Predstavitev normalizacije

V tem poglavju je strnjeno predstavljen potek normalizacije nestandardnih besed. Podrobnejšo predstavitev najdete v smernicah v poglavju Označevalne smernice.

Normalizacija tvitov, v tabeli razdeljenih na pojavnice, je potekala hkrati s tokenizacijo.
Pri ročnem pregledu je bilo odkritih 5 vrst popravkov:

- Beseda, ki ji je bilo treba zgolj pripisati normalizirano ustreznico, npr. **tukó**, ki je bil ročno normaliziran v **tako**.
- Več besed, ki so bile nestandardno zapisane skupaj in jih je bilo treba razdružiti ter po potrebi še normalizirati, npr. **nauta** v **ne** in **bosta**, pri čemer je presledek med pojavnicama označen z navpičnico | (gl. Tabelo 1).
- Beseda, ki je bila nestandardno zapisana v več pojavnicah in jo je bilo treba združiti v eno in združek po potrebi normalizirati, npr. **o ga bn** v **ogabno**; odvečne vrstice (gl. Tabelo 1) so bile označene s sosledjem znakov $0.
- Beseda, ki jo je tokenizator avtomatsko razdružil; npr. **s'm** (narečni zapis oblike »sem« pomožnega glagola) v tri pojavnice: **s** + **'** + **m**. Tovrstne pojavnice so bile najprej ročno združene in po potrebi normalizirane.
- Beseda, ki jo je tokenizator avtomatsko združil, npr. **5km** kot ena pojavnica, kar je bilo treba ročno razdružiti in po potrebi normalizirati.

Pojavnica	tokenizacija	normalizacija
zato		
tukó		tako
nauta		ne l bosta
s	s’m	sem
‘	$0	$0
m	$0	$0
pršva		prišla

Tabela 1: Normalizacija in tokenizacija tvita.

Revision #11
Created 2023-03-05 21:30:52 UTC by Tina Munda
Updated 2023-11-30 12:12:30 UTC by Tina Munda