

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
1 Stičnost	2
2 Tokenizacija (<i>Corrections</i>).....	2
2.1 Razdruževanje pojavnic.....	2
2.2 Združevanje pojavnic.....	3
3 Stavčna segmentacija (<i>Sentences</i>).....	4
3.1 Popravki segmentacije	4
3.2 Izbris tvita	5
4 Normalizacija (<i>Normalisations</i>)	5
4.1 Normalizacija pojavnic.....	5
4.2 Normalizacija ene pojavnice v več besed	6
4.3 Normalizacija več pojavnic v eno besedo.....	7
5 Kombinirani popravki	7
5.1 Več popravkov na sosednjih pojavnicah	7
5.2 Označevanje pojavnic s popravljeno tokenizacijo.....	8
6 Pregled.....	9
6.1 Ravni označevanja	9
6.2 Posebni znaki.....	9

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 izvirnega 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:

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

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/sluzbe.

Prikaz v WebAnno:

s posli in delom ustvarjajo delovna mesta\ / službe\ .

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:

več njih\ ? :\ \$0 brrr\ ...

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

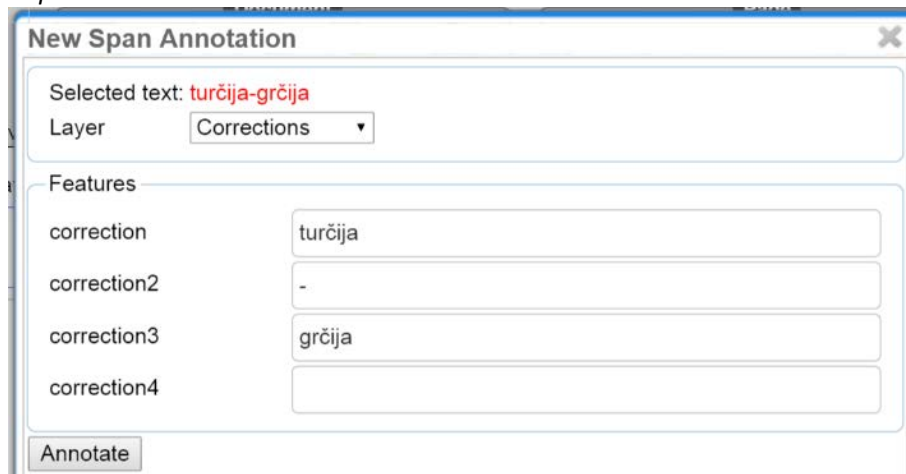
Napake v tokenizaciji popravljamo na ravni *Corrections*, ki ima, poleg privzete lastnosti *correction*, še dodatne lastnosti *correction2*, *correction3*, *correction4*. Slednje uporabimo, ko je treba neko pojavnico razdružiti:

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

Izvirni zapis v WebAnno:

El classico turčija-grčija planiram že ful cajta\ ,

Popravek:



Rezultat:

turčija | - | grčija
El classico turčija-grčija planiram že ful cajta\ ,

2.2 Združevanje pojavnic

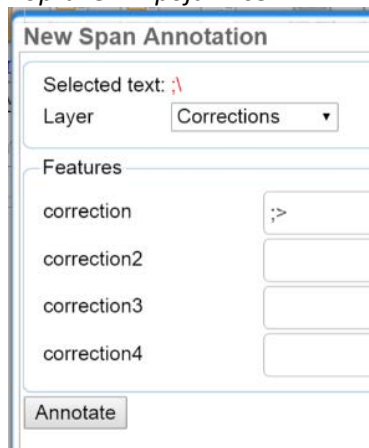
Obratni problem je, kadar je v WebAnno ena pojavnica razdeljena na več pojavnic. V tem primeru združeno pojavnico vpišemo v *Corrections* prve izmed pojavnic, preostalim pa pripišemo poseben znak za izbris »\$0«:

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

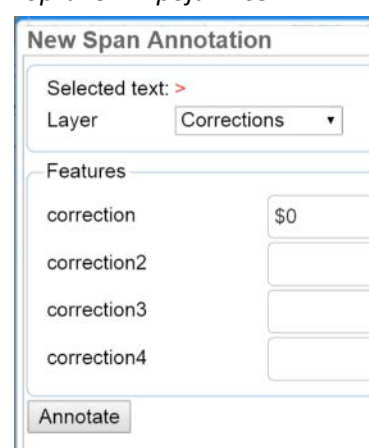
Izvirni zapis v WebAnno:

Ni mi mar za zasebnost ;\ >

Popravek 1. pojavnice:



Popravek 2. pojavnice:



Rezultat:

;\ \$0
:\ >

3 Stavčna segmentacija (*Sentences*)

3.1 Popravki segmentacije

Za segmentacijo tvita na stavke (oz. povedi) uporabljamo raven *Sentences*, v katero vpišemo (ali iz nje zberišemo) 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:

jah oni se majo že fino\ \$. otroc nimajo kej dost od tega :/ sj bo job sj bo\ \$. js sm po 5 mescih najdu :) !

