JANES

Smernice za označevanje računalniško posredovane komunikacije: tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladenjsko označevanje

v1.0

Avtorji: Jaka Čibej, Špela Arhar Holdt, Tomaž Erjavec, Darja Fišer, Katja Zupan

Datum zadnje spremembe: 2016-12-21

Kazalo vsebine

0 Uvod

Smernice podajajo jezikoslovna načela za tokenizacijo, stavčno segmentacijo in normalizacijo besed v računalniško posredovani komunikaciji, kot jo najdemo v tvitih, blogih, spletnih komentarjih na novice itd. Tehnična izvedba teh oznak v orodju WebAnno je opisana v svojem priročniku.

1 Splošna načela

1. Preverjamo in po potrebi dodamo oz. popravimo oznake vseh pojavnic ne glede na to, ali imajo avtomatsko pripisano oznako ali ne.
2. Pri popravkih upoštevamo kontekst. Kadar smo v dvomu, se posvetujemo z referenčnimi viri (glej Dodatek 1).
3. Če se kljub temu ne moremo odločiti, ali besedo normaliziramo ali ne, je ne normaliziramo.
4. Če je tvit v celoti v tujem jeziku, avtomatsko generiran ali pa popolnoma nerazumljiv, tvit izbrišemo (glej tehnična navodila). V takšnem tvitu ne označujemo ničesar drugega.

Zapis ponazoritvenih primerov:

Primeri so zapisani ležeče in v narekovajih, npr. "*jst*". Če je pojavnic (ali njihovih normalizacij) več, so med seboj ločene s presledkom, ponavadi brez pripisanega znaka za stičnost, npr. "*življ .*".

Kjer ponazorimo napako v oznakah in njen popravek, je na levi strani puščice (→) napačen zapis v WebAnnu, na desni strani pa popravek. Če je npr. v izvornem tvitu zapisano "*IBM-ja*" in bi tokenizator to napačno ločil na tri pojavnice (*IBM*, - in *ja*), je to v WebAnnu izpisano kot "*IBM\ -\ ja*", v pričujočem priročniku pa to zapišemo kar "*IBM - ja*". Ker je ta zapis treba združiti v eno pojavnico, zapišemo cel primer popravka kot "*IBM - ja*" → "*IBM-ja*".

2 Stavčna segmentacija

Cilj:
Besedila so pravilno ločena na stavke,[1] 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[2] (*"@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 !*").

[1] "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).
[2] V smernicah konec stavka za lažjo predstavo označujemo s simbolom ¶.

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), emotikoni (\o/) ali emojiji (👦) in heštegi (#hešteg)
    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 tvita, jih obravnavamo kot del prvega stavka (*"@TadejTrcekTITO @lucijausaj @JJansaSDS titek, ne seri. odv. častno razsodišče je JE zgolj za odvetnike."*).
    c. Če nadomeščajo končno ločilo, zaznamujejo konec stavka (*"kot da je to važn :)) ¶ nobenga to ne briga vec sploh"*).
    d. Če sledijo končnemu ločilu, jih obravnavamo kot samostojen stavek (*"Sonce, sneg in pot pod noge! ¶ :) ¶ Gremo v hribe!"*).
    e. Če je pri koncu stavka nanizanih več imen, emotikonov ali heštegov, za konec stavka velja zadnji element (*"itak ne morm sploh keša dvignt :) @tibonalta #broke"* → konec stavka je hešteg #broke).

3 Tokenizacija

Cilj:
Tvit je pravilno ločen na pojavnice (besede ali ločila).

Smernice:

1. Na ravni tokenizacije samo razdružujemo ali združujemo tiste pojavnice, ki jih je napačno združil ali ločil tokenizator. Napake se pojavljajo zaradi ločil in posebnih znakov (simbolov), npr. tokenizator loči besedo, vezaj in končnico na tri pojavnice ali ne loči številke od znaka za odstotke ("*IBM - ja*" → "*IBM-ja*", "*5%*" → "*5 %*").
2. Na ravni tokenizacije ničesar ne popravljamo v zapisu (šumnikov, obrazil ipd.), temveč samo združujemo ali razdružujemo ("*življ .*" → "*življ.*", "*Nemčija-Grčija*" → "*Nemčija - Grčija*").
3. Pri tokenizaciji ne popravljamo namenoma ali pomotoma skupaj oz. narazen zapisanih besed ("hodildomov", "porka duš"). Te razdružimo ali združimo na ravni normalizacije.

