DIGITALIZACIJA V JEZIKOSLOVJU – NOV POGLED NA JEZIK IN NOVE MOŽNOSTI POUČEVANJA JEZIKA

DARINKA VERDONIK

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija.
E-pošta: darinka.verdonik@um.si

Povzetek V jezikoslovju se digitalizacija kaže predvsem skozi področji korpusnega jezikoslovja in jezikovnih tehnologij. V prvi polovici prispevka zato pregledamo korpusne in jezikovnotehnološke začetke in današnje stanje ter izpostavimo, kako se je hkrati razvijalo in spreminjalo tudi razumevanje jezika. V drugem delu skušamo odgovoriti na vprašanje, kako lahko korpusnojezikoslovne in jezikovnotehnološke pristope učinkovito uporabimo pri razvoju digitalnih učnih e-okolij za učenje (maternega) jezika. Po kratkem pregledu, kako to izvedejo v Pedagoškem slovničnem portalu, in nekoliko obširnejšem pregledu na podlagi našega lastnega sodelovanja pri razvoju Slovenščine na dlani povzamemo pet načinov. Ti vključujejo definiranje učnih vsebin za učno e-okolje, analizo jezikovnih pojavov, priklic primerov za vaje in naloge, vrednotenje uporabnikove uspešnosti ter vodenje uporabnika skozi učno e-okolje.

Ključne besede:
korpusno jezikoslovje,
jezikovne tehnologije,
učno e-okolje,
učni e-pripomočki,
korpus

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/978-961-286-521-4.1
ISBN 978-961-286-521-4

DIGITALISATION IN LINGUISTICS – NEW PERSPECTIVES ON LANGUAGE AND NEW POTENTIALS IN LANGUAGE TEACHING

DARINKA VERDONIK
University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia.
E-mail: darinka.verdonik@um.si

Abstract Digitalisation in linguistics refers to corpus linguistics and language technologies. First, we overview the beginnings of corpus linguistics and language technologies in Slovenia and outline the state-of-the-art today. We then point out how our understanding of language has changed along the way. In the second part, we try to answer the question of how to exploit text corpora and natural language processing methods in developing e-learning environments for (native) language teaching. After providing a brief overview of the "Pedagoški slovnični portal", we undertake a more detailed examination of the "Slovenščina na dlani" e-learning environment, in which we have cooperated ourselves. Finally, we summarise five potential areas in which we can exploit corpus linguistic and language technologies approaches for developing e-learning environments: definition of content for learning or training; analysis of language phenomena; a source of samples for exercises; evaluation of user performance; and leading the user through an e-learning environment.

Keywords: corpus linguistics, language technologies, e-learning environment, e-learning tools, corpus

https://doi.org/10.18690/978-961-286-521-4.1 DOI
978-961-286-521-4 ISBN
University of Maribor Press

1 Uvod

S široko uporabo računalnikov je konec prejšnjega stoletja nove tehnologije začelo raziskovati tudi jezikoslovje. Nastajali so prvi korpusi besedil, za slovenščino na primer korpus FIDA. Morda se je sprva zdelo, da je to samo hitrejši in preprostejši način zbiranja jezikovnih primerov, ki so si jih jezikoslovci pred tem zapisovali ali izpisovali na papir. Toda del jezikoslovcev, zlasti v krogih, povezanih z birminghamsko univerzo, je na novo orodje za raziskave jezika gledal veliko bolj prelomno: zavrgli so predhodna teoretska izhodišča in začeli postavljati hipoteze izključno na podlagi rezultatov korpusnih analiz. Zanje korpus ni bil več samo metoda, ampak tudi teorija. To je prineslo tudi spremembe teoretsko-metodoloških načel in razumevanja jezika. Čeprav so bili ti premiki sprva specifični za ozek jezikoslovni krog, pa lahko danes rečemo, da so implicitno precej široko prisotni. Korpus ali iz njega izpeljani jezikovni viri in priročniki so namreč postali osnovni pripomoček skoraj vsakogar, ki se ukvarja z jezikoslovjem, in so – naj se tega zavedamo ali ne – spreminjali naš pogled na jezik.

Počasi si korpus (bodisi neposredno, kot pripomoček pri učenju, ali posredno, kot zaledni vir, na katerem temeljijo pripomočki za učenje) utira svojo pot tudi v učilnice. V svetu se že dolgo razvija področje računalniško podprtega učenja jezikov (Beatty 2013), ki zajema najrazličnejše načine uporabe računalnikov: od tega, da imamo namensko sprogramirano e-okolje za učenje jezika, prek tega, da računalnik uporabljamo za učno raziskovanje jezika (najpogosteje prav s pomočjo korpusov), do multimedijsko podprtega učenja jezika (npr. avdio ali video vodeni programi učenja) ter ne nazadnje do virtualnih učilnic in uporabe interneta kot medija interakcije in hkrati nosilca atraktivnih vsebin (npr. posnetkov na Youtubu). Med temi različnimi pristopi se bomo v tem prispevku osredotočili predvsem na uporabo korpusov in jezikovnotehnoloških pristopov. Naše osrednje raziskovalno vprašanje bo: Kako lahko besedilne korpuse in jezikovnotehnološke pristope učinkovito uporabimo pri razvoju digitalnih učnih e-okolij za učenje (maternega) jezika?

V nadaljevanju najprej pregledamo začetke digitalizacije v slovenistiki in njeno stanje danes (poglavje 2.1); nato predstavimo, kako je digitalizacija vplivala tudi na razumevanje in pristope v raziskovanju jezika (poglavje 2.2). V tretjem poglavju se osredotočimo na digitalizacijo pri pouku slovenskega jezika s poudarkom na uporabi besedilnih korpusov in jezikovnotehnoloških pristopov ter predstavimo digitalno

učno e-okolje *Slovenščina na dlani*[1] in načine avtomatizacije v njem. Diskusija z zaključkom sledi v četrtem poglavju.

