Fast High-Accuracy Part-of-Speech Tagging by Independent Classifiers

Грачев Даня, Кошелева Даша, Медведева Маша

ТЕКСТЫ

Genre	Tokens	Percent of the whole corpus
press	5,912,746	94.15%
blogs	146,856	2.34%
religious texts	134,210	2.14%
Wikipedia articles	57,376	0.91%
essays	29,174	0.46%

Table 1. Genre composition of the Udmurt corpus.

ТЕКСТЫ

- Удмурт Дунне
- Иднакар
- Мынам Удмуртия (статьи с сайта телеканала)
- сайт правительства Удмуртии
- 150 статей из Википедии
- 7 блогов
- Библия и другие религиозные тексты
- и др.