Težavni primeri:

1. Emotikoni
    a. Če je tokenizator emotikon razdelil na več pojavnic, ga združimo (“: |” → “:|”).
    b. Več zaporednih emotikonov obravnavamo kot eno pojavnico in jih ne delimo (*“:):):)”*). Če jih je tokenizator razdelil, v besedilu pa so stični, jih združimo v eno pojavnico (*“:)\ :**” → ”:):**”*).
2. Okrajšave
    a. Okrajšava in njena pika naj bosta ena pojavnica (*“dr .” → “dr.”*).
    b. Če tokenizator okrajšave ni prepoznal, bo njeno piko ločil v posebno pojavnico in jo tudi obravnaval kot konec stavka. V tem primeru popravimo tako tokenizacijo kot segmentacijo (če pa okrajšava dejansko konča stavek, z njo stavek tudi zaključimo).
3. Skupaj zapisani nizi
    a. Zaradi ločil napačno skupaj zapisane nize ločimo na več pojavnic (*“turčija-grčija” → “turčija - grčija”*).
    b. Pri tokenizaciji ne popravljamo skupaj zapisanih nizov z manjkajočimi presledki (“*hodildomov*”). Te popravljamo pri normalizaciji.
4. Narazen napisani nizi
    a. Nize, ki jih je tokenizator napačno ločil na več pojavnic (npr. tropičje na tri posamezne pike), združimo. To *ne* velja za besede, ki jih je napačno narazen napisal avtor (npr. *“porka duš”*). Te združimo na ravni normalizacije (*“porka duš” → “porkaduš”*).
5. Nizi iz številk in končnih črk ali simbolov
    a. Napačno tokenizirane nize tipa *“2 x”*, *“3 x”*, *“13 - ih”*, *“12 - ih"*, združujemo v eno pojavnico (*“2x”, “3x”, “13-ih”, “12-ih”*). To ne velja za enote in druge simbole (*“20 km”, “40 €”, “50 %”, “12 +”*), ki naj bodo ločene pojavnice ne glede na to, ali so zapisane skupaj ali narazen.
    b. V primerih tipa “*27\ ih*” na ravni tokenizacije pojavnici samo združimo (“*27ih*”), vezaj pa dodamo šele pri normalizaciji (“*27-ih*”).

4 Normalizacija

Cilj:

Vsaka nestandardna beseda ima pripisano svojo normalizirano ustreznico.

Smernice:

1. V celotnem tvitu preverimo, ali so besede zapisane v skladu s standardom (gl. Dodatek 1), in jim v primeru, da od njega odstopajo, pripišemo normalizirane ustreznice.
2. Normaliziramo samo na nivoju besed: ne spreminjamo besednega reda, skladenjskih razmerij, ločil, stičnosti ali izbora besedišča.
3. Heštegov, uporabniških imen, emotikonov, emojijev in elips ne normaliziramo (“#lepa-drevesa”, “@janez”, “:)))”, “”, “pi***”). Besede normaliziramo samo ortografsko in

jim ne pripisujemo standardnih sopomenk ("*una", "guna"* → *"ona"* in ne **"tista"; "pocahnu"* → *"pocahnal"* in ne **"označil"*).

4. Besedam ne pripisujemo standardnih sopomenk (npr. "*pocahnu*" → "*pocahnal*" in ne **"označil*", "*pofarbat*" → "*pofarbati*" in ne **"pobarvati*", "*pucajne*" → "*pucanje*" in ne **"čiščenje*").
5. Nestandardno zapisane besede (npr. zatipke in regionalne različice) normaliziramo ("*polgedal*" → "*pogledal*", "*knižnica*" → "*knjižnica*", "*hodildomov*" → "*hodil domov*", "*hodu*" → "*hodil*").
6. Šumnike normaliziramo ("*macka*" → "*mačka*").
7. Medmete normaliziramo v dve ponovitvi enakih zlogov ("*hahahahaha*" → "*haha*"), ponovljene črke pa skrajšamo na največ tri ponovitve ("*grr*" → "*grr*", "*grrr*" → "*grrr*", "*grrrr*" → "*grrr*", vendar "*hahhaaahaa*" → "*haha*").
8. Polnopomenske besede s ponovljenimi črkami skrajšamo na nepodaljšano različico ("*Mooojcaaa*" → "*Mojca*").
9. Besed, za katere ni mogoče ugotoviti, ali je normalizacija potrebna, *ne* normaliziramo ("*Vcerajsnji problem je resen*" → "*resen*").
10. Besednih zvez z nesklonljivim levim prilastkom ne normaliziramo z vezajem ("SD kartica" → "SD kartica", ne *"SD-kartica").

