Technical guidelines for annotating non-standard language corpora in WebAnno

Tomaž Erjavec

Jožef Stefan Institute

UniBg, 10. 12. 2015

Overview

1. Goals and method
2. Tokenisation
3. Sentence segmentation
4. Normalisation
5. Combinations and table

Goals

„Kaće nam podele ruske zastavice“

1. Difficult to search in corpus: „kaće “
2. Low quality tagging and lemmatisation

Therefore:

- Normalisation (standardisation) of word forms: *„Kada će“*

Additional goal: what is a word, what is a sentence?

„utakmica ***Dinamo-Partizan***“

„moj ***deb .*** *brat me opet gnjavi“*

- Automatic tokenisation sometimes makes errors in assigning boundaries between tokens
- Similarly, there are errors in segmentation into sentences

- We need a „gold“ corpus, so we can improve tokenisation and sentence segmentation programs.

Manual annotation

- We have chosen tweets of different non-standardness levels
- The tweets are already automatically tokenised and sentence segmented
- Manual annotation:
  - **Technical guidelines**
  - Linguistic guidelines
  - Practical
  - Annotation…

Overview

Annotation Layer	Use
Corrections	Tokenisation corrections
Sentences	Sentence segmentation
Normalisations	Normalised word

Special chars	Use
\	No space after word
$0	Deletion
$.	End of sentence

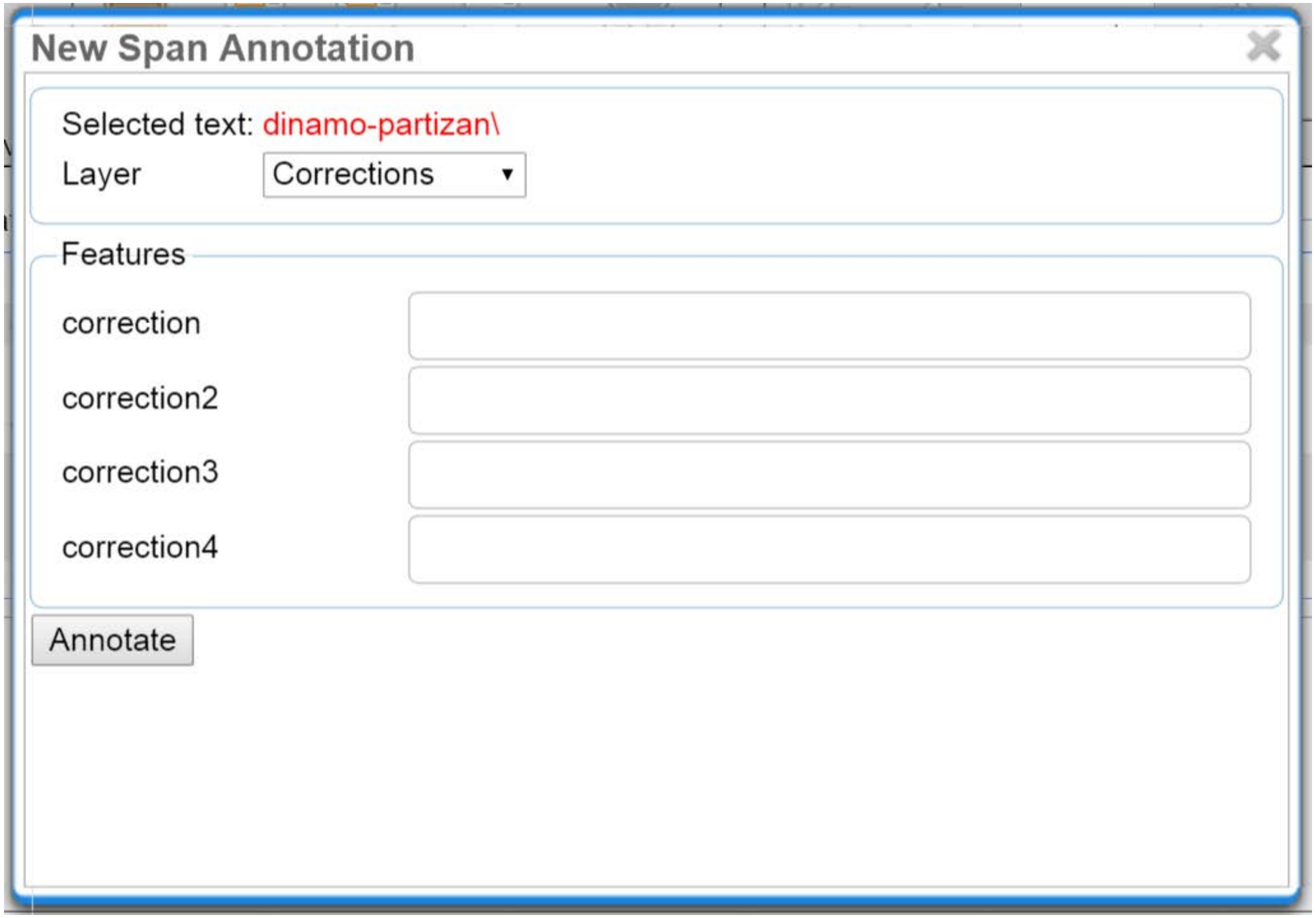
Spacing: \

- Tokens in WebAnno end in backslash if there is no space after it in the source tweet
- This char helps us to reconstruct how the original tweet was written, including spacing
- Added only for information; we do not need to write it in our annotations

- *Da, trebalo bi.* → *Da\ , trebalo bi \.*
- *Javim ti se danas/sutra.* → *Javim ti se danas\ /\ sutra .*
- *više njih? :\ brrr...* → *više njih\ ? :\$0 brrr\ ...*

