Razvoj učne množice za izboljšano označevanje spletnih besedil

Jaka Čibej,* Špela Arhar Holdt,* Tomaž Erjavec,† Darja Fišer*†

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
jaka.cibej@ff.uni-lj.si, spela.arharholdt@ff.uni-lj.si, darja.fiser@ff.uni-lj.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

Jezik spletne komunikacije se v marsikaterem vidiku razlikuje od standardnega jezika. Obstoječa orodja za označevanje besedil se z njim težje spopadajo, saj je učenje označevalnih postopkov do zdaj potekalo predvsem na standardnih besedilih. Za čimbolj natančno računalniško obdelavo računalniško posredovane komunikacije je torej treba ustrezno nadgraditi označevalne metodologije in orodja. V prispevku zato predstavljamo izdelavo učnega korpusa slovenske spletne komunikacije, ki bo uporabljen kot učna množica za izboljšano jezikoslovno označevanje slovenskih spletnih besedil. Korpus je bil vzorčen iz korpusa spletne slovenščine JANES in sprva avtomatsko označen, nato pa ročno popravljen na petih ravneh: tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladnja.

The Development of a Training Set for an Improved Annotation of Internet Texts

In many ways, the language of internet communication differs from standard language. Internet texts are more difficult to process for existing tools, which have predominantly been trained on standard language. To improve their accuracy when processing internet texts, the tools and annotation methodologies need to be upgraded. In this paper, we present the compilation of a training corpus of Slovene internet communication to be used as a training set to improve the automatic annotation of Slovene internet texts. The training corpus was sampled from the JANES corpus of Internet Slovene, then automatically annotated and manually corrected on five annotation levels: tokenisation, sentence segmentation, normalisation, lemmatisation, and morphosyntax.

1 Uvod

Jezik spletnih žanrov, kot so tviti, forumi in komentarji, se v marsikaterem vidiku razlikuje od standardnega jezika (Baldwin et al., 2013). Med pogosteje omenjanimi razlikami so npr. pogovorni zapisi besed, pogostejša raba regionalizmov in tujejezičnih prvin ter raba okrajšav in jezikovnih oz. grafičnih elementov, specifičnih za spletno komunikacijo (Crystal, 2001). Z naštetimi specifikami se obstoječa orodja za označevanje besedil težje spopadajo, saj je učenje označevalnih postopkov do sedaj potekalo predvsem na standardnih besedilih (Ljubešić et al., 2014). Če želimo kvalitetno obdelavo jezika zagotoviti tudi za računalniško posredovano komunikacijo, je torej potrebna ustrezna prilagoditev oz. nadgradnja označevalne metodologije in orodij.

V prispevku kot enega od korakov v tej smeri predstavimo izdelavo učnega (in testnega) korpusa spletne komunikacije, ki je bil vzorčen iz korpusa JANES (Fišer et al., 2015). Učni korpus je bil nato avtomatsko označen na petih ravneh (tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladnja), avtomatsko pripisane oznake pa so bile ročno popravljene. V prvem delu prispevka predstavimo cilje in delotok označevalnega procesa, v drugem pa se osredotočimo na tisti del apliciranih rešitev, ki se razlikuje od dosedanjih praks jezikoslovnega označevanja slovenščine.

2 Sorodne raziskave

Predstavljena raziskava sodi v obširno in dejavno področje normalizacije nestandardnih prvin v besedilih. Čeprav nekatere raziskovalce zanimajo specializirani problemi, npr. normalizacija zapisa velikih začetnic (Nebhi et al., 2015), rediakritizacija besed (Ljubešić et al., 2016) ali normalizacija ključnikov (Declerck in Lendvai, 2015), se večina skupnosti ukvarja s splošno normalizacijo nestandardnega besedišča v spletnih besedilih, ki izboljšuje nadaljnjo obdelavo besedil, npr. oblikoskladenjsko označevanje in lematizacija. Ker gre za razmeroma novo raziskovalno področje, za večino jezikov normalizirane učne in testne množice še ne obstajajo. Prav tako ni vzpostavljenih enotnih evalvacijskih meril, zato je orodja težko razvijati, evalvirati in primerjati. Prva prizadevanja v to smer so že obrodila sadove za angleščino v okviru delavnice in skupne naloge WNUT (Baldwin et al., 2015), za nemščino v okviru skupne naloge EmpiriST (Beißwenger et al., 2015a) in za španščino v okviru delavnice in skupne naloge Tweet-Norm (Alegria et al., 2014). Poleg prosto dostopnih učnih množic so bile za te jezike izdelane tudi smernice za ročno označevanje (npr. Beißwenger et al., 2015b). Za podobno si prizadevamo tudi s označevalsko kampanjo za slovenščino, ki jo predstavljamo v pričujočem prispevku.

Cilj kampanje je vzpostaviti načela in razviti učne množice za učenje avtomatske normalizacije nestandardnega besedišča v šumnih spletnih besedilih kot predhodni postopek jezikoslovnega označevanja teh besedil z orodji, sicer razvitimi za standardni jezik (Sproat et al., 2001). S tem bomo nadgradili in izboljšali osnovni pristop k normalizaciji slovenskih tvitov (Ljubešić et al. 2014), ki temelji na statističnem strojnem prevajanju na nivoju znakov, naučenem na ročno preverjenem leksikonu izvornih in normaliziranih parov 1000 neznanih najbolj ključnih besed v korpusu tvitov glede na referenčni korpus Gigafida. Čeprav so bili rezultati, doseženi z osnovnim modelom, spodbudni (69% točnost), je njegova pomanjkljivost ta, da pri iskanju najverjetnejše normalizirane oblike ne upošteva sobesedila, kar bomo

omogočili z ročno normaliziranim korpusom, ki ga predstavljamo v prispevku.

3 Priprava podatkov in označevalna platforma

Besedila za označevanje smo vzorčili iz korpusa JANES ter izdelali dva vzorca: Kons1, ki vsebuje tvite, in Kons2, ki vsebuje forumska sporočila ter komentarje na blogovske zapise in spletne novice. V nadaljevanju predstavimo kriterije za vzorčenje teh podkorpusov, označevalno platformo WebAnno in postopek pretvorbe, uvoza in izvoza podatkov.

3.1 Vzorčenje

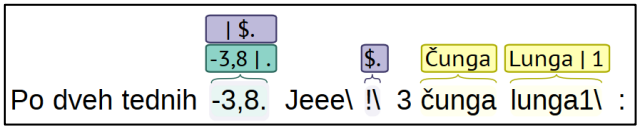
Vzorec Kons1 zajema 4.000 tvitov, ki so bili vzorčeni naključno, a z upoštevanjem nekaterih dodatnih omejitev. Izločili smo tvite, daljše od 120 znakov[1], in tvite z uradnih računov organizacij (npr. agencij in podjetij). Pri vzorčenju smo upoštevali tudi stopnjo tehnične (T1-T3) in jezikovne (L1-L3) standardnosti tvita, ki smo jo merili s posebej za to razvito avtomatsko metodo (Ljubešić et al., 2015). Ker se nismo želeli osredotočiti le na nestandardne tvite (3), temveč so nas zanimale tudi splošne specifike žanra, smo v Kons1 vključili po 1.000 tvitov iz vsake od kategorij T1L1, T3L1, T1L3 in T3L3.[2]

Na podoben način smo izdelali vzorec Kons2, ki vsebuje 4.000 besedil, razdeljenih po stopnjah (ne)standardnosti. Pri žanrih, vključenih v Kons2, ni znakovne omejitve, kakršno ima Twitter, zato smo zaradi primerljivosti z vzorcem Kons1 v Kons2 vključili samo besedila z dolžino med 20 in 280 znaki.

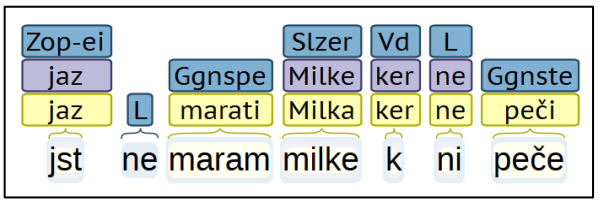
Pred ročnim označevanjem sta bila vzorca z obstoječimi orodji (Erjavec, 2011; Ljubešić et al., 2014) avtomatsko označena na vseh petih obravnavanih ravneh jezikoslovnih oznak.

3.2 Označevalna platforma

Vzorčena besedila smo razdelili na datoteke, ki so vsebovale po 10 besedil, in naložili na WebAnno (Eckart de Castilho et al., 2014), prosto dostopno spletno platformo, ki omogoča večravninsko označevanje besedila. Primeri oznak za normalizacijo, tokenizacijo in stavčno segmentacijo so prikazani na Sliki 1, primeri oblikoskladenjskih oznak in lem pa na Sliki 2.



Slika 1: Oznake za normalizacijo (rumena), tokenizacijo (zelena) in stavčno segmentacijo (vijolična) v WebAnnu.



Slika 2: Oznake za oblikoskladnjo (modra) in lematizacijo (rumena) v WebAnnu. Prikazane so tudi normalizirane oblike (vijolična).

WebAnno smo prilagodili tako, da je omogočal označevanje besedil na vseh petih ravneh, relevantnih za našo učno množico. V platformo je vgrajena tudi funkcija razsojanja, ki se jo lahko uporabi, če iste podatke označi več označevalcev. Pri tem razsodnik primerja večkratne oznake iste datoteke in izbere dokončno različico.

3.3 Pretvorba podatkov

Posebno pozornost smo namenili formatu podatkov, da bi vse ročno preverjene oznake združili v enovit zapis. Pri zapisu korpusa JANES uporabljamo priporočila za kodiranje besedil TEI (Text Encoding Initiative), ki so v uporabi pri večini slovenskih korpusov. Ker WebAnno formata TEI ne podpira, smo med razpoložljivimi formati izbrali TSV, tabelarični format, v katerem je vsaka pojavnica zapisana v svoji vrstici, ki ji je pripisan njen identifikator ter vse oznake.

