Smernice za označavanje nestandardnog hrvatskog i srpskog jezika

8. 12. 2015

Ovaj dokument sadrži pregled principa tokenizacije, segmentacije na rečenice i normalizacije reči u tvitovima. Tehničko sprovođenje ovih principa u sistemu WebAnno opisano je u posebnom dokumentu.

(hr) Jezična napomena: Većina je ovog dokumenta pisana srpskim jezikom. Pojmovi za koje se očekuje da ih neki od anotatora možda ne razumiju prevedeni su u zagradi na hrvatski s oznakom hr. Odlomci teksta pisani hrvatskim jezikom (poput ovog) počinju oznakom jezika u zagradi.

Opšti principi

1. Potrebno je proveriti sve pojavnice (i po potrebi im dodati ili ispraviti oznake).
2. 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.
3. 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.

Objašnjenje primera:

Primeri su u ovom dokumentu napisani kurzivom i pod navodnicima, npr. “*tvit*”. Ako pojavnica (ili njihovih normalizovanih oblika) ima više, odvojene su razmakom, bez znaka za spojeno pisanje, npr. “*npr .*”.

Prilikom ilustrovanja greške i odgovarajuće ispravke, sa leve strane strelice (→) navodi se pogrešan zapis, a sa desne strane ono u šta grešku treba ispraviti. Na primer, ako je u izvornom tvitu bilo napisano “*IBM-a”* i tokenizator je ovaj niz pogrešno podelio na tri pojavnice, zapis u WebAnno je “*IBM*\ -\ *a*”, a u smernicama stoji “*IBM - a”.* Ispravljanje na jednu pojavnicu navodi se kao “*IBM - a*” → “*IBM-a”.*

Segmentacija na rečenice

Cilj:

Ispravna podela tvitova na rečenice, tako da je kraj svake rečenice označen.[1]

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.[2] (“*@mademoiselle_np .... reeeEEeci mu , prijateeeelj samoOoOoo , ne moora sve da znaaAaa .... ¶ :D ¶ Ju ' r velkm .*”)

[1] Termin “rečenica” ovde koristimo kao oznaku za jedinicu koja bi u standardnom jeziku počela velikim slovom i završila se interpunkcijskim znakom poput tačke, uzvičnika ili upitnika.

[2] U smernicama je kraj rečenice radi lakšeg uočavanja označen simbolom ¶ .

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.

Složeniji primeri:

1. Tri tačke (ili više tačaka)
    a. Ponekad označavaju kraj rečenice ("*Jedu me komarci celo veče .... ¶ poludeću*").
    b. Ponekad označavaju elipsu ili pauzu u sredini rečenice – u tom slučaju ***nisu*** završni znak interpunkcije ("*Stomak kao da sam ... treci mesec trudnoce*").
2. Imena (@jovan), emotikoni (\o/ ili ☺) i heštagovi (#politika)
    a. Ako se pojavljuju u sredini rečenice, deo su te rečenice ("*Kada sam bila mala #Bajaga je bio sinonim za muziku .*", "*Ja ću sad sanjati ej kako @Moljac_Volonter s Angelom jedeš trešnje ...*").
    b. Ako se pojavljuju na početku tvita, računaju se kao deo prve rečenice ("*@KatarinaGlumac2 Idemo idemooooo mene nije sramota !*", "*#utisak je da su nam zabranili da imamo utisak .*")
    c. Ako stoje na mestu završnog znaka interpunkcije, označavaju kraj rečenice ("*@Ortodoksni hahaha pa druze to je i bila poenta :) ¶ poz*").
    d. Ako stoje iza završnog znaka interpunkcije, treba ih tretirati kao posebnu rečenicu ("*@CuvajMojeSrceee Uuu … ¶ :-) ¶ Tako znaci ? ¶ :-D ¶ Od sad cemo tako , a ? ¶ :-)*", "*" Nisam , braćala , stizala Upravljačko da spremim , izlazila sam zadnjih pet vikenada . " ¶ #VolimDaPrisluskujemUBusu*").
    e. Ako na kraju rečenice stoji niz od više imena, emotikona ili heštagova, kao kraj rečenice računa se poslednji element ("*Prijavi je tamo gde treba ;-) @tatauboji @Nase_Novine ¶ Prosledite!*" → kraj rečenice je ime @Nase_Novine).
3. Odsustvo odvojene oznake kraja rečenice
    a. Ukoliko između dve rečenice ne stoji bilo kakva oznaka (znak interpunkcije, emotikon…) kraj rečenice treba označiti na poslednjoj reči prve rečenice ("*home alone sam i cuje se nesto napolju ¶ LEA POJEDI IH*" → kraj rečenice je "*napolju*")

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 %*").

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.

Složeniji primeri:

1. Emotikoni
    a. Ako je tokenizator podelio emotikon na više pojavnica, treba ih spojiti u jednu ("*: |*" → "*:|*") .
2. Skraćenice sa tačkom
    a. Skraćenica i njena tačka predstavljaju jednu pojavnicu ("*dr .*" → "*dr.*") .
    b. Ako tokenizator nije prepoznao da je u pitanju skraćenica, tačku će tretirati kao kraj rečenice. U tom slučaju treba ispraviti i tokenizaciju i segmentaciju (ako skraćenica stoji na kraju rečenice, cela pojavnica dobija oznaku kraja rečenice).
3. Pojavnice napisane zajedno
    a. Pojavnice koje su zbog prisustva interpunkcijskih znakova pogrešno spojene treba razdvojiti na više pojavnica ("*Federer-Djokovic*" → "*Federer - Djokovic*").
    b. Ukoliko u nekom tvitu postoji pojavnica koja bi zahtevala podelu na više od četiri posebne pojavnice, čitav tvit treba označiti za brisanje ("*stvari-koje-necu-raditi-sledece-godine*").
    c. Na nivou tokenizacije ***ne*** treba ispravljati pojavnice koje bi trebalo pisati odvojeno, ali su (namerno ili slučajno) napisane zajedno ("*neznam*"). Takve pojavnice ispravljaju se na nivou normalizacije.
4. Pojavnice napisane odvojeno
    a. Pojavnice čiji su sastavni elementi pogrešno označeni kao odvojeni treba spojiti (npr. "*. . .*" → "*...*").
    b. Pravilo spajanja ***ne*** važi za reči koje bi trebalo pisati zajedno, ali su (namerno ili slučajno) napisane odvojeno ("*od vra tan*" → "*odvratan*"). Takve pojavnice se ispravljaju na nivou normalizacije.
5. Nizovi sastavljeni od brojeva, crtica i nastavaka, ili od brojeva i simbola
    a. Pogrešno tokenizovane nizove poput "*2 x"*, "*3 x"*, "*13 - i"*, "*12 - og"* treba spojiti u jednu pojavnicu ("*2x*", "*3x*", "*13-i*", "*12-og*"). To ***ne*** važi za merne jedinice i druge simbole ("*20 km", "40 €", "50 %", "12 +"*), koji predstavljaju posebne pojavnice i koji treba da budu odvojeni od brojeva.
    b. U primerima tipa "*60 ih*" na nivou tokenizacije pojavnice treba samo spojiti ("*60ih"*), dok se crtica dodaje prilikom normalizacije ("*60-ih"*).
    c. Numeričke sekvence poput datuma, vremena, sportskih rezultata i brojeva smatramo jednom pojavnicom ("*23. 12.*", "23:45", "2 : 5")

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.
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*").
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 nevlastitih imenica normalizuju se u puni oblik ("*nmg*" → "*ne mogu*", "*msm*" → "*mislim*"). Nestandardne skraćenice vlastitih imenica ("YT", "fb", "*Bgd*") ostaju nepromijenjene.
9. Punoznačne reči u kojima se ponavljaju slova pri normalizaciji treba skratiti na neproširenu varijantu ("*noooć*" → "*noć*").
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 !*").

Složeniji primeri:

1. Varijante zapisa
    a. Reči koje se mogu interpretirati na više načina treba razjasniti uz pomoć konteksta ("*ko*" → "*ko*", "*tko*", "*kao*"). Ako je to nemoguće, treba ih ostaviti kako jesu.

b. U slučaju postojanja dubleta treba dopuštati oba oblika (“*ujutru”* → *“ujutru”, “ujutro”* → *“ujutro”*, “*stricem”* → *“stricem”, “stricom*” → “*stricom*”)

