JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino

Darja Fišer,* Tomaž Erjavec,† Ana Zwitter Vitez, *‡ Nikola Ljubešić₩†

* Oddelek za prevajalstvo, Filozofska Fakulteta
Aškerčeva 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si
‡ Trojina, Zavod za uporabno slovenistiko
Dunajska 116, 1000 Ljubljana
ana.zwitter@guest.arnes.si
₩ Odsek za informacijske znanosti, Fakulteta za humanistične in družbene vede, Univerza v Zagrebu
Ivana Lučića 3, HR-10000 Zagreb
nikola.ljubesic@ffzg.hr

Povzetek

V prispevku predstavljamo vire, orodja in metodologijo, ki jih razvijamo za analizo nestandardne pisne spletne slovenščine. Ti so nujni za izdelavo sodobnih leksikografskih, normativnih in pedagoških priročnikov, ki brez podatkov o dejanski jezikovni rabi ni mogoča. Jezikovne modele, ki so dovolj robustni za obdelavo nestandardne pisne slovenščine, potrebujemo tudi za procesiranje spletnih besedil. Opisujemo gradnjo obsežnega korpusa pisne spletne slovenščine, izdelavo slovarja nestandardnih besed, tipičnih za pisno spletno komunikacijo, vrsto jezikoslovnih raziskav in razvoj metod za izboljšanje avtomatskega procesiranja nestandardne pisne spletne slovenščine. Razviti jezikovni viri bodo, primerno anonimizirani, ponujeni v odprt dostop pod licenco Creative Commons. Tako bodo omogočili prenos znanj na vsa področja, ki uporabljajo spletne vsebine, ki jih ustvarjajo uporabniki.

The JANES Project: methods, tools and resources for nonstandard Slovene

The paper presents an infrastructure and methodology under development for the analysis of user-generated content written in non-standard Slovene. They are indispensible in contemporary lexicographic, normative and pedagogic work, which cannot be comprehensive without information about real language use. Robust language models that can deal with nonstandard written Slovene are also needed for automatic text processing. A large and representative corpus of publicly available user-generated content and a web dictionary of non-standard Slovene will be compiled, comprehensive linguistic analyses will be performed and methods for automatic processing of non-standard text will be developed. The developed resources will be suitably anonymised and made openly available for download under the Creative Commons license. The developed resources, tools and methods will thus enable the transfer of knowledge to R&D in language technologies, lexicographic work and linguistic research.

1. Uvod

V času, ko računalniško posredovana komunikacija (ang. *computer-mediated communication*) in količina spletnih vsebin, ki jih na blogih in družbenih omrežjih ustvarjajo uporabniki, tako strmo naraščata, da je 90% tovrstnih besedil nastalo samo v zadnjih dveh letih (IBM 2013), postajajo njihove vsebine vse pomembnejši vir človeškega znanja in mnenj. Posledično se je povečala potreba po poznavanju in razumevanju t.i. internetnega jezika (*netspeak*), v katerem so te vsebine ustvarjene. Pisno spletno komunikacijo določajo okoliščine, kot so (ne)interaktivnost, (a)sinhronost, fizična (ne)prisotnost sogovornika in drugi situacijski dejavniki (Noblia 1998). Bolj kot je izbrana oblika komuniciranja interaktivna, poteka v realnem času in ima na drugi strani prisotnega sogovornika, več prvin spontanega govorjenega jezika vsebuje, vključno s (za računalniško komunikacijo prilagojenimi) paralingvističnimi in prozodičnimi elementi (Crystal 2001).

Za jezik pisne spletne komunikacije je značilna pogosta raba nestandardnih jezikovnih oblik, kot je nestandarden (bolj fonetičen) zapis besed, (npr. izključno male tiskane črke, opuščanje večine ločil in večkratno ponavljanje črk za čustveno poudarjanje zapisane izjave), in pogoste specifične okrajšave. Zaradi tega je jezikoslovna analiza in posledično tudi avtomatska obdelava tovrstnih vsebin otežena (Sproat idr. 2001), prizadevanja za premostitev teh ovir pa so trenutno ena bolj vročih tem na področju računalniškega jezikoslovja.