Izdelali smo program, ki izvorni TEI izvozi v format TSV. Tega je nato mogoče uvoziti v WebAnno, po označevanju pa lahko popravljeni TSV znova izvozimo in združimo z izvornim TEI, tako da rezultat vsebuje vse oznake izvornega TEI, a dopolnjene s popravki ročnega označevanja. Postopek je razmeroma zapleten, saj smo pretvorbo zasnovali tako, da bo uporabna tudi za morebitne nadaljnje označevalske kampanje z bistveno drugačnimi oznakami. Težava je tudi v tem, da naša označevalska metodologija predvideva prvine, za katere WebAnno ni predviden, predvsem popravljanje pojavnic in mej med njimi. Poleg tega je z vidika pretvorbe podatkov problematično, da lahko eni pojavnici ustreza več normaliziranih oblik ali obratno.

4 Smernice za označevanje

Na podlagi preliminarnega ročnega pregleda uravnoteženega vzorca 200 tvitov sta bili izdelani dve zbirki smernic za označevanje[3]. Tehnične smernice so označevalce seznanile z označevalsko shemo v WebAnnu in s splošnimi napotki za delo s platformo (npr. (ra)združevanje pojavnic, brisanje nerelevantnih ali avtomatsko generiranih besedil, delo z večplastnimi oznakami), jezikoslovne smernice pa so obravnavale kriterije za sprejemanje jezikoslovnih odločitev pri označevanju. Za zagotavljanje kompatibilnosti podatkovnih množic so smernice v največji možni meri upoštevale navodila za označevanje slovenskih korpusov (JOS,[4] ssj500k,[5] GOS,[6] IMP[7]) in referenčnih virov za slovenščino (Fran,[8] Sloleks[9]).

[1] Daljši tviti se zaradi splošne omejitve do 140 znakov pogosto končajo z odrezanimi besedami, ki bi predstavljale šum.

[2] Kategorij s stopnjama T2 in L2, pri katerih so značilnosti nestandardnih tvitov manj izražene, nismo vključili.

[3] Smernice so prosto dostopne na naslovu http://nl.ijs.si/janes/viri/

[4] http://nl.ijs.si/jos/msd/html-sl/

[5] http://www.slovenscina.eu/tehnologije/ucni-korpus

V nadaljevanju predstavimo poglavitne specifike označevanja prvin, ki so značilne za spletno komunikacijo, ter pri vsaki ravni označevanja razpravljamo o težavnih primerih in rešitvah zanje.

4.1 Stavčna segmentacija

V standardnem jeziku meje med povedmi najpogosteje zaznamujejo končna ločila, v spletnih besedilih, vključenih v našo učno množico, pa smo na ravni stavčne segmentacije kot signal za konec povedi poleg klasičnih končnih ločil (pika, klicaj, vprašaj) upoštevali tudi druga ločila, ki lahko delujejo kot končna (npr. večpičje, vezaj in narekovaj), oz. druge prvine, ki se lahko pojavljajo na tem mestu:

a) emotikoni in emojiji (=D ☺),
b) ključniki (#sampovem),
c) URL- ali e-naslovi (http://youtube.com, avtor@domena.com),
d) sklici na uporabniška imena (@avtor).

Te prvine zaključujejo stavke zlasti v besedilih brez končnih ločil. Če se stavek konča z nizom prvin, se za konec stavka[10] šteje zadnja prvina v nizu:

> Liverpool zasluženo owna Twitter, ampak na vrhu je pa fucking Iago Aspas hahaha :) #nogomet #LFC #SOULIV http://t.co/LCyEvyoVD7 ¶

V primerih, ko se tovrstne prvine sopojavljajo s končnimi ločili (bodisi samostojno ali kot niz), jih obravnavamo kot nov stavek, četudi se nanašajo na prejšnjega:

> Življenje Je Cirkus. js sm pa čefur. Luka Stigl js sm se poscal v hlače k sm se vidu. bolano. ¶ :) ... ¶ http://t.co/QyzKRZqZnS ¶

4.2 Tokenizacija

Specifične označevalne izzive pri spletnih besedilih predstavlja tudi raven tokenizacije. Določene vrste pojavnic, ki vsebujejo ločila, je tokenizator najpogosteje napačno ločil, zato jih je bilo treba združiti ročno. Pogost primer so okrajšave (npr. slov.), pri katerih je tokenizator piko interpretiral kot končno ločilo in jo obravnaval kot ločeno pojavnico (ter jo na ravni stavčne segmentacije označil tudi kot konec stavka), kar je moral označevalec popraviti ročno.

