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.

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.