V sodobnem jezikoslovju so paradigme, ki na rabo nestandardnih jezikovnih različic v internetni pisni komunikaciji gledajo kot na odraz nepopolnosti ali osiromašenosti komunikacijskih zmožnosti, preživete, saj številne analize jezikovne rabe na internetu demonstrirajo sposobnost uporabnikov, da se prilagodijo računalniškemu mediju oziroma da zmožnosti medija izrabijo za zadovoljevanje svojih komunikacijskih potreb (glej npr. Tagg 2012), da si prizadevajo skrajšati in poenostaviti pisanje, predvsem pa da pisanje približajo svoji identiteti in govoru (Herring 2001). Večkrat je bilo dokazano tudi, da izpostavljenost nestandardnemu jeziku in njegova pogosta raba ne zmanjšujeta jezikovne zmožnosti (npr. Baron 2010).

Razkoraka med živostjo jezika in statičnostjo njegovega opisa ter iz tega izhajajočo nujno potrebo po raziskavah nestandardnega jezika se zavedajo tudi nekateri vodilni slovenski jezikoslovci, ki so že proučevali jezik SMS sporočil, spletnih forumov in elektronske pošte (npr. Logar 2003, Kalin Golob 2008, Dobrovoljc 2008, Jakop 2008, Michelizza 2008), kljub vsemu pa so tovrstne študije pri nas še vedno na obrobju interesa jezikoslovcev, zaradi česar je slovenski jezikoslovni prostor s tovrstnimi raziskavami izrazito podhranjen.

Zaenkrat se tudi najsodobnejša jezikovnotehnološka orodja, ki jih uporabljamo za procesiranje besedil, zelo slabo spopadajo z elementi nestandardnega jezika. Z njim imajo težave že povsem temeljna orodja, kot so na primer oblikoslovni označevalniki. Stanford tagger, eden najboljših označevalnikov na svetu, na standardnih angleških besedilih dosega 97 % natančnost, pri označevanju tvitov pa le 85 % (Gimpel idr. 2011).

Zato je cilj predstavljenih raziskav zapolniti eno največjih vrzeli slovenskega jezikoslovja: pomanjkanje virov, orodij in metodologij za jezik, ki se vedno bolj uporablja v vsakodnevni pisni komunikaciji in ki ga ustvarjajo vsi govorci slovenščine, ne zgolj novinarji, prevajalci, pisatelji ipd.

Tudi razvoj računalniškega jezikoslovja je odvisen od dostopnosti jezikovnih virov in orodij za obdelavo nestandardnega jezika. Pričakovani rezultati presegajo znanstveno relevantnost, saj bodo omogočili tudi razvoj najrazličnejših spletnih servisov in mobilnih aplikacij za slovenščino. Ti bodo imeli neposreden vpliv na zmanjševanje e-izključenosti govorcev slovenščine, ki trenutno iz pragmatičnih razlogov posegajo po tujejezičnih spletnih in mobilnih aplikacijah.

V nadaljevanju prispevka predstavljamo vire, ki jih bomo zgradili (razdelek 2), korpusnojezikoslovne analize, ki jih bomo opravili (razdelek 3), in orodja za računalniško obdelavo spletnih besedil, ki jih bomo razvili (razdelek 4). Prispevek sklenemo z razmislekom o razsežnostih in pomenu rezultatov raziskav za slovensko jezikoslovje in družbo.

2. Razvoj virov za proučevanje nestandardne pisne spletne slovenščine

Zgradili bomo reprezentativen korpus spletnih besedil, tipično zapisanih v nestandardnem jeziku, ki bo vseboval vsaj 20 milijonov pojavnic. Zajem besedil bo potekal avtomatsko, za kar bomo razvili namenska orodja. Osredotočili se bomo na javno objavljene pisne spletne vsebine in besedilne vrste, ki so tako po količini kot vplivu med najpomembnejšimi predstavniki nestandardnega jezika in zato najbolj relevantni za jezikoslovne raziskave. V korpus bodo vključeni:

- tviti (50 %)
- blogi (30 %)
- sporočila na forumih (10 %)
- komentarji na novice (5 %)
- komentarji na slovenski Wikipediji (5 %)

Z metodo za izbor bomo skušali zaobjeti čim bolj realno podobo tega dela slovenskega svetovnega spleta, da bo izdelan korpus zanj reprezentativen. Korpus bo vseboval natančne oznake besedilnih zvrsti, zato jih bo glede na konkretne raziskovalne potrebe mogoče proučevati tudi individualno in jih primerjati med seboj ter z drugimi korpusi.