Podobno je bilo z emotikoni, ki so se pogosto pojavljali v strnjenih nizih (npr. :):):**), tokenizator pa jih je (napačno) razdelil na posamezne pojavnice. V takih primerih smo celoten niz združili v eno samo pojavnico:

> :) ¶ :) ¶ : ¶ * ¶ * → :):):**

Poleg že znanih zadreg združevanja in ločevanja elementov pisnega jezika se v naši učni množici pojavlja tudi višji delež primerov, v katerih avtor pri zapisu besed narazen oz. skupaj ne sledi standardu (*nebi*), ne uporablja presledkov ob ločilih (*fruktoza-glukoza*) ali vpenja ločila v besede na manj predvidljiv način (*iTunes-ih*, *žen(sk)am*, *politike/o*). Tovrstne pojavnice smo združevali:[11]

> TV ¶ - ¶ ja → TV-ja
> sms ¶ - ¶ i → sms-i
> žen ¶ (¶ sk ¶) ¶ am → žen(sk)am
> politik ¶ (¶ e ¶ / ¶ o ¶) → politik(e/o)

4.3 Normalizacija

Pri normalizaciji smo upoštevali načelo minimalne intervencije in besedam nismo pripisovali standardnih sopomenk (npr. *pofarbat* → *pofarbati* in ne **pobarvati*). Normalizirane so bile besede v nestandardnem zapisu (*priemerjavi* → *primerjavi, sovascana* → *sovaščana, mamo* → *imamo*) ali z nestandardno morfologijo (*na Ptuji* → *na Ptuju*), v izvirni obliki pa so ostale tviterske prvine (*#krneki*, *@RTV_Slovenija*, *www.youtube.com*), samocenzurirane besede (*p*****, *pof**am*) in jezikovne napake na ravni skladenjskih razmerij (*pri Harry Potterju*, *ne rabim knjigo*), četudi so zelo verjetno naključne (*morajo delajo*). Prav tako pri normalizaciji nismo popravljali izbire besedišča (menjave glagolov *moči-morati*) ali napak na ravni sloga ali registra (*rabiti-potrebovati*).

Pri normalizaciji sta se za najbolj problematični izkazali dve kategoriji besed: nestandardne besede brez neposredne standardne ustreznice in z več različicami zapisa (*orng*, *ornk*, *oreng*, *orenk* ali *fovš*, *favš*, *fouš*, *fauš*, *fowš*) ter tujejezične prvine z različnimi stopnjami prevzetosti na ravneh zapisa in oblikoslovja (*updateati*, *updajtati*, *updejtati*, *apdejtati*), ki jim zgolj s pomočjo referenčnih virov ni bilo mogoče določiti normalizirane ustreznice.

Pri nestandardnih besedah z več različicami zapisa smo normalizirano obliko določili tako, da smo v korpusu JANES s pomočjo regularnih izrazov poiskali vse različice zapisa in izbrali najpogostejšo (v zgornjih primerih sta to *ornk* in *fouš*).

V primeru tujejezičnih prvin bi bila normalizacija v izvirno obliko problematična, saj bi s tem v korpus vnesli umetne oblike, ki jih v realni jezikovni rabi ne najdemo (npr. *poapdejtati* → *po-update-ati*). Tujejezične prvine smo zato obravnavali po naslednjih kriterijih:

a) če je bila beseda zapisana povsem fonetizirano (npr. *dankešn* ‘danke schön’, *aprišiejt* ‘appreciate’), smo jo obravnavali kot slovensko nestandardno besedo z več različicami zapisa (glej *fouš* in *ornk* zgoraj);

b) če je beseda še vedno izkazovala značilnosti tujejezičnega zapisa, npr. neslovenske črke (*wau*) oz. ostanke izvirnega zapisa (*meil*), smo normalizirano obliko določili tako, da smo iz korpusa JANES izbrali najpogostejšo različico med tistimi, ki so še vsebovale značilnosti tujejezičnega zapisa (npr. *updateati*, *updajtati*, *updejtati* → *updejtati*).

[6] http://www.korpus-gos.net/Support/About
[7] http://nl.ijs.si/imp/
[8] http://fran.si/
[9] http://www.slovenscina.eu/sloleks
[10] Konec stavka oz. mejo med pojavnicami v tem prispevku označujemo s simbolom ¶.
[11] Pri tem je treba omeniti, da napačno zapisanih nizov z manjkajočimi ali odvečnimi presledki (*hodildomov*, *porka duš*) ne popravljamo na nivoju tokenizacije, temveč pri normalizaciji.

4.4 Lematizacija

Pripisovanje lem je v največji možni meri sledilo smernicam za označevanje korpusa ssj500k (Holozan et al., 2008), ki je v vmesniku SketchEngine služil kot referenčni vir za označevalce. Razlike ali dopolnitve označevalnega sistema zadevajo odločitve, vezane na specifike označevanih besedil. Pri tem gre izpostaviti tujejezične prvine in raznovrstne kratice, ki se v spletni slovenščini pojavljajo mnogo pogosteje in oblikovno bolj raznorodno kot v standardnem jeziku.

