ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Научный семинар Школы лингвистики НИУ ВШЭ

с участием лингвистической лаборатории по корпусным технологиям и лаборатории языков Кавказа

Борис Орехов
доцент Школы лингвистики

Людмила Зайдельман, Ирина Крылова, Иван Попов, Екатерина Степанова
студенты Школы лингвистики

Языки России из Интернета: какие, сколько и где взять?

Наиболее дорогой этап создания корпусов -- оцифровка печатных текстов. Поэтому самые большие современные корпуса составляются из того, что находится в Интернете. Корпуса (а также зависимые от наличия текстов компьютерные инструменты) для малых языков либо отсутствуют, либо недостаточно развиты.

Длящийся уже несколько лет проект (сначала под крышей РЭШ, а теперь в Школе лингвистики НИУ ВШЭ) поставил целью найти и скачать все имеющиеся в Интернете тексты на всех языках России, то есть таких, на которых говорят проживающие в России народы, но которые не являются титульными в других странах. Пока что объёмы этих интернет-сегментов небольшие и в целом позволяют это сделать, но, возможно, скоро для некоторых языков наших мощностей будет уже не хватать.

Мы разработали технологию поиска сайтов с текстами на интересующих нас языках, нашли и скачали всё, что могло бы пригодиться, из социальной сети VK.com, сложили всё это на сайте в открытом доступе и попутно проанализировали и посчитали долю присутствия каждого языка в Интернете.

Старая Басманная 21/4,
аудитория 518
25 марта (пятница) в 10:30

Сайт семинара:
ling.hse.ru/SciSeminar