Smernice za označevanje kodnega preklapljanja v korpusu slovenskih tvitov JANES

Špela Reher, 2017-04-10

Kazalo vsebine

Smernice podajajo jezikoslovna načela za označevanje lastnosti kodnih preklopov oziroma tujejezičnih besed v slovenskih tvitih.

1 Splošna načela

1. Označujemo samo tujejezične besede v tvitih, v katerih je uporabljena tudi slovenščina. Pri presoji, ali je določena beseda slovenska, uporabimo načelo vključenosti v Slovar slovenskega knjižnega jezika, Slovenski pravopis ali Sloleks – če besedo najdemo v katerem od teh priročnikov, jo štejemo za slovensko. Podrobneje v 4. razdelku Jezik.

2. Če je tvit v celoti v tujem jeziku oz. več tujih jezikih, avtomatsko generiran ali popolnoma nerazumljiv, tvit izbrišemo. V takšnem tvitu ne označujemo ničesar drugega.

3. Pri označevanju stavčnih mej upoštevamo *Smernice za označevanje korpusa slovenskih tvitov JANES* (v nadaljevanju Osnovne smernice). Relevantna pravila so vključena v naslednjem razdelku.

4. Na prvem nivoju vsakemu preklopu pripišemo jezik in vrsto preklopa. Na drugem nivoju za vsako besedno zvezo (gl. 3. razdelek Večbesedni preklopi) označimo zapis, morfologijo in besedno vrsto. Večinoma se oba nivoja prekrivata, torej se oznake v WebAnnu raztezajo čez isti sklop besed; do razlike prihaja pri določenih večbesednih preklopih, kar je podrobneje pojasnjeno v nadaljevanju.

5. Kadar se tujejezična prvina pojavi kot hešteg oziroma kot del heštega, kodni preklop dodatno označimo z oznako hashtag.

6. Seznam vseh oznak za označevanje je vključen v 8. razdelku.

Zapis ponazoritvenih primerov:

Primeri so zapisani ležeče in v narekovajih, npr. *»fair enough«*. Pri večbesednih preklopih oz. več ločenih zaporednih preklopih so skupaj z neprekinjeno črto podčrtane besede, ki jih v WebAnnu označimo z eno skupno oznako, če pa je med njimi nepodčrtan presledek, to pomeni, da gre za dve ločeni enoti.

2 Stavčna segmentacija[1]

Cilj:

Besedila so pravilno ločena na stavke,[2] konec vsakega stavka pa je označen.

Smernice:

1. V celotnem tvitu preverimo, ali je avtomatska stavčna segmentacija pravilna.
2. Če del tvita deluje kot samostojen stavek, ga tako tudi obravnavamo[3] (*»@multikultivator Najbrž ne . ¶ :) ¶ Kot rečeno : bolje BO . ¶ Zrihtamo , ko utegnemo . ¶ (PS : tudi v veselje " konkurence " ;)«*).
3. Merilo za konec stavka je predvsem ločilo, ki deluje kot končno v stavku, npr. pika, klicaj, vprašaj, narekovaj ali večpičje (»*Kaj praviš ? ¶ Aha !*«).
4. Če ni dobrega razloga, da nekaj obravnavamo kot dva stavka, naj ostane eden (*»@urosgruber pri meni naloži CSS .. kar pa ne pomeni , da stran zgleda lepo :)»* → en stavek, ker večpičje deluje bolj kot vejica, ne kot pika).
5. Konec tvita je avtomatično tudi konec stavka, zato tega ne označujemo.

Težavni primeri:

1. Večpičje

a. Ponavadi je končno ločilo (*»@SLO_Super_Visor po moje se jo izogiba kot hudič križa. ¶ Glavn da on spet laja … ¶ :-)))))«*).

b. Včasih označuje zgolj elipso ali zamolk sredi stavka – v takšnem primeru ni končno ločilo *(»To se mi zdi ... neumno.«*).

2. Imena (@ime)

a. Če se pojavljajo sredi stavka, so del stavka (*»neka baka :) uleti pa praša če loh gre kr naprej«*, *»sej #tarca je pa dons kr ok«*, *»sej je rekla @Sandra d je treba to drgac«*).

b. Če se pojavljajo na začetku ali koncu tvita, jih ne označujemo oz. upoštevamo pri določanju, ali je preklop znotraj- ali medstavčni (npr. v primeru *»@Donfarfezi Lol. Raj za pivo zber :P«* besedo »lol« označimo kot medstavčni preklop, ime pa ignoriramo).

3. Emotikoni

a. Če nadomeščajo končno ločilo, zaznamujejo konec stavka (*»kot da je to važn :)) ¶ nobenga to ne briga vec sploh«*).

b. Če sledijo končnemu ločilu, jih obravnavamo kot samostojen stavek (*»Sonce, sneg in pot pod noge! ¶ :) ¶ Gremo v hribe!«*).