Med večjimi izzivi označevanja je določanje meje med tujejezičnim in slovenskim besediščem. V tvitih je tujejezičnih prvin veliko, pojavljajo pa se kot posamezne besede različnih besednih vrst in variant zapisa (*share/shareati/share-ati/šerati*), kot besedne zveze ali daljši segmenti. Zadnje smo označevali kot niz pojavitev v tujem jeziku, pri čemer so leme enake oblikam, oblikoskladenjska oznaka po sistemu JOS pa je *Nj*. Podobno velja za občnoimenske besedne zveze (*bonus score, sugar rush*) in posamezne besede, ki so v besedilu zapisane citatno, brez jasno razvidnih prilagoditev slovenskemu zapisu oz. pregibanju (*jailbreak, hrvatskog*). Pri besedah, ki prilagoditev izražajo, smo lemo določili v skladu s slovenskimi oblikoslovnimi načeli (*benchmarki → benchmark, chatala → chatati*). Pri odločanju, ali besedo obravnavati kot tujejezično ali prevzeto, so bili uporabljeni tudi referenčni leksikalni viri, predvsem SSKJ in SNB ter leksikon besednih oblik Sloleks.

Vprašanja uvrščanja kratičnih poimenovanj med kratice in okrajšave na eni strani ter občna (*lol, drž.*) in lastna imena (*Sds, Slo.*) na drugi so bila rešena že na ravni normalizacije. V teh primerih so označevalci pri pripisovanju lem (in oblikoskladenjskih oznak) sledili normaliziranim oblikam.

Projektnospecifična je še odločitev, da se URL-naslovi lematizirajo v domeno (*http://t.co/ZaVQdnaN5p* → t.co), s čimer omogočimo preglednejše prikazovanje korpusnih podatkov v vmesniku. Pri ostalih tviterskih prvinah (uporabniška imena, ključniki, emotikoni) je lema enaka obliki.

4.5 Oblikoskladenjsko označevanje

Tudi na oblikoskladenjski ravni so bile osnovno izhodišče za označevanje smernice korpusa ssj500k. Med razlikami gre v prvi vrsti omeniti prilagoditev oz. širitev sistema za oblikoskladenjsko označevanje JOS, ki so mu bile dodane naslednje nove oznake: *Nh* za ključnike; *Nw* za URL- in e-naslove; *Na* za sklice na uporabniška imena; in *Ne* za emotikone in emojije. Z naštetimi oznakami in načelom lematizacije, pri katerem lema sledi izvorni obliki, smo na enostaven in sledljiv način rešili vprašanje označevanja tvitersko specifičnih prvin.

Pri označevanju niso bile uporabljene oznake *Nt* (zatipkana beseda) ter *Np* (tokenizacijska napaka), saj so bile tovrstne težave ročno odpravljene že na ravni normalizacije.

Zaradi specifik tviterske komunikacije se je pri označevanju pojavljalo večje število pomensko nejasnih oz. dvoumnih primerov (npr. *dobr* kot pridevnik ali prislov). Kot je to veljalo za označevanje ssj500k, so označevalci take primere interpretirali in označili po principu najverjetnejše možnosti. Podobno načelo je veljalo za označevanje samocenzuriranih besed (*v p***i* → Sozem). V primeru odstopov od norme na skladenjski ravni so bile oznake pripisane skladno z dejansko (in ne pričakovano) pojavitvijo. Tipični tovrstni primeri so na ravni rabe sklonov (*nisem oblikovala intergalaktično brisačo* → Sozet, ne Sozer), števila (*Z Martino smo se tekmovale* → Ggnd-mz, ne Ggnd-dz) in rabe kategorije živosti (*jaz vem za kvalitetnega centra z nba izkušnjami* → Sometd, ne Sometn).

Nazadnje je treba omeniti še označevanje zaprtih besednih vrst, ki je skladno z načeli ssj500k v izhodišču potekalo leksikonsko pogojeno, a z možnostjo dodajanja nestandardnega besedišča. Kategorija, ki je na ta način dobila največ novih elementov, je členek (npr. *eto, evo, ajde, naka, kao, glih/lih* in *ta* v primerih tipa *ta star*). Pri drugih kategorijah se potreba po dopolnitvi pojavlja redkeje, npr. z veznikom *samo* (*Nism še vidu, sam so rekl da je dobr*).

5 Označevalska kampanja

V tem razdelku predstavljamo pregled in opis različnih stopenj označevalske kampanje, ki je zajemala tri stopnje:

a) NTS-Kons1 – normalizacijo, tokenizacijo in stavčno segmentacijo vzorca Kons1 (od decembra 2015 do marca 2016);
b) LO-Kons1 – lematizacijo in oblikoskladenjsko označevanje vzorca Kons1 (od marca 2016 s predvidenim zaključkom avgusta 2016); in
c) NTS-Kons2 – normalizacijo, tokenizacijo in stavčno segmentacijo vzorca Kons2 (od marca 2016 do maja 2016).

5.1 Usposabljanje označevalcev

Ob začetku prvega dela označevalske kampanje (NTS-Kons1) smo priredili dvodnevno delavnico, na kateri so se označevalci seznanili z delom v WebAnnu in s smernicami za označevanje. Na delavnici je sodelovalo 11 študentov jezikoslovnih smeri na magistrski stopnji. Teoretičnemu uvodu v WebAnno s praktičnim delom in predstavitvi smernic je sledila uvajalna označevalska faza, med katero so udeleženci označili manjše število tvitov. Cilji označevanja so bili naslednji:

a) vsak tvit mora biti pravilno razdeljen na stavke;
b) vsak tvit mora biti pravilno razdeljen na pojavnice; in
c) vse pojavnice morajo imeti pripisano normalizirano obliko; dvoumne pojavnice ohranijo izvirno, nenormalizirano obliko.