2.1 Zajem besedil

Pri zajemu tvitov bomo razvili metodo, s katero bomo identificirali čim več slovenskih uporabnikov in sledili njihovo besedilno produkcijo. To bomo izvedli v naslednjih korakih:

- izdelava seznama visokofrekventnih polnopomenskih slovenskih besed, ki se ne pojavljajo v drugih jezikih,
- zajem množice tvitov s slovenskimi besedami,
- identifikacija dodatnih avtorjev in njihovih tvitov s pomočjo seznama sledilcev,
- razvoj natančnejših metod za identifikacijo slovenščine in izločanje tujih jezikov.

Za zajem blogov, forumov in komentarjev bomo nadgradili metodo gradnje splošnega spletnega korpusa slovenskih besedil (Ljubešić in Erjavec 2011). Fokusirano bomo pajkali samo domene, ki so bogate s temi tremi zvrstmi besedil. Pajkanje bo upoštevalo ime domene z ročno izdelanim seznamom bolj znanih jezikovnozvrstno specifičnih domen in s pomočjo spremljanja agregatorjev slovenskih blogov.

Enciklopedija Wikipedija je odprtodostopna, tako da je mogoče prevzeti celotno bazo, kar bo močno olajšalo identifikacijo komentarjev na posamezne strani in njihovo nadaljnje procesiranje.

2.2 Obdelava besedil

Avtomatsko zajeta besedila s spleta vsebujejo precejšnjo mero šuma (tudi do 80 %), kot je posredovanje nejezikovnih sporočil (fotografije, hiperpovezave ipd.), ki ga je treba odstraniti, da dobimo uporaben korpus besedil. Viri nestandardnega pisnega jezika na spletu pogosto vsebujejo mešanico slovenskih in tujih besed in črk, besedila so velikokrat napisana brez uporabe šumnikov, predvsem pa so posamezna besedila lahko zelo kratka, kar vse oteži delo programom za detekcijo jezika in njegovo označevanje.

Uporabnost korpusa je mnogo večja, če so besedila v njem jezikoslovno označena. Za slovenščino so bila zaenkrat razvita predvsem orodja za označevanje standardnega jezika, in sicer ToTaLe (Erjavec idr. 2005) in Obeliks (Grčar idr. 2012) za oblikoskladenjsko označevanje in lematizacijo, program DependencyParser (Dobrovoljc 2012) za skladenjsko analizo ter slNER (Štajner idr. 2013) in StandfordNER s slovenskim modelom (Ljubešić idr. 2013) za prepoznavanje imenskih entitet. Vendar predvidevamo, da bodo za obdelavo nestandardne pisne spletne slovenščine potrebne številne prilagoditve. Zato bomo avtomatsko zajeta besedila obdelali v naslednjih korakih:

1. čiščenje in deduplikacija spletnih vsebin (Ljubešić in Erjavec 2011),
2. identifikacija jezika prek seznama visokofrekventnih polnopomenskih slovenskih besed,
3. identifikacija preklapljanja med različnimi jeziki s statističnimi metodami,
4. identifikacija in poenotenje metapodatkov,
5. pretvorba v format XML po priporočilih TEI P5,
6. jezikoslovno označevanje (tokenizacija, oblikoskladenjsko označevanje, lematizacija, identifikacija imenskih entitet).

V tej fazi bomo z ročno evalvacijo identificirali najpogostejše napake obstoječih orodij za obdelavo standardnega pisnega jezika. Te napake bomo kasneje odpravili z izdelavo leksikona najbolj pogostih nestandardnih besed.

2.3 Izdelava spremljevalnega korpusa

Za splet je značilna velika dinamika produciranja besedil in hitro spreminjajoče se besedišče. Zato bomo vzpostavili prototipni sistem, ki bo sproti zajemal nove vsebine, jih občasno pretvoril, označil, indeksiral in ponudil v uporabo skozi konkordančnik. S tem bomo vzpostavili prvi slovenski spremljevalni korpus, ki bo omogočal sprotno spremljanje pisne spletne slovenščine ter zaznavanje novosti in sprememb na ravni leksike (neologizmi, naraščanje in upadanje rabe, vključevanje tujejezičnih prvin).

