02 Segmentacija

Segmentacija je postopek deljenja besedila na povedi. Pri strojnem označevanju korpusov v slovenskem prostoru trenutno uporabljamo označevalnik CLASSLA-Stanza oz. vanj vključeni segmentator Obeliks. Pravilom, na katerih je osnovan strojni označevalnik, sledi tudi ročni pregled.

- Predstavitev segmentacije
- Označevalne smernice
- Reference in povezave

Predstavitev segmentacije

V tem poglavju je strnjeno predstavljena stavčna segmentacija.

Glavno vodilo za razmejevanje povedi je kombinacija končnega ločila, presledka in besede, zapisane z veliko začetnico. Temu se pridružujejo dodatna pravila, ki zajemajo okrajšave. Te se namreč zapisujejo s piko, ki je lahko hrati tudi končno ločilo (kadar okrajšava stoji na koncu povedi, npr. 'itd.') ali pa ne (kadar okrajšava stoji sredi povedi, recimo 'tj.'). Končen nabor okrajšav, ki spadajo v eno in v drugo kategorijo, je vključek v orodje Obeliks.

Za segmentacijo nestrandardnih besedil veljajo še dodatna pravila:

 V celotnem tvitu preverimo, ali je avtomatska stavčna segmentacija pravilna. V smernicah konec stavka (oz. povedi v slovenistični terminologiji) za lažjo predstavo označujemo s simbolom ¶.

 Če del tvita deluje kot samostojen stavek, ga tako tudi obravnavamo (“@multikultivator Najbrž ne . ¶ :) ¶ Kot rečeno : bolje BO . ¶ Zrihtamo , ko utegnemo . ¶ (PS : tudi v veselje " konkurence " ;)”).

 Merilo za konec stavka je predvsem ločilo, ki deluje kot končno v stavku, npr. pika, klicaj, vprašaj, narekovaj ali večpičje (“Kaj praviš ? ¶ Aha !”).

 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).

 Konec tvita je avtomatično tudi konec stavka, zato tega ne označujemo.

Težji primeri:

 Večpičje:

    Ponavadi je končno ločilo (“@SLO_Super_Visor po moje se jo izogiba kot hudič križa. ¶ Glavn da on spet laja … ¶ :-)))))”).
    Včasih označuje zgolj elipso ali zamolk sredi stavka – v takšnem primeru ni končno ločilo (“To se mi zdi … neumno.”).

 Imena (@ime), emotikoni (\o/) ali emojiji (□□) in heštegi (#hešteg):

    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”).

 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.”).

 Če nadomeščajo končno ločilo, zaznamujejo konec stavka (“kot da je to važn :)) ¶ nobenga to ne briga vec sploh”).

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

 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).

Segmentacija govorjene slovenščine je zaenkrat izvedena ročno na podlagi prozodično oz. semantično zaključenih enot.

Označevalne smernice

V tem poglavju so zbrane označevalne smernice za segmentacijo, ki so na voljo.

Različica 1.0 za nestandardno slovenščino

projekt Janes - Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine
ČIBEJ, Jaka, ARHAR HOLDT, Špela, ERJAVEC, Tomaž, FIŠER, Darja, ZUPAN, Katja, 2016: *Smernice za označevanje računalniško posredovane komunikacije: tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladenjsko označevanje.* Različica 1.0. Rezultat projekta Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine. [PDF]

Reference in povezave

V tem poglavju so zbrane relevantne reference in povezave na projekte, v katerih se je postopek segmentacije razvijal in uporabljal.

Projekti, na katerih se je označevalni sistem razvijal oz. uporabljal

JOS - Jezikoslovno označevanje slovenskega jezika: metode in viri

Sporazumevanje v slovenskem jeziku

Universal Dependencies

Janes - Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine

Razvoj slovenščine v digitalnem okolju

Orodje Obeliks za tokenizacijo in stavčno segmentacijo

https://github.com/clarinsi/obeliks

Reference

Krek, Simon; Erjavec, Tomaž; Dobrovoljc, Kaja; Gantar, Polona; Arhar Holdt, Špela; Čibej, Jaka in Brank, Janez. The ssj500k training corpus for Slovene language processing. V: Fišer, D. in Erjavec, T. Jezikovne tehnologije in digitalna humanistika: zbornik konference: 24.-25. september 2020, Ljubljana, Slovenija. Ljubljana: Inštitut za novejšo zgodovino, 2020. Str. 24–33.

http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene-Language-Processing.pdf [PDF]