Uvajalni označevalski fazi je sledila diskusija, na kateri smo z označevalci razpravljali o njihovih odločitvah in razhajanjih med njihovimi oznakami, podali pa smo tudi pravilne rešitve in razloge zanje, da bi čim bolj uskladili odločitve označevalcev in izboljšali njihovo ujemanje. V drugem delu kampanje (LO-Kons1) smo na enodnevni delavnici označevalce seznanili s konceptom oblikoskladenjskih oznak in lem ter jim predstavili smernice. Tudi tej delavnici je sledila uvajalna faza, cilj pa je bil tokrat vsaki pojavnici v tvitu (z izjemo ločil) pripisati ustrezno lemo in oblikoskladenjsko oznako JOS. Odločitve smo skupaj prediskutirali in utemeljili z načeli iz smernic.

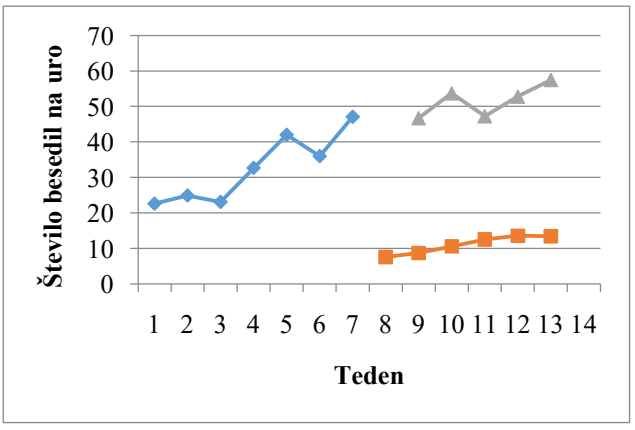
5.2 Preizkušanje označevalcev

Obema delavnicama je sledila preizkusna faza. V delu NTS-Kons1 smo označevalce razdelili v dve skupini po 5 oz. 6 označevalcev, vsaki skupini pa smo dodelili 100 tvitov iz preizkusne množice, pri katerih so morali popraviti avtomatsko pripisane oznake in dodati nove, kjer je bilo to potrebno. Pri oblikoskladenjskih oznakah in lematizaciji smo označevalce razdelili v pare, vsak par pa je označil po 50 tvitov iz preizkusne množice. Oznake sta nato ročno preverila razsodnika, ki sta ocenila tudi natančnost označevalcev. Na podlagi rezultatov sta bila v delu NTS-Kons1 iz kampanje izključena dva nezanesljiva označevalca, v obeh delih pa sta razsodnika po začetni evalvaciji dopolnila smernice za označevanje še s primeri, ki so se v preizkusni seji izkazali za problematične.

5.3 Delotok označevanja

Delotok označevanja je vključeval skupino označevalcev in dva razsodnika z dobrim poznavanjem smernic za označevanje. Razsodnika, ki sta bila zadolžena tudi za vodenje označevalske kampanje, sta v tedenskih fazah posamezni skupini označevalcev[12] dodelila določeno število datotek, po koncu vsake faze pa sta oznake ročno preverila in, če je bilo potrebno, označevalcem podala konstruktivno povratno informacijo ter na ta način odstranila najpogostejše oz. najresnejše napake. Če so označevalci med delom naleteli na posebno problematično dilemo, so bile z njo dopolnjene tudi smernice za označevanje. Ustvarjen je bil tudi e-poštni seznam, na katerem so lahko označevalci razsodnikoma zastavljali vprašanja in razreševali problematične ali dvoumne primere, ki niso bili vključeni v smernice.

Med delom smo spremljali učinkovitost označevalcev, tako da smo v vsaki fazi merili razmerje med časom označevanja in številom označenih besedil (glej Sliko 3).



Slika 3: Učinkovitost označevalcev pri označevanju vzorcev Kons1 in Kons2. Modri del predstavlja NTS-Kons1, sivi NTS-Kons2 in oranžni LO-Kons1.

Iz grafa je razvidno, da normalizacija, tokenizacija in stavčna segmentacija potekajo mnogo hitreje od lematizacije in oblikoskladenjskega označevanja. Dobro usposobljeni označevalci lahko v eni uri normalizirajo med 45 in 55 besedil, lematizirajo in oblikoskladenjsko označijo pa le nekaj nad 10 besedil.

6 Rezultati in diskusija

V tem razdelku podamo strnjeno kvantitativno analizo označenih vzorcev in razpravljamo o najpogostejših razhajanjih med označevalci na vseh nivojih označevanja.

6.1 Kvantitativna analiza rezultatov

Tabela 1 prikazuje velikosti do zdaj označenih podatkovnih množic. V prvi vrstici so navedeni podatki za oblikoskladenjsko označeni del Kons1, v drugi in tretji za celotna Kons1 in Kons2, v zadnji pa za njuno vsoto.

