Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2022»

Москва, 15–18 июня 2022 г.

Prototype of a next-generation corpus platform for the RNC[1]

S. Gladilin, V. Sizov, A. Kazennikov, D. Morozov, P. Dyachenko, O. Don, A. Kozerenko, S. Piskounova, A. Makhova, N. Bujlova
Institute for Information Transmission Problems (Kharkevich Institute)
{gladilin@iitp.ru, victor.sizov@gmail.com, kazennikov@gmail.com, morozov@ruscorpora.ru, pavelvd@iitp.ru, oleg.don.it@yandex.ru, akozerenko@ruscorpora.ru, saruwatari.lara@gmail.com, discourse@yandex.ru, nbujlova@ruscorpora.ru}

Abstract

The paper presents the prototype of a next-generation corpus platform for the RNC, developed in 2020-2021. Internet search engine systems and relational database management systems have been considered as potential tools for designing corpus platforms. Also, the paper analyzes some of currently existing corpora and reveals their shortcomings, which show that a new kind of platform has to be developed specifically for the RNC. The article describes a functional specification for such a platform and lists the key features of the developed prototype. The evaluation of the prototype showed that it's interactivity, reliability and scalability do not decrease in comparison with the existing RNC platform, while the accuracy and functionality increase. Several prototype modules have been implemented in the RNC, and some examples of how these modules work are included in the paper.

Keywords: corpus linguistics; corpus platforms; full-text search engines; lexico-grammatical search engines; RNC; homonymy

DOI: 10.28995/2075-7182-2022-21-1043-1054

Прототип корпусной платформы нового поколения для НКРЯ

Гладилин С. А., Сизов В. Г., Казенников А. О., Морозов Д. А., Дяченко П. В., Дон О. Р., Козеренко А. Д., Пискунова С. В., Махова А. А., Буйлова Н. Н.
Институт проблем передачи информации им. А. А. Харкевича
{gladilin@iitp.ru, victor.sizov@gmail.com, kazennikov@gmail.com, morozov@ruscorpora.ru, pavelvd@iitp.ru, oleg.don.it@yandex.ru, akozerenko@ruscorpora.ru, saruwatari.lara@gmail.com, discourse@yandex.ru, nbujlova@ruscorpora.ru}

Аннотация

В статье представлен прототип корпусной платформы НКРЯ нового поколения, разработанный в 2020–2021 гг. Произведен анализ систем поиска по сайтам в сети Интернет и реляционных систем управления базами данных как инструментов для создания корпусных платформ. Также произведен анализ существующих корпусных платформ. Выявленные недостатки существующих программных продуктов приводят к необходимости создания собственной платформы. Сформулированы функциональные требования к такой платформе. Перечислены ключевые особенности разработанного прототипа. Выполненные оценки показали, что интерактивность, надежность и масштабируемость прототипа не снижаются по сравнению с существующей корпусной платформой НКРЯ, в то время как точность и функциональность повышаются. Приведены доступные для пользователей результаты внедрения модулей прототипа в существующую программную платформу НКРЯ.

Ключевые слова: корпусная лингвистика; корпусные платформы; полнотекстовые поисковые системы; лексико-грамматические поисковые системы; НКРЯ; омонимия

[1] Работа выполнена при поддержке гранта Министерства науки и высшего образования Российской Федерации № 075-15-2020-793.

1 Введение

Национальный корпус русского языка (НКРЯ)[2] — разрабатываемый на протяжении уже почти 20 лет компьютерно-лингвистический ресурс общим объемом около 1,5 миллиарда словоупотреблений, доступный в виде сайта сети Интернет [10]. На сайте развернута корпусная платформа, обеспечивающая возможность поиска и визуализации корпусных данных. Эта корпусная платформа основана на специализированном программном обеспечении Яндекс.Сервер/Яндекс.Поиск, предназначенном для поиска в корпоративных сетях и сети Интернет [1]. Действующая корпусная платформа обеспечивает поддержку ряда уникальных особенностей НКРЯ, однако имеет также ряд принципиальных ограничений, не позволяющих, в частности, реализовать на ее основе современные корпусные инструменты, такие как, например, коллокационный анализ, поиск по синтаксическим связям и др. В настоящей работе описывается прототип корпусной платформы нового поколения для НКРЯ, разработанный в 2020–2021 гг.

2 Ключевые отличия корпусных платформ от других информационно-справочных систем

Корпусные платформы — подвид информационно-справочных систем, предназначенных для организации работы с лингвистическими корпусами текстов. Основной операцией в таких системах является поиск слов или словосочетаний, удовлетворяющих заданным условиям. Это делает корпусные платформы схожими с системами поиска по сайтам в сети Интернет. И те, и другие, как правило, основаны на полнотекстовых инвертированных индексах для быстрого выполнения пользовательских запросов [8].

Несмотря на схожесть решаемой задачи, между поисковыми системами и корпусными платформами есть ряд существенных различий. Во-первых, поисковые системы ориентированы на поиск целых документов (текстов), в то время как корпусные платформы — на поиск отдельных примеров в текстах. В одном и том же тексте могут содержаться тысячи примеров, удовлетворяющих условиям поиска. Как следствие, возникают различия в требуемых стратегиях исполнения поисковых запросов и подсчета количества найденных вхождений. Так, например, в поисковых системах требуется подсчет только найденных документов без анализа числа вхождений в каждом из них и числа предложений, в которые входят найденные вхождения.

Во-вторых, основной целью поисковых систем является нахождение малого количества наиболее релевантных документов, в то время как в корпусных платформах важно точное соответствие результатов запросу и возможность последующего статистического анализа и визуализации всего множества найденных вхождений.

В-третьих, в то время как задача поисковой системы — «угадать» по неструктурированному запросу пользователя, что он хотел найти, язык запросов в корпусной системе является математически строгим, т.е. запрос полностью описывает требования к искомым вхождениям. Это реализуется через наложение поисковых условий с привязкой к способу разметки корпуса.

Другим подходом, реализованным, например, в системе PML-TQ [11] является построение корпусной платформы на основе реляционной системы управления базами данных (РСУБД), поддерживающей стандартизированный язык запросов SQL. Достоинство языка SQL — его декларативность, то есть возможность строить произвольные запросы через описание того, что требуется найти, а не того, как должен работать алгоритм поиска. При этом SQL очень выразителен — он позволяет строить запросы из очень широкого класса, что делает его применимым к почти любой задаче, возникающей в информационно-справочных системах. Кроме того, существующие РСУБД содержат развитые механизмы оптимизации плана исполнения запросов.

Однако язык SQL требует описания запросов в виде сложных операций соединения (*англ.* join), что в применении к корпусной лингвистике обосновано только для корпусов с разметкой лингвистических связей (*англ.* treebank), а в остальных случаях создает излишнюю вычислительную нагрузку, делая РСУБД существенно менее эффективным инструментом, чем полнотекстовые поисковые системы. Обратим также внимание на то, что языки реляционной алгебры (в т.ч. SQL)

[2] https://ruscorpora.ru/

далеки от естественной для корпусной лингвистики формы задания поисковых запросов. Поэтому их применение возможно только в случае включения в корпусную платформу массивного модуля автоматического перевода лингвистических запросов на язык SQL, как это сделано в системе PML-TQ.

Исходя из вышесказанного, можно заключить, что ни существующие системы поиска по сайтам в сети Интернет, ни РСУБД не пригодны для использования в качестве корпусной платформы, а могут быть использованы лишь как компоненты для её построения.

3 Поддержка нескольких омонимичных разборов в корпусных данных

Важной проблемой разметки корпусных текстов является омонимия — неоднозначность определения характеристик (обычно — грамматических) размечаемых единиц текста. Можно выделить несколько типов неоднозначности: снимаемая (омонимия может быть снята при помощи контекста внутри предложения); контекстуальная (омонимия может быть снята только с учетом контекста/прагматики); неснимаемая (возможны разные интерпретации того, какими характеристиками обладает конкретная единица) [6]. В зависимости от решаемых при помощи корпуса задач предпочтительным может быть как максимально возможное снятие омонимии (вплоть до присвоения каждой единице текста строго одного набора характеристик), так и сохранение в корпусе нескольких разборов одной словоформы.

С самого начала в Национальном корпусе русского языка была внедрена автоматическая грамматическая разметка текстов, подразумевающая сохранение всех неоднозначных разборов слов [7]. В то же время, в Основном корпусе НКРЯ существует подкорпус со снятой вручную омонимией (небольшой по сравнению с объемом всего корпуса — 6 из 375 млн словоупотреблений). Также омонимия снята в корпусе СинТагРус в составе НКРЯ и в ряде исторических корпусов. Кроме того, авторам известно несколько проектов, ставящих задачу автоматического снятия лексической и грамматической омонимии в некоторых корпусах НКРЯ. Можно предположить, что в будущем количество корпусов НКРЯ со снятой омонимией будет расти.

При наличии развитой системы с тремя типами корпусных данных (со снятой вручную омонимией, с автоматически снятой омонимией, с неснятой омонимией), к разрабатываемой корпусной платформе нового поколения для НКРЯ очевидно предъявляются требования поддержки как снятой, так и неснятой омонимии.