Težavni primeri:

1. Velike začetnice
    a. Nepravilno rabo velike začetnice normaliziramo ("miha" → "Miha", ("*On je Ameriški predsednik*" → "On je ameriški predsednik").
    b. Prvo besedo v stavku normaliziramo z veliko začetnico samo, če je lastno ime. V ostalih primerih začetnice na začetku stavka ne spreminjamo ("Šel je v ljubljano" → "Šel je v Ljubljano", "šel je v ljubljano" → "šel je v Ljubljano", "ljubljana je lepa" → "Ljubljana je lepa").
    c. Če pri večbesednih lastnih imenih ne moremo ugotoviti, ali se vsi elementi pišejo z veliko začetnico, dvoumne pustimo pri miru ("*pr'Kovac*" → "*pri Kovaču*").
    d. Zapisov z velikimi črkami (ZASTONJ, SREČNO) ne popravljamo, razen če gre za lastno ime ("DUNAJ" → "Dunaj") ali pa besedo na začetku stavka, ki jo je treba normalizirati ("VIDM DA SI PAMETEN" – "vidim DA SI PAMETEN").
2. Variantnost zapisa
    a. Nestandardne besede, ki imajo več kot eno interpretacijo, razdvoumimo s pomočjo sobesedila ("*k*" → "*ker*", "*ki*", "*ko*"; "*ko*" → "*ko*", "*kot*"). Če to ni mogoče, jih ne normaliziramo.
    b. Besedam, ki nimajo standardne ustreznice, a se zapisujejo v več variantah, kot normalizirano obliko pripišemo najpogostejšo različico v korpusu, gl. Dodatek 1 (npr. "*fouš*", "*fauš*", "*favš*" – najpogostejša različica je "*fouš*").
    c. Pri nekaterih nestandardnih besedah si obliko v standardni slovenščini sicer lahko zamislimo, a ni v uporabi. V takšnih primerih besede ne normaliziramo v

namišljeno standardno obliko (“*krigl*” → *“*krigelj*”, “*reglc*” → *“*regelc*” / *”*regeljc*”, “*Prešerc*” → *“*Prešerec*”).

d. Glagole z variantnim zapisom predpone *z-iz*, *z-za* ipd. normaliziramo v najbližjo obliko (“je zgledu” → “je zgledal”; “sm zvedu” → “sem zvedel”; “je izgubla” → “je izgubila”; “se mi je zluštal” → “se mi je zluštalo”).

e. V primeru dvojnic dopuščamo obe obliki (“*bojo*” → “*bojo*”, “*bodo*” → “*bodo*”, “*zadanem*” → “*zadanem*”, “*prizadanem*” → “*prizadanem*”, “*softverja*” → “*softverja*”).

3. Posebni primeri

a. Ne normaliziramo napačne rabe predlogov “*s*”/“*z*”, besede “*en*”, glagolov “*moči*”/“*morati*”, “*rabiti*”/“*potrebovati*” ipd. (*“z slonom”, “en je reku”, “tu bi se mogla strinjat”, “danes ne rabim laptopa”*).

b. Napačne rabe sklona ne normaliziramo, kadar ne gre za nestandardno obrazilo, temveč za napako na ravni skladnje (“ne uporabljam rodilnik”, “z 240 milijonov”, “tem pnevmatike pa itak da morš prek neta nabavt”, “a to je zdej klasika v starim firmam”).