	Besedila	Stavki	Besede	Pojavnice	Norm.	Norm. %
Kons1-MSD	880	2.365	20.537	23.958	4.888	20,4
Kons1	3.940	9.976	86.593	102.719	11.881	11,6
Kons2	1.927	4.473	34.583	41.056	4.728	11,5
Kons	5.867	14.449	121.176	143.775	16.609	11,6

Tabela 1: Velikost označenih vzorcev.

Stolpci od leve proti desni podajajo število označenih besedil, število ročno preverjenih stavkov, število preverjenih besednih pojavnic v izvornem besedilu, število vseh pojavnic ter število (in nazadnje delež) pojavnic, ki jim je bila pripisana normalizirana oblika. Vseh ročno preverjenih besed je več kot 120.000. Več kot desetina je bila potrebna normalizacije.

Oblikoskladenjski del je bistveno manjši, saj je trenutno ročno označenih le približno 20.000 besed. Predpostavljamo pa, da je že ta količina zadostna za preverjanje točnosti označevanja z razvitimi orodji in za dopolnjevanje učne množice, da lahko oblikoskladenjski označevalniki bolje označujejo uporabniške spletne vsebine. Besedila, ki so bila oblikoskladenjsko označena do zdaj, so v povprečju tudi bolj nestandardna kot preostanek vzorca Kons1 - normalizacije je bilo namreč potrebnih dobrih 20 % pojavnic.

Omeniti je treba tudi leksikon oblikoskladenjsko označenega vzorca Kons1, ki vsebuje vsega skupaj 8.033 različnih izvornih pojavnic (vključno z ločili). Od tega je normaliziranih 7.305 pojavnic (skoraj 91 %). Normalizirane pojavnice imajo 5.548 različnih lem, označene pa so bile s 592 različnimi oblikoskladenjskimi oznakami.

6.2 Najpogostejša razhajanja pri označevanju

Na ravni stavčne segmentacije je do razhajanja prišlo predvsem pri stavkih, ki niso vsebovali nobenih klasičnih ločil in so bili npr. razdeljeni z večpičji, ki jih je bilo mogoče interpretirati bodisi kot zamolk bodisi kot konec stavka.

Pri normalizaciji so bile glavni vir razhajanja besede, ki jih je bilo mogoče normalizirati v več različnih oblik (npr. *k* → *ker/ko/ki/kjer/kot* itd. ali *sm* → *sem*/*samo*), v

[12] Na začetku so bili označevalci razdeljeni v skupine po 3, pozneje pa v pare. Zelo natančni označevalci so v nekaterih fazah označevali tudi posamezno.

omejenem kontekstu pa je bila interpretacija stvar posameznega označevalca.

Na ravni lematizacije in oblikoskladenjskega označevanja je mogoče razhajanja med označevalci in njihove napake pripisati več razlogom. V prvo skupino sodijo objektivno odpravljive probleme, ki jih označevalci razumejo, a v praksi pogosto spregledajo, npr. enakopisne oblike (npr. *si* kot oblika glagola *biti* ali povratni zaimek v dajalniku; *da* kot oblika glagola *dati* ali podredni veznik). V drugi skupini so razhajanja, ki so posledica dveh različnih, a znotraj sistema legitimnih interpretacij (npr. *Džizs, to bi b'lo fajn,* kjer je prvo besedo mogoče uvrstiti med občnoimenske ali lastnoimenske samostalnike ali pa med medmete).

Zadnja skupina razkriva težave s smernicami za označevanje, bodisi ker so slednje nejasno napisane ali pa ker predvidevajo rešitve, ki so manj intuitivne ali odstopajo od siceršnjih načel sistema. Med pogostimi napakami tega tipa so denimo pozabljeni popravki stopnjevanih prislovov tipa *večji/največji* (izjema, pri kateri je lema enaka stopnjevani in ne osnovni obliki, oblikoskladenjska oznaka pa izraža kategorijo stopnjevanosti) ali napake pri besednovrstnem uvrščanju povedkovega določila tipa *je bilo lepo*, ki se po smernicah ssj500k označuje kot pridevnik srednjega spola, označevalci pa ga razumejo kot prislov. Pri tem je nujno omeniti, da smernice temeljijo na referenčnih virih za slovenščino in v določeni meri preslikavajo kategorizacijske težave pri primerih, ki lahko nastopajo v različnih vlogah (npr. *nič, ves, kaj, prav* ipd.). Na ravni označevanja kratic in lastnih imen (predvsem tujih) je opremljenost še toliko bolj pomanjkljiva, saj tovrstno besedišče praviloma v vire ni zajeto. Težavam, identificiranim v tej skupini, bi se bilo pri nadaljnjem razvoju označevanja slovenščine smiselno natančneje posvetiti.

