Smernice za ročno normalizacijo *Janes Norm 3.0*

Jakob Lenardič in Darja Fišer
Inštitut za novejšo zgodovino

1. Cilj in smernice

Pričujoče smernice temeljijo na smernicah Čibej et al. (2016: 4–8) in jih tudi nadgrajujejo.

Normalizacija **predvsem vključuje sledeče:**

1. Popravljanje zatipkov in variantnih zapisov obrazil

    polgledal → pogledal

    knižnica → knjižnica

    hodu → hodil

2. Normalizacija šumnikov

    macka → mačka

3. Popravljanje velikih začetnic pri lastnih imenih

    jankovic → Janković

4. Normalizacija medmetov v dve ponovitvi zlogov

    hahahahaha → haha

Ponovljene črke skrajšamo na največ tri ponovitve

grr → grr, grrr → grrr, grrrr → grrr; vendar *hahhaaahaa → haha*

5. Polnopomenske besede s ponovljenimi črkami skrajšamo na nepodaljšano različico

Mooojcaaa → Mojca

Glavno načelo normalizacije je, da normaliziramo **zgolj na nivoju besedišča**, kar pomeni, da:

1. **ne** spreminjamo skladenjskih razmerij (npr. nestandardne rabe tožilnika v skladenjskem okolju zanikanja ne popravljamo)

ne uporabljam rodilnik

2. **ne** spreminjamo izbora besed, kar pomeni, da besede **ne** zamenjujemo s knjižnimi sopomenkami

pocahnu → pocahnal in ne **označil*

pofarbat → pofarbati in ne **pobarvati*

pucajne → pucanje in ne **čiščenje*

cajt → cajt in ne **čas*

špegu → špegel in ne **ogledalo*

3. **ne** spreminjamo napačne stave ločil
4. Ključnikov (heštegov), uporabniških imen, emotikonov, emojijev in elips **ne** normaliziramo

*#lepa-drevesa, @janez, :))), 😀, pi****

5. Besed, za katere ni mogoče ugotoviti, ali je normalizacija potrebna, **ne** normaliziramo.

*Vcerajsnji problem je **resen** → resen*

6. Besednih zvez z nesklonljivim levim prilastkom **ne** normaliziramo z vezajem:

SD kartica → SD kartica in ne **SD-kartica*

2. Težavni/posebni primeri

2.1. Velike začetnice

2.1.1. Prvo besedo v stavku normaliziramo z veliko začetnico samo, če je lastno ime:

biden je Ameriški predsednik → Biden je ameriški predsednik

2.1.2. 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

2.1.3. Vendar, 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.

2.1.4. Če pri večbesednih lastnih imenih ne moremo ugotoviti, ali se vsi elementi pišejo z veliko začetnico, dvoumne pustimo pri miru:

pr'Kovac → pri Kovaču

2.1.5. 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").

2.2. Variantnost zapisa

2.2.1. 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.

2.2.2. Narečne različice nikalnic normaliziramo v *ne*.

nej, *nje → ne*

2.2.3. Besedam, ki nimajo standardne ustreznice, a se zapisujejo v več variantah, kot normalizirano obliko pripišemo najpogostejšo različico v korpusu, gl. tudi razdelek 4.

fouš, fauš, favš → fouš, ki je v celotnem JANES korpusu najpogostejša različica

ornk, orng, orenk, oreng → ornk, ki je najpogostejša

armeja, armija → armija, ki je najpogostejša

fertik, fertig → fertik

frej, fraj → fraj

cajtng, cajteng, cajtung → cajtng

podkast, podkest → podkast

auspuh, avspuh → auspuh

taužent, taužnt, tavžent, tavžnt → tavžent

džoint, džojnt, đoint, đojnt → džoint

oštija, oštja → oštja

kelner, kelnar → kelnar

2.2.4. 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*

2.2.5. 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

2.2.6. Nestandardnih vidskih predpon, kot je *z-* v *zinštalirati* ali *s-* v *sfotošopati*, ne odstranjujemo

2.2.7. Nestandardni zapis predloga *u* normaliziramo v *v*

Velbek bo pršu u arsenal → Velbek bo prišel v Arsenal

2.2.8. V primeru dvojnic dopuščamo obe obliki

bojo → bojo

bodo → bodo

2.3. Obrazila

2.3.1. Nestandardna obrazila normaliziramo v standardna

na Ptuji → na Ptuju

se spomne → se spomni

2.3.2. Nestandardna obrazila, ki tvorijo novo besedno vrsto iz besedotvorne podstave, **ne** normaliziramo

Včer je snegovalo → Včeraj je snegovalo in ne **Včeraj je snežilo*

2.3.3. Napačno rabljene nedoločnike in namenilnike popravljamo.

šla je pogledati → šla je pogledat

2.3.4. Akronime, ki imajo obrazila pripisana ali ločena na nestandarden način, normaliziramo z vezajem

KUDu → KUD-u

tv.ju → tv-ju

2.4. Tujejezične prvine

2.4.1. Tujejezične besede, ki so se poslovenile po zapisu, obravnavamo kot druge variantne nestandardnih besed

knekšna → konekšna

mučas hvalas → mučas hvala

2.4.2. 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.