c. Pri pogostih besedah *pol*, *kok*/*kolk*, *tok*/*tolk*, *tko* upoštevamo naslednje normalizacije: “*pol*” → “*potem*”, *štajerski* “*te*” → “*potem*”, “*kok*” / “*kolk*” → “*koliko*”, “*tok*” / “*tolk*” → “*toliko*”, “*tko*” → “*tako*”.

d. Pri zaimkih in prislovih s členico le (npr. *tale, tele, tule, tukajle*) členico upoštevamo tudi v normalizirani obliki *(“tehle” → “tehle”, “tukile” → “tukajle”*).

e. Če je prva beseda v stavku kapitalizirana in jo je treba normalizirati (in ne gre za lastno ime), ji pripišemo normalizirano oznako z malo začetnico (*“Zlo je bedno tuki” → “zelo je bedno tukaj”*).

f. Pri akronimih[3] pustimo izvorno kapitalizacijo ne glede na to, ali so pisani z velikimi, malimi ali mešanimi črkami (*“rt”, “RT”, “Lp”, “lp”, “LP”*).

g. Izjema so akronimi, ki so v celoti pisani z malimi črkami in so rabljeni kot lastno ime. Te pišemo z veliko začetnico (*“Pridi v kud” → “Pridi v Kud”*).

h. Nepopolno zapisanih simbolov ne popravljamo (*38 C → 38 C*, ne *38 °C*).

4. Obrazila

a. Nestandardna obrazila normaliziramo v standardna (*“na Ptuji” → “na Ptuju”, “se spomne” → “se spomni”*).

b. Napačno rabljene nedoločnike in namenilnike popravljamo.

c. Kratice, ki imajo obrazila pripisana ali ločena na nestandarden način, normaliziramo z vezajem (*“KUDu” → “KUD-u”, “tv.ju” → “tv-ju”*).

5. Tujejezične prvine

[3] Akronime razumemo kot besede, sestavljene iz prvih črk večbesednih zvez (lep pozdrav – lp) oziroma iz izbranih črk daljše besede (retweet – rt).

a. Tujejezične besede, ki so se poslovenile po zapisu, obravnavamo kot druge variantne nestandardne besede (*"knekšna"* → *"konekšna",* kot bolj pogosta oblika; *"mučas hvalas"* → *"mučas hvala"*).
b. Pregibane tujejezične besede, ki ohranjajo elemente izvornega zapisa (torej niso v celoti fonetično zapisane), normaliziramo v najpogostejšo med različicami s tujimi prvinami zapisa v korpusu JANES ("*sharati*" → "*sherati*" in ne *"*šerati*", "*fittnessa*" → "*fitnessa*" in ne *"*fitnesa*", "*pogoogglati*" → "*pogooglati*" in ne *"*poguglati*").
c. Tujejezične prvine, ki so zapisane citatno (*share*, *like*), pustimo pri miru. Zatipkane (*chessburger*) normaliziramo v standardne ustreznice ("*chessburger*" → "*cheeseburger*").
d. Pri normalizaciji pregibanih tujih lastnih imen se držimo Slovenskega pravopisa (npr. "*Godoja*" → "*Godota*").
e. Napačno zapisana lastna imena popravljamo ("*Tweeter*" → "*Twitter*").

6. Tujejezične okrajšave
   a. Tujejezičnim okrajšavam tipa thx, srsly kot normalizirano obliko pripišemo najpogostejšo obliko v korpusu JANES ("*thx*" → "*tnx*", "*srly*" → "*srsly*").
7. Tuje črke v slovenskih besedah
   a. Slovenske besede, ki so zapisane s tujimi črkami, normalizirano v različico, ki je zapisana v skladu s splošno veljavnim standardom (*"za faxom"* → *"za faksom", "qrba"* → *"kurba", "kaxi"* → *"kako si"*).
8. Skupaj zapisane besede
   a. Napačno skupaj zapisane besede pri normalizaciji ločimo ("*hodildomov*" → "*hodil domov*", "*nebi*" → "*ne bi*").
   b. Variantne oblike (npr. *donedavna, do nedavna; karkoli, kar koli*) pustimo pri miru.
   c. Skupaj zapisane besede, ki vsebujejo nestandardne različice, ločimo na več pojavnic in jih hkrati tudi normaliziramo ("*čes*" → "*če si*", "*čej*" → "*če je*").
9. Okrajšave
   a. Normaliziramo le očitne okrajšave, in sicer z dodajanjem pike (*"devalv"* → *"devalv."*).
   b. Okrajšave, ki vsebujejo cifre, normaliziramo v njihove standardne ustreznice ("*ju3*" → "*jutri*", "*gr8*" → "*great*").