7 Zaključek

V prispevku smo povzeli ključne vidike smernic za označevanje učnega korpusa na različnih ravneh in na kratko predstavili rešitve za označevanje prvin, specifičnih za spletno komunikacijo. Cilj priprave korpusa, ki bo prosto dostopen na repozitoriju CLARIN.SI, je dvojni: uporaben bo kot učna množica za izboljšanje označevanja spletnih besedil, smernice za označevanje učnega korpusa pa bodo ponudile prvi celoviti vpogled v problematiko jezikoslovnega označevanja slovenske računalniško posredovane komunikacije ter rešitve za najbolj problematične primere. Izdelava učnega korpusa ponuja tudi možnosti za dopolnitev obstoječih leksikonov besednih oblik z nestandardnim besediščem (npr. z besedami *evo*, *eto*, *ajde*, *naka*, *kao*, *glih* ipd. v kategoriji členkov).

8 Zahvala

Avtorji se najlepše zahvaljujejo Kaji Dobrovoljc, Simonu Kreku in Katji Zupan za konstruktivne pripombe pri izdelavi smernic za označevanje, ter vsem označevalcem, ki so sodelovali v označevalski kampanji: Teji Goli, Melaniji Kožar, Vesni Koželj, Poloni Logar, Klari Lubej, Dafne Marko, Barbari Omahen, Eneji Osrajnik, Predragu Petroviću, Poloni Polc, Aleksandri Rajković in Izi Škrjanec.

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, ter s podporo Slovenske raziskovalne infrastrukture za jezikovne vire in tehnologije (CLARIN.SI).

9 Literatura

Iñaki Alegria, Nora Aranberri, Pere R. Comas, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo in Arkaitz Zubiaga. 2014. TweetNorm_es Corpus: an Annotated Corpus for Spanish Microtext Normalization. V: *Zbornik konference Ninth International Conference on Language Resources and Evaluation (LREC2014)*. ELRA, Reykjavik-Paris.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay in Li Wang. 2013. How Noisy Social Media Text, How Diffrnt Social Media Sources. V: *Sixth International Joint Conference on NLP*, str. 356–364.

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter in Wei Xu: Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. V: *Workshop on Noisy User-generated Text at ACL 2015, July 31, 2015*. Peking, Kitajska.

Michael Beißwenger, Thomas Bartz, Angelika Storrer in Swantje Westpfahl. 2015a. Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. V: *Guideline document from the Empirikom shared task on automatic linguistic annotation of internet-based communication (EmpiriST 2015)*.

Michael Beißwenger, Sabine Bartsch, Stefan Evert in Kay-Michael Würzner. 2015b. Richtlinie für die manuelle Tokenisierung von Sprachdaten aus Genres internetbasierter Kommunikation. V: *Guideline document from the Empirikom shared task on automatic linguistic annotation of internet-based communication (EmpiriST 2015)*.

David Crystal. 2001. *Language and the Internet*. Cambridge: Cambridge University Press.

Thierry Declerck in Piroska Lendvai. 2015. Processing and Normalizing Hashtags. V: *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference, 7–9 September 2015*, str. 104–109, Hissar, Bolgarija.

Richard Eckart de Castilho, Chris Biemann, Iryna Gurevych in Sied Muhie Yimam. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. V: *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands.

Tomaž Erjavec. 2011. Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. V: *5th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011)*. Portland: Association for Computational Linguistics, 2011, str. 33–38. http://aclweb.org/anthology-new/W/W11/W11-1505.pdf.

Darja Fišer, Nikola Ljubešić in Tomaž Erjavec. 2015. The JANES corpus of Slovene user generated content: construction and annotation. V: *International Research Days: Social Media and CMC Corpora for the*

eHumanities: Book of Abstracts, 23–24 October 2015, str. 11, Rennes, Francija.

Peter Holozan, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman in Aleš Velušček. 2008. *Specifikacije za učni korpus (kazalnik 2): projekt Sporazumevanje v slovenskem jeziku*. Kamnik. Dostopno na: http://projekt.slovenscina.eu/Vsebine/Sl/Kazalniki/K2.aspx.

Nikola Ljubešić, Tomaž Erjavec in Darja Fišer. 2014. Standardizing tweets with character-level machine translation. V: *Computational linguistics and intelligent text processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6–12, 2014: proceedings: part II, (Lecture notes in computer science, 8404)*, str. 164–175, Heidelberg: Springer, 8404.

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. V: *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference, 7–9 September 2015*, str. 371–378, Hissar, Bolgarija.

Nikola Ljubešić, Tomaž Erjavec in Darja Fišer. 2016. Corpus-Based Diacritic Restoration for South Slavic Languages. V: *Zbornik konference Tenth International Conference on Language Resources and Evaluation (LREC2016)*. ELRA. Portorož, Slovenija, str. 3613–3616.

Kamel Nebhi, Kalina Bontcheva in Genevieve Gorrell. 2015. ResToRinG CaPitaLiZaTion in #TweeTs. V: *Zbornik konference 24th International Conference on World Wide Web (WWW '15 Companion)*. ACM, New York, NY, USA, str. 1111–1115.

Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf in Christopher Richards. 2001. Normalization of non-standard words. V: *Computer Speech and Language, 15 (3)*, str. 287–333.