4 Существующие корпусные платформы

В рамках выполнения проекта был проанализирован ряд распространенных корпусных платформ, предназначенных для поддержки корпусов большого объема: SketchEngine/NoSketchEngine [5], Corpus Workbench (CWB) [4], BlackLab [3] и MTAS [2]. Изучался вопрос поддержки данными платформами разметки с неснятой омонимией. Все рассмотренные корпусные платформы основаны на полнотекстовых поисковых системах. В поисковом индексе таких систем для каждого *терма*, то есть словоформы, леммы или лингвистической пометы (например, падежа, числа, рода, …) записываются все его позиции в каждом тексте [8]. Терм уникально определяется координатами в виде идентификатора текста и позиции в нем. Омонимия реализуется через запись нескольких термов, соответствующих разным атомарным разборам, на одной и той же позиции в тексте. Но в случае составных разборов, как, например, в Национальном корпусе русского языка, такая реализация может давать ложные срабатывания при поиске, когда искомая комбинация термов принадлежит разным разборам, но отсутствует такой разбор, который бы содержал все термы искомой комбинации. В работе [1] эта проблема получила название «проблема перемешивания характеристик слов» (так же известная как «проблема мешка»). Например, если у некоторого слова есть 2 разбора (ЛЕММА1, S) и (ЛЕММА2, V), то в такой реализации это слово будет ошибочно найдено по запросу (ЛЕММА2, S). Таким образом, прямая поддержка неснятой омонимии для составных грамматических разборов отсутствует во всех перечисленных платформах.

Существуют варианты ограниченного использования составных омонимичных разборов, если в корпусной платформе отсутствует их прямая поддержка. Первый подход — преобразование составных разборов в атомарные. Все составные части разбора склеиваются в одну строку, образуя таким образом атомарное значение. Такой вариант поддержки омонимии реализован в существующей версии корпусной платформы Национального корпуса русского языка отдельно для грамматических и отдельно для семантических признаков. Существеннейшим недостатком такого подхода является необходимость хранения в индексе всевозможных сочетаний термов, что ведет к многократному увеличению индекса [1].

Другим подходом к реализации частичной поддержки омонимии является введение дополнительных позиционных пространств термов. Такой подход применяется, например, в корпусной платформе SketchEngine/NoSketchEngine. Однако, обеспечивая компактное хранение, такой подход требует поиска по подстроке терма, что может существенно снижать эффективность за счет отказа от прямого поиска терма в пользу поиска всех термов, удовлетворяющих заданному паттерну (к примеру — регулярному выражению).

Впрочем, оба приведенных подхода не решают проблему перемешивания характеристик полностью. Так, в обоих случаях лемма и словоформа не включаются в составной терм, поэтому в результате поиска одновременно по лемме/словоформе и грамматическим характеристикам в результаты поиска могут попадать позиции, в которых искомое сочетание не встречается ни в одном разборе, но по отдельности искомые лемма/словоформа и грамматические характеристики присутствуют в разных разборах.

Таким образом, можно заключить, что ни одна из существующих корпусных платформ не удовлетворяет требованиям в части полноценной поддержки неснятой омонимии, предъявляемым к разрабатываемой корпусной платформе НКРЯ нового поколения.

5 Корпусная платформа нового поколения для НКРЯ

К разрабатываемой платформе нового поколения предъявлялись следующие функциональные требования:

- поддержка поиска по корпусу как в режиме автоматически снятой омонимии, так и в режиме поиска всех вариантов омонимичных разборов;
- поддержка современных инструментов статистики и визуализации, таких как расчет n-грамм как по формам, так и по леммам, и как по всему корпусу, так и по подкорпусу; коллокаций; портретов («скетчей») слова и т.д.;
- поддержка поиска по синтаксическим отношениям в корпусах большого объема;
- корректная работа с большими поисковыми выдачами: сортировка как по мета-атрибутам, так и по атрибутам найденного вхождения, прореживание, отображение результатов на большую глубину поиска, расширенная выгрузка фрагмента в файл и т.д.;
- поддержка корпусов с несколькими разнотипными видами разметки (параллелизованные исторические корпуса, мультимедийный параллельный поэтический корпус и т.д.);
- поддержка межкорпусного и панхронического поиска.

В рамках выполняемого проекта разработан прототип корпусной платформы нового поколения для НКРЯ. Под созданием прототипа понималась разработка алгоритмического и программного обеспечения и его интеграция с НКРЯ, позволяющая экспериментально исследовать предварительные версии модулей будущей платформы в близком к запланированному применению окружении. Перечислим ключевые технические особенности созданного прототипа.

1. Созданный прототип спроектирован для одновременной поддержки нескольких различных технологий баз данных и полнотекстового поиска. В состав НКРЯ входят как большие корпуса с преимущественно морфологической разметкой (основной, газетный и т.д.), так и глубоко аннотированный корпус с разметкой синтаксических связей и лексических функций СинТагРус. Для поддержки первого вида корпусов оптимально использовать полнотекстовые поисковые системы, для поддержки второго вида — РСУБД. Фактически, в настоящее время НКРЯ обслуживается двумя различными корпусными платформами

(одна — для корпусов большого размера, другая — для СинТагРуса), заведенными под общий пользовательский интерфейс. Функциональность этих платформ частично дублируется, но значительная часть функциональности оказывается просто не поддерживаемой на одной из платформ. Прототип корпусной платформы нового поколения обеспечивает единую реализацию всей функциональности в ядре системы и возможность подключения к ядру различных поисковых систем и РСУБД через абстрактную модель запроса, ранее предложенную в работе [9].

2. Обеспечена поддержка запланированного подхода к построению пользовательского запроса более общего вида, чем в существующей корпусной платформе НКРЯ. Данный подход ранее был предложен для корпуса СинТагРус [12].