5 Lematizacija

Cilj:

Vsaka beseda ima pripisano svojo lemo (osnovno obliko).

Smernice:

1. Načeloma sledimo smernicam za označevanje ssj500k.
2. Vsaki besedi v tvitu osnovno obliko določimo na podlagi normalizirane (in ne izvirne!) oblike ("*Btc-ja*" → "*Btc*", "*Ff-ju*" → "*Ff*").
3. Pri lematizaciji odstranimo odvečne vezaje ("*iTunes-ih*" → "*iTunes*", "*sem share-al*" → "*biti shareati*"; "*shar-am*" → "*sharati*").
4. Funkcijske besede (predvsem zaimki) imajo včasih nepričakovano lemo (npr. "*nikogar*" → "*nihče*", "*čemer*" → "*kaj*", "*midva*" → "*jaz*", "*vidva*" → "*ti*"). Čeprav bodo večinoma že avtomatsko pravilno označene, so možne tudi napake. Če smo v dvomu, pogledamo v katerega od referenčnih virov (gl. Dodatek 1).
5. Pri emotikonih, emojijih, uporabniških imenih in ključnikih je lema enaka izvorni obliki.
6. Preverimo, ali so URL-naslovi lematizirani s svojo domeno, in jih po potrebi popravimo (*"https://www.youtube.com/watch?v=ED4UGIJTvV8"* → *"youtube.com"*).
7. Elipse poskušamo lematizirati glede na končnice (*"v p***i"* → *"v p***a"*). Če to ni izvedljivo, jih lematiziramo v normalizirano obliko.
8. Medmete s ponovitvami črk, ki so bili normalizirani na različice z največ tremi ponovitvami črk (*"looooool"* → *"loool"*), lematiziramo v normalizirano obliko ("*loool*" → "*loool*", "*grr*" → "*grr*").

Težavni primeri:

1. Tujejezične prvine
   a. Pri tujejezičnih prvinah, ki so pisane povsem citatno in obenem niso lastna imena, iz njihove zapisane oblike pa niso razvidna slovenska obrazila, je lema enaka izvorni obliki ("*jailbreak*" → "*jailbreak*", "*hrvatskog*" → "*hrvatskog*").
   b. Pri lemah tujejezičnih lastnih imen upoštevamo začetnice v normalizirani obliki ("*Candy Crushu*" → "*Candy Crush*", "*Flappy Birda*" → "*Flappy Bird*").
   c. Pri ostalih tujejezičnih prvinah pripišemo lemo v skladu s slovenskimi oblikoslovnimi načeli ("*benchmarki*" → "*benchmark*", "*shaderi*" → "*shader*", "*chatala*" → "*chatati*", "*chetaš*" → "*chetati*", "*stejtmenta*" → "*stejtment*", "*dosadne*" → "*dosaden*", ne *"*dosadan*"). O dvoumnih primerih se posvetujemo.
2. Lastna imena
   a. Imena programov in firm obravnavamo kot lastna imena. Pri lematizaciji drugih stvarnih lastnih imen (pesmi, zakonov ipd.) upoštevamo smernice za označevanje učnega korpusa ssj500k.

3. Kratice in okrajšave

a. Občnoimenske kratice (*LP*, *Rtm*, *lol*, *Rofl*) lematiziramo v različico brez velikih črk ("*LP*" → "*lp*", "*Rtm*" → "*rtm*", "*lol*" → "*lol*", "*Rofl*" → "*rofl*").
b. Pri lastnoimenskih kraticah in okrajšavah sledimo normalizirani obliki ("*Csd-ja*" → "*Csd*", "*devalv.*" → "*devalv.*", "*drž.*" → "*drž.*", "*N. Zelandije*" → "*N. Zelandija*", "*v J. Torturru*" → "*v J. Torturro*").
c. Če se kratica pregiba, lemo določimo v skladu s slovenskimi oblikoslovnimi načeli ("*Sds-u*" → "*Sds*", "*rd-ja*" → "*rd*").