Popravek:

New Span Annotation	
Selected text: :/	
Layer	Sentences
Features	
sentence	\$.
sentence2	
sentence3	
sentence4	
<input type="button" value="Annotate"/>	

Rezultat:

@_Inja_ @GobaFunk jah oni se majo že fino\ \$. otroc nimajo kej dost od tega :/ sj bo job sj bo\ \$. js sm po 5 mescih najdu :) !

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

Izvirni zapis v WebAnno:

pri nas dobivajo denar za filme levaki\ \$. V.\ M. in tist nj\ \$. nov film\ \$. ????

Popravki (pri prvem samo kliknete »Delete«):

Edit Span Annotation	
Selected text: .	
Layer	Sentences
Features	
sentence	\$.
sentence2	
sentence3	
sentence4	
<input type="button" value="Annotate"/> <input type="button" value="Delete"/>	

New Span Annotation	
Selected text: .	
Layer	Corrections
Features	
correction	\$0
correction2	
correction3	
correction4	
<input type="button" value="Annotate"/>	

New Span Annotation	
Selected text: nj\	
Layer	Corrections
Features	
correction	nj.
correction2	
correction3	
correction4	
<input type="button" value="Annotate"/>	

Rezultat:

pri nas dobivajo denar za filme levaki\ ... V.\ M. in tist nj\ . nov film\ ... ????

3.2 Izbris tvita

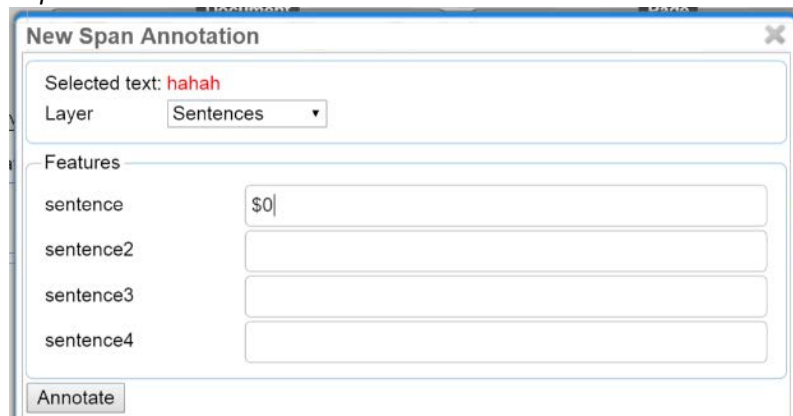
V redkih primerih se tudi zgodi, da tvita nima smisla označevati, ker je npr. popolnoma v tujem jeziku ali pa vsebuje avtomatsko generirano sporočilo. To označimo tako, da prvi pojavnici v tvitu v raven *Sentences* zapišemo posebni znak »\$0«:

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

Izvirni zapis v WebAnno:

hahah\ ..\ u 5 sam došo doma\ ..\ bili smo do fajrunda u melinu\ ,\ dobijo sam pivu od AR :)

Popravek:



Rezultat:

\$0 \$.\ u 5 sam došo doma\ ..\ bili smo do fajrunda u melinu\ ,\ dobijo sam pivu od AR :)

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:

@nejcsimsic lej\ ..\ res da je vasa strategija že zazgala enkrat\ ,

Rezultat:

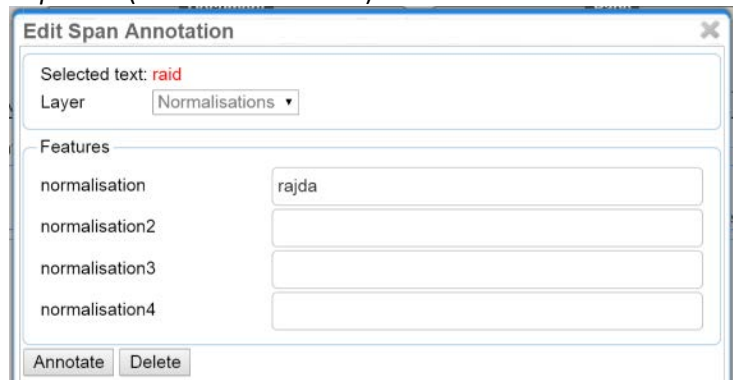
glej \$.\ res da je vasa strategija že zazgala enkrat\ ,

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:

rajda
en serious raid corruption\ ,

Popravek (kliknemo na *Delete*):



