


Эволюция поисковых технологий НКРЯ и их перспективы в приложении к тюркским языкам


Борис Орехов (НИУ ВШЭ, ИРЛИ РАН, НКРЯ)





Поисковые движки

НАЦИОНАЛЬНЫЙ
КОРПУС
*русского
языка*





Требования

Способность обрабатывать **миллионные** объемы
(практические соображения)

Делать это с **минимумом** отказов и за разумное
время (репутационные соображения)



Из середины 2000-х в 2020-е

Яндекс.Сервер → Яндекс.Поиск (SaaS) → ElasticSearch





Прошлое и настоящее

Яндекс.Сервер

- Решение для локального поиска
- Построение обратного индекса
- Средняя гибкость в кастомизации (нет атрибутов для единиц между словом и текстом)
- Больше не поддерживается

Яндекс.Поиск (Search as a service)

- Черный ящик
- Малая гибкость в кастомизации
- Трудности с подсчетом в результатах поиска
- Все еще используется в ряде корпусов

Статья



А. А. Аброскин Поиск по корпусу: проблемы и методы их решения. СПб.: Нестор-История, 2009.

С. 353–373

<https://ruscorpora.ru/media/uploads/2022/04/21/r60.pdf>



Настоящее и будущее

Подстроенный под наши нужды Elasticsearch.

На нем же работает tsakorpus

Позволяет строить разнообразные статистики

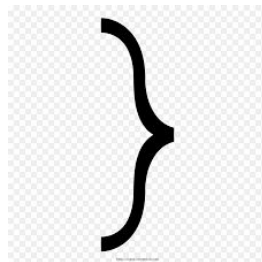
Внедряется одновременно с новыми инструментами морфологического анализа



Размер индекса сейчас

Основной (374 млн)

Газетный (850 млн)



300-400 Гб

Поиск по морфемам

Словообразование

Морфема

с учетом чередования

Тип

префикс корень суффикс флексия

Позиция

-ИНК-

8. Владимир Ляпоров. Молодая гвардия. Искусство быстрого завосбыта // «Бизнес-журнал», 2003.10.23

И пить их в клубах (ведь не секрет, что значительная часть продаж дорогих алкогольных оптовые закупки ночных клубов, дискотек и больших *вечеринок*).

9. Илья Петрусенко. Я вдыхаю ветер воли... // «Народное творчест

И пусть ты *песчинка* в мироздании, но что-то вдруг разом всколыхнет все твоё сущест порадуешься своей причастности к этому трудному, но притягательному земному бы

10. Беззащитная братва // «Криминальная хроника», 2003.07.08

11 часов — Говоров уже в мантии, секретарь Анечка заботливо сдула с нее *пылинки*, м телефон.



Морфология





От правил к машинному обучению

MyStem → Рубик



Рубик

Разметка на основе нейросетевой модели

Несколько этапов дополнительной коррекции

Требует вычислений на GPU

Дает возможность избавиться от паразитических разборов

Статья



Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Дони́на О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания, 2024. № 2. С. 7–34.

<https://ruscorpora.ru/media/publications/6e9ac7da0736967e4d9750e865eb91b3.pdf>

Тюркские языки в НКРЯ





Тюркские параллельные

башкирский, 550 тыс.



хакасский, 1 194 тыс.

чувашский, 24 168 тыс.

башкирский:

Беззә асфальтбетонға кушыу өсөн якшы катнашмалар юк, улар һауа торошоноң үзгәрештәренә каршы торорға һәм юлдарзы һакларға ярзам итер ине. Махсус онталған вак *таш* урынына кәзимге кырсынташ кулланалар, балсыкты еткерә тотонмайзар, урлайзар.  

башкирский:

90-сы йылдарза йөк *ташыу* тимер юлдан автомобиль юлдарына күсте, ә был тәңгәлдә тейешле контроль әлегәсә юк.  