6 Oblikoskladenjsko označevanje

Cilj:

Vsaka beseda (z izjemo posebnih primerov, opisanih spodaj) ima pripisano oblikoskladenjsko oznako JOS.

Smernice:

1. Vsaki normalizirani obliki besede glede na smernice, referenčne vire (gl. Dodatek 1) in kontekst pripišemo njeno oblikoskladenjsko oznako JOS. Oznak **Nt** (zatipk) in **Np** (tokenizacijska napaka) **ne** uporabljamo!
2. Pri dvoumnih interpretacijah pripišemo najnevtralnejšo obliko glede na (predvidevani) kontekst.
    a. "*nerazdružljivo sem mislil :)*" – *nerazdružljivo* kot prislov/pridevnik ženskega ali srednjega spola
3. Prvinam, značilnim za spletno komunikacijo, pripisujemo naslednje oznake:
    a. **Nw**: spletni in e-poštni naslovi (*www.prevajalstvo.net*, *lizika85@yahoo.com*), delčki URL-jev (*.si*, *.de*, *.uk*)
    b. **Nh**: ključniki (*#gremodelat*)
    c. **Na**: uporabniška imena (*@uporabnik*)
    d. **Ne**: emotikoni in emojiji (=D, ☺)
4. Klepetalniške krajšave (npr. *lol*, *rofl*, *OMFG*, *gg*) obravnavamo kot medmete (M).

Težavni primeri:

1. Tujejezične prvine
    a. Če je celoten stavek ali celovita oz. ločena skladenjska enota v tujem jeziku, vse besede znotraj te enote označimo kot **Nj** (v spodnjih primerih označeno s krepkim tiskom):
        i. zakaj se še vedno z nekimi belkami ukvarjaš :) **that shit is played out** :)
        ii. In bo vse zaštekal, **I guess**.
        iii. sorazmernost (zahtevaj toliko kot rabiš) ali izsiljevanje (**take it or leave it**)
        iv. hehe, ne :) **Borči ain't got time for that**
    b. Tujejezične prvine, ki so zapisane citatno, označimo kot Nj, razen če gre za lastno ime (*Candy Crush*, *Chuck Norris*) ali če so iz oblike razvidna slovenska

obrazila (*page-a*, *benchmarki*, *shaderi*) oz. se prvina v zapisu fonološko prilagaja (*kjut*, *kveščn*, *updejtati*).

i. Izjema so besede, ki so že prisotne v Sloleksu ali Franu. Te označimo kot običajne slovenske besede (npr. *printer*, *online*, *mail*).

ii. Izjema so tudi besede tipa *comp*, ki so sicer zapisane v skladu z načeli izvirnega jezika, a v izvirniku v tej obliki ne obstajajo. Obravnavamo jih kot slovenske besede.

iii. Če v tujejezičnem kontekstu nastopa beseda, ki bi jo lahko interpretirali kot slovensko, jo kljub temu označimo kot Nj (npr. *Oh* v primeru *Oh, you British peeps!*).