Tokenisation correction

- Annotation layer: *Corrections* (4 values)



New Span Annotation

Selected text: dinamo-partizan\

Layer Corrections

Features

correction

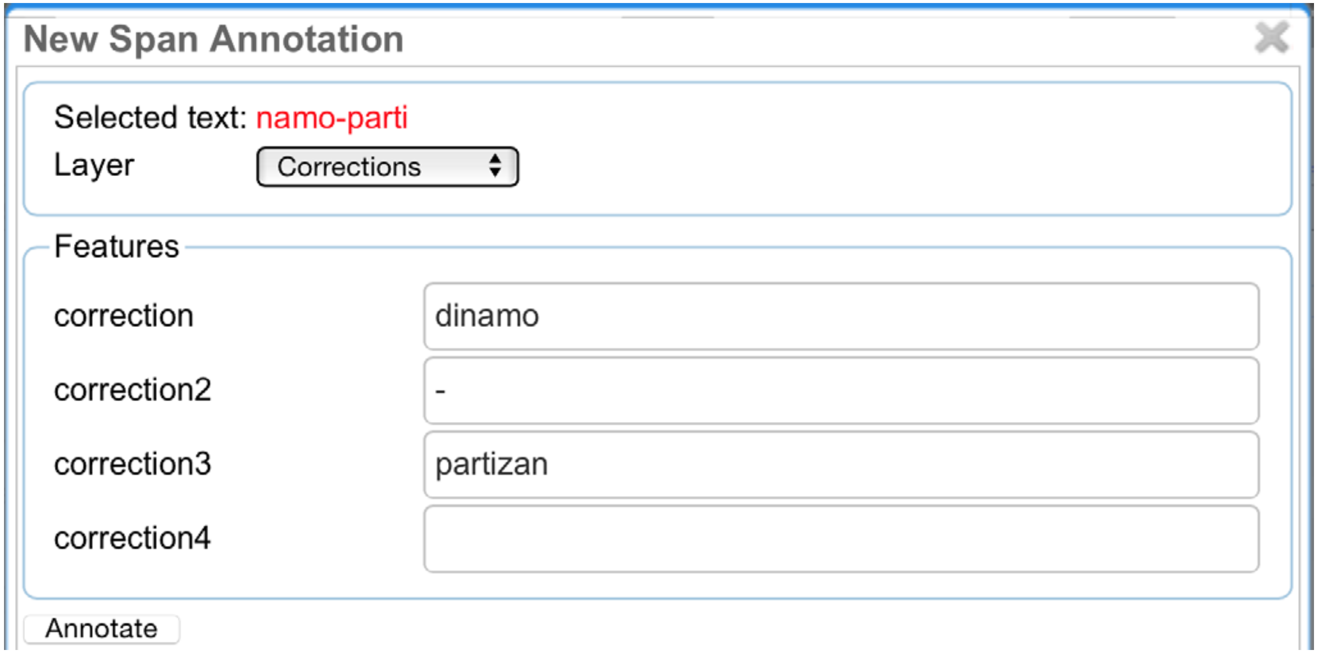
correction2

correction3

correction4

Tokenisation: splitting tokens

Ne mogu gledati ovaj dinamo-partizan\ ...

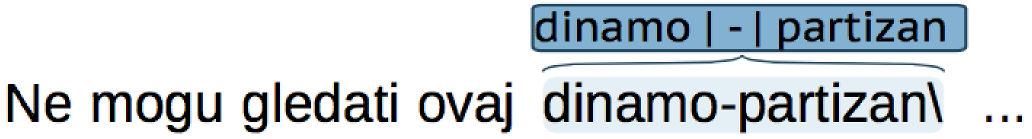


Selected text: namo-parti

Layer: Corrections

Features

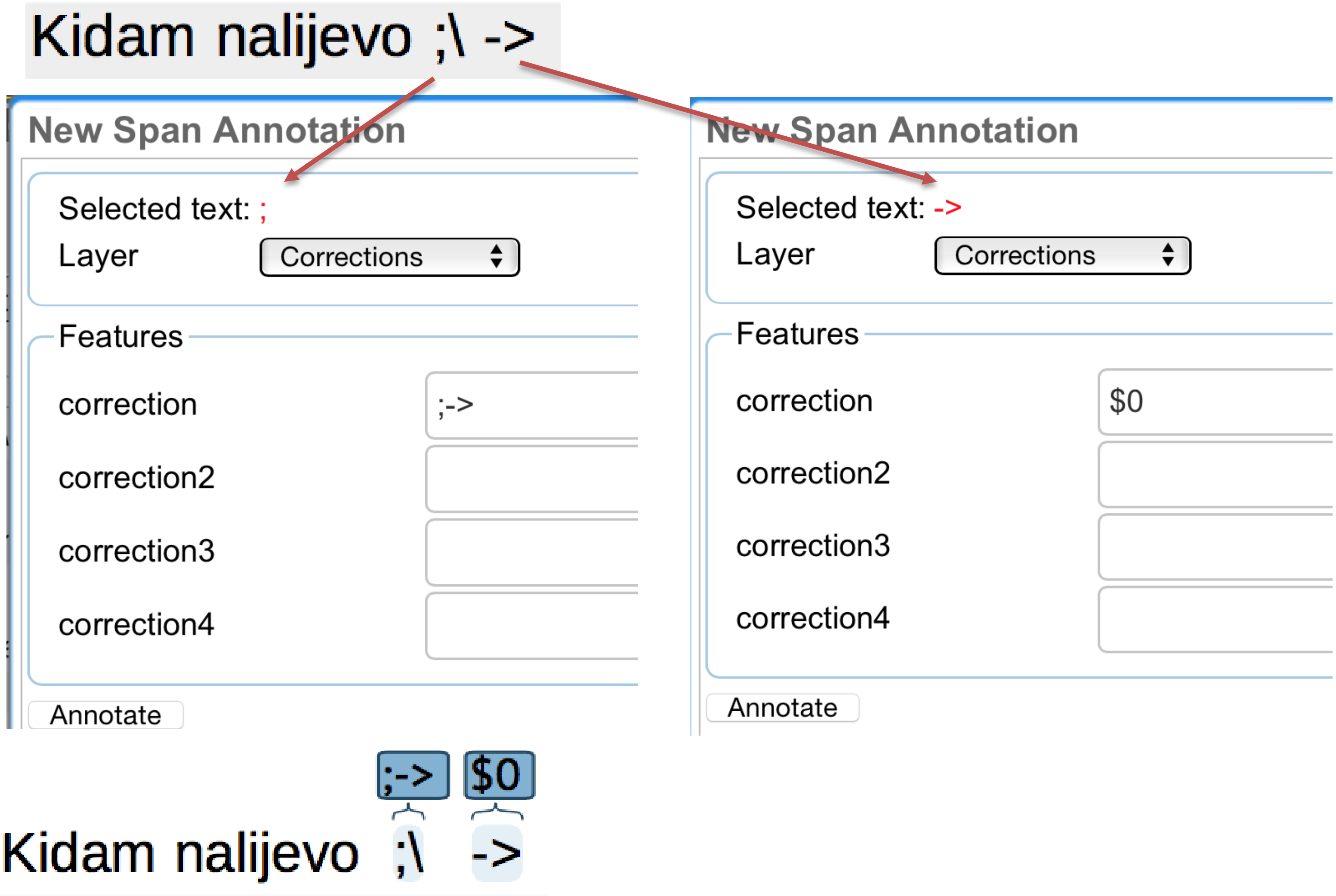
correction	dinamo
correction2	-
correction3	partizan
correction4	