3. Создана специализированная полнотекстовая поисковая система, поддерживающая эффективную модель представления омонимичных словарных разборов. В частности, разработаны основные поисковые предикаты, позволяющие комбинировать как однозначные, так и неоднозначные разборы на данной позиции в тексте, основные предикаты на расстояние, а также возможность сортировки результатов не только по атрибутивным признакам текста, но и по атрибутам словоформ в найденных примерах. Данная поисковая система основана на открытом программном обеспечении ElasticSearch и реализована как встраиваемый модуль (plug-in) к нему, обеспечивающий необходимое расширение функциональности.

4. В разработанной специализированной полнотекстовой поисковой системе обеспечена возможность вычислительно эффективной реализации инструментов корпусной статистики. Разработана агрегация поисковых запросов по мета-атрибутам (атрибутам, приписанным к текстам) с ведением статистики не только по количеству найденных текстов, но и по количеству примеров. Реализована динамическая агрегация по n-граммам поисковых запросов, что позволяет гибко комбинировать поисковые критерии как по содержательной части текста, так и по атрибутивной.

5. Спроектированная архитектура корпусной платформы нового поколения обеспечивает изоляцию ядра платформы от подсистемы пользовательского интерфейса. Это создает возможность заменять пользовательский интерфейс системы без коррекции её ядра, что представляется важным, поскольку скорость изменения технологий пользовательского интерфейса в сети Интернет значительно превышает скорость эволюции НКРЯ. Еще одной возможностью такой архитектуры является создание программного интерфейса (API) для обращения к системе из пользовательских программ, минуя графический интерфейс.

6 Результаты разработки прототипа Корпусной платформы НКРЯ нового поколения

В рамках выполненной работы создан и исследован программный прототип Корпусной платформы НКРЯ нового поколения. Прототип не является единым программным продуктом, готовым к эксплуатации как единое целое. Часть разработанных модулей были интегрированы в существующую программную платформу НКРЯ и, таким образом, уже доступны пользователю через сайт НКРЯ[3]. Другая часть испытывалась на специально подготовленных программных стендах.

6.1 Исследование разработанного прототипа

В рамках исследования созданный прототип оценивался в соответствии со следующими требованиями:

1) интерактивность, то есть комфортная для интернет-пользователя скорость ответа сайта на заданные запросы;

2) надежность, то есть правильное функционирование прототипа при всех сценариях использования системы;

[3] https://ruscorpora.ru/

3) точность, то есть соответствие результатов поиска заданным условиям;
4) масштабируемость, то есть готовность системы к загрузке в нее растущих объемов текстов;
5) функциональность, то есть реализация возможностей для решения задач пользователя.

Рассмотрим каждую из этих характеристик в отдельности.

1. В настоящее время выдача в формате KWIC для основного и газетных корпусов на сайте НКРЯ реализована модулем, созданным в рамках разработанного прототипа, в то время как обычная выдача выполняется существующей программной платформой НКРЯ. Проведенные многочисленные экспериментальные сравнения для разных поисковых запросов показали, что поиск в формате KWIC обеспечивает субъективно такую же «почти мгновенную» выдачу, как и запрос в обычном формате, то есть разработанный прототип обладает не худшей интерактивностью, чем существующая система НКРЯ.
2. В процессе разработки прототипа каждая выпускаемая версия (как доступная на сайте НКРЯ, так и функционирующая на специальных тестовых стендах) подвергается всестороннему ручному тестированию. Тестирование включает порядка пятидесяти сценариев, для каждого из которых функционирование системы проверяется на правильность. Результаты тестирования позволяют утверждать, что разработанные модули прототипа обеспечивают большую надежность, чем существующая система НКРЯ, поскольку в процессе разработки был исправлен ряд ранее существовавших в системе ошибок, но сохранено правильное поведение в остальных сценариях.
3. Точность поиска разработанного прототипа выше точности существующей системы, поскольку разработанная специализированная полнотекстовая поисковая система обеспечивает раздельное хранение омонимичных разборов, что полностью решает проблему перемешивания характеристик. Для исследования точности поиска производилось сравнение результатов в режиме обычной выдачи (реализованной через существующую корпусную платформу НКРЯ) и режиме KWIC (реализованном через разработанный прототип).