2. Lastna imena
   a. Alfanumerični dodatki ob lastnih imenih niso lastna imena, temveč občna.
      i. Xperia [**Sl**zei] G3600 [**So**mei] 4S [**So**mei]
      ii. Gigabyte [**Sl**mei] GTX660 [**So**mei] 2 [Kag] GB [**So**mei]
3. Nestandardna ali nepričakovana raba sklonov
   a. Pri pripisovanju oblikoskladenjskih oznak upoštevamo sklon, ki je uporabljen v izvirniku, ne glede na to, ali je skladenjsko ustrezen ali ne.
      i. nisem oblikovala *intergalaktično brisačo* [tožilnik, ne rodilnik]
      ii. največji šovmejker v *zgodovina* [imenovalnik, ne mestnik]
      iii. tko kt pr *Harry Potterju* [imenovalnik - mestnik; ne mestnik - mestnik]
      iv. Naloga OVS skrb za varnost vojakov v tujini in le VOJAŠKA kontra *obveščevalno* [Ppnze**t**] dejavnost doma.
      v. V eni uri shoppinga sem od *prodajalka* [Soze**i**] in prodajalcev nabrala: love, sweetheart, hun, darling in doll.
4. Živost
   a. Če se samostalnik moškega spola v tožilniku pregiba, kot da ima kategorijo živosti, to upoštevamo tudi pri oblikoskladenjski oznaki, ne glede na to, ali lema to predvideva ali ne.
      i. **daljinca** [Somet**d**] od vodne pojstle pofotkaj;
      ii. jaz vem za **kvalitetnega centra** [Somet**d**] z nba izkusnjami
5. Kratice in okrajšave
   a. Besede, ki so bile normalizirane s piko (*devalv. drž. tož. slo.*), označimo kot okrajšave (O).
   b. Kratice – npr. *SDS*, *BTC*, *rd* (kot rojstni dan), *JJ* – obravnavamo kot samostalnike in jim pripišemo oznako v skladu z njihovo vlogo v stavku.
      i. Ni član SDS [Slze**r**] (rodilnik, ne imenovalnik)
6. Nepopisani členki
   a. V spletnih besedilih naletimo na besede, ki sicer niso popisane v referenčnih virih, a jih obravnavamo kot členke (L). To so npr. besede *glih*, *eto*, *evo* in *ajde*. O dvoumnih primerih se posvetujemo.

7 Dodatek: Referenčni viri za normalizacijo

7.1 Oblikoskladenjsko označevanje in lematizacija

1. Iskanje po korpusu ssj500k
2. Oblikoskladenjske specifikacije projekta JOS in seznam oblikoskladenjskih oznak
3. Specifikacije projekta SSJ

Normalizacija

Normalizacija oz. standardizacija po definiciji predpostavlja neko normo, ki ji želimo približati (nestandardno) besedo. V večini primerov bo jasno, kakšno obliko naj ima normalizirana beseda, v ostalih primerih pa uporabimo referenčne vire, ki nam pomagajo pri odločitvi (v tem prioritetnem vrstnem redu):

1. Spletni portal Fran, predvsem SSKJ in Pravopis (v tem prioritetnem vrstnem redu);
2. Leksikon Sloleks, predvsem za pregibanje (plus Gigafida);
3. Konkordančnik Gigafida, če besede tam ni oz. je zelo redka, pa korpus *JANES v0.4* v konkordančniku noSkE. Konkordančnik je koristen predvsem za nestandardne besede, ki nimajo standardne ustreznice in jih moramo zato normalizirati v najpogostejšo nestandardno različico.

V primeru kontradiktornih rezultatov ne izgubljajte preveč časa in problematični primer raje zapišite v Težavne primere.

Za iskanje variantnih zapisov neke besedne oblike lahko uporabimo konkordance in štejemo, kolikokrat se posamezna varianta pojavi, v NoSketchEnginu pa obstaja tudi opcija *Seznami* (*Word list*). V okence "*Filter wordlist by:Regular expression:*" vnesemo regularni izraz, ki najbolje opisuje vse variantne zapise. V nadaljevanju podamo uvod v regularne izraze.

7.2 Regularni izrazi

- omogočajo prepoznavanje množice nizov, ne samo konkretnih primerov
- sestavljeni so iz literalov (a–ž, 0–9), nadomestnih znakov (.) in operatorjev (?, *, +, |)
- z operatorji kombiniramo posamezne dele iskalnega pogoja

Osnovna sintaksa:

- konkatenacija: abc prepozna abc
- disjunkcija: ab|bc prepozna ab, bc
- ponavljanje ničkrat ali enkrat: ab? prepozna a, ab
- ponavljanje ničkrat ali večkrat: ab* prepozna a, ab, abb, …

- ponavljanje enkrat ali večkrat: ab+ prepozna ab, abb, abbb, …
- združevanje in disjunkcija: (ab?)|c prepozna a, ab, c

Razširitev osnovne sintakse:

- katerikoli znak: . npr. abc.
- poljubno število katerega koli znaka: .* npr. abc.*
- skupine znakov: […] npr. [fgm]iga prepozna {figa, giga, miga}
- negirana množica znakov [^…] npr. abc[^def]ghi prepozna {abcgghi, abchghi, abcighi, …, abczghi, abcžghi}
- ponavljanje: {n,m} npr. a{2,5} prepozna {aa, aaa, aaaa, aaaaa}