dinamo | - | partizan

Ne mogu gledati ovaj dinamo-partizan\ ...

Tokenisation: merging tokens: ;->



Kidam nalijevo ;\ ->

New Span Annotation

Selected text: ;

Layer Corrections

Features

correction ;->

correction2

correction3

correction4

New Span Annotation

Selected text: ->

Layer Corrections

Features

correction $0

correction2

correction3

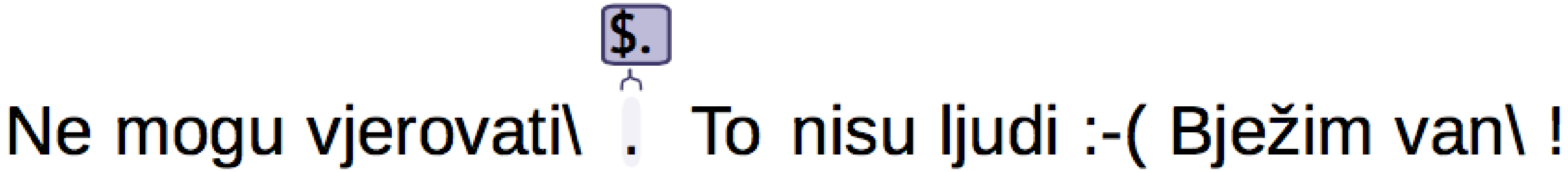
correction4

;-> $0

Kidam nalijevo ;\ ->

Sentence segmentation

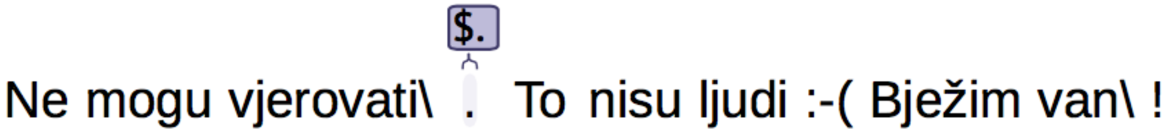
- Layer *Sentences*
- Value: only *$.* (end of sentence)
- Not marked at end of tweet



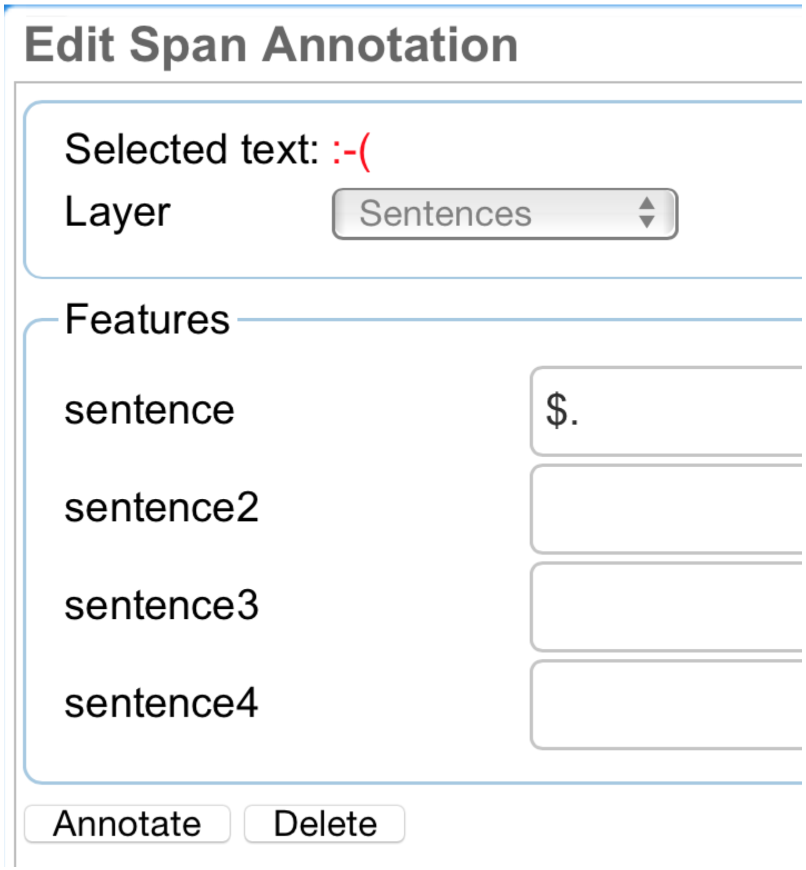
$.

