Russian Minority Languages on the Web: Descriptive Statistics

Orekhov B., Krylova I., Popov I., Stepanova E., Zaydelman L.

Higher School of Economics

ru-lang@yandex.ru

4 June 2016

Contents

Contents

Lack of computer linguistic tools

- about a hundred national languages in Russia;
- lack of linguistic tools for these languages;
- lack of digitized texts;
- Wikipedia is a low-quality resource for these languages;
- lack of information about web representation of these languages;

Aim

- collect corpora of minority languages;
- collect some statistics about these languages on the web.

Contents

Pipeline

- Search for lexical markers in grammars and phrasebooks.
- Search for websites with Yandex.XML.
- Remove ambiguous domains from further processing.
- Download texts
  - VK API for social network
  - Crowler for the Internet
- Identify language (letter n-grams)

Contents

Social networks

- VK.com;
- 1735 communities were downloaded;
- with 1633 of them containing at least one relevant text;
- Udmurt and Bashkir are on the top of the list.

Internet

- clean list of urls and domains for 49 minor languages.
- 379 different *download-whole* language domains.
- see detailed data on our poster.

According to our data:

- offline life of a language is completely different from its online existence.
- representation of the language on the Web is not limited to the number of its speakers.

Available online colections

http://web-corpora.net/minorlangs/