2 Jezikoslovje in jezikovne tehnologije

Digitalizacija je v jezikoslovje vstopala predvsem skozi korpuse in skozi na korpusih temelječe jezikovne vire in orodja, ki sta jih omogočila razvoj in široka dostopnost računalnikov. Osrednji zagon za vpeljavo jezikovnih virov in orodij v jezikoslovje je na mednarodni ravni prihajal po eni strani iz posamičnih jezikoslovnih centrov, kot so angleške University College London, Lancaster University, University of Birmingham, University of Nottingham, ameriška Northern Arizona University ter češka Karlova univerza v Pragi, po drugi strani pa tudi iz tehnoloških ved, ki so pri razvoju tehnologij, kot so razpoznavanje in sinteza govora, strojno prevajanje idr., pritegnile k sodelovanju jezikoslovje. V nadaljevanju bomo na kratko pregledali razvoj v slovenskem prostoru ter teoretske premike v jezikoslovju kot vedi, ki so jih spodbudili (tudi) razviti viri in orodja.

2.1 Od zametkov do danes

Večji razvoj korpusov in jezikovnih tehnologij se je v Sloveniji začel v devetdesetih letih 20. stoletja. V jezikoslovju je bil ta razvoj spodbujen zlasti z leksikografijo in posameznimi za nove poglede na jezik odprtimi jezikoslovci (Stabej in Vitez 1990; Erjavec 1990; Krek 1999). Erjavec idr. (1998) ob predstavitvi korpusa FIDA, prvega referenčnega korpusa slovenščine, tako zapišejo:

> Že nekaj let se je na marsikaterem področju dejavnosti, povezanih z raziskovanjem in opisovanjem slovenskega jezika, vse očitneje kazala potreba po dovolj obsežnem, reprezentativnem in dostopnem korpusu, ki bi zagotavljal objektiviziran pogled na jezik in omogočal uporabo sodobne računalniške tehnologije tako pri temeljnih jezikoslovnih in drugih raziskavah kot pri razvijanju najrazličnejših programskih orodij za obdelavo besedil, predvsem v tistih delih, kjer morajo biti prilagojeni posameznemu naravnemu jeziku. (Erjavec idr. 1998)

[1] Projekt *Slovenščina na dlani* sofinancirata Republika Slovenija in Evropska unija iz Evropskega socialnega sklada. Izvaja se na Filozofski fakulteti, Pedagoški fakulteti in Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru v obdobju 2017–2021.

Tako je sledilo povezovanje institucij, ki so iz takega ali drugačnega razloga videle to kot pomemben korak v razvoju infrastrukture za slovenščino: Filozofske fakultete Univerze v Ljubljani, Instituta Jožef Stefan, založbe DZS in podjetja Amebis. Nastal je korpus FIDA, predhodnik današnje Gigafide. Poleg korpusa FIDA se je v okviru Inštituta za slovenski jezik Frana Ramovša pri ZRC SAZU začel razvijati še en podoben, vendar vsebinsko manj uravnotežen korpus, Beseda, kasneje nadgrajen v Novo besedo (Jakopin 2003). Aktivnosti okrog tega korpusa so danes zamrle. Na daljši rok je namreč vedno bolj eksplicitno postajalo jasno dvoje: prvič, da je treba za izdelavo digitalne jezikovne infrastrukture združevati moči, saj gre za časovno in finančno zajetne projekte, ki jih težko izvede en sam partner, in drugič, da je edina pot k napredku na tem področju odpiranje virov za vse zainteresirane uporabnike, saj se sicer vedno znova vračamo k izdelavi istovrstnih temeljnih virov in nikoli ne pridemo do tega, da bi lahko začeli razvijati vire, orodja in jezikovne priročnike višjih nivojev, ki temeljijo na bazičnih virih.

Medtem ko je zgoraj opisana veja izhajala v veliki meri iz jezikoslovja, zlasti v povezavi s slovarji in leksiko, pa se je v tehničnih centrih začel razvoj govornih tehnologij, ki so prav tako predstavljale enega od začetnih stebrov digitalizacije v jeziku. Tako na Fakulteti za elektrotehniko Univerze v Ljubljani kot na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru so nastajali prvi doktorati, številne znanstvene objave (npr. Kačič idr. 1990; Mihelič idr. 1992; Pavešič idr. 1996; Sepesy Maučec 1997; Imperl idr. 1997) in tudi prvi govorni viri (npr. SNABI – Kačič 1994, danes dostopen na povezavi http://hdl.handle.net/11356/1051, in GOPOLIS – Dobrišek idr. 1998, dostopen na povezavi http://hdl.handle.net/11356/1125), vezani na razvoj strojne sinteze govora, stojnega razpoznavanja govora in sorodnih tehnologij.

V tem zgodnjem obdobju so nastajali tudi zametki digitalizacije na področju prevajalstva: pri Institutu Jožef Stefan je nastal prvi vzporedni korpus, IJS-ELAN (Erjavec 1999), danes del vzporednega korpusa TRANS5 (dostopen prek povezave https://www.clarin.si/noske/para.cgi/). Ne nepomemben razvojni dejavnik pa so bile tudi končne jezikovnotehnološke aplikacije, zanimive za trg, kot so črkovalnik, digitalizacija slovarjev, strojni prevajalnik, sintetizator govora, kjer sta svojo nišo našli dve še danes aktivni jezikovnotehnološki podjetji (Amebis, d. o. o.; Alpineon, d. o. o.).