Ne mogu vjerovati\ . To nisu ljudi :-(Bježim van\ !

Adding a sentence boundary



Ne mogu vjerovati\ . To nisu ljudi :-(Bježim van\ !



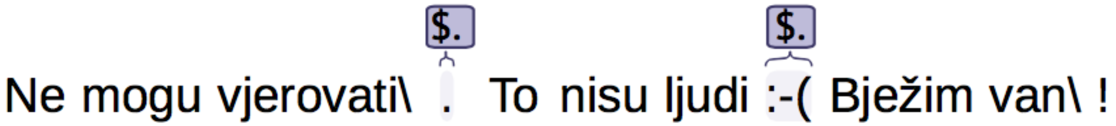
Edit Span Annotation

Selected text: :-(

Layer: Sentences

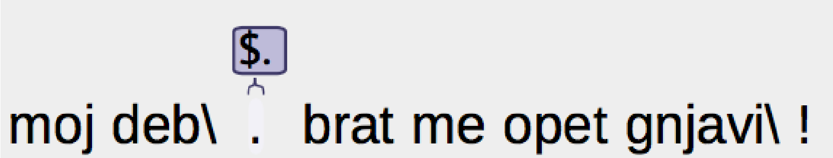
Features

sentence	$.
sentence2	
sentence3	
sentence4	

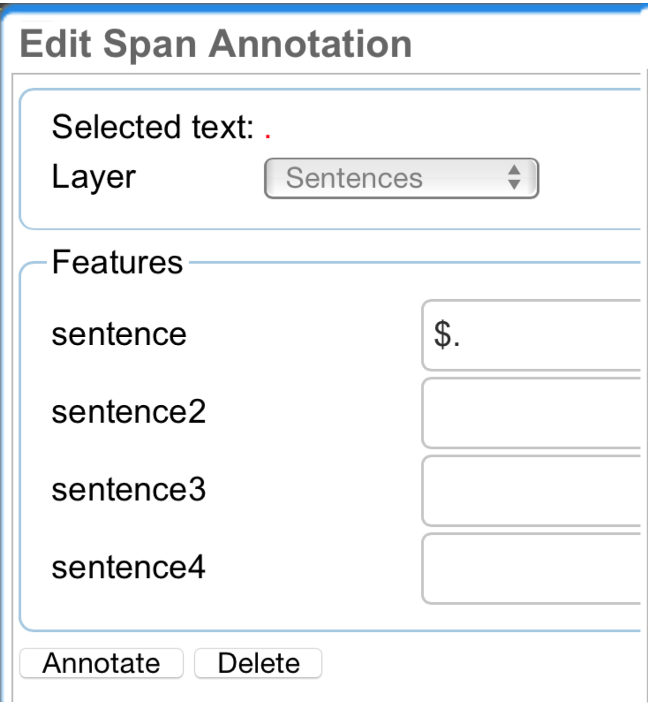


Ne mogu vjerovati\ . To nisu ljudi :-(Bježim van\ !

Deleting a sentence boundary

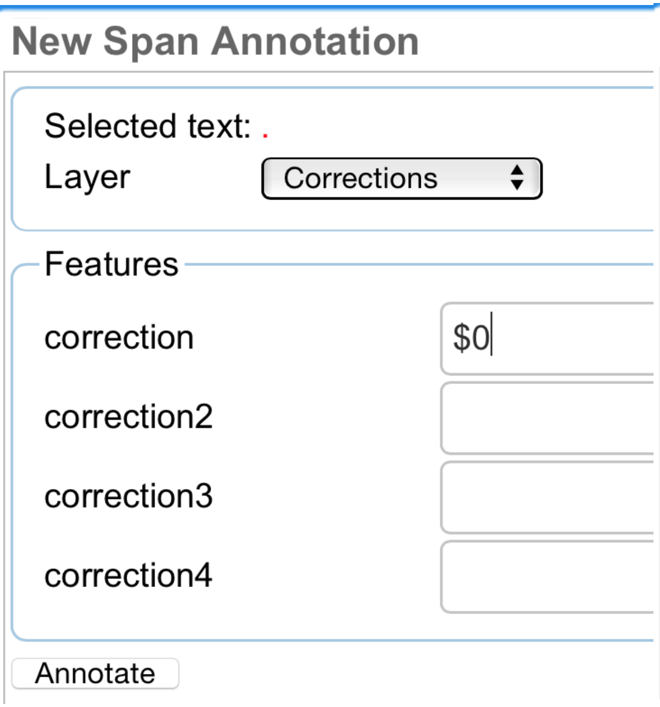


moj deb\ . brat me opet gnjavi\ !