Taki postopki so že vključeni v delovni proces pri najsodobnejših leksikografskih projektih v tujini (Atkins idr. 2010), zanimivi pa so tudi za druge raziskave, ki se nanašajo na (ne)ustaljenost variant zapisa besed skozi čas, prilagajanje sloga in registra uporabnikov, spreminjanje diskurzivnih praks ipd. To je pri tako mladem in hitro razvijajočem se mediju zelo pomembno, saj se po eni strani uporabniki šele privajajo nanj, po drugi pa z razvojem tehnologije medij ponuja vedno nove funkcionalnosti, ki vplivajo tudi na rabo jezika.

3. Korpusna analiza nestandardne pisne spletne slovenščine

Poleg gradnje virov in orodij za nestandardno slovenščino je zelo pomembno osvetliti tudi rabo pisnega jezika na spletu iz različnih zornih kotov. Posebej se bomo posvetili sedmim jezikoslovnih raziskavam, ki bodo vsaka s svojega zornega kota osvetlile rabo pisne slovenščine na spletu. Rezultati raziskav bodo neposredno uporabni že pri razvoju orodij za računalniško obdelavo besedil, koristen pa bo tudi za številne druge jezikoslovne raziskave in jezikovnotehnološke aplikacije.

3.1 Primerjalna raziskava s pisnim standardom

Zbran in označen korpus spletne slovenščine bomo primerjali z referenčnim in uravnoteženim korpusom sodobnega slovenskega jezika KRES (Logar Berginc idr. 2012) s 100 milijoni pojavnic. Zasnovo za analizo smo na korpusu tvitov pripravili že v pilotni študiji (Erjavec in Fišer 2013), ki jo bomo sedaj razširili na celoten obseg korpusa JANES, ki poleg posodobljenega nabora tvitov vsebuje tudi štiri druge pomembne spletne besedilne zvrsti. V analizi bomo preučili:

- posebnosti zapisa (vzporedna raba večjega števila različic zapisa iste besede, npr. *itak/itaq, lahko noč/ln, počitniceeee*),
- leksikalne značilnosti (z metodo frekvenčnega profila (Rayson in Garside 2000), ki omogoča zaznavanje neologizmov in besedišča, ki je najbolj specifično za enega od korpusov),
- skladenjske značilnosti (kompleksnost povedi, besedni red in raba sklonov).

3.2 Primerjalna raziskava z govorom

Besedila tvitov, forumov in komentarjev nastajajo v okoliščinah, močno podobnih tistim, ki zaznamujejo govorno jezikovno produkcijo: avtor besedilo formulira kot neposreden odziv na družbeno dogajanje znotraj tesnih časovnih omejitev, poleg tega pa s strani naslovnikov pričakuje neposreden odziv na svoje jezikovno udejstvovanje. Govorne prvine nestandardne slovenščine bomo raziskali v primerjavi s korpusom govorjene slovenščine Gos (Verdonik, Zwitter Vitez 2011), pri tem pa bomo pozorni na:

- oblikoslovne posebnosti (primerjava deležev posameznih besednih vrst in analiza različic, ki se uporabljajo za eno standardno obliko),
- skladenjsko kompleksnost govornih in pisnih enot, stalne besedne zveze, besedni red in rabo sklonov,
- leksikalne specifike posameznih spletnih in govorjenih zvrsti,
- značilnosti avtorjev različnih profilov na podlagi označenih metapodatkov korpusa Gos, ki vsebuje podatke o spolu, starosti, izobrazbi in geografski pripadnosti.

3.3 Kolokacije v nestandardni pisni spletni slovenščini

Do razlik med standardno in nestandardno jezikovno rabo pogosto prihaja tudi na ravni kolokacij. Tovrstna razhajanja (npr. *Kaj dogaja! / Ful dogaja! / Tebi pa dogaja!* ipd.) so za jezikoslovje pomembna, ker je vezljivost v jeziku navadno relativno stabilna in zato spremembe v tem segmentu nakazujejo smer razvoja jezika. Če se dovolj ustalijo, sčasoma lahko postanejo del standarda (npr. *rabiti* v pomenu *potrebovati: Ne rabiš nobene dodatne opreme.*). Luščenje kolokacij bomo izvedli z naslednjima postopkoma:

- identifikacija nadpovprečno pogoste sopojavitve besed glede na njihovo siceršnjo frekvenco v korpusu v orodju Sketch Engine (Kilgarriff idr. 2010) na podlagi vnaprej pripravljenih leksikogramatičnih vzorcev za slovenščino (Krek in Kilgarriff 2006),
- analiza napak pri ekstrakciji kolokacij in prilagoditev orodja CollTerm, ki smo ga razvili v prejšnjih raziskavah (Pinnis idr. 2012).