Posebej navedimo še jezikovne vire in orodja, ki so se nanašali na oblikoslovne (in tudi fonetične) informacije o jeziku. Razvoj teh je bil prepleten z zgoraj navedenimi stebri, saj je zagon za razvoj prihajal iz vseh smeri: slovarji in besedila z označenimi informacijami o osnovni obliki, spolu, sklonu, številu, osebi itd. so se kazali pomembni na vseh področjih, od korpusnih analiz prek črkovalnikov in leksikografije do prevajanja, razpoznavanja in sinteze govora. Posledično so nastajali prvi oblikoslovni (za potrebe pri procesiranju govora pa tudi fonetični) jezikovni viri, npr. MULTEXT-EAST (Erjavec 1998); SiLEX (Verdonik idr. 2002).

Od opisanih zametkov v devetdesetih letih 20. stoletja do danes je sledil razmah korpusnega jezikoslovja, jezikovnih virov ter govornih in jezikovnih tehnologij, kot nazorno pričajo številne monografije (npr. Gantar 2008; Zemljarič Miklavčič 2008; Vintar 2010; Arhar Holdt 2011; Logar idr. 2012; Logar 2013; Šorli 2020), zborniki konference Slovenskega društva za jezikovne tehnologije, tj. konferenc Informacijska družba/Jezikovne tehnologije oz. njenih naslednic, Jezikovne tehnologije in digitalna humanistika, ter 234 vnosov jezikovnih virov ali orodij v repozitorij slovenskega konzorcija CLARIN.SI ob pisanju tega prispevka, od katerih jih je večina tudi odprto dostopnih.

Področje digitalizacije v jezikoslovju danes lahko nazorno opišemo prek vsebinskih sklopov trenutno največjega projekta na tem področju v Sloveniji, Razvoj slovenščine v digitalnem okolju. Razporeditev vsebinskih sklopov tega projekta odlično prikazuje razvejenost področja in hkrati medsebojno prepletenost in soodvisnost posameznih vej, obenem pa še vedno odslikava vodilne motive, prek katerih se je razvoj digitalizacije v jezikoslovju začel. Najbolj bazični sklop so temeljni jezikovni viri: od referenčnih korpusov (npr. Gigafida, GOS) in specializiranih korpusov (Janes, Šolar, KOST) prek leksikonskih virov (Sloleks) do orodij za avtomatsko označevanje besedil (tokenizacija, lematizacija, oblikoslovno označevanje, skladenjsko razčlenjevanje …). Drugi sklop, ki je po manjšem vmesnem zatišju v zadnjih letih pridobil nekdanjo aktualnost, so govorne tehnologije: sintetizator govora, razpoznavalnik govora, govorne baze in korpusi govorjenega jezika … Tretji, danes verjetno najbolj aktualen in živahno razvijajoč se sklop so semantični viri in semantične tehnologije: semantične mreže, korpusi za izvajanje semantičnih analiz, orodja za razdvoumljanje pomenov in prepoznavanje semantičnih premikov, avtomatsko povzemanje besedil, avtomatsko odgovarjanje na vprašanja … Svoje stabilno, vedno izzivov polno mesto ohranjata strojno

prevajanje in prevodoslovje s potrebami po vzporednih korpusih in drugih jezikovno poravnanih virih, podpornih orodjih, ki olajšajo zbiranje in obdelavo besedil ... Posebno mesto pripada izzivom in potrebam terminologije: področno specifični korpusi, avtomatsko luščenje terminoloških kandidatov, terminološki slovarji, terminološki portal za uporabnike … Da vsa številna digitalna jezikovna infrastruktura ostane varno shranjena in dostopna, pa se je skozi čas pokazala tudi potreba po vzpostavitvi skupnega centra, ki zagotavlja hranjenje, dostopnost, standardizacijo, tj. repozitorij konzorcija CLARIN.SI.

2.2 Premiki v razumevanju jezika

Hkrati z razvojem korpusnega jezikoslovja in jezikovnih tehnologij so se spreminjali tudi nekateri temeljni jezikoslovni pogledi na jezik. Teh sprememb ne smemo gledati črno-belo v smislu dihotomije pred digitalizacijo – po uvedbi digitalizacije. Spremembe tudi niso nujno vseprisotne, znanstveni pristopi in pogledi na jezik so bili in so raznovrstni, in ta pluralizem je pomembno ohranjati. Digitalni viri, orodja in tehnologije ne smejo postati edini postulat razvoja. A vendarle želimo tukaj opozoriti na nekaj razmerij med prej – potem, kot jih sami vidimo in ki se pomembno odslikavajo tudi v slovenskem jezikoslovnem prostoru, če primerjamo bolj »tradicionalne«, s strukturalizmom zaznamovane pristope k jeziku, ki jih povezujemo predvsem z vplivnim Toporišičevim pogledom na jezik, in novejše korpusne pristope k jeziku (npr. Krek 2013; Gantar 2008; Šorli 2020), pri katerih tudi pri nas prepoznamo vplive birminghamske šole.

Birminghamski krog je znan po t. i. popolnem korpusnem pristopu (Tognini-Bonelli 2001) in vrsti precej vplivnih teoretskih smernic in del v korpusnem jezikoslovju. Sinclair (1991; 2004) kot vodilni jezikoslovec tega kroga je izhajal iz stališča, da je treba s pojavom korpusov začeti opazovati jezik na novo, neobremenjeno s predhodnimi teorijami. Opozarjal je na tesno povezavo med slovnico in skladnjo. Vplivni sta dve njegovi načeli, in sicer načelo idiomatskosti, ki: »pomeni, da ima jezikovni uporabnik na voljo veliko količino napol vnaprej sestavljenih fraz, ki predstavljajo zanj eno samo izbiro, čeprav se zdi, da jih lahko ločimo na segmente« (Sinclair 1991: 110; prev. avt.), in načelo proste izbire, ki nastopi, ko načelo idiomatskosti odpove: »To je način gledanja na besedilo kot rezultat velikega števila kompleksnih izbir. Na vsaki točki, kjer je enota zaključena (beseda, fraza ali stavek), se odpre širok nabor izbir in edina omejitev je slovničnost« (Sinclair 1991: 109; prev.