Edit Span Annotation

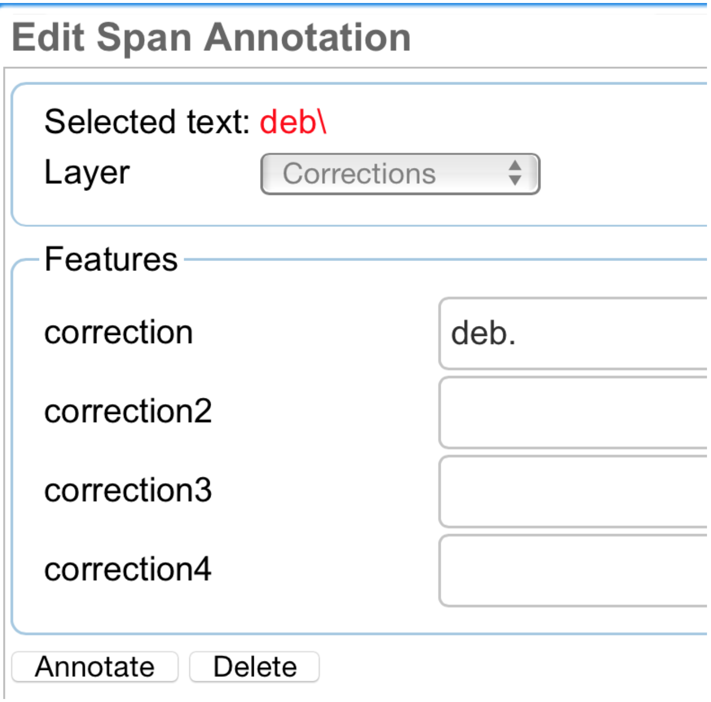
Selected text: .
Layer Sentences

Features
sentence $.
sentence2
sentence3
sentence4


New Span Annotation

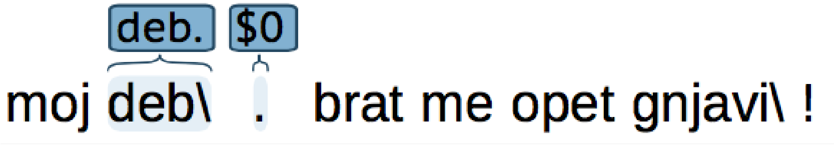
Selected text: .
Layer Corrections

Features
correction $0
correction2
correction3
correction4


Edit Span Annotation

Selected text: deb\
Layer Corrections

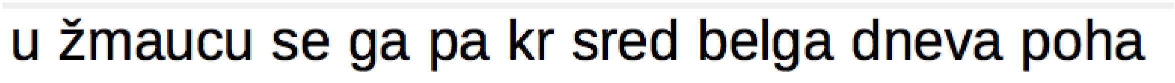
Features
correction deb.
correction2
correction3
correction4



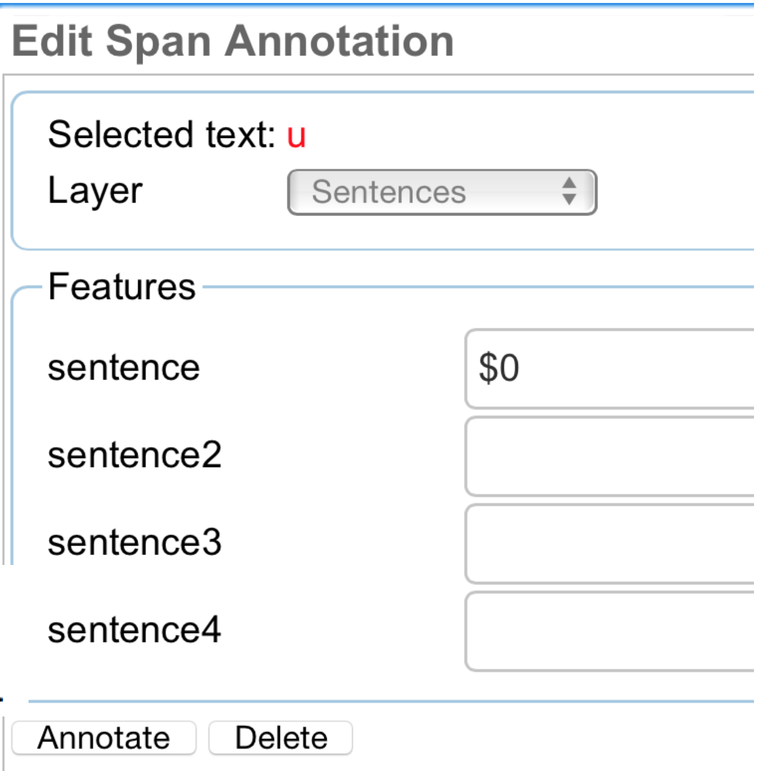
moj deb\ . brat me opet gnjavi\ !

Deleting a tweet

- If it makes no sense to annotate a tweet
- The first token should be annotated with *$0* on the layer „Sentences“



u žmaucu se ga pa kr sred belga dneva poha



Edit Span Annotation

Selected text: u

Layer: Sentences

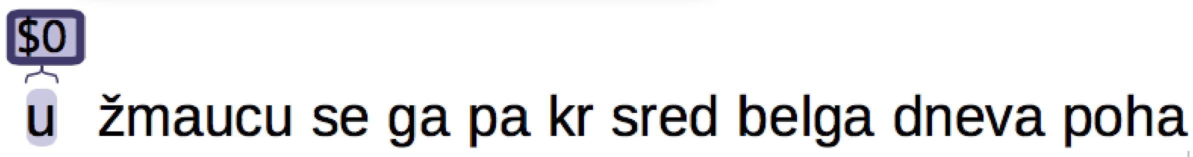
Features

sentence: $0

sentence2:

sentence3:

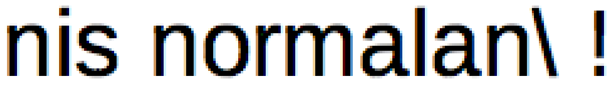
sentence4:



u žmaucu se ga pa kr sred belga dneva poha

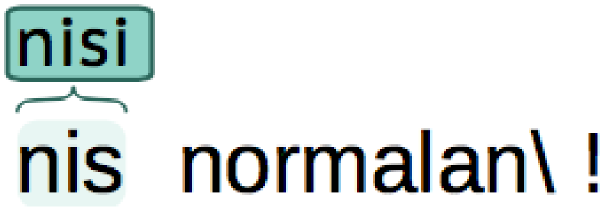
Normalisation

- Layer *Normalisations*
- Original:



nis normalan\ !