Rezultat:

en serious raid corruption\ ,

4.2 Normalizacija ene pojavnice v več besed

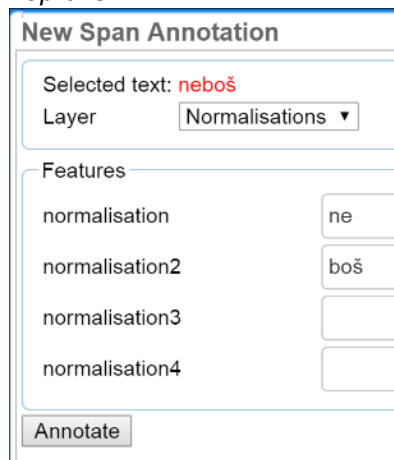
V primeru, da mora biti ena pojava normalizirana v dve ali več, uporabimo *normalisation2* ... *normalisation4*, podobno kot pri popravkih tokenizacije (*Corrections*):

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

Izvirni zapis v WebAnno:

me neboš našemu Robertu zacinkaril

Popravek:



Rezultat:

ne | boš
me neboš našemu Robertu zacinkaril

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:

poroka duše
Porka duš\ , v začetku\ , ko sem ga

Popravek 1. in 2. pojavnice:

New Span Annotation	New Span Annotation
Selected text: Porka	Selected text: duš
Layer: Normalisations	Layer: Normalisations
Features	Features
normalisation: porkaduš	normalisation: \$0
normalisation2:	normalisation2:
normalisation3:	normalisation3:
normalisation4:	normalisation4:
Annotate	Annotate

Rezultat:

porkaduš \$0
Porka duš\ , v začetku\ , ko sem ga

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; zberišemo stavčno mejo in popravimo v eno pojavnico.

Izvirni zapis v WebAnno:

.si
nimam nič proti skromnosti in v .si je veliko skromnih

Rezultat:

.si \$0
nimam nič proti skromnosti in v .si je veliko skromnih

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:

[přišel](#)
dolgo je *hodu.pol* je pa le *pršu* .

Popravki:

The image shows three sequential screenshots of the WebAnno interface for editing a span annotation. Each window has a title bar and a 'Selected text' field containing 'hodu.pol'.
1. **Edit Span Annotation:** The 'Layer' dropdown is set to 'Corrections'. The 'Features' section contains four input fields: 'correction' with 'hodu', 'correction2' with '.', 'correction3' with 'pol', and 'correction4' which is empty. There are 'Annotate' and 'Delete' buttons at the bottom.
2. **New Span Annotation:** The 'Layer' dropdown is set to 'Normalisations'. The 'Features' section contains four input fields: 'normalisation' with 'hodil', 'normalisation2' with '.', 'normalisation3' with 'potem', and 'normalisation4' which is empty. There is an 'Annotate' button at the bottom.
3. **New Span Annotation:** The 'Layer' dropdown is set to 'Sentences'. The 'Features' section contains four input fields: 'sentence' which is empty, 'sentence2' with '\$.', 'sentence3' which is empty, and 'sentence4' which is empty. There is an 'Annotate' button at the bottom.

Rezultat:

[hodu | . | pol](#)
[hodil | . | potem](#)
[| \\$.](#)
[hodu.pol](#)

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:

[glejga!nehi mi pamet solit\ !](#)

Popravki:

The image shows three sequential screenshots of the WebAnno interface for editing a span annotation. Each window has a title bar and a 'Selected text' field containing 'glejga!nehi'.
1. **New Span Annotation:** The 'Layer' dropdown is set to 'Corrections'. The 'Features' section contains four input fields: 'correction' with 'glejga', 'correction2' which is empty, 'correction3' with '!', and 'correction4' with 'nehi'. There is an 'Annotate' button at the bottom.
2. **New Span Annotation:** The 'Layer' dropdown is set to 'Normalisations'. The 'Features' section contains four input fields: 'normalisation' with 'glej', 'normalisation2' with 'ga', 'normalisation3' with '!', and 'normalisation4' with 'nehaj'. There is an 'Annotate' button at the bottom.
3. **New Span Annotation:** The 'Layer' dropdown is set to 'Sentences'. The 'Features' section contains four input fields: 'sentence' which is empty, 'sentence2' which is empty, 'sentence3' with '\$.', and 'sentence4' which is empty. There is an 'Annotate' button at the bottom.

Rezultat:

glejga!!!nehi
glej|ga!!!nehai
||\$.
glejga!nehi mi pamet soliti!

6 Pregled

6.1 Ravni označevanja

Raven	Uporaba
<i>Corrections</i>	popravki tokenizacije
<i>Sentences</i>	stavčna segmentacija
<i>Normalisations</i>	normalizacija pojavnice

6.2 Posebni znaki

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