sherati → sherati in ne **šerati*

fittnessa → fitnessa in ne **fitnes*

pogoogglati → pogooglati ne **poguglati*

2.4.3. Tujejezične prvine, ki so zapisane citatno, pustimo pri miru.

share, *like*

2.4.4. Zatipkane (*chessburger*) normaliziramo v standardne ustreznice

chessburger → cheeseburger

2.4.5. Pri normalizaciji pregibanih tujih lastnih imen se držimo Slovenskega pravopisa

Godoja → Godota

2.4.6. Napačno zapisana lastna imena popravljamo

Tweeter → Twitter

2.5. Okrajšave

2.5.1. Normaliziramo le očitne okrajšave, in sicer z dodajanjem pike

devalv → devalv.

oz → oz.

2.5.2. Okrajšave, ki vsebujejo cifre, normaliziramo v njihove standardne ustreznice

ju3 → jutri

gr8 → great

2.5.3. **Tujejezičnim okrajšavam** tipa *thx*, *srsly* kot normalizirano obliko pripišemo najpogostejšo obliko v korpusu JANES

thx → *tnx*; glej tukaj

plz → *pls*; glej tukaj

2.5.4. Slovenske besede, ki so zapisane s tujimi črkami, normaliziramo v različico, ki je zapisana v skladu s splošno veljavnim standardom

faxom → *faksom*

qrba → *kurba*

2.6. Posebni oblikoskladenjski primeri

2.6.1. **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

2.6.2. Napačne rabe sklona **ne** normaliziramo, kadar ne gre za nestandardno obrazilo, temveč za napako na ravni skladnje:

ne uporabljam rodilnik

z 240 milijonov

a to je zdej klasika v starim firmam

2.6.3. 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*

2.6.4. Prekmurski večpomenski besedi *ge* in *ka* normaliziramo glede na kontekst; *ge* namreč lahko ustreza besedam *jaz*, *kjer* in *kje* , medtem ko *ka* lahko ustreza besedam *kaj, kar*, *da* in *ker*.

2.6.5. Pri zaimkih in prislovih s členico *le* (npr. *tale, tele, tule, tukajle*) členico upoštevamo tudi v normalizirani obliki

tehle → tehle

tukile → tukajle

tle → tule

2.6.6. Pri akronimih[1] pustimo izvorno kapitalizacijo ne glede na to, ali so pisani z velikimi, malimi ali mešanimi črkami

rt, RT, Lp, lp, LP

2.6.7. 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

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

3. Referenčni viri

3.1. Splošni referenčni viri

Normalizacija 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:

1. Spletni portal Fran, predvsem SSKJ in Pravopis;
2. Leksikon SloLeks 2.0, predvsem za pregibanje
3. Referenčni korpus Gigafida 2.0, ki je na voljo bodisi preko konkordančnika CVJT bodisi preko konkordančnika noSketch Engine (ki nudi naprednejše iskalne možnosti)

Če je beseda redka, uporabimo kot referenčni vir korpus JANES in korpus JANES-Norm, ki sta opisana v sledečih razdelkih.

3.2. Korpus JANES

V korpusu JANES, ki je dostopen preko konkordančnika *noSketch Engine*, moremo preverjati dejansko rabo nestandardne spletne slovenščine. Preverba v tem korpusu je zlasti primerna za normalizacijo, pri kateri se moramo odločiti med več možnimi variantami zapisa in izbrati najpogostejšo različico.

Pri tem je dobro poznati nekaj **regularnih izrazov** za iskanje po korpusu:

- katerikoli znak: **.** npr. iskalni niz **l.pa** prepozna variante *lipa, lepa, lopa*
- poljubno število katerega koli znaka: **.*** npr. **fant.*** prepozna *fant, fantje, fantoviščina* itd.
- skupine znakov: **[…]** npr. **[fgm]iga** prepozna **{figa, giga, miga}**
- poljubni znak: **?** npr. **ore?nk** prepozna **{orenk, ornk}**
- ponavljanje: **{n,m}** npr. **a{2,5}** prepozna **{aa, aaa, aaaa, aaaaa}**