- Corrected:

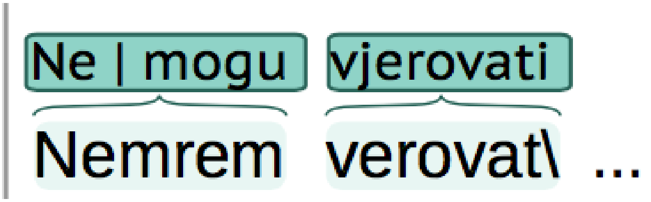


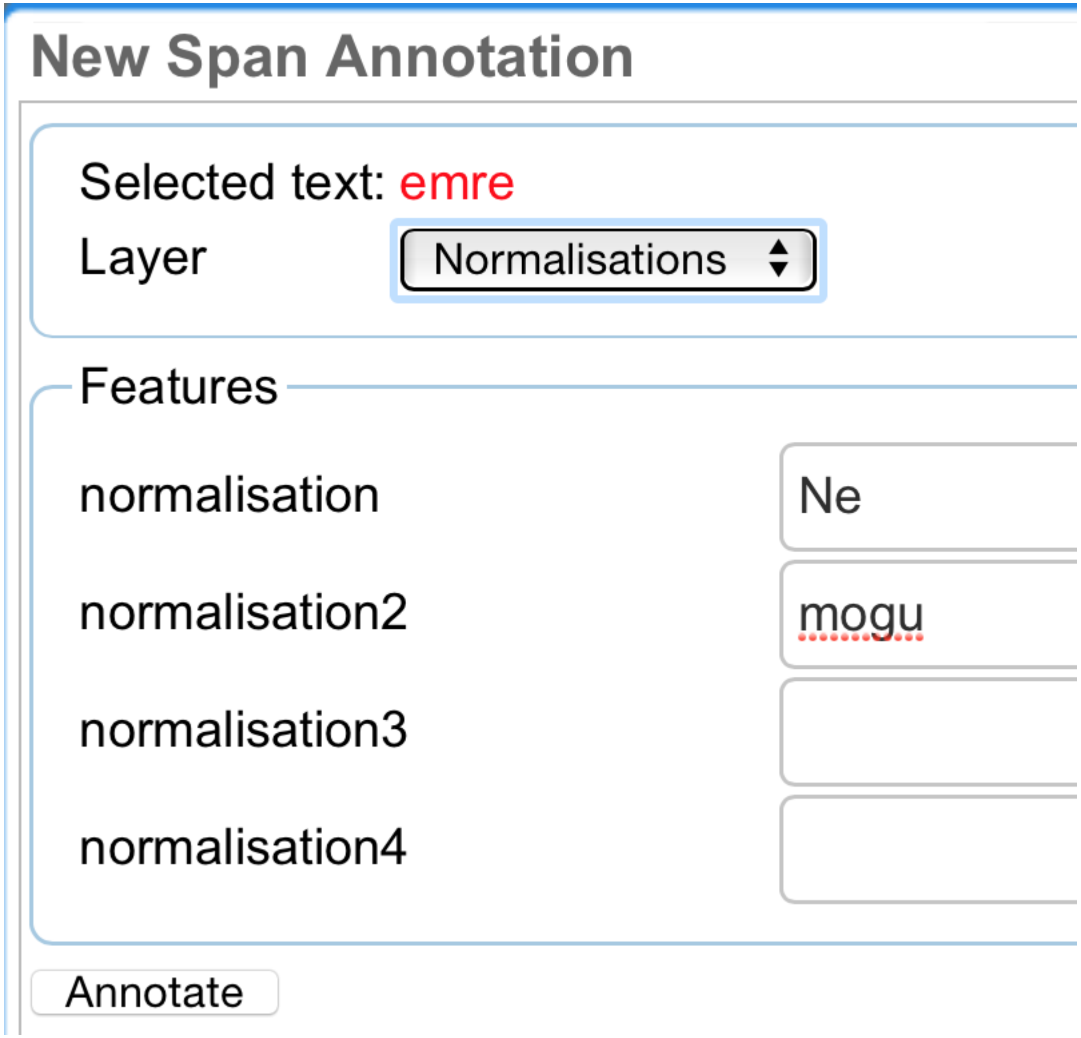
nisi

nis normalan\ !

Normalisation: splitting tokens

Nemrem verovat\ ...


Ne | mogu vjerovati
Nemrem verovat\ ...