   Рассмотрим, например, запрос «**пирога, m**», обеспечивающий поиск словоупотреблений с леммой «пирога» (индейская лодка) мужского рода. Существующая система НКРЯ находит в основном корпусе 6936 вхождений, ни одно из которых на самом деле не является искомым, поскольку лемма «пирога» женского рода. Все эти вхождения находятся в подкорпусе с неснятой омонимией. Причина такой неточности поиска — перемешивание характеристик леммы и грамматики: у данных словоупотреблений есть омонимичный разбор с леммой «пирог» мужского рода. В то же время, разработанный прототип дает правильный результат: «по запросу ничего не найдено».
4. В процессе разработки прототипа Корпусной платформы нового поколения данные НКРЯ продолжали пополняться. Так, за два года реализации проекта суммарный объем основного и газетного корпусов НКРЯ вырос более чем в два раза. Каждое пополнение загружалось параллельно как в существующую корпусную платформу НКРЯ, так и в разрабатываемый прототип. При этом производился мониторинг параметров работы сервера (загрузки процессора, дискового ввода/вывода и т.д.) на предмет перегрузок и экстремальных значений. Мониторинг показал, что разработанный прототип успешно масштабируется.
5. В настоящее время разработанный прототип реализует не всю функциональность существующей корпусной платформы. Однако реализованные модули покрывают весь спектр внутренних операций специализированной полнотекстовой поисковой системы, что позволяет утверждать, что вся еще не реализованная функциональность может быть успешно реализована в рамках предложенного подхода.

   В то же время, реализован ряд новых функций, отсутствовавших в существующей корпусной платформе НКРЯ. Таким образом, прототип демонстрирует возможность существенного расширения функциональности в Корпусной платформе нового поколения.

6.2 Новая функциональность разработанного прототипа

К новой функциональности прототипа относится:

1. Сортировка по правому/левому контексту в основном и газетных корпусах с опциональным прореживанием. Существующая корпусная платформа НКРЯ ранее позволяла осуществлять сортировку прореженной выдачи (1000 случайных примеров) по правому или левому контексту искомого слова. Разработанный прототип позволяет осуществлять сортировку по контексту как с прореживанием, так и без прореживания, ср. пример отсортированных по правому контексту слова результатов поиска на Рис. 1. В настоящее время данная возможность доступна через сайт только в режиме с прореживанием, режим без прореживания испытан на стендах.

стали корить англичан Ватерлооским мостом,	или	австрийцы -- французов Вандомскою колонною. ←…→
осуществляется разделение в географическом, технологическом	или	административном отношении, система разрывается. ←…→
которого все зовут Капитоном Павловичем	или	адмиралом Макаровым. ←…→
лучшего качества, чем тогдашний американский	или	английский. ←…→
которую в зависимости от симпатий	или	антипатий к нему характеризовали как ←…→
Результатом мог бы гордиться археолог	или	антрополог, или реставратор высочайшего класса ←…→
пригласили меня быть третейским судьей	или	арбитром: бросать ли ему семью ←…→
города и у потенциального покупателя	или	арендатора всегда есть возможность выбора ←…→
контрсилы -- будь то луки, стрелы	или	атомные бомбы, которые делали бы ←…→
кадиллак» с мощью пневматического ружья	или	аэродинамической трубы. ←…→
с созданием принципиально новой крылатой	или	баллистической ракеты, требует участия десятков ←…→
Бока»,	или	Барятинский, страстно любил удить и ←…→
так он всё время идёт	или	бежит. ←…→
и целой кучей, с детьми	или	без, бухались в бухту и ←…→
быть красивым, привлекательным, величественным, интересным	или	безобразным, но он никогда не ←…→
в самом деле надобно быть	или	безрассудным, или просто механическим маляром ←…→

Рис. 1. Отсортированные по правому контексту искомого слова результаты поиска

2. Расчет n-грамм по пользовательскому подкорпусу. Существующая корпусная платформа НКРЯ выполняла расчет n-грамм при индексации корпуса. При этом расчет n-грамм по пользовательскому подкорпусу (например, только по текстам одного автора) был невозможен, поскольку на момент индексации подкорпус неизвестен. Разработанный прототип осуществляет подсчет n-грамм «на лету» (для больших запросов используется не вся выдача, а случайный миллион результатов из нее). Это делает возможным подсчет n-грамм с учетом ограничений на подкорпус. Данная возможность доступна на сайте НКРЯ, ср. результаты поиска по запросу 2-грамм для леммы «отец» с прилагательным в текстах Толстого на Рис. 2.