3.4 Terminologija v nestandardni pisni spletni slovenščini

Številni blogi in forumi obravnavajo zelo specifično tematiko (npr. *medicina*), zato bo v njih pogosta tudi raba terminologije. Podobno velja za določene komentarje o urejanju specifičnih gesel na Wikipediji in tematsko specifične uporabniške račune na Twitterju. Ker gre v številnih primerih za neformalni sporočanjski položaj, lahko pričakujemo, da se bo raba terminologije razlikovala od tiste, ki je uporabljena v bolj formalnih registrih. Zato nas bosta pri analizi zanimali raba terminoloških dvojnic (npr. *slikovna pika-piksl*) in stopnja razhajanja nestandardne terminologije od standardne (npr. *HTML format* namesto *format HTML*).

S temeljito analizo terminologije nestandardnega pisnega jezika na spletu bomo izboljšali avtomatski luščilnik terminov LUIZ (Vintar 2010) in ga prilagodili tudi za luščenje terminologije s spletnih besedil za tri različne izbrane domene: medicino, računalništvo in gastronomijo.

3.5 Analiza pomenskih premikov v nestandardni pisni slovenščini

V jeziku se stalno razvija in spreminja tudi pomen že obstoječih besed. Detekcija novih pomenov je velik in pomemben izziv za leksikografijo in posodabljanje slovarskih gesel. Spletne publikacije, blogi in družbena omrežja pa so zaradi množične priljubljenosti in živahne jezikovne rabe idealen vir tovrstnih informacij. Aktualen popis semantičnega inventarja potrebujejo tudi različne jezikovnotehnološke aplikacije, kot sta npr. odgovarjanje na vprašanja in strojno prevajanje.

V ta namen bomo razvili algoritem, ki na podlagi vnaprej določenih pomenov iz semantičnega leksikona sloWNet (Fišer idr. 2012) v skladu z načeli distribucijske semantike v korpusu detektira tiste pojavitve določene besede, ki glede na sobesedilo ni dovolj podobna nobenemu od že obstoječih pomenov v sloWNetu. Seznam kandidatov bomo nato ročno pregledali in z morebitnimi novimi zaznanimi pomeni sloWNet tudi razširili.

3.6 Prepoznavanje žaljivega govora na spletu

V korpusu spletnih besedil bomo identificirali elemente žaljivega govora, saj anonimnost "omogoča posameznikom, da se obnašajo na načine, ki so zelo različni od njihovega vsakdanjega predstavljanja v vsakdanjem svetu" (Praprotnik 2003). Rezultati raziskave bodo zanimivi za institucije, odgovorne za zagotavljanje kulture dialoga (varuh človekovih pravic, spletni portali novinarskih hiš, družbena omrežja ipd.). Elemente žaljivega govora bomo identificirali v naslednjih korakih:

- izdelava manjšega učnega korpusa besedil, ki so jih uporabniki zaznali kot neprimerne, žaljive ali sovražne,
- označevanje eksplicitnih elementov žaljivega govora,
- identifikacija načel odklona žaljivih elementov od standarda (npr. *hebite se, čfeurji* ipd.),
- opredelitev značilk za avtomatsko zaznavanje potencialno žaljivih segmentov v celotnem korpusu spletnih besedil.

3.7 Izdelava slovarja nestandardne pisne spletne slovenščine

Poleg korpusa spletnih besedil in spremljevalnega korpusa bomo izdelali tudi leksikalno bazo, ki nam bo služila kot osnova za izdelavo spletnega slovarja. Ta pomemben jezikovni vir bo vseboval gesla, tipična in specifična za nestandardno jezikovno rabo v pisnih besedilih na spletu. Povezan bo tudi z drugimi viri, ki omogočajo uvid v lemo, obliko ali varianto skozi konkordančnik, druge spletne slovarje, kot je npr. SSKJ ali iskalnike najdi.si in Google.