---

[1] Povzeto in prilagojeno po: http://nl.ijs.si/janes/wp-content/uploads/2014/09/JANES-jezikoslovne-smernice-v0.9.pdf

[2] "Stavek" uporabljamo v širšem pomenu, torej kot najmanjšo samostojno enoto jezikovnega sporočila, ki bi se v standardni slovenščini začela z veliko začetnico in zaključila s končnim ločilom ("poved" v slovenistični terminologiji).

[3] V smernicah konec stavka za lažjo predstavo označujemo s simbolom ¶.

4. Znak *
a. Kot mejo stavka označimo tudi znak * v primerih, kot je »*starts thinking * *hurts just a little bit *« (torej gre za medstavčni preklop).

3 Večbesedni preklopi

Cilj:

Pri vsakem kodnem preklopu sta označena jezik in vrsta preklopa (znotraj- ali medstavčni), pri posameznih besednih zvezah pa označimo še obliko zapisa, morfologijo in besedno vrsto.

Smernice:

1. Pri večbesednih preklopih skupno (tj. z eno oznako) označimo JEZIK in VRSTO PREKLOPA. Za seznam znak in njihov pomen gl. tabelo na koncu Smernic.
2. Za vsako besedno zvezo posebej označimo ZAPIS, MORFOLOGIJO in BESEDNO VRSTO, pri čemer za posamezno besedno zvezo štejemo:

a. **stavek ali daljši sklop besed, ki jim ni mogoče določiti ene besedne vrste (N)**

i. *»@InaMcMina I need this so bad right now . Ne mačke . Svaljkanje«* (MS)

ii. *»@GobaFunk ohhh i know predobr kako je po 72 h urah nespanja ...«* (ZS)

b. **ključnik (hashtag)**

i. *»Tisto , ko si pred tvojiim komadom fentas koleno :S #fml #jobparty«* (označimo skupaj MS+ANG, nato zapis + morfologijo + bes. vrsto za vsak hešteg posebej)

c. **sklop ponovljenih besed:** *»omg omg omg«* ali *»and and and«* ali *»jp, jp«*

d. **samostalniško zvezo** skupaj s prilastkom (premodifier ali postmodifier)

i. pridevnik + samostalnik: *»old boysov«, »scary shit«*

ii. samostalnik + samostalnik: *»buffalo burger«*

iii. daljši prilastek: *»take your kid to the work day«; »' i dont give a fuck ' attitude«*

iv. samostalnik + PostM: *»a breath of fresh air«*

v. predlog + samostalnik = predložna zveza

e. **pridevniško zvezo** skupaj s prilastkom

i. pridevnik + prislov: *»fair enough«*

ii. prislov + pridevnik: *»#oddlysatisfying«*

f. **predložno zvezo**: *»on a regular basis«*

g. pri drugih kombinacijah zapis, morfologijo in besedno vrsto označimo za vsako besedo oziroma besedno vrsto posebej

i. *»sej pravm , francozi ownajo horrorland .«* (glagol + samostalnik)

ii. *»awwww kjuuut si napisala ^^«* (medmet + pridevnik)

Posebnosti pri določanju meje oz. vrste preklopa:

1. Upoštevanje meje stavka

a. Če je med znotrajstavčnim preklopom in medstavčnim preklopom oziroma dvema znotrajstavčnima preklopoma meja stavka, takšne preklope označimo vsakega zase.

i. *»Moj HTC Cha Cha umira in spet bom pristala na tipkanju po touch screenu .¶ I hate that.«*

ii. *»Jočem . @MateNemo there will be blood ! ¶ ;) btw jaz se kar pustim čakati za tisto kavo , čaj , pivo ... ;))«*

iii. *»adejno , a zdej še tviterji težijo z add your birthday ..¶ kinky , lubčki , kinky ...«*

b. Stavčne meje pa ne upoštevamo med dvema ali več medstavčnimi preklopi, ki jih na tem nivoju označimo skupno kot en medstavčni preklop.

i. »Pa smo tam. http://t.co/BvNxgwKiDH. 10min zamude. Evo je. Lušna punca. Tri punce. Fast driver. ¶ Smells like teen spirit. #romanjevljubljano«

2. Preklope znotraj heštegov, kjer je več besed napisanih skupaj, označimo, kot da bi bili zapisani razvezano (saj v WebbAnnu mogoče označiti samo dela besede).

i. Primere, kot je *»#dojenjemyass«*, označimo kot znotrajstavčni preklop.

ii. Primere, kot je *»#ilovemyjobs«*, označimo kot medstavčni preklop.

4 Jezik

Cilj:

Vsak kodni preklop ima pripisano oznako za jezik, iz katerega je.