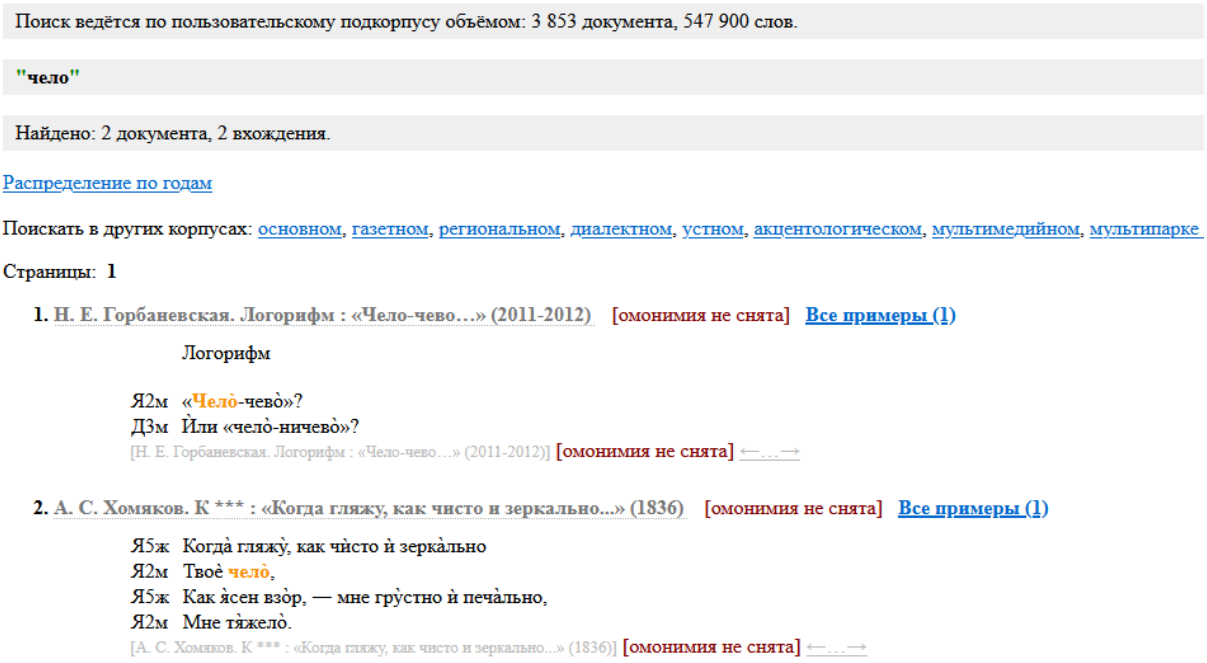
Поиск ведётся по пользовательскому подкорпусу объёмом: 176 документов, 3 010 952 слова.

Слово 1: **отец**
Слово 2: **А** на расстоянии от **-1** до **1** от **Слова 1**

№	Документы	Вхождения	ipm	Фрагмент
1	6	7	2.32	отца небесного
2	4	6	1.99	покойного отца
3	1	5	1.66	плотского отца
4	3	5	1.66	святых отцов
5	4	4	1.33	отец родной
6	2	3	1.00	старый отец
7	2	3	1.00	больше отца
8	2	2	0.66	святые отцы
9	1	2	0.66	истинно отец
10	1	2	0.66	голодных отцов
11	2	2	0.66	отец небесный
12	2	2	0.66	совершенен отец

Рис. 2. Результат поиска по 2-граммам в подкорпусе

3. Поддержка построчной метрической разметки в поэтическом корпусе позволяет задавать пользовательский подкорпус по условию на метрическую разметку отдельных строк. Это существенно повышает точность поиска. Данная возможность доступна на сайте НКРЯ, например, можно найти лемму «чело» только в строках с формулой Я2м, ср. результаты на Рис. 3. При этом лемма «чело» в строках с другой метрической разметкой не попадет в результаты поиска (результаты поиска подсвечены оранжевым).


Поиск ведётся по пользовательскому подкорпусу объёмом: 3 853 документа, 547 900 слов.

"чело"

Найдено: 2 документа, 2 вхождения.

Распределение по годам

Поискать в других корпусах: основном, газетном, региональном, диалектном, устном, акцентологическом, мультимедийном, мультипарке

Страницы: **1**

1. Н. Е. Горбаневская. Логорифм : «Чело-чево…» (2011-2012) [омонимия не снята] **Все примеры (1)**

Логорифм

Я2м «Чело̀-чево̀»?
Дзм Йли «чело̀-ничево̀»?
[Н. Е. Горбаневская. Логорифм : «Чело-чево…» (2011-2012)] [омонимия не снята] ←…→

2. А. С. Хомяков. К *** : «Когда гляжу, как чисто и зеркально...» (1836) [омонимия не снята] **Все примеры (1)**

Я5ж Когда̀ гляжу̀, как чи́сто ѝ зерка̀льно
Я2м Твоѐ чело̀,
Я5ж Как я́сен взо̀р, — мне гру́стно ѝ печа̀льно,
Я2м Мне тя́жело̀.
[А. С. Хомяков. К *** : «Когда гляжу, как чисто и зеркально...» (1836)] [омонимия не снята] ←…→

Рис. 3. Результат поиска в подкорпусе из строк с заданной метрической формулой

4. Поиск по условиям, накладываемым по-отдельности на слова в двух языках, в параллельном корпусе. Существующая корпусная платформа НКРЯ позволяла задавать условия поиска, которые в дальнейшем накладывались на слова обоих языков языковой пары. Разработанный прототип позволяет задавать условия для поиска в каждом из языковой пары языке по-отдельности. Так, например, можно задать поиск по лемме «cat» в английском языке и лемме «кот» в русском и найти все такие пары предложений, где cat переводится как кот (но не как кошка). Данная возможность доступна на сайте НКРЯ, ср. форму для поискового запроса на двух языках в параллельном корпусе на Рис. 4. и результаты поиска на Рис. 5.

