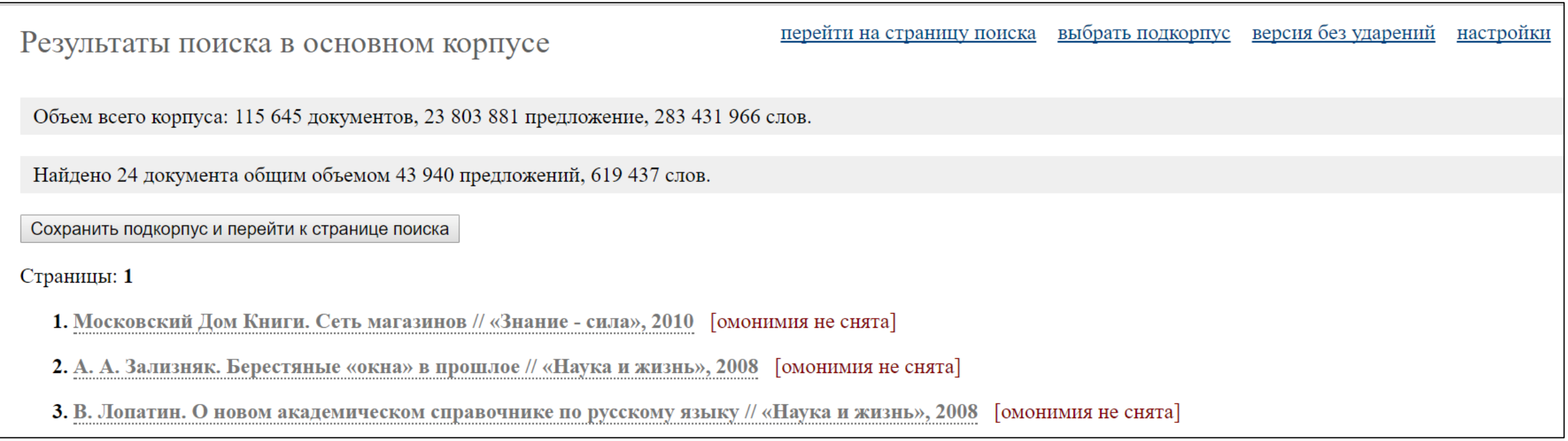
Русский академический корпус, или *CAT and the kittens*

Михаил Копотев (Хельсинкский университет)

CAT and the kittens

Главная задача — создание онлайн-сервиса, который сравнивает студенческий текст (“kitten”) с эталонным представительным корпусом (CAT) и помечает «подозрительные» токены и их сочетания, которые **резко** отличают студенческий текст от эталонных.



Результаты поиска в основном корпусе

перейти на страницу поиска | выбрать подкорпус | версия без ударений | настройки

Объем всего корпуса: 115 645 документов, 23 803 881 предложение, 283 431 966 слов.

Найдено 24 документа общим объемом 43 940 предложений, 619 437 слов.

Сохранить подкорпус и перейти к странице поиска

Страницы: **1**

1. Московский Дом Книги. Сеть магазинов // «Знание - сила», 2010 [омонимия не снята]
2. А. А. Зализняк. Берестяные «окна» в прошлое // «Наука и жизнь», 2008 [омонимия не снята]
3. В. Лопатин. О новом академическом справочнике по русскому языку // «Наука и жизнь», 2008 [омонимия не снята]

CAT and the kittens

Первая задача (простая) — создание корпуса CAT:

- выработка критериев отбора текстов
- выработка структуры метаданных (с учетом уже имеющихся на сайтах cyberleninka.ru и elibrary.ru)
- получение текстов (crawling) и метаданных (page parsing) на основе выработанных критериев
- собственно создание корпуса (включая лемматизацию, аннотирование и т.д.)

CAT and the kittens

Вторая задача (сложная) — создания онлайн сервиса:

- выработка параметров для сравнения kitten относительно CAT (например, длина предложения; частоты знаменательных токенов / n-грам; близость текста к эталонным по моделям doc2vec / word2vec; вхождение (ключевых) токенов в семантический класс соответствующего тематического домена и др.)
- Создание системы сравнения по выработанным параметрам: а) создание сервиса, который «на лету» создает для загруженного kitten: семант. вектора, n-граммы, лемматизацию и т.д. и б) создание алгоритмов сравнения этих параметров с данными из CAT «эталонными».