

Navodila za označevanje računalniško posredovane komunikacije v WebAnno

v1.0

Datum zadnje spremembe: 2016-12-21

Avtorji: Tomaž Erjavec, Cyprian Laskowski, Jaka Čibej, Darja Fišer, Kaja Dobrovoljc

Kazalo

[0 Uvod 1](#_Toc470024768)

[1 Stičnost 2](#_Toc470024769)

[2 Tokenizacija (*Corrections*) 2](#_Toc470024770)

[2.1 Razdruževanje pojavnic 2](#_Toc470024771)

[2.2 Združevanje pojavnic 3](#_Toc470024772)

[3 Stavčna segmentacija (*Sentences*) 4](#_Toc470024773)

[3.1 Popravki segmentacije 4](#_Toc470024774)

[3.2 Izbris tvita 5](#_Toc470024775)

[4 Normalizacija (*Normalisations*) 5](#_Toc470024776)

[4.1 Normalizacija pojavnic 5](#_Toc470024777)

[4.2 Normalizacija ene pojavnice v več besed 6](#_Toc470024778)

[4.3 Normalizacija več pojavnic v eno besedo 7](#_Toc470024779)

[5 Kombinirani popravki 7](#_Toc470024780)

[5.1 Več popravkov na sosednjih pojavnicah 7](#_Toc470024781)

[5.2 Označevanje pojavnic s popravljeno tokenizacijo 8](#_Toc470024782)

[6 Pregled 9](#_Toc470024783)

[6.1 Ravni označevanja 9](#_Toc470024784)

[6.2 Posebni znaki 9](#_Toc470024785)

# 0 Uvod

Pričujoča navodila razložijo, kako so v WebAnno predstavljeni besedila računalniško posredovane komunikacije in kako popraviti avtomatsko določeno tokenizacijo, stavčno segmentacijo in normalizacijo. Navodila niso namenjena splošnemu uvodu v delo z WebAnno, niti kako se pravilno odločiti v posameznih primerih; slednje je razloženo *v Smernicah za označevanje računalniško posredovane komunikacije: tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladenjsko označevanje* projekta Janes.

# 1 Stičnost

Izvirnim pojavnicam je v WebAnno pripisan znak za stičnost z naslednjo pojavnico, za kar uporabljamo levo poševnico, »***\***«. Oznaka nam pomaga rekonstruirati zapis izvornega tvita s presledki vred. Stičnost je podana samo v informativni namen; v oznake pojavnic znaka za stičnost ne vnašamo.

**Primer 1:** Pred vejico ni presledka, pred tropičjem pa je.

*Izvirni zapis tvita:*

No ja, svojega omiljenega kriminalca so pa le izvlekli iz zapora pod pretvezo bogi-bogi-politik sranja ...

*Prikaz v WebAnno:*



**Primer 2**: Pred levo poševnico in za njo ni presledka, ravno tako ne pred piko.

*Izvirni zapis tvita:*

s posli in delom ustvarjajo delovna mesta/službe.

*Prikaz v WebAnno:*


**Primer 3**: Poseben primer je, ko se že sama pojavnica konča z levo poševnico, saj je v tem primeru ne smemo vzeti kot znak stičnosti. Zato je v tem primeru v WebAnno levi poševnici pripisan posebni niz ”**$0**”.

*Izvirni zapis tvita:*

več njih? :\ brrr...

*Prikaz v WebAnno:*


**Opomba:** če bi bila v izvornem zapisu leva poševnica stična z naslednjo besedo, bi bila v WebAnno pojavnica “:\$0\”.

# 2 Tokenizacija (*Corrections*)

## 2.1 Razdruževanje pojavnic

Napake v tokenizaciji popravljamo na ravni *Corrections*, ki ima, poleg privzete lastnosti *correction,* še dodatne lastnosti *correction2, correction3, correction4*. Slednje uporabimo, ko je treba neko pojavnico razdružiti:

**Primer 3:** »*turčija-grčija*« je ena pojavnica; popravimo v tri.

*Izvirni zapis v WebAnno:*



*Popravek:*



*Rezultat:*



## 2.2 Združevanje pojavnic

Obratni problem je, kadar je v WebAnno ena pojavnica razdeljena na več pojavnic. V tem primeru združeno pojavnico vpišemo v *Corrections* prve izmed pojavnic, preostalim pa pripišemo poseben znak za izbris »**$0**«:

**Primer 4:** Emotikon je bil razdeljen v dve pojavnici; popravimo v eno (in dodamo znak za konec stavka, gl. naslednji razdelek).

*Izvirni zapis v WebAnno:*

 

*Popravek 1. pojavnice: Popravek 2. pojavnice:*

 

*Rezultat:*



# 3 Stavčna segmentacija (*Sentences*)

## 3.1 Popravki segmentacije

Za segmentacijo tvita na stavke (oz. povedi) uporabljamo raven *Sentences*, v katero vpišemo (ali iz nje zbrišemo) posebni znak »***$.***«: če neki pojavnici sledi konec stavka, naj ima v *Sentences* pripisan »*$.*«, če ji ne sledi konec stavka, pa te oznake ne sme imeti. Izjema je zadnja pojavnica v tvitu, saj ta vedno zaključuje stavek in zato tja *$.* ne pišemo.