avt.). Nekatera nadaljnja vplivna dela birminghamskega kroga so slovnica vzorcev (Hunston in Francis 2000), osredotočena na iskanje slovničnih vzorcev in njihovo povezovanje z leksikalnimi enotami, s katerimi se pogosto družijo; Hoeyjeva (2005) teorija leksikalnega proženja, ki razlaga, da beseda s tem, ko jo srečujemo v različnih okoliščinah in sobesedilih, postane povezana s sobesedili in okoliščinami, v katerih je bila rabljena, in naše védenje o tej besedi vključuje informacijo, da se sopojavlja z določenimi besedami v določenih vrstah konteksta; Hanksova (2013) teorija konvencij in invencij, ki skuša vzpostaviti sistematično ločevanje med običajnimi, konvencionalnimi vzorci in inventivnimi rabami teh vzorcev; ter področje semantične oz. diskurzne prozodije, ki se nanaša na opažanje, da mnoge besede oz. večbesedne enote izražajo določeno vrsto pomena, ki ga lahko pogosto označimo kot pozitivnega ali negativnega in ki se ustvari skozi sistematično zaporedje kolokacij, njegova osnovna vloga pa je, da izraža avtorjev odnos do določene pragmatične situacije (Louw 2000).

V navedenih delih lahko prepoznamo nekaj pomembnih premikov v pogledu na jezik kot predmet raziskovanja, če gledamo z vidika nekoga, ki se je o jeziku učil skozi »tradicionalno« slovensko, pretežno strukturalistično obarvano jezikoslovno šolo. Ti premiki so bili podrobneje predstavljeni v Verdonik (2015), zato jih tukaj samo povzemamo. Prvi premik se nanaša na predmet raziskovanja. To ni več jezikovni sistem, ampak jezikovna raba: korpusni izpis v obliki konkordance usmeri našo pozornost na besede v sobesedilu. Drugi premik se nanaša na raziskovalni fokus: jezikoslovja ne zanima več, kaj je sistemsko možno, ampak kaj je v jezikovni rabi tipično/običajno. Tretji premik se nanaša na vlogo raziskovalčeve intuicije: nesprejemljivo postane, da bi si jezikoslovec izmišljal primere, ki se zdijo mogoči in verjetni, ter analiziral takšne primere. Jezikoslovčeva intuicija se lahko uporablja samo za kvalitativno analizo korpusnih primerov, ne pa za ustvarjanje primerov za analizo. Četrti premik se nanaša na to, kako gledamo na dve osrednji jezikovni osi: slovar in slovnico. Namesto ločenega raziskovanja in opisovanja ene in druge (v slovarju in slovnici) postane prepletanje obeh ravni (npr. skozi pojme kolokacij, koligacij ipd.) eden osrednjih predmetov jezikoslovnega raziskovanja. Peti, izredno pomemben premik se nanaša na naše dojemanje jezikovnega sistema: jezik ni več dojet kot sistem, ki se realizira v vsakokratnih rabah, ampak je vrsta vsakokratnih rab in vzorci teh rab se nalagajo ter tako sestavljajo le okviren, ohlapen, večno spreminjajoč se, visoko dinamičen sistem. Zadnji premik je morda najmanj izrazit, morda celo bolj kot ne indiciran s strani avtorice tega pregleda (Verdonik 2015). Gre

za premik osrednje enote raziskovanja, ki ni več »jezikovno pravilo«, ampak »jezikovni vzorec« z naslednjimi lastnostmi: gre za neko jezikovno strukturo; pojavlja se bolj pogosto, kot bi bilo naključno; elementi vzorca so medsebojno pomensko povezani; ni stabilen, ampak je variabilen; pojavlja se lahko na različnih jezikovnih ravneh.

3 Jezikovne tehnologije in poučevanje jezika

Računalniki so v poučevanju jezika, zlasti kot tujega jezika, aktivno prisotni več kot 40 let. V zvezi s tem so se uveljavili termini, kot so računalniško podprto učenje jezika (angl. *computer assisted language learning*), podatkovno vodeno učenje jezika (angl. *data-driven learning*) ali mobilno podprto učenje jezika (angl. *mobile-assisted language learning*). Pri učenju jezika kot maternega jezika se kaže manj aktivna vpeljava digitalnih okolij in orodij kot za učenje jezika kot tujega jezika, a je vseeno dokaj široko prisotna. V tem poglavju se osredotočamo na tovrstna okolja in orodja za slovenščino, s poudarkom na korpusnojezikoslovnih in jezikovnotehnoloških pristopih in orodjih.

3.1 E-priročniki in e-okolja za slovenščino

Pregled e-priročnikov in e-okolij za slovenski jezik razporedimo v tri skupine. Za prvo skupino prehoda v digitalni medij je značilno, da je še močno zaznamovana s pristopi iz papirnega medija: e-priročniki so narejeni po istem postopku kot tiskani priročniki, dodani pa so videoposnetki, animacije, interaktivne vaje in drugi elementi, ki jih papirni medij ne podpira (npr. interaktivni delovni zvezki, učbeniki, berila na portalu Lilibi;[2] iUčbeniki[3]). Drugo skupino predstavljajo e-okolja, ki poleg grafike, animacije in zvočnih učinkov izkoriščajo igrifikacijo z virtualnim okoljem, nagrajevanjem in animiranimi junaki (tak tipičen primer je portal UČIMse[4]). Tretjo skupino predstavljajo portali, ki izkoriščajo uporabo jezikovnih tehnologij v didaktične namene. Tak primer za slovenščino je Pedagoški slovnični portal.[5] Ta skupina je predmet našega zanimanja v tem prispevku, zato ji posvetimo nekaj več prostora.

[2] https://www.lilibi.si/
[3] http://eucbeniki.sio.si
[4] https://www.ucimse.com/
[5] http://slovnica.slovenscina.eu/

Uporaba jezikovnih tehnologij je na primeru Pedagoškega slovničnega portala opazna v štirih korakih. V prvem koraku se uporabi korpusni pristop za definiranje učnih vsebin. Tako je bil zbran korpus šolskih besedil in v njem označene napake, ki jih učitelji popravljajo v pisnih izdelkih učencev (Rozman idr. 2020). Na podlagi tega je bila narejena analiza najpogostejših napak pri pisanju (Kosem idr. 2012), ki je bila podlaga za definiranje vsebin portala. V drugem koraku so bili uporabljeni besedilni korpusi in korpusna analiza za dodatno raziskavo jezikovnih pojavov, ki so potem razloženi v učne namene. V tretjem koraku so za vaje izbrani primeri v celoti vzeti iz obstoječih besedilnih korpusov, so torej avtentični, taki, kot jih je nekdo nekje dejansko zapisal. V četrtem koraku se vzpostavlja tudi izobraževanje o obstoječih jezikovnih virih in priročnikih, s pomočjo katerih se spodbuja uporabnike, da se naučijo sami najti odgovore na svoja jezikovna vprašanja, ter spodbuja raba teh priročnikov v šolskem okolju.

3.2 Avtomatizacija vaj in nalog v e-okolju *Slovenščina na dlani*

Slovenščina na dlani je interaktivno učno e-okolje za podporo jezikovnemu pouku slovenščine od 6. razreda osnovne šole naprej. Pokriva štiri vsebinske sklope: pravopis, slovnico, frazeme in pregovore, besedila. Pravopis in slovnica sta namenjena urjenju knjižnih pravopisnih in slovničnih vzorcev, ki šolarjem pri pisanju pogosto povzročajo težave. Pri definiranju tem teh dveh sklopov vaj smo izhajali iz že omenjene analize napak pri pisanju (Kosem idr. 2012), narejene na podlagi korpusa Šolar (Rozman idr. 2020), dodatno pa tudi iz informacij, pridobljenih od učiteljev 14 osnovnih in srednjih šol, sodelujočih v projektu.[6] Sklop frazemov in pregovorov zapolnjuje vrzel v znanju na tem področju. Osnovno- in srednješolski učitelji namreč opažajo, da si učenci oz. dijaki »pogosto napačno interpretirajo frazeme« (Voršič 2018: 91), kar kaže na potrebo po dodatnem spoznavanju slovenskih frazemov in pregovorov ter njihovega pomena in rabe. V sklopu besedil so razvite vaje za razvijanje kritičnega branja, pridobivanje znanja o jezikovni rabi in prvinah besedil kot predstavnikov določenih besedilnih skupin. Brez poznavanja in učinkovite rabe žanrov namreč ne moremo doseči ustreznega nivoja sporazumevalne pismenosti.

[6] http://projekt.slo-na-dlani.si/sl/o-projektu/

V vseh štirih sklopih smo sledili cilju čim večje avtomatizacije. V tem prispevku se osredotočamo na opis avtomatizacije s treh vidikov:

- avtomatiziran priklic ustreznih avtentičnih primerov za posamezno vajo,
- avtomatizirano vrednotenje pravilnosti uporabnikovih rešitev ali odgovorov,
- individualno prilagojeno avtomatizirano vodenje uporabnika skozi e-okolje.

Avtomatiziran priklic primerov nam omogoči, da lahko za posamezno vajo ponudimo veliko število različnih primerov. S tem lahko uporabnik usvoji in utrdi knjižni jezikovni vzorec, ki mu povzroča težave, tudi če morda ne pozna teorije in razlage, po kateri je neki jezikovni vzorec knjižni, drugi pa ne (teoretsko se tak didaktični pristop naslanja na teorijo leksikalnega proženja – Hoey 2005). Da smo lahko zagotovili avtomatski priklic velikega števila primerov k vajam, je bil prvi korak v izdelavi e-okolja *Slovenščina na dlani* izdelava zalednih jezikovnih virov. Za sklopa pravopisa in slovnice smo potrebovali za primere vir povedi: to je bil besedilni korpus MAKS, v katerem smo skušali zbrati besedila, ki so vsebinsko zanimiva za mlade in so bila jezikovno pregledana (Dobrovoljc 2018). Korpus smo tokenizirali in lematizirali ter označili z oblikoslovnimi oznakami, skladenjskimi razmerji in imenskimi entitetami, kar nam je omogočilo usmerjen avtomatski priklic točno takih povedi, ki so vsebovale jezikovne entitete, potrebne za izvajanje določene vaje, npr. dvostavčne povedi z določeno vrsto odvisnika ali prirednega razmerja. Avtomatsko priklicane primere je bilo treba sicer še natančno ročno pregledati in izločati neustrezne primere, vendarle pa smo na tak način definirali v povprečju po 500 primerov za posamezno vajo, dodajanje novih, aktualnih primerov v prihodnosti pa je lahko dokaj hitro in enostavno z uporabo dodatnih besedilnih korpusov. Za frazeme in pregovore je bil vir primerov v projektu izdelani e-slovar frazemov in pregovorov FRIDA (Ulčnik 2019; Ulčnik in Meterc 2019). Slovar opisuje različne nivoje lastnosti za 200 izbranih, didaktično relevantnih frazeoloških in paremioloških enot. Iz tega vira so bile v vaje vključene enote in primeri, s pomočjo katerih uporabniki usvajajo podobo, pomen in rabo frazemov in pregovorov. Vir za avtomatski priklic primerov v sklopu besedil je bila zbirka besedil praktičnega sporazumevanja BERTA (Krajnc Ivič 2018). Ta vključuje več kot 200 avtentičnih pisnih, govorjenih in večmodalnih besedil iz različnih besedilnih skupin in o različnih, za mlade aktualnih tematikah. Tudi ta zbirka je izdelana v e-obliki, za namene avtomatiziranja vaj pa so bili poleg osnovnih podatkov o besedilih, kot so

avtor, čas nastanka, kanal ipd., k vsakemu besedilu dodani tudi bolj specifični podatki, npr. o slogovnem postopku, družbenem razmerju med udeleženci, jezikovni zvrstnosti itd., potrebni za izvajanje vaj, s katerimi uporabniki razvijajo sporazumevalno pismenost. Izdelavi virov je sledil razvoj programskega orodja za priklic primerov (Verdonik idr. 2021: 205–209).

Avtomatizirano preverjanje pravilnosti uporabnikovih rešitev in odgovorov se v e-okolju *Slovenščina na dlani* izvaja za veliko večino vaj in nalog. Večinoma je pravilna rešitev znana že iz zalednih jezikovnih virov: oblika povedi v korpusu MAKS, vnos v FRIDI oz. podatek v BERTI je referenčna vrednost, po kateri se ocenjuje uporabnikova rešitev ali odgovor, zato ročno dodajanje pravilnih rešitev ni bilo potrebno. V redkih primerih je bilo treba upoštevati, da je možnih pravilnih rešitev več in da v zalednem viru zapisana oblika ali odgovor morda ni edini ustrezen. Tak primer so recimo vaje utrjevanja jezikovnih vzorcev s predlogi, pri katerih smo ustrezno prilagodili navodilo in omogočili večkraten vnos, tako da noben uporabnikov odgovor ni ovrednoten kot napačen, samo en (tisti iz korpusa, ki predstavlja zaželen jezikovni vzorec) pa se potrdi kot pravilen. Del nalog pa je takšen, da (popolno) avtomatizirano vrednotenje ni mogoče. Take so na primer naloge, ki imajo različne možne rešitve (npr. razlaga pomena frazema ali pregovora), oz. naloge, vezane na uporabnikovo ustvarjalnost (raba frazemov in pregovorov, pisanje povzetka besedila ali ustvarjanje novega besedila). Pri take vrste nalogah so možne rešitve samovrednotenje (uporabnik sam ovrednoti svoj odgovor na podlagi vpogleda v referenčne rešitve), medvrstniško vrednotenje ali vrednotenje s strani učitelja.

Individualno prilagojeno avtomatizirano vodenje uporabnika skozi e-okolje se nanaša na to, v kakšnem zaporedju se uporabniku pojavljajo vaje. V drugih obstoječih e-okoljih za učenje slovenščine mora uporabnik sam izbirati teme vaj ali pa ga računalnik vodi skozi vaje na linearen način, po vnaprej določenem zaporedju. Tudi v e-okolju *Slovenščina na dlani* je mogoč način uporabe, pri katerem uporabnik sam izbira teme, pri čemer pa jih lahko določa poljubno splošno ali podrobno ter jih neomejeno kombinira. Poleg te možnosti je predviden tudi način, da teme izbira in določa učitelj. Tretji način uporabe pa predvideva, da teme in vaje povsem avtomatizirano izbira računalniški algoritem. V nobenem od teh primerov, tudi ko teme vaj izbere uporabnik sam, pa zaporedje vaj in primerov ni linearno in vnaprej določeno, ampak odvisno od mnogih dejavnikov. Sistem na podlagi statističnih

modelov (kjer upošteva npr. uporabnikovo uspešnost pri preteklih poskusih reševanja) in nekaj vnaprej določenih pravil (npr. vezanih na zahtevnost vaj glede na stopnjo šolanja uporabnika) adaptivno izbira vaje in primere, ki jih rešuje uporabnik. Tak pristop je mogoč in smiseln, ker je na voljo veliko število vaj in veliko število primerov za vsako vajo. Njegova pomembna prednost je, da uporabnik več vadi tisto, kar mu povzroča težave, in manj to, kar že dobro obvlada.

4 Diskusija in zaključek

V prispevku smo se osredotočali na vprašanje digitalizacije pri poučevanju jezika predvsem z vidika korpusnega jezikoslovja in jezikovnih tehnologij. Pregledali smo korpusnojezikoslovne in jezikovnotehnološke začetke in današnje stanje, premislili, kako so se hkrati spreminjali tudi jezikoslovni pogledi na jezik, osrednjo pozornost pa smo namenili vprašanju, kakšne potenciale imajo korpusno jezikoslovje in jezikovnotehnološki pristopi za učenje jezika kot maternega jezika. Pri tem smo se osredotočili predvsem na izkušnje dveh vidnejših tovrstnih projektov za slovenščino, Pedagoškega slovničnega portala in Slovenščine na dlani, z večjim poudarkom na slednjem, saj predstavlja (skupaj z drugimi soavtorji in člani projekta) tudi naše lastno delo. Če ugotovitve združimo, lahko sklenemo, da lahko korpusno jezikoslovje in jezikovne tehnologije uporabimo vsaj na naslednje načine pri snovanju učnih e-okolij za učenje jezika:

- za definiranje učnih vsebin, ki jih e-okolje pokriva,
- za dodatno analizo in raziskavo jezikovnih pojavov, ki jih pojasnjujemo,
- za avtomatski priklic velike količine (potencialnih) primerov, ki jih uporabimo v vajah in nalogah, in (pol)avtomatizirano dodajanje novih primerov prek večanja zalednih jezikovnih virov,
- pri avtomatiziranem vrednotenju pravilnosti uporabnikovih rešitev ali odgovorov, kjer pravilnih rešitev ni treba posebej ročno določati, ampak se prepoznajo iz podatkov v zalednih jezikovnih virih,
- za vključevanje korpusov in drugih e-jezikovnih priročnikov v sicer osebno voden pouk o jeziku,
- za individualno prilagojeno avtomatizirano vodenje uporabnika skozi učno e-okolje.

Seveda pa ta seznam ni dokončen.

Če sklepamo na primeru premikov jezikoslovja v razumevanju jezika, tudi širša vpeljava korpusnega jezikoslovja in jezikovnih tehnologij v pouk jezika lahko prinese premike v miselnosti v zvezi z jezikom. Želeli bi si, da bi bili ti premiki v smer doživljanja jezika kot polja odprtih možnosti za izražanje, kot jih določata načelo idiomatskosti in načelo proste izbire, in ne kot polja jezikovnih pravil, ki omejujejo naše možnosti izražanja. A to ni odvisno samo od tehnoloških možnosti in razvoja, ampak tudi od tistih, ki v tem razvoju sodelujejo oz. ki o jeziku poučujejo.

Čeprav smo se v prispevku ukvarjali z digitalizacijo pri poučevanju jezika zlasti iz specifičnega korpusnojezikoslovnega in jezikovnotehnološkega vidika, pa je v šolskem letu, ki se zaključuje ob pisanju tega prispevka (2020/21), šolstvo pod vplivom izkušnje šolanja na daljavo v času epidemije. Rezultat tega je med drugim tudi množično in obsežno podaljšan čas, ki ga otroci preživijo pred ekrani ne samo v šolske, ampak tudi druge namene (gl. npr. DAK-Studie 2020). V medijih je vse več opozoril o digitalnih zasvojenostih. V luči teh uvajanja digitalnih tehnologij v pouk ne moremo izvajati brez premislekov in tudi pomislekov. Prav aktualna izkušnja namreč kaže, da je za mnoge otroke, ko so enkrat pred ekranom, skušnjava, da pobrskajo še po kakšnih drugih vsebinah, ne samo šolsko predpisanih, prehuda, da bi se ji uprli. Verjetno zato prav nobeno vpeljevanje digitalnih učnih pripomočkov ne more potekati brez sočasnega izobraževanja o digitalni in medijski pismenosti, prek katerih se bodo otroci in mladostniki učili ščititi se pred negativnimi vzorci uporabe in negativnimi vplivi, ki jih doživljajo prek ekranskih tehnologij.

Literatura

Špela ARHAR HOLDT, 2011: *Luščenje besednih zvez iz besedilnega korpusa z uporabo dvodelnih in tridelnih oblikoskladenjskih vzorcev*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Ken BEATTY, 2013: *Teaching and Researching Computer-Assisted Language Learning.* Second edition. London, New York: Routledge (Taylor & Francis Group).

DAK-Studie, 2020: *Mediensucht 2020 – Gaming, Social-Media and Corona.* Dostop 9. 6. 2021 na https://www.dak.de/dak/gesundheit/dak-studie-gaming-social-media-und-corona-2295548.html#.

Simon DOBRIŠEK, Jerneja ŽGANEC GROS, Ivo IPŠIĆ, Karmen PEPELNJAK, France MIHELIČ in Nikola PAVEŠIĆ, 1998: Gopolis: Slovenska Podatkovna Zbirka Govorjenih Poizvedovanj. *Jezikovne tehnologije za slovenski jezik: zbornik konference.* Ur. Tomaž Erjavec in Jerneja Žganec Gros. Ljubljana: Institut Jožef Stefan. 105–108

Kaja DOBROVOLJC, 2018: Kako nastajajo korpusi. *Slovenščina na dlani 1*. Ur. Natalija Ulčnik. Maribor: Univerzitetna založba Univerze. 61–64. Dostop 9. 6. 2021 na

https://press.um.si/index.php/ump/catalog/book/341.

Tomaž ERJAVEC, 1990: Razvoj in opis dvonivojskega modela morfološke analize in sinteze. *Informatica: an international journal of computing and informatics* 14/3, 59–62.

Tomaž ERJAVEC, 1998: The MULTEXT-East Slovene lexicon. *Zbornik sedme Elektrotehniške in računalniške konference ERK '98*. Ur. Baldomir Zajc. Ljubljana: IEEE Region 8, Slovenska sekcija IEEE. 189–192.

Tomaž ERJAVEC, 1999: Slovensko-angleški korpus ELAN. *Slavistična revija* 47/4, 515–522.

Tomaž ERJAVEC, Vojko GORJANC in Marko STABEJ, 1998: Korpus FIDA. *Zbornik Jezikovne tehnologije za slovenski jezik*. Ur. Tomaž Erjavec in Jerneja Gros. Ljubljana: Institut Jožef Stefan. 124–127.

Polona GANTAR, 2008: (Slovenska) leksika med leksikonom in slovnico. *Jezik in slovstvo* 53/5, 19–35.

Patrick HANKS, 2013: *Lexical Analysis: Norms and Exploitations*. Cambridge, London: The MIT Press.

Michale HOEY, 2005: *Lexical Priming: A New Theory of Words and Language*. London, New York: Routledge.

Susan HUNSTON in Gill FRANCIS, 2000: *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam, Philadelphia: John Benjamins.

Bojan IMPERL, Zdravko KAČIČ in Bogomir HORVAT, 1997: A study of harmonic features for the speaker recognition. *Speech communication* 22/4, 385–402.

Primož JAKOPIN, 2003: Nekaj zanimivosti iz besedilnega korpusa Nova beseda. *Jezikoslovni zapiski* 9/2, 145–152.

Zdravko KAČIČ, Bogomir HORVAT in Igor URLEP, 1990: Govorna komunikacija človek-stroj – tehnologija bližnje prihodnosti. *Elektrotehniški vestnik* 57/4 , 274–280.

Zdravko KAČIČ in Bogomir HORVAT, 1994: Zasnova baze izgovarjav slovenskega jezika SNABI. *Zbornik tretje Elektrotehniške in računalniške konference ERK '94*. Ur. Franc Solina in Baldomir Zajc. Ljubljana: Slovenska sekcija IEEE. 327–330.

Iztok KOSEM, Mojca STRITAR, Sara MOŽE, Ana ZWITTER VITEZ in Špela ARHAR HOLDT, 2012: *Analiza jezikovnih težav učencev: korpusni pristop*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Mira KRAJNC IVIČ, 2018: Kdo ali kaj je BERTA. *Slovenščina na dlani 1*. Ur. Natalija Ulčnik. Univerzitetna založba Univerze. 65–72. Dostop 9. 6. 2021 na https://press.um.si/index.php/ump/catalog/book/341.

Simon KREK, 1999: Računalniški korpusi v slovaropisju. *Razgledi: tako rekoč intelektualni tabloid 13 (*23. jun. 1999), 8–9.

Simon KREK, 2013: Korpusne metode in njihov odsev v jezikoslovnih teorijah 20. stoletja. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave* 1/1, 4–23.

Nataša LOGAR, 2013: *Korpusna terminografija: Primer odnosov z javnostmi*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Nataša LOGAR BERGINC, Miha GRČAR, Marko BRAKUS, Tomaž ERJAVEC, Špela ARHAR HOLDT in Simon KREK, 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: Gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Bill LOUW, 2000: Contextual prosodic theory: Bringing semantic prosodies to life. *Words in Context: A tribute to John Sinclair on his Retirement*. Ur. C. Heffer, H. Sauntson. Birmingham: University of Birmingham. 48–94.

France MIHELIČ, Ivo IPŠIĆ, Simon DOBRIŠEK in Nikola PAVEŠIĆ, 1992: Feature representations and classification procedures for Slovene phoneme recognition. *Pattern recognition letters: an official publication of the International Association for Pattern Recognition* 13/12, 879–891.

Nikola PAVEŠIĆ, Jerneja ŽGANEC GROS, France MIHELIČ in Simon DOBRIŠEK, 1996: Text-to-speech synthesis for the Slovenian language: an overview. *Conference on software in telecommunications and computer networks, SoftCOM '96*. Split: Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture: Croatian Post and Telecommunications. 99–109.

Tadeja ROZMAN, Irena KRAPŠ VODOPIVEC, Mojca STRITAR in Iztok KOSEM, 2020: *Empirični pogled na pouk slovenskega jezika*. Ljubljana: Znanstvena založba FF UL. Dostop 9. 6. 2021 na https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/227/327/5303-1.

Mirjam SEPESY MAUČEC, 1997: Statistično jezikovno modeliranje pri razpoznavanju govora. *Elektrotehniški vestnik* 64/2–3, 123–128.

John SINCLAIR, 1991: *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

John SINCLAIR, 2004: *Trust the Text: Language, Corpus and Discourse*. London, New York: Routledge.

Marko STABEJ in Primož VITEZ, 2000: KGB (korpus govorjenih besedil) v slovenščini. *Informacijska družba IS'2000: zbornik 3. mednarodne multi-konference*. Ur. Cene Bavec idr. Ljubljana: Institut Jožef Stefan.

Mojca ŠORLI, 2020: *Semantična prozodija : leksikalni in besedilno-diskurzivni vidiki*. Ljubljana: Založba ZRC.

Elena TOGNINI-BONELLI, 2001: *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Natalija ULČNIK, 2019: Izbor frazemov za bazo FRIDA. *Slovenščina na dlani 2*. Ur. Natalija Ulčnik. Maribor: Univerzitetna založba Univerze. 37–45. Dostop 9. 6. 2021 na https://press.um.si/index.php/ump/catalog/book/447.

Natalija ULČNIK in Matej METERC, 2019: Izbor pregovorov za bazo FRIDA. *Slovenščina na dlani 2*. Ur. Natalija Ulčnik. Maribor: Univerzitetna založba Univerze. 47–55. Dostop 9. 6. 2021 na https://press.um.si/index.php/ump/catalog/book/447.

Darinka VERDONIK, 2015: Jezikovnoteoretska načela v korpusnem jezikoslovju. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave* 3/1, 1–27. Dostop 9. 6. 2021 na http://www.trojina.org/slovenscina2.0/arhiv/2015/1/Slo2.0_2015_1_01.pdf, https://dk.um.si/IzpisGradiva.php?id=67123.

Darinka VERDONIK, Matej ROJC, Zdravko KAČIČ in Bogomir HORVAT, 2002: Zasnova in izgradnja oblikoslovnega in glasovnega slovarja za slovenski knjižni jezik. *Jezikovne tehnologije: zbornik konference*. Ur. Tomaž Erjavec in Jerneja Žganec Gros. Ljubljana: Institut Jožef Stefan. 44-48.

Darinka VERDONIK, Simona MAJHENIČ, Špela ANTLOGA, Sandi MAJNINGER, Marko FERME, Kaja DOBROVOLJC, Simona PULKO, Mira KRAJNC IVIČ in Natalija ULČNIK, 2021: Učno e-okolje Slovenščina na dlani: izzivi in rešitve. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave* 9/1, 181–215.

Špela VINTAR (ur.), 2010: *Slovenske korpusne raziskave*. Ljubljana: Znanstvena založba Filozofske fakultete.

Ines VORŠIČ, 2018: Prvi odzivi učiteljic in učiteljev. *Slovenščina na dlani 1*. Ur. Natalija Ulčnik. Maribor: Univerzitetna založba Univerze. 89–91. Dostop 9. 6. 2021 na http://press.um.si/index.php/ump/catalog/book/341.

Jana ZEMLJARIČ MIKLAVČIČ, 2008: *Govorni korpusi*. Ljubljana: Znanstvena založba Filozofske fakultete.