Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2025»

23–25 апреля 2025 г.

Russian National Corpus 2.0: corpus platform, analysis tools, neural network models of data markup

Bonch-Osmolovskaya A. A.
Vinogradov Russian Language Institute of the Russian Academy of Sciences
abonch@gmail.com

Gladilin S. A.
IITP (Kharkevich Institute), FRC CSC
gladilin@iitp.ru

Kozerenko A. D.
Vinogