Smernice za označavanje nestandardnog hrvatskog i srpskog jezika

JANES + ReLDI

HR/SR: Nikola Ljubešić, Maja Miličević

SLO: Darja Fišer, Tomaž Erjavec, Jaka Čibej

JANES

ReLDI

Opšti principi

Potrebno je proveriti sve pojavnice (i po potrebi im dodati ili ispraviti oznake).

Prilikom provere treba se oslanjati na kontekst. Kada je uprkos kontekstu nemoguće odlučiti da li da se neka reč normalizuje ili ne, ne treba je normalizovati.

Ukoliko je ceo tvit na stranom jeziku, automatski generisan ili potpuno nerazumljiv, treba ga obrisati. U tvitovima koji se brišu ne treba označavati ništa drugo.

Tri sloja anotacije

1. segmentacija na rečenice
2. tokenizacija
3. normalizacija

Segmentacija na rečenice

Cilj:

Ispravna podela tvitova na rečenice, tako da je kraj svake rečenice označen.

Smernice:

1. U celom tvitu treba proveriti da li je automatska segmentacija ispravna.
2. Ako je deo tvita samostalna rečenica, tako ga treba i označiti. (“@mademoiselle_np .... reeeEEeci mu , prijateeeelj samoOoOoo , ne moora sve da znaaAaa .... ¶ :D ¶ Ju ' r velkm .”)

Segmentacija na rečenice

3. Merilo za određivanje kraja rečenice jeste pre svega znak interpunkcije kakav se uobičajeno koristi za označavanje kraja rečenice, npr. tačka, uzvičnik (hr uskličnik), upitnik i tri ili više tački (“Sta je ovo sunce ti poljubim ? ¶ Uvodna spica za smak sveta .... ¶ O zivote !!!”).
4. Ako ne postoji dobar razlog da nešto smatramo dvema rečenicama, treba ostaviti jednu (“Ne zavaravaj sebe ... jer navika ljubav nije .“, “Slucajno ?!?! duvam u pepeljaru punu pepela” → u oba primera ostaje jedna rečenica, jer tačke u sredini pre vrše funkciju zareza nego tačke, a ?!?! se odnose na prvu reč, a ne na čitavu rečenicu)
5. Kraj tvita je automatski ujedno i kraj rečenice, pa ga ne treba posebno označavati.

Tokenizacija

Cilj:

Ceo tvit je ispravno podeljen na pojavnice (reči ili znakove interpunkcije).

Smernice:

1. Na nivou tokenizacije treba spajati ili razdvajati one pojavnice koje je tokenizator pogrešno razdvojio ili spojio. Greške u tokenizaciji se najčešće javljaju zbog znakova interpunkcije i posebnih simbola, npr. ukoliko tokenizator podeli reč, crticu i flektivni nastavak na tri pojavnice, ili ne odvoji broj od oznake za procenat (“IBM - a” → “IBM-a”, “5%” → “5 %”).

Tokenizacija

2. Na nivou tokenizacije ne treba ispravljati bilo šta unutar pojavnica (npr. dijakritike ili nestandardne oblike), već treba samo spajati ili razdvajati pojavnice (“Federer-Djokovic” → “Federer - Djokovic”, “pred ' o” → “pred’o”).

3. Na nivou tokenizacije ne treba ispravljati one slučajeve pogrešnog sastavljenog i rastavljenog pisanja koje tokenizator nema na osnovu čega da prepozna (“neznam”, “od vra tan”). Takve slučajeve treba ispraviti na nivou normalizacije.

Normalizacija

Cilj:

Svakoj nestandardnoj reči pripisan je normalizovan oblik.

Smernice:

1. U celom tvitu treba proveriti da li su pojedinačne reči u skladu sa standardnim jezikom, a u slučaju da odstupaju od standarda, treba im pripisati normalizovane verzije.
2. Normalizovati treba samo na nivou reči: ne treba ispravljati red reči, sintaksičke odnose, interpunkciju, ili izbor leksike.

Normalizacija

3. Heštagove, korisnička imena, emotikone i elipse ne treba normalizovati (“#samokazem”, “@vikendholicarka”, “:)))”, “pi***”).
4. Reči ne treba normalizovati na sinonime iz standardnog jezika (npr. “ufotkao” ostaje “ufotkao”, ne ispravlja se u “fotografisao”).
5. Upotrebu velikih i malih slova ne treba ispravljati, bez obzira na to da li se radi o ličnim imenima, početku rečenice, akronimima, ili nečem drugom (“kako nas je zajebao putin” → “kako nas je zajebao putin”, “On je Američki predsednik” → “On je Američki predsednik”, “rt” → “rt”, “RT” → “RT”, “Lp” → “Lp”).
6. Reči u kojima nedostaju dijakritici treba normalizovati (“macka” → “mačka”, “medjutim” → “međutim”).

Normalizacija

7. Nestandardno napisane reči (npr. očigledne greške u kucanju, namerna ili slučajna fonetska prilagođavanja, regionalne varijante) treba normalizovati (“pocenjem” → “počinjem”, “svecki” → “svetski”, “numem” → “ne umem”, “kaće” → “kad će”).

8. Nestandardne skraćenice ne normalizuju se u puni oblik (“nmg”, “msm”, “Bgd”).

9. Punoznačne reči u kojima se ponavljaju slova pri normalizaciji treba skratiti na neproširenu varijantu (“noooć” → “noć”).

Normalizacija

10. Uzvike treba normalizovati na dva ponavljanja jednakih slogova (“hahahahaha” → “haha”), dva ili tri ponovljena pojedinačna slova treba ostaviti, a više ponovljenih slova skratiti na tri ponavljanja (“grr” → “grr”, “grrr” → “grrr”, “grrrr” → “grrr”; međutim “hahhaaahaa” → “haha”).

11. Reči za koje je nemoguće utvrditi da li je normalizacija potrebna ne treba normalizovati (“ne vise !” → “ne vise !”).