**Primer 5:** Za emotikonom :/ bi se moral začeti nov stavek; dodamo oznako za konec prejšnjega stavka.

*Izvirni zapis v WebAnno:*



*Popravek:*



*Rezultat:*



**Primer 6:** »*nj.*« (kot okrajšava za *njegov*) ne konča stavka; zbrišemo stavčno mejo in piko pripišemo k pojavnici nj.

*Izvirni zapis v WebAnno:*



*Popravki*(pri prvem samo kliknete »Delete«)**:**

  

*Rezultat:*



## 3.2 Izbris tvita

V redkih primerih se tudi zgodi, da tvita nima smisla označevati, ker je npr. popolnoma v tujem jeziku ali pa vsebuje avtomatsko generirano sporočilo. To označimo tako, da prvi pojavnici v tvitu v raven *Sentences* zapišemo posebni znak »***$0***«:

**Primer 7:** Ker je celotni tvit v hrvaščini, ga označimo za odstranitev.

*Izvirni zapis v WebAnno:*



*Popravek:*



*Rezultat:*

­

# 4 Normalizacija (*Normalisations*)

## 4.1 Normalizacija pojavnic

Pojavnice normaliziramo (standardiziramo) v ravni *Normalisations*.

**Primer 8:** »*ze*« in »*vasa*« sta že pravilno normalizirani, »*lej*« in »*razlaga*« pa ne, zato ju normaliziramo.

*Izvirni zapis v WebAnno:*



*Rezultat:*



**Pozor:** Če je pojavnice ni potrebno normalizirati, avtomatsko pripisano (in napačno) normalizacijo odstranimo, torej pri označevanju ravni *Normalisation* pritisnemo *Delete*, in **ne** vpisujemo pojavnice v vrednost *normalisation*:

*Izvirni zapis v WebAnno:*



*Popravek*(kliknemo na *Delete*)**:**



*Rezultat:*



## 4.2 Normalizacija ene pojavnice v več besed

V primeru, da mora biti ena pojavnica normalizirana v dve ali več, uporabimo *normalisation2 … normalisation4*, podobno kot pri popravkih tokenizacije (*Corrections*):

**Primer 9:** »*neboš*« je zapisano skupaj; normaliziramo v dve besedi.

*Izvirni zapis v WebAnno:*



*Popravek:*



*Rezultat:*



## 4.3 Normalizacija več pojavnic v eno besedo

Kot pri popravkih tokenizacije imamo tudi obratne primere, ko se dve ali več pojavnic normalizirajo v eno besedo. V teh primerih prvi pojavnici pripišemo normalizacijo celotnega zaporedja, ostalim pa, podobno kot za popravke tokenizacije, za normalizacijo pripišemo posebni znak »***$0***«:

**Primer 10:** »*porka duš*« naj bi se pisal skupaj; normaliziramo v eno besedo.

*Izvirni zapis v WebAnno:*



*Popravek 1. in 2. pojavnice:*

 

*Rezultat:*



# 5 Kombinirani popravki

## 5.1 Več popravkov na sosednjih pojavnicah

Pogost pojav je, da ena napaka v avtomatskem označevanju povzroči še druge. V teh primerih popravimo vse napake v skladu z zgornjimi navodili:

**Primer 11:** Domena »*.si*« je zapisana kot dve pojavnici, kjer naj bi prva končala stavek; zbrišemo stavčno mejo in popravimo v eno pojavnico.

*Izvirni zapis v WebAnno:*

*Rezultat:*



## 5.2 Označevanje pojavnic s popravljeno tokenizacijo

V redkih primerih se bo zgodilo, da bodo popravki tokenizacije povzročili, da je treba normalizirati pojavnice, ki so v *Corrections* zapisane v *correction2 … correction4*, ali pa končati stavek na eni od teh popravljenih pojavnic. To je tudi razlog, da imajo tako *Normalisations* kot *Sentences* tudi *normalisation2 … normalisation4* oz. *sentence2 … sentence4*:

**Primer 12:** Če je »*hodu.pol*« ena pojavnica, popravimo v tri (na ravni tokenizacije še pustimo nenormalizirano!), jih nato istoležno normaliziramo in vstavimo stavčno mejo.

*Izvirni zapis v WebAnno:*



*Popravki:*

  

*Rezultat:*



**Primer 13:** Če je »*glejga!nehi*« ena pojavnica, popravimo v tri, ob tem pa moramo »*glejga*« normalizirati v dve pojavnici, zato pazimo, da si že v Corrections pustimo prostor za to normalizacijo.

*Izvirni zapis v WebAnno:*



*Popravki:*

  

*Rezultat:*



# 6 Pregled

## 6.1 Ravni označevanja

|  |  |
| --- | --- |
| **Raven** | **Uporaba** |
| *Corrections* | popravki tokenizacije |
| *Sentences* | stavčna segmentacija |
| *Normalisations* | normalizacija pojavnice |

## 6.2 Posebni znaki

|  |  |  |
| --- | --- | --- |
| **Znak** | **Raven** | **Uporaba** |
| *\* | *Pojavnice* | stičnost z naslednjo pojavnico |
| *$0* | *Pojavnice* | »zaščita« znaka »\« na koncu same pojavnice |
| *$0* | *Corrections* | izbris (združene) pojavnice  |
| *$0* | *Normalisations* | združena normalizacija na predhodni pojavnici |
| *$0* | *Sentences* | izbris celotnega stavka (na prvi pojavnici) |
| *$.* | *Sentences* | konec stavka (na pojavnici, ki zaključuje stavek) |