez, BEZJAK, Sonja, LUŽAR, Sanja, 2013: *Odprti podatki: načrt za vzpostavitev sistema odprtega dostopa do raziskovalnih podatkov v Sloveniji*. Ljubljana: FDV.

TEUTSCH, Philippe, PIAT, Frédéric, REFFAY, Christophe, 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.