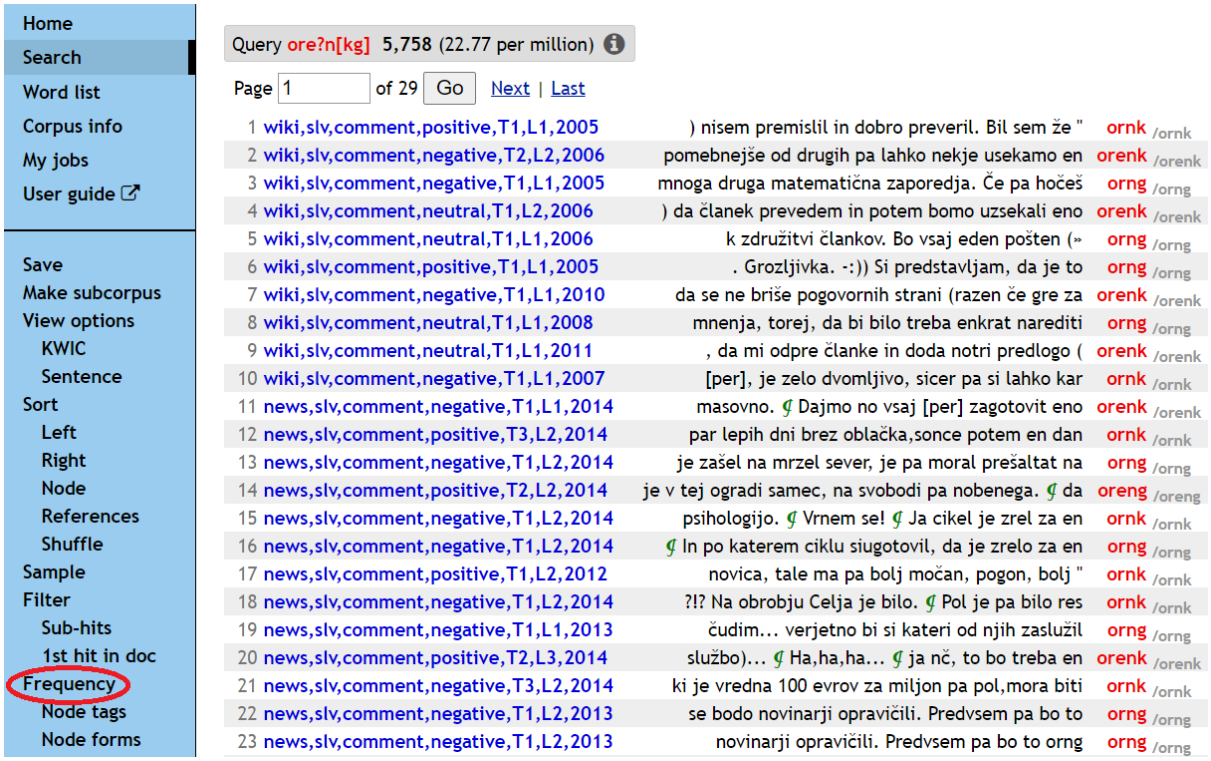
[1] Akronime razumemo kot besede, sestavljene iz prvih črk večbesednih zvez (lep pozdrav - lp) oziroma iz izbranih črk daljše besede (retweet - rt).

Recimo, da moramo normalizirati besedo *oreng* (v prislovnem pomenu “zelo”), ki se pojavlja v tvitu *Če pa hočeš **orng** zakomplicirati, si izmisliš svoje številčne vrednosti za črke.* V dejanski rabi (torej, v korpusu JANES) se izkaže, da je ta beseda zapisana v več različicah: *orenk, orng, ornk* in *oreng*. Da ugotovimo, katera varianta je najpogostejša, lahko v iskalni niz pri **Word form** zapišemo sledeče:[2]

ore?n[kg]

Z znakom ? opredelimo, da je črka *e* poljubna, z oglatima oklepajema pa določimo, da je zadnja črka v besedi bodisi *k* bodisi *g*. Klikni tukaj za rezultate s tem iskalnim nizom.[3]

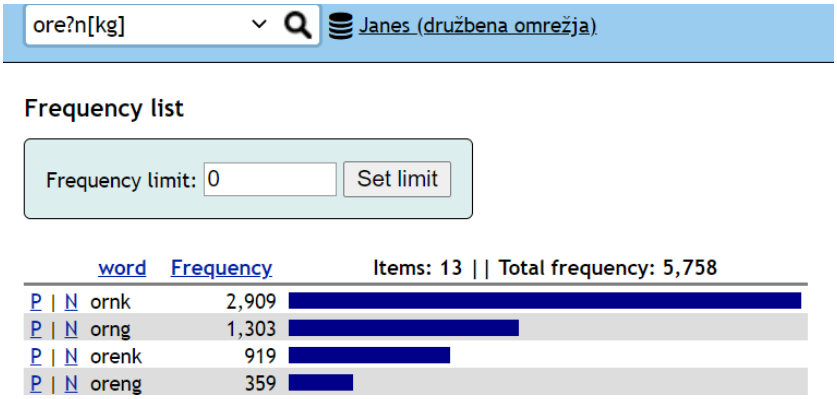
Do najpogostejše različice pridemo z uporabo funkcije Frequency, ki se nahaja v navpičnem modrem traku na levi strani konkordančnika:



S klikom na “Make frequency list” dobimo frekvenčni seznam, ki kaže, da je različica *ornk* najpogostejša. **Nasvet:** kot frekvenčni atribut izberemo “word (lowercase)”, s čimer se izognemo, da konkordančnik posebej navaja frekvence za različno kapitalizirane variante istih besednih oblik.

[2] V primeru pregibnih besednih vrst namesto z **Word Form** raje iščemo z **Lemma.**

[3] Alternativno, vse variante bi lahko tudi upoštevali z iskalnim nizom *orenk|oreng|orng|ornk*, kjer navpičnica | ustreza operaciji disjunkcije. Različica tega ukaza v iskalnem jeziku CQL je [*word* = *“orenk|oreng|orng|ornk”*]



Podobno smo preverjali različice *thx/tnx/thnx* (z iskalnim nizom *t[hn]n?x*) ter *pls/plz* (z iskalnim nizom *pl[sz]*).

3.3. Korpus JANES-Norm

Janes-Norm vsebuje besedila družbenih omrežij, ki so bila ročno normalizirana v prvotnem projektu JANES (2014--2017), in je tako predhodnik korpusa besedil, ki jih normaliziramo v trenutni kampanji. V korpusu lahko preverimo, kako so bile besede ročno normalizirane v pretekli kampanji označevanja. Janes-Norm je torej primeren referenčni vir za normalizacijo težavnejših zgledov.

Zgled uporabe JANES-Norm kot referenčni vir za trenutno kampanjo
V korpusu Janes-Norm so besedila označena glede na tehnično standardnost, pri čemer so besedila z oznako T3 smatrana kot tehnično nestandardna (kar predvsem vključuje izpust šumnikov in nestandardno stavo ločil), besedila z oznako T1 pa tehnično standardna, ter glede na lingvistično standardnost, pri čemer so besedila z oznako L3 smatrana kot lingvistično nestandardna (veliko število nestandardnih besed v enem stavku), L1 pa lingvistično standardna (malo število nestandardnih besed v stavku).

Recimo, da nas zanima, kako so **vse besede** normalizirane v tem korpusu, zlasti v manj standardnih besedil. V **Word Form** pod **Query Type** zapišemo sledeči iskalni niz

.*

Z njim določimo, da korpus išče za pojavnice, ki vsebujejo katerikoli znak, ki se lahko v pojavnici poljubno ponavlja; z drugimi besedami, konkordančnik išče za katerokoli besedo. V **Text Types** izberemo možnost T3 pod TECHNICAL STANDARDNESS ter možnost L3 pod LINGUISTIC STANDARDNESS.

Konkordance, ki jih vrne gornji iskalni niz, lahko opremimo z vizualizacijo strukturalnih atributov/metapodatkov, ki so v korpusu dodeljeni posamezni pojavnici (npr. njena lema, oblikoskladenjske lastnosti ter, kar je ključno, normalizirana oblika).

S klikom na **View options** lahko izberemo atribut **norm**; s tem dosežemo, da je pod vsako rdeče odebeljeno pojavnico (ti. KWIC, "KeyWord In Context; se pravi, pojavnice, ki jih vrne prvotni iskalni niz) v sivem zapisana njena normalizirana oblika:

Tukaj mdr. vidimo, da je nestandardno zapisana beseda *sm* normalizirana v *sem*, *ustrašu* v *ustrašil* ter *Jankoviča* v *Jankovića*, *oment* v ustrezno nedoločniško obliko *omeniti*. Vidimo tudi, da vezni glagol *sm* v normalizirani obliki ni kapitaliziran (čeprav je prva beseda v povedi), kar je v skladu s pravili za normalizacijo velikih začetnic (glej razdelek 3.1).

4. Reference

Jaka Čibej – Špela Arhar Holdt – Tomaž Erjavec – Darja Fišer – Katja Zupan. Smernice za označevanje računalniško posredovane komunikacije: tokenizacija, stavčna segmentacija, lematizacija in oblikoskladenjsko označevanje v1.0. Na spletu (2016). https://nl.ijs.si/janes/wp-content/uploads/2014/09/Janes-smernice-v1.0.pdf.