Introduction to Normalization

This chapter summarizes the process of normalizing non-standard Slovene words. A more detailed presentation can be found in the guidelines in the Annotation Guidelines chapter.

In the case of Slovene tweets, normalization was carried out simultaneously with tokenization, as shown in Table 1.
During the manual revision, five types of corrections were identified:

- A word requiring only a direct normalized form, e.g. **tukó** was manually normalized to **tako**.
- Words incorrectly written as one and requiring split into several tokens and, if neccessary, subsequent normalization, e.g. **nauta** (dialectal spelling of "you two will not" in Slovene) was manually normalized to **ne** ("not") and **bosta** ("you two will"), with the space between the two tokens indicated by a pipe | (see Table 1).
- A word incorrectly segmented into multiple tokens that needed merging and possibly further normalization, such as changing **o ga bn** to **ogabno**; redundant tokens were marked with $0 (see Table 1).
- A word erroneously split by the tokenizer, such as **s'm** (dialectal spelling of the word form 'sem' ("am") of the verb 'biti' "to be") being incorrectly tokenized as **s** + **'** + **m**, which were then manually merged and normalized as needed.
- Words erroneously merged by the tokenizer, for instance **5km** as one token, which had to be manually split and normalised as needed.

Token	Tokenizataion	Normalization
zato		
tukó		tako
nauta		ne l bosta
s	s’m	sem
‘	$0	$0
m	$0	$0
pršva		prišla

Table 1: Normalization and tokenization of a tweet.

Revision #2
Created 2023-11-13 12:56:45 UTC by Tina Munda
Updated 2023-11-30 12:12:49 UTC by Tina Munda