New Span Annotation

Selected text: emre

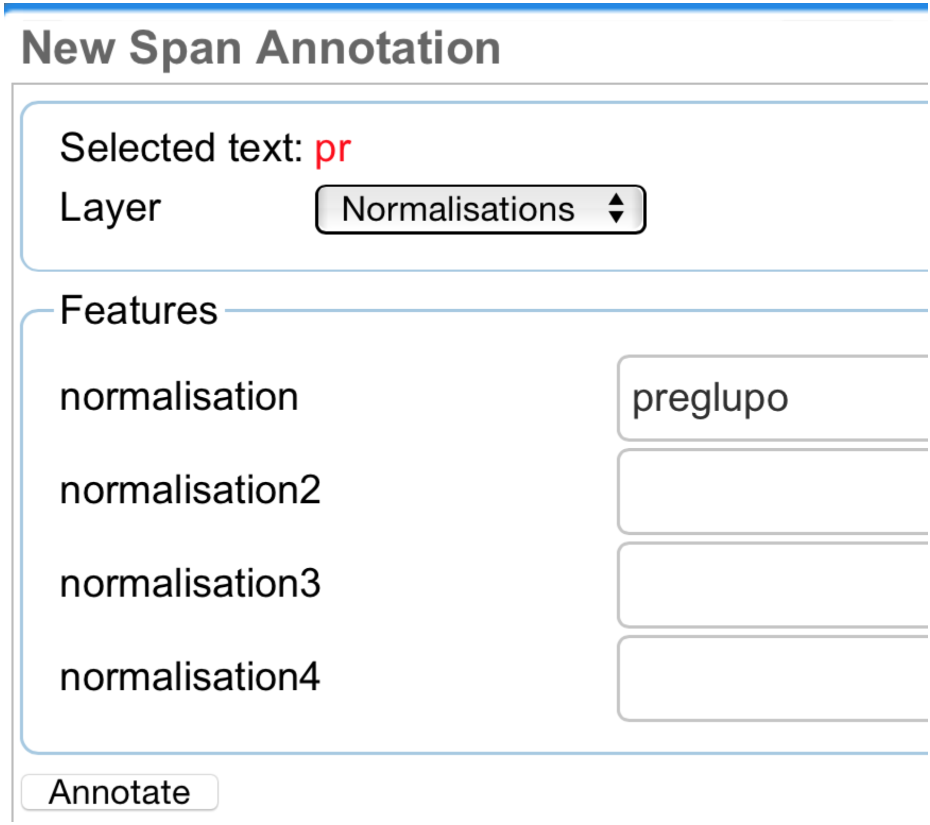
Layer: Normalisations

Features

normalisation	Ne
normalisation2	mogu
normalisation3	
normalisation4	

Normalisation: merging tokens

pre glupo buraz\ !



New Span Annotation

Selected text: pr

Layer: Normalisations

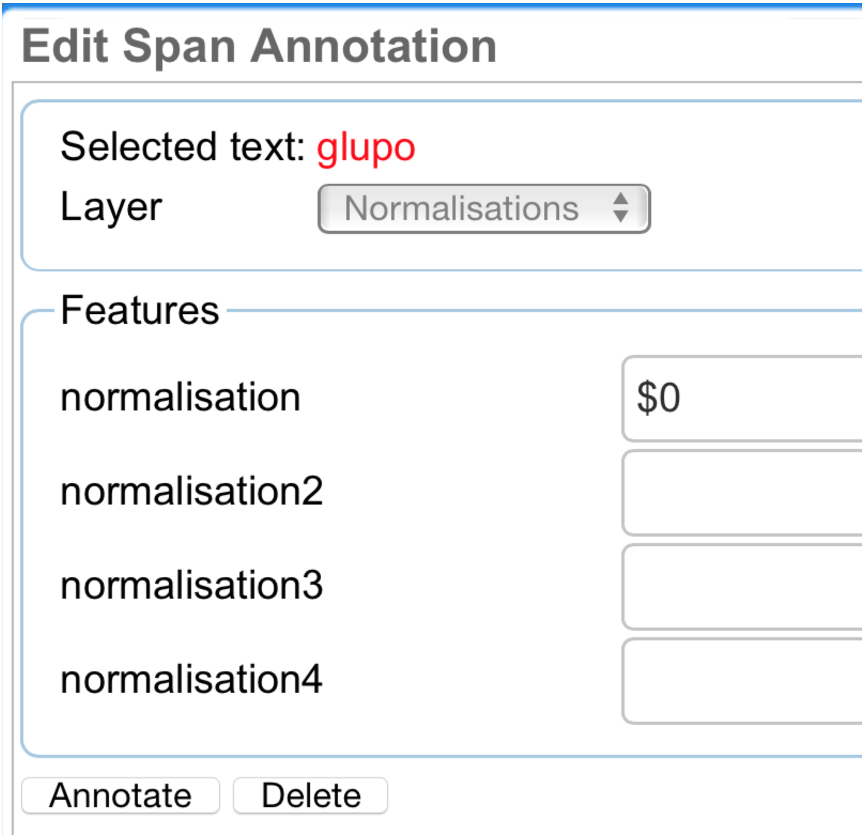
Features

normalisation: preglupo

normalisation2:

normalisation3:

normalisation4:



Edit Span Annotation

Selected text: glupo

Layer: Normalisations

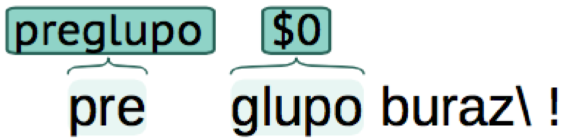
Features

normalisation: $0

normalisation2:

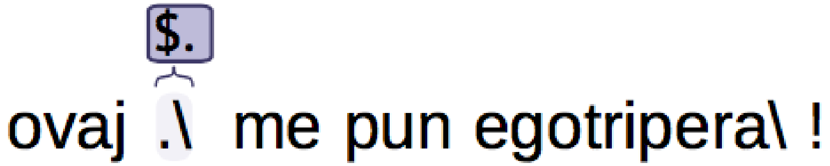
normalisation3:

normalisation4:

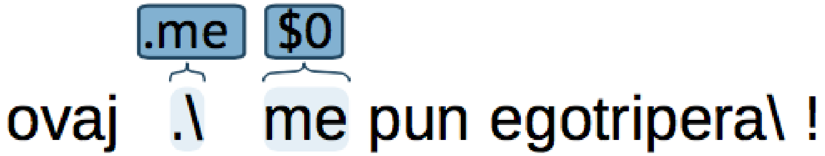


Combined corrections

- One error in automatic annotation causes others:



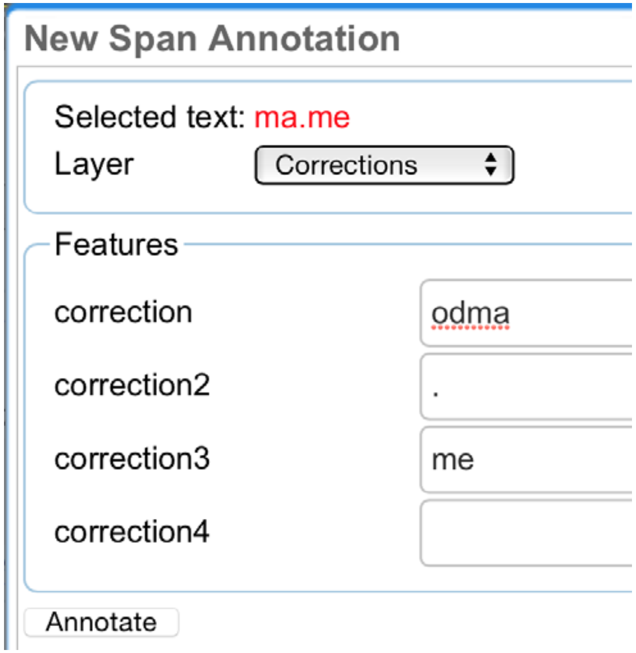
ovaj .\ me pun egotripera\ !



ovaj .\ me pun egotripera\ !

Annotating split tokens

ne odma.me kasnije pingni\ ...



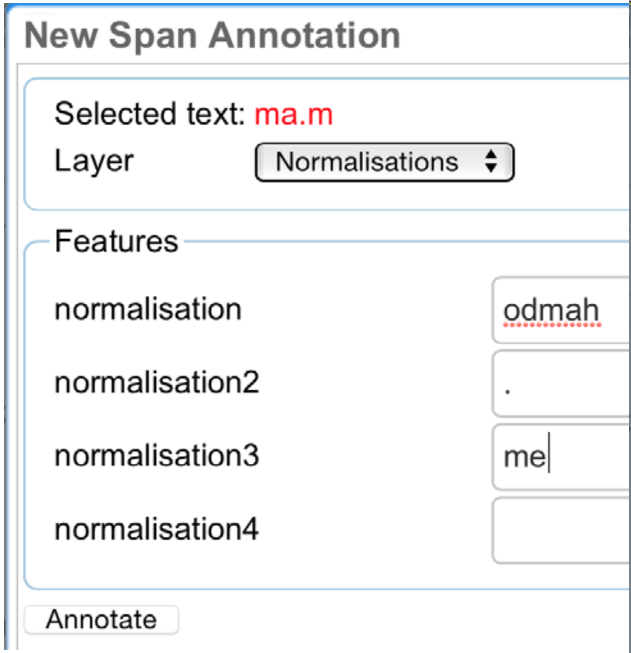
New Span Annotation

Selected text: ma.me

Layer: Corrections

Features

correction	odma
correction2	.
correction3	me
correction4	



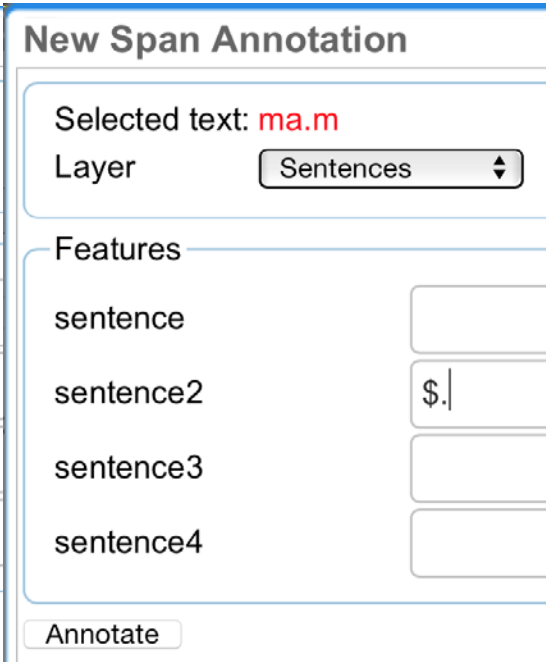
New Span Annotation

Selected text: ma.m

Layer: Normalisations

Features

normalisation	odmah
normalisation2	.
normalisation3	me
normalisation4	



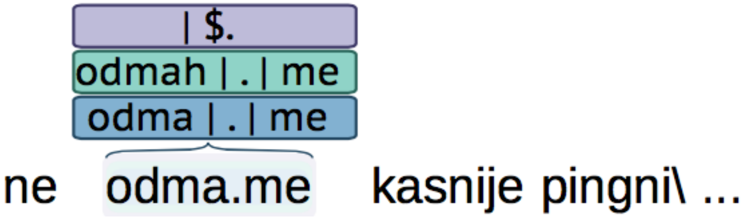
New Span Annotation

Selected text: ma.m

Layer: Sentences

Features

sentence	
sentence2	$.
sentence3	
sentence4	



| $.

odmah | . | me

odma | . | me

ne odma.me kasnije pingni\ ...

Overview

Annotation Layer	**Use**
Corrections	Tokenisation corrections
Sentences	Sentence segmentation
Normalisations	Normalised word

Char	**Layer**	**Use**
\	Tokens	No space after word
$0	Corrections	Deletion of (merged) token
$0	Sentences	Deletion of complete tweet (on first token)
$0	Normalisations	Normalisation merged on previous token
$.	Sentences	End of sentence (on token that ends the sentence)