



Rekonštrukcia a normalizácia slovenských textov

Jakub Gederá

Pod vedením Mariána Šimka

Čo robíme? Čo by sme chceli?



- Orientácia na slovenský jazyk
- Problematické spracovanie informácií z textu na Webe
- Rekonštrukcia – korekcia chýb, spell-checker
- Normalizácia – prevedenie neformálneho textu do formálnej podoby

Príklady normalizácie



- Boom! Ya ur website suxx bro
(Sarah Silverman)
- ...dats why pluto is pluto it can neva b a star
(Shaquille O'Neil)
- Problém. Nie až tak typické pre slovenský jazyk

Normalizácia v slovenčine



Typ chyby	Počet príspevkov obsahujúci danú chybu	Percento príspevkov z 200 analyzovaných príspevkov
Chýbajúca diakritika (kon)	49	24,50%
Emotikony (:D)	48	24.00%
Pridávanie interpunkcie (... !!!)	31	15.50%
Interpunkcia (. , ? !)	26	13.00%
Gramatická chyba (vyderam)	19	9.50%
Opakovanie písmen (chceeem, supeeer, hahahahaha)	6	3.00%
Skratky (FB, lol, sk)	5	2.50%
Url	3	1.5%
Shouting (ORANGE)	2	1.00%
Chýbajúce slovo do kontextu (Neskôr sa zastavím pre ...)	2	1.00%

200
slovenských
príspevkov

Rekonštrukcia



- Korekcia chýb, spell-checker
- S využitím kontextu dokážeme opravovať aj reálne chyby – slovo, ktoré je v slovníku ale nepatrí do kontextu

*Veľmi **pekný** chlapci*

- Reálne chyby paradoxne vznikajú pri automatickej oprave textu.

Zhodnotenie



- Problém normalizácie slovenských internetových príspevkov je výrazne odlišný od normalizácie anglických.
- Pre normalizáciu využitie pravidlového prístupu (*doplňovanie čiarok, odstraňovanie emotikonov, ...*)
- Pre rekonštrukciu (spell-checking) využitie jazykového a chybového modelu. Opravovanie nie len slov, ktoré nie sú v slovníku ale aj reálnych chýb.