2. Posebni primeri
   a. Treba ispravljati nepravilne oblike glagola *biti* u 1. licu jednine i množine i 2. licu množine aorista (“*bi*” → “*bih*”, “*bi*” → “*bismo*”, “*bi*” → “*biste*”).
   b. Treba unositi potrebne glasovne promene tamo gde ih nema (“*predposlednji*” → “*pretposlednji*”, “*stanbeni*” → “*stambeni*”, “*burekdžinica*” → “*buregdžinica*”, “*mislioc*” → “*mislilac*”)
   c. Treba normalizovati sažete samoglasnike (“došo” → “došao”, “k’o” → “kao”)
   d. U srpskom ***ne*** treba ispravljati pogrešnu upotrebu glagola *trebati* (“*za sledeci vikend trebao bi da sam u Bg*” → “*za sledeći vikend trebao bi da sam u Bg*”, ne “*za sledeći vikend trebalo bi da sam u Bg*”), prisvojnih zamenica (“*Zelim moj glas nazad*” → “*Želim moj glas nazad*”, ne “*Želim svoj glas nazad*”), prideva *zadnji* i *poslednji*, priloga *mnogo* i *puno*, i sl.
   e. (hr) U hrvatskom treba poštivati pravilo krnjeg infinitiva. Tako u “*raditi će*” “*raditi*” treba normalizirati na “*radit*”, a u “*ne mogu vjerovat*” “*vjerovat*” treba normalizirati na “*vjerovati*”.
   f. (hr) U hrvatskom sintetički futur treba normalizirati na nesintetički.
   g. (hr) U hrvatskom “*Ne bum išel*” i sl. normaliziramo na “*Ne budem išao*”, bez obzira na činjenicu da se normalizirana sekvenca ne koristi.
   h. (hr) U hrvatskom ne treba normalizirati oblik prijedloga *s/sa* kao ni druge nenavedene fenomene.
3. Flektivni nastavci
   a. Nestandardne nastavke treba normalizovati na standardne (“*oni volu*” → “*oni vole*”).
   b. Skraćenice koje imaju nastavke pripisane ili odvojene na nestandardni način treba normalizovati crticom (*“TVu*” → “*TV-u*”, “*tv.u*” → “*tv-u*”).
4. Elementi iz stranih jezika
   a. (hr) U hrvatskom se za elemente stranog jezika prate pravila iz Pravopisa IHJJ-a.
   b. U srpskom strane reči koje su fonetski transkribovane treba ostaviti onako kako su napisane, osim ako se ne proceni da je u pitanju greška u kucanju (“*knekšna*” → “*konekšna*”).
   c. Flektivne oblike stranih reči koji čuvaju elemente izvornog pisanja (nisu u celini fonetski transkribovane) treba ostaviti onako kako su napisani, osim ako se ne proceni da je u pitanju greška u kucanju (“*fitnessa*” → “*fitnessa*”, *“googlati*” → “*googlati*”, *“googglati*” → “*googlati*”).
   d. U srpskom strane reči koje su napisane izvorno (*share, like*) treba ostaviti u izvornom zapisu. Očigledne greške u kucanju treba normalizovati u standardni (strani) oblik (“*chessburger*” → “*cheeseburger*”).

e. U srpskom transkribovana lična imena treba ostaviti onako kako su napisana. Treba ispraviti jedino očigledne greške u kucanju ("*Čomsky*" → "*Čomski*").
f. Očigledno pogrešno napisana strana lična imena treba ispraviti ("*Tweeter*" → "*Twitter*").
g. Punoznačne strane reči u kojima se ponavljaju slova pri normalizaciji treba skratiti na neproširenu varijantu ("lovveeeeee" → "love").
h. Strani uzvici normaliziraju se kao i nestrani.

5. Strane skraćenice
   a. Strane skraćenice tipa *thx, tnx*, *srsly* treba ostaviti onako kako su napisane.
6. Strana slova
   a. Domaće reči koje su delom napisane stranim slovima ili kombinacijama slova treba normalizovati u standardni oblik ("*na faxu*" → "*na faksu*", "*qq lele*" → "*kuku lele*", "*loodilo*" → "*ludilo*", "*shkolitza*" → "*školica*").
7. Reči napisane spojeno
   a. Reči (slučajno ili namerno) pogrešno napisane zajedno treba odvojiti na više pojavnica ("*nebi*" → "*ne bi*", "*jel*" → "*je li*").
   b. Reči koje su napisane zajedno i sadrže nestandardne varijante treba odvojiti na više pojavnica i normalizovati (*"kaksi → "kako si",* "*kaće*" → "*kad će*"; "*ćmić ?*" → "*hoćemo ići ?*").
   c. Varijantne oblike (npr. "*zauzvrat*", "*za uzvrat*") treba ostaviti kako jesu.
8. Skraćenice
   a. Očigledne skraćenice treba normalizovati dodavanjem tačke (*"tj" → "tj."*).
   b. Reči koje sadrže brojeve treba normalizovati u njihove slovne ekvivalente ("*lu3ja*" → "*lutrija*", "*is3povati*" → "*istripovati*", "*gr8*" → "*great*").