русский:

Ну, самое главное, что башкирский язык есть в Яндекс.
Переводчике, правда в бета-версии. 📄 ↔

русский:

Ну, самое главное, что башкирский язык есть в Яндекс.
Переводчике, правда в бета-версии. Марийский язык тоже пока
бета. Мы начали работу по включению марийского языка в
Яндекс. Переводчик в июне этого года. 📄 ↔

башкирский:

Иң мөһиме, *башкорт теле* Яндекста бар. *Тәржемәлә бета*
версияла ғына. 📄 ↔

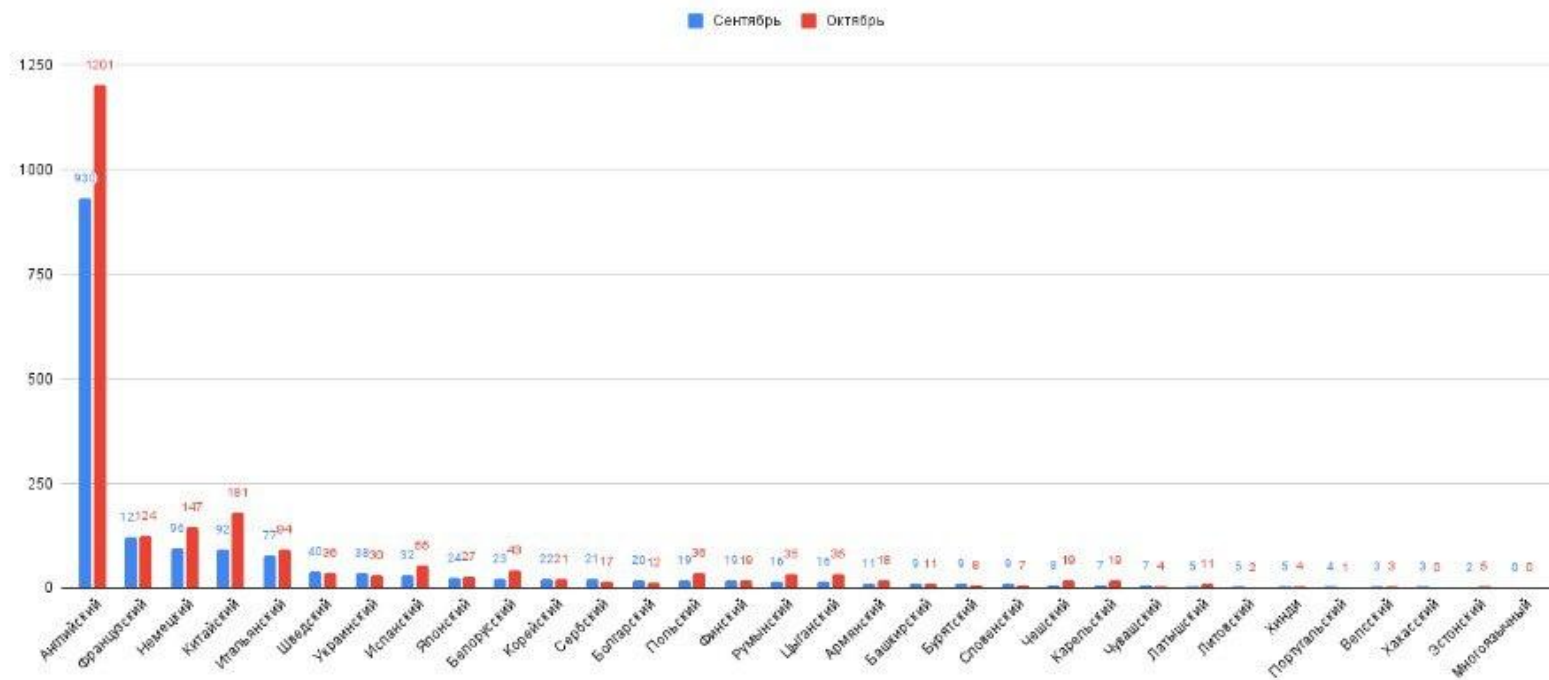
башкирский:

Иң мөһиме, башкорт теле Яндекста бар. Тәржемәлә бета
версияла ғына. *Мари теле* лә әле *бетала* ғына Без
Яндекс.Переводчикка мари телен индереүзе быйылғы йылдың
июнендә генә башлағайнык 📄 ↔



<https://ruscorpora.ru/s/dLo1r>

Уникальные пользователи по отдельным параллельным корпусам, чел.



Неожиданно высокие показатели башкирского



Число пользователей на миллион словоупотреблений (параллельные)

Корейский 300,0

Японский 54,0

Китайский 41,14

Английский 24,02

Бурятский 20,00

Башкирский 18,33

Французский 16,32

Хакасский 1,25

Чувашский 0,17

Статья



Сичинава Д. В. Параллельные тексты в составе Национального корпуса русского языка: новые языки и новые задачи // Труды Института русского языка им. В.В. Виноградова. 2019. 3 (21). С. 41-61.

<https://ruscorpora.ru/media/publications/0b53ed6b886e099dca3aee3767151452.pdf>



<https://nevmenandr.github.io/>