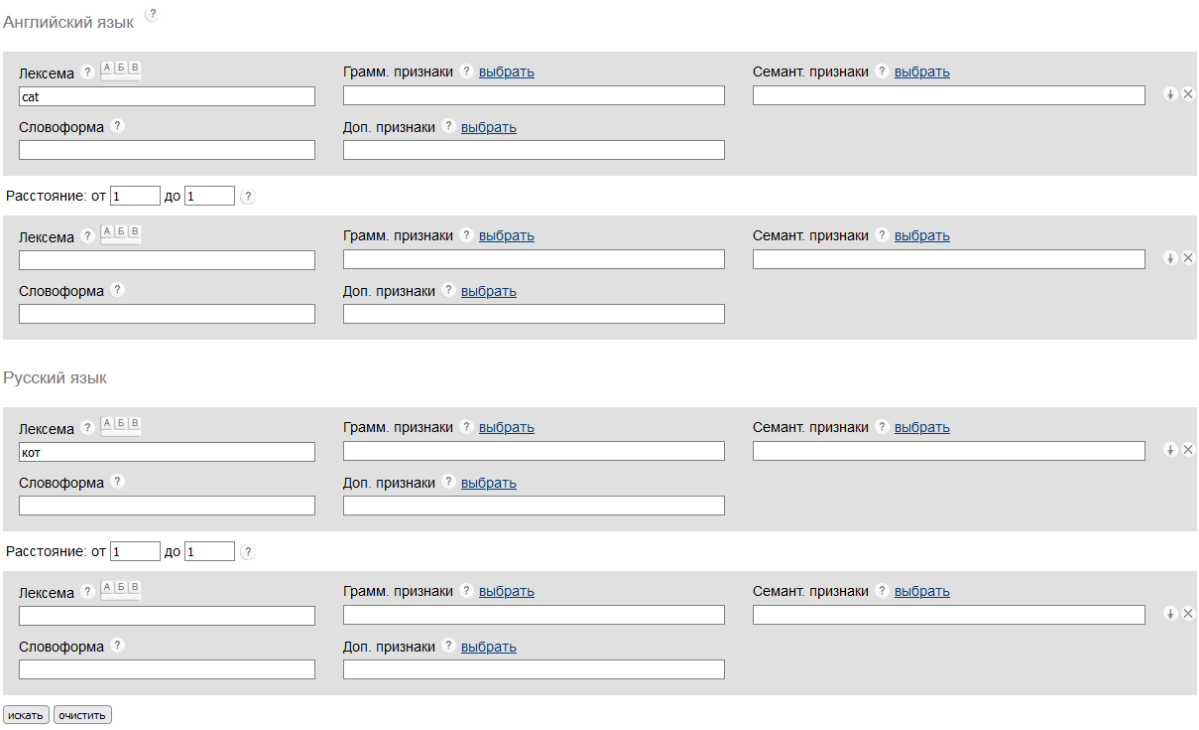


Рис. 4. Форма для поискового запроса на двух языках в параллельном корпусе

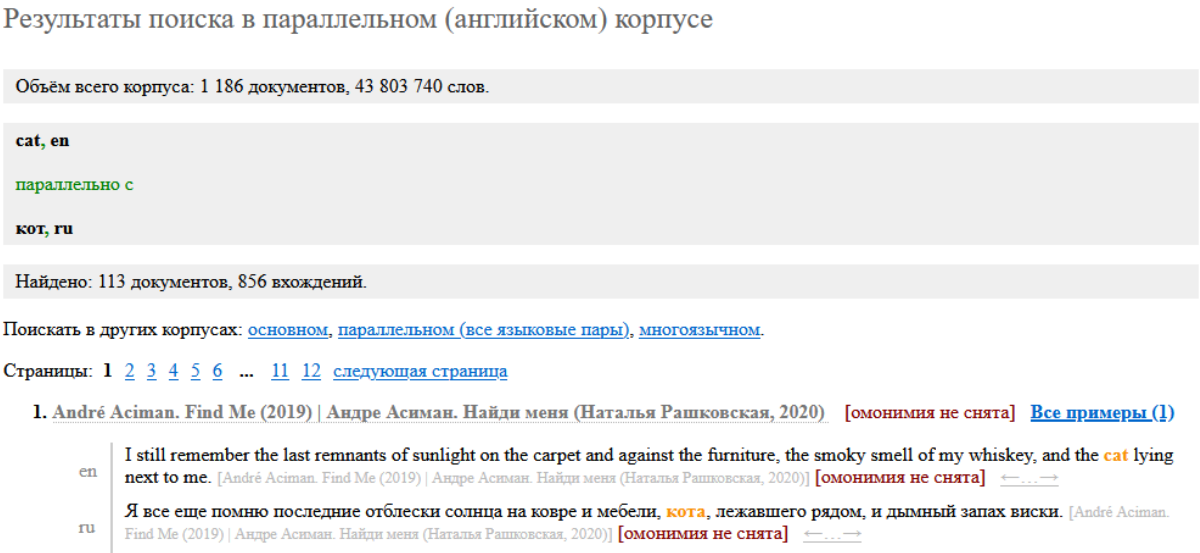


Рис. 5. Результаты поиска по условиям, накладываемым на слова в двух языках

5. Разработанный прототип программного интерфейса приложения (application programming interface, API) позволил отделить реализацию пользовательского интерфейса от реализации ядра корпусной платформы нового поколения. Это, в свою очередь, позволило заменить устаревший интерфейс сайта и отдельных модулей НКРЯ на современный и реализовать мобильную версию. В настоящее время на сайте НКРЯ в новом интерфейсе доступна главная страница, новости, текстовые страницы «о корпусе», модуль отображения статистики НКРЯ и модуль построения графиков, ср. Рис. 6. Прототип нового интерфейса поиска в основном корпусе испытан на стендах.