Smernice:

1. Pri določanju jezika smiselno upoštevamo Osnovne smernice v delu, ki se nanaša na tujejezične prvine. Če v tujejezičnem kontekstu nastopa beseda, ki bi jo lahko interpretirali kot slovensko, štejemo, da je del prevladujočega jezika v tvitu, torej ni preklop

a. V primeru *»Dobro jutro uz pjesmu " Dobro jutro«* vse skupaj upoštevamo kot BHS, torej tvit ni relevanten za raziskavo in ga izbrišemo.

b. Primer *»@AlesValenko Read my lips : p-f-f-f-f .«* upoštevamo kot v celoti angleški.

2. Smiselno enako velja za besedne zveze, kjer je en element sicer v SSKJ/SP/Sloleks, vendar je glede na kontekst treba besedno zvezo razumeti kot zaključeno enoto: *buffalo burger, blog post, party pooper, scam email, usb hubi* (vse angleške besedne zveze)

3. Latinsko okrajšavo VS (vs., v.) označimo kot ANG.

4. Medmeti: tudi tu upoštevamo načelo vključenosti v SSKJ/SP/Sloleks.

a. Primeri slovenskih: *uf, eh, hm, pst, ah, mhm, no, vau, vav, mnja, jao, grr*

b. Primeri angleških: *nah, ugh, ouje, meh, pfft, khm, ahm, uhm, blah, erm, ehm, aw, huh*

5. Tujejezične prvine in primeri, ki jih **ne** označujemo kot kodni preklop:

a. **Besede, ki niso v SSKJ/SP/Sloleks, a jim ni mogoče določiti izvora**: *jupi, orng (orenk)*

b. **Neologizmi**: biznisira, slochi

c. **Imena**, npr. medijev, mest, glasbenih skupin, lokalov, prireditev, športnih tekmovanj

i. Izjema so države, saj imajo slovensko ustreznico:

ii. Če se ime pojavi znotraj preklopa, ga pustimo označenega kot del preklopa. V primeru *»too much Bieber all around«* denimo celotno zvezo označimo kot en znotrajstavčni preklop (ne kot *»too much Bieber all around«*).

d. **Avtomatski tviti oz. deli tvitov**

i. Samodejno generirani predlogi pri deljenju strani: via @youtube/@RevijaReporter

ii. Besedila, generirana v različnih aplikacijah ali na spletnih straneh, npr. *»I'm at @7_Burger_kamnik in Kamnik https://t.co/NdJbAPdAqA«* = Swarmapp; *»Just completed a 10.71 km run«* = Runkeeper; *»Drinking an Asahi Super Dry @ Shambala«* = Untappd.com

e. **Naslovi pesmi, člankov, spletnih strani**

i. *»In zadonela je iz zvocnikov :)) Iron Maiden - Phantom of the Opera http...«*

ii. *»Dobro jutro vrtičkarji ! Se zbira ideje za novo sezono ? Še ena : An Urban Micro-Farm in a Mobile Box http://t.co/UY0D11Ieo2 via @zite«*

f. **V celoti tujejezični tvit oz. preklopi med tujimi jeziki** (tviti brez slovenščine niso relevantni)

i. Tvite kot je *»Breaking : Diego Costa zove Marijanu . Vrati pare !«* izbrišemo.

g. **Tujejezične prvine v citiranem tvitu**

i. **»***No , tok da veste . Ne se zajebavat z mano . “ @conspiracyimage : Fact http://t... ”*«

ii. *»Lej, @freeeky, to je pa čist zate . RT @FascinatingVids: How to make your own turtle«* Tukaj besedo RT označimo kot preklop, saj jo uporabnik napiše sam, drugi del pa ni relevanten, saj je citiran tuj avtor, zato ga pustimo neoznačenega.

h. **Interference na nivoju skladnje**: *»Rešitelj od Svete trojice ?«* ali *»tut t-2 je bl shit v mojih izkusnjah«*

i. **Znak &**

j. Besede, kjer zapis ni tuja beseda, ampak samo s tujo črko zapisana slovenska beseda.

i. *»full pester promet«* ali *»Žal . Še huje je to , da raste exponentno . So pred dnevi krožli neki grafi . Poglejte , blizu te krivulje«*

5 Zapis

Cilj:

Vsak kodni preklop je označen kot povsem podomačen, delno podomačen ali nepodomačen.

Smernice:

1. Pri presoji zapisa upoštevamo osnovno obliko oz. jedro besede, ne morebitnih obrazil in končnic. Primeri, kot so *»downloadala«* ali *»trailerji«*, so torej označeni kot nepodomačeni.

a. Kot nepodomačene označimo tudi zapise, ki so zapisani nestandardno glede na izvirni jezik, a sprememba ne odraža »poslovenjenja«, temveč neformalni slog pisanja: *waaaaaaay better, pls, fcukers, I luv bieber, ima se, moze se*

2. (Delno) podomačen zapis označimo pri besedah, kjer je opazna sprememba izvirnega zapisa, sicer jih označimo kot nepodomačene (npr. pri besedah BSH). Delno prilagojen je zapis v naslednji primerih:

a. besede z izpuščenim končnim e-jem: *simpl, nop*

b. primeri, kjer kljub podomačitvi ostanejo tuji elementi: *gutenacht, Wuuhuu*

c. besede, ki so prilagojene (npr. izgubijo dvojni soglasnik), a ne odražajo v celoti »slovenske« izgovorjave: *Haštag, helo, Uber alles, jup*

d. »mešani primeri«, kjer je del besedne zveze povsem prilagojen, del pa je zapisan v skladu z izvirnikom (po logiki PP+NP=DP): *ofišl app, velkom back, sori to dissapoint you, dbest trailer*

6 Morfologija

Cilj:

Kodne preklope označimo glede na to, ali je tujejezična beseda uporabljena z dodano slovensko končnico (spreganje/sklanjanje), dodanim obrazilom (besedotvorje) ali z obema.

Smernice:

1. Večbesedne zveze (gl. 3. razdelek) obravnavamo kot celoto, in sicer dodamo ustrezno oznako glede na jedro besedne zveze. V primeru *»old boysov«* je denimo razvidna končnica.
2. Primeri za ROK (razvidna končnica in obrazilo):

a. *ofarbane, cheerleaderca, potegal, zalaufi, Prešaltu, Skenslaš, douchebagovskega*

b. Primere, kot sta »*laufer«* in »*marketingar«* v imenovalniku obravnavamo kot primere z ničto končnico, zato jih označimo z ROK

7 Besedna vrsta

Cilj:

Vsak kodni preklop ima pripisano ustrezno oznako za besedno vrsto.

Smernice:

1. Smiselno upoštevamo Osnovne smernice (Oblikoskladenjsko označevanje) in napotke v 3. razdelku teh Smernic.

a. Cele stavke ali daljše besedne zveze, ki jim ne moremo jasno določiti besedne vrste, označimo z »neuvrščeno« oz. N.

i. »*It was close*.« (MS) ali »*a zdej še tviterji težijo z add your birthday*« (ZS)

ii. Enako označimo primere, kot so *feeling comfortable, installing updates, watching basketball*

b. Samostalniške zveze z desnim ali levim prilastkom označimo s S: *selfie queen, your personal assistant*

c. Predložne zveze iz predloga in samostalnika označimo z D: *for your information; down under*

d. Pridevniške zveze z desnim ali levim prilastkom označimo s P: *free like a birdy ali waaaay better*

2. Pri problematičnih problemih, kjer pride do »mešanja kodov«, označimo samo tujo besedo in njeno besedno vrsto, ne glede na to, da se dejansko pojavi v predložni zvezi s slovenskim predlogom: *na repeat* (S), *v offline* (R), *na easy* (P)

3. Kot medmete (M) glede na njihovo funkcijo v besedilu označujemo tudi naslednje:

a. *please, sori, ofak, holy fuck, fakof, fuck yeah, damn, bummer*

b. Klepetalniške krajšave: lol, rofl, OMFG, BTW, WTF, FFS.

4. Kot členke (L) označujemo npr.:

a. da (*jp, jap, jep, jup*), ne (*nope, nah*) = členek (L);

b. **Izjema**: primere, kot sta »*jesssss*« ali »*Yes!!*« označimo kot medmete (upoštevamo funkcijo).

c. Besedi *kao* in *val(j)da*.

8 Oznake za označevanje in njihov pomen

Plast	Oznake	Opis
1. Jezik (odprta kategorija)	EN DE HBS SP LA AR FR IT PT	- angleščina - nemščina - srbohrvaščina - španščina - latinščina - arabščina - francoščina - italijanščina portugalščina
2. Vrsta preklopa	MS ZS	- medstavčni preklop - znotrajstavčni preklop
3. Zapis	NP DP PP	- nepodomačen - delno podomačen - povsem podomačen
4. Morfologija	NR RK RO ROK	- ni razvidna (brez slovenske končnice/obrazila) - razvidna iz končnice (spreganje/sklanjanje) - razvidna iz obrazila (besedotvorje) - razvidna iz obrazila in končnice
5. Besedna vrsta	G S P R O M L N D V K Z	- glagolska zveza - samostalniška zveza - pridevniška zveza - prislovna zveza - okrajšava - medmet - členek - neuvrščeno - predložna zveza - veznik - številnik - zaimek
6. Hešteg	hashtag	To oznako za razliko od ostalih označimo samo pri heštegih.