Izdelan slovar bo uporaben za učitelje in učence, prevajalce ter zainteresirano javnost, pa tudi kot vir informacij o nestandardni leksiki za izdelavo novega slovarja slovenskega jezika.

4. Orodja za računalniško obdelavo

Priprava virov in prilagajanje orodij za avtomatsko obdelavo spletnih besedil bo temeljil na izkušnjah izdelave virov za starejši slovenski jezik (Erjavec 2012), kar mdr. predvideva cikličen pristop k izdelavi in izboljšavi orodij, v katerem ročno preverjeni podatki služijo za izboljšanje avtomatskega označevanje, to pa omogoča kvalitetnejšo osnovo za nadaljnje ročno označevanje.

4.1 Izdelava ročno označenega učnega podkorpusa

Za vse jezikovnotehnološke raziskave je zelo koristen ročno označeni korpus, saj služi kot učna množica za induktivno generiranje jezikoslovnih modelov, uporaben pa je tudi kot testna množica, na kateri je moč ovrednotiti kvaliteto razvitih avtomatskih postopkov za jezikoslovna označevanja. Tak korpus je koristen tudi za jezikoslovce, saj se na oznake v njem lahko bolj zanesejo kot na avtomatske.

Iz zajetega korpusa bomo po vnaprej določenih kriterijih za reprezentativnost in uravnoteženost vzorčili posamezna besedila oz. dele daljših besedil, dobljeni podkorpus avtomatsko označili in s tem dobili osnovo za izdelavo zlatega standarda, pri čemer je predvidena velikost korpusa 100.000 pojavnic. Vsaka besedna oblika v korpusu bo označena s svojo ustreznico in lemo iz standardnega jezika, z oblikoslovno oznako in predvidoma tudi s površinskoskladenjsko odvisnostno povezavo.

4.2 Izdelava učnega leksikona

Na podlagi primerjave besedišča iz referenčnega korpusa KRES in izdelanega korpusa bomo izdelali učni leksikon nestandardne pisne spletne slovenščine, ki bo vsebovala vsaj 1.000 gesel in 10.000 besednih oblik. Leksikon bo vseboval gesla, definirana s standardno zapisano lemo (osnovno obliko), besedno vrsto in, kjer knjižni jezik nima ustreznice, najbližje knjižne sinonime ali razlago. Geslo bo vsebovalo vse identificirane besedne oblike te leme in zglede iz korpusa, opremljene z metapodatki.

Tako kot korpus bo tudi leksikon zapisan v XML / TEI, kar pomeni, da ga bo možno povezati ali po želji vključiti v druge leksikalne vire, kot npr. v novi slovar

sodobnega slovenskega jezika. Takšnega leksikona tudi ni težko pretvoriti v leksikon za raznovrstne jezikovnotehnološke aplikacije.

4.3 Prilagajanje jezikoslovnega označevanja

Ker vsebujejo spletne vsebine, ki jih ustvarjajo uporabniki, jezik, ki se razlikuje od standardnega, je točnost označevanja standardnih jezikoslovnotehnoloških orodij tu bistveno slabša. To se je pokazalo tudi v naši preliminarni raziskavi označevanja tvitov (Erjavec in Fišer 2013), v kateri je bil delež napak v tokenizaciji, oblikoskladenjskem označevanju in lematizaciji občutno višji kot za standardno slovenščino, saj jih je bilo med pregledanimi 500 lemami 22 % napačnih, medtem ko je za korpus ccKRES bilo napačnih samo 4 %. Največ težav je bilo z lematizacijo pogovorno zapisanih besed (npr. *jst, js, nism*), s tokenizacijo emotikonov, ki jih program obravnava kot ločene pojavnice (npr. *:d, :p*) in označevanjem dvoumnih besed, kot so pridevniki in prislovi (npr. *oblačno*).

Naš cilj je prilagoditi obstoječe metode in tehnologije, da bodo sposobne obdelovati tudi nestandardni pisni jezik. Izboljšanje bomo dosegli na podlagi izdelanih jezikovnih virov (ročno označenega učnega podkorpusa in leksikona) in izsledkov ročne evalvacije avtomatsko pripisanih oznak.

Kot osnova nam bodo služile metode, ki smo jih že razvili za avtomatsko označevanje starejših besedil (Erjavec 2013), kjer po (prilagojeni) tokenizaciji posodobimo besedne pojavnice, nato pa nad tako normaliziranim besedilom uporabimo standardne modele za oblikoskladenjsko označevanje in lematizacijo. Posodabljanje besed je potekalo s pomočjo učnega leksikona, za neznane besede pa na podlagi ročno napisanih pravil transkripcije, ki se jih izvaja v formalizmu končnih avtomatov.

Vendar smo v novejših raziskavah (Scherer in Erjavec 2013) dosegli boljše rezultate z metodo transkripcije, ki temelji na statističnem strojnem prevajanju. Metoda, ki v primeru standardizacije besed kot enoto ne uporablja besed, temveč posamezne črke, se nauči modela preslikav iz parov *nestandardna beseda : standardna beseda*, ki jih bomo zajeli iz leksikona.

Nad besedili s standardizirano leksiko lahko uporabimo modele za standardno slovenščino in že s tem izboljšamo oblikoskladenjsko označevanje, lematizacijo in skladenjsko označevanje. Vendar ima trenutni pristop dve slabosti: posamezne besede standardizira neodvisno od konteksta, te pa so lahko dvoumne (npr. *jest*, v *jaz* oz. *jesti*), po drugi strani pa pri nestandardnem jeziku ne prihaja do razlik samo v leksiki, temveč tudi v skladnji. Zato bomo raziskali tudi druge, kompleksnejše metode označevanja, kjer označevalnike dodatno učimo tudi na ročno označenem podkorpusu ali pozamezne korake označevanja na različne načine povezujemo.

5 Zaključek

V prispevku smo predstavili metode, vire in orodja za analizo nestandardne pisne spletne slovenščine, ki jih razvijamo v okviru nacionalnega projekta JANES. Rezultati projekta bodo korpus pisne spletne slovenščine z več kot 20 milijoni pojavnic, slovar nestandardne spletne slovenščine, korpusno podprt jezikovni opis pisne spletne slovenščine na ortografski, oblikoslovni, leksikalni, pomenski in skladenjski ravni ter viri in metode za izboljšanje avtomatskega procesiranja nestandardne slovenščine.

Korpus in slovar bosta omogočila sodobnejšo in celovitejšo izdelavo empirično zasnovanih leksikografskih, normativnih in pedagoških priročnikov, s čimer želimo prispevati k dvigovanju samozavesti govorcev pri uporabi slovenščine. Metode za procesiranje nestandardne slovenščine bodo po eni strani olajšale prodor empiričnih raziskav v jezikoslovje, prav tako pa bodo olajšale postopke poizvedovanja po informacijah, rudarjenja po besedilih in povzemanja besedil, kar bo govorcem slovenščine omogočilo dostop do produktov, ki bolje podpirajo slovenski jezik.

Posredni doprinos raziskave predstavlja jezikovno neodvisna metodologija izgradnje virov in orodij, zaradi katere bodo pristopi neposredno uporabni tudi za sorodne jezike, kot so hrvaščina, srbščina in bosanščina, ki tovrstnih virov in orodij še nimajo razvitih.

Ob upoštevanju varovanja osebnih podatkov bodo izdelani viri ponujeni v odprt dostop pod licenco Creative Commons (CC BY-SA – Priznanje avtorstva, deljenje po enakimi pogoji). Izdelane vire in orodja bomo prenesli na slovensko raziskovalno infrastrukturo jeziko(slo)vnih podatkov in servisov za raziskave v humanistiki in družbenih vedah CLARIN.SI[1], kjer se bodo tudi trajno vzdrževali.

Literatura

B. T. S. Atkins, A. Kilgarriff, M. Rundell. 2010. Database of Analysed Texts of English (DANTE): the NEID database project. *Zbornik konference Euralex*.

N. S. Baron. 2010. *Always On: Language in an Online and Mobile World.* Oxford University Press.

D. Crystal. 2001. *Language and the Internet.* Cambridge University Press.

H. Dobrovoljc. 2008. Jezik v e-poštnih sporočilih in vprašanja sodobne normativistike. *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19*.

K. Dobrovoljc, S. Krek, J. Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. *Zbornik Osme konference Jezikovne tehnologije,* str. 42-47.

T. Erjavec. 2013. Posodabljanje starejše slovenščine. *Uporabna informatika*, 21/4, str. 186-195.

T. Erjavec, C. Ignat, B. Pouliquen, R. Steinberger. 2005. Massive multi lingual corpus compilation : acquis communautaire and totale. *Archives of Control Sciences* 15, str. 529-540.

T. Erjavec. D. Fišer. 2013. Jezik slovenskih tvitov: korpusna raziskava. *Družbena funkcijskost jezika: (vidiki, merila, opredelitve). Obdobja 33*, str. 109-116.

[1] Common Language Resources and Technology Infrastructure: http://clarin.eu/

T. Erjavec. 2012. Jezikovni viri starejše slovenščine IMP : zbirka besedil, korpus, slovar. *Zbornik Osme konference Jezikovne tehnologije*, str. 52-56.

D. Fišer, T. Erjavec, J. Novak. 2012. sloWNet 3.0: development, extension and cleaning. *Zbornik konference Global Wordnet Conference*, str. 113-117.

K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, N. A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments *Zbornik konference Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, str. 42–47.

M. Grčar S. Krek, K. Dobrovoljc. 2012. Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene. *Zbornik Osme konference Jezikovne tehnologije*, str. 42–47.

S. C. Herring. 2001. Computer-Mediated Discourse. *The Handbook of Discourse Analysis*. Oxford: Blackwell Publishers, str. 612–634.

IBM (2013) http://www.ibm.com/smarterplanet/us/en/business_analytics/article/it_business_intelligence.html [22.03.2014]

N. Jakop. 2008. Pravopis in spletni forumi – kva dogaja?. *Slovenščina med kulturami. Zbornik Slavističnega društva Slovenije 19*, str. 315–327.

M. Kalin Golob. 2008. SMS-sporočila treh generacij. *Slovenščina med kulturami. Zbornik slavističnega društva Slovenije 19*, str. 283–294.

A. Kilgarriff, S. Reddy, J. Pomikalek, P. V. S. Avinesh. 2010. A Corpus Factory for Many Languages. *Zbornik konference LREC'10*.

S. Krek, A. Kilgarriff. 2006. Slovene word sketches. *Zbornik konference Jezikovne tehnologije*.

N. Ljubešić, M. Stupar, T. Jurić, Ž. Agić. 2013. Combining available datasets for building named entity recognition models of Croatian and Slovene. *Jezikovne tehnologije, Slovenščina 2.0*, 1/2, str. 35–57

N. Ljubešić, T. Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. *Zbornik konference Text, Speech and Dialogue*.

N. Logar. 2003. Kratice in tvorjenke iz njih - aktualna poimenovalna možnost. *Współczesna polska i słoweńska sytuacja językowa*, str. 131–149.

N. Logar., M. Grčar, M. Brakus, T. Erjavec, Š. Arhar Holdt, S. Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina.

M. Michelizza. 2008. Jezik SMS-jev in SMS-komunikacija. *Jezikoslovni zapiski* 14/1, str. 151–166.

M. V. Noblia. 1998. The Computer-Mediated Communication: A New Way of Understanding The Language. *Zbornik konference Internet Research and Information for Social Scientists*, str. 10–12.

M. Pinnis, N. Ljubešić, D. Ştefanescu, I. Skadina, M. Tadić, T. Gornostay. 2012. Term extraction, tagging, and mapping tools for under-resourced languages. *Zbornik konference Terminology and Knowledge Engineering*, str. 193–208.

T. Praprotnik. 2003. Pragmatični vidiki žaljive komunikacije v računalniško posredovani komunikaciji – multipla perspektiva. *Teorija in praksa*, 40/3, str. 515–540.

P. Rayson, R. Garside. 2000. Comparing corpora using frequency profiling. *Zbornik konference Comparing Corpora*, str. 1–6.

R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, C. Richards. 2001. Normalization of non-standard words. *Computer Speech and Language*, 15(3), str. 287–333.

T. Štajner, T. Erjavec, S. Krek. 2013. Razpoznavanje imenskih entitet v slovenskem besedilu. *Jezikovne tehnologije, Slovenščina 2.0*, 1/2, str. 58-81.

C. Tagg. 2012. *Discourse of Text Messaging*. London: Continuum.

D. Verdonik, A. Zwitter Vitez. 2011. *Slovenski govorni korpus Gos*. Ljubljana: Trojina.

Š. Vintar. 2010. Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology* 16(2).