Рис. 6. Новый интерфейс построения графиков распределения по годам

7 Заключение

Для построения корпусной системы НКРЯ нового поколения был проведен анализ существующих корпусных платформ, а также сравнительный анализ корпусных платформ с поисковыми системами и РСУБД. Выявлено, что поисковые системы и РСУБД сами по себе не пригодны для использования в качестве корпусной платформы, а ни одна из существующих корпусных платформ не удовлетворяет предъявляемым требованиям в части поддержки неснятой омонимии.

Сформулированы требования к функциональности разрабатываемой платформы. С учетом этих требований разработан прототип, позволяющий экспериментально исследовать предварительные версии программных модулей будущей системы. Созданный прототип отличается поддержкой нескольких различных технологий баз данных и полнотекстового поиска и включает в себя специализированную полнотекстовую поисковую систему, оптимизированную для вычислительно эффективной реализации инструментов корпусной статистики.

Исследование разработанного прототипа показало, что интерактивность, надежность и масштабируемость прототипа не ниже чем у существующей корпусной платформы НКРЯ. В то же время точность прототипа выше, а функциональность существенно шире.

Несколько модулей прототипа включены в существующую программную платформу НКРЯ и реализуют доступный пользователям новый функционал.

В 2022 году работа по созданию корпусной платформы НКРЯ нового поколения будет продолжена.

References

[1] Abroskin A. Corpus Search: Problems and Solutions [Poisk po korpusu: problemy i metody ih resheniya]. // Russian National Corpus: 2006-2008. New results and perspectives [Nacional'nyj korpus russkogo yazyka: 2006-2008. Novye rezul'taty i perspektivy]. Ed. by V. A. Plungyan. — Sankt-Peterburg: Nestor-Istoriya, 2009. — P. 277–282.

[2] Brouwer P., Brugman H., Kemps-Snijders M. MTAS: A Solr/Lucene based Multi Tier Annotation Search solution. — Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure. Issue 136. — Linköping: Linköping University Electronic Press, 2016. — P. 19–37.

[3] de Does J., Niestadt J., Depuydt K. Creating Research Environments with BlackLab. // CLARIN in the Low countries. Odijk, J and van Hessen, A. (eds.). — London: Ubiquity Press, 2017. — P. 245–257.

[4] Evert S., Hardie A. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. — Proceedings of the Corpus Linguistics 2011 conference. — Birmingham: University of Birmingham, 2011.

[5] Kilgarriff, A., Baisa, V., Bušta, J. et al. The Sketch Engine: ten years on. — Lexicography ASI-ALEX. Issue 1. — New York: Springer, 2014. — P. 7–36.

[6] Kopotev M. Introduction to corpus linguistics [Vvedenie v korpusnuyu lingvistiku]. — Praga: Animedia Company, 2014. — P. 78–83.

[7] Lyashevskaya O., Plungyan V., Sichinava D. On the morphological standard of the Russian National Corpus [O morfologicheskom standarte Nacional'nogo korpusa russkogo yazyka]. // Russian National Corpus: 2003–2005. Results and perspectives [Nacional'nyj korpus russkogo yazyka: 2003-2005. Rezul'taty i perspektivy]. — Moscow: Indrik, 2005. — P. 111–135.

[8] Manning C., Schutze H., Raghavan P. Introduction to Information Retrieval. — Cambridge: Cambridge University Press, 2009.

[9] Morozov D., Gladilin S. An abstract model of search index query in the Russian National Corpus. — Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference “Dialogue 2020”. Issue 19. — Moscow: RSHU, 2020. — P. 1109–1116.

[10] Plungyan V., Reznikova T., Sichinava D. National corpus of the Russian language: general characteristics [Nacional'nyj korpus russkogo yazyka: obshchaya harakteristika. —Scientifical and technical information [Nauchno-tekhnicheskaya informaciya]. Issue 2(3). — Moscow: VINITI, 2005. — P. 9–13.

[11] Štěpánek, J., Pajas P. Querying Diverse Treebanks in a Uniform Way. — Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). — Malta: Valetta, 2010. — P. 1828–1825.

[12] Timoshenko S., Iomdin L., Gladilin S., Inshakova E. SynTagRus as part of the NKRYA: new opportunities [SinTagRus v sostave NKRYA: novye vozmozhnosti]. — Proceedings of the International Conference «Corpus Linguistics». Saint-Petersburg: Skifia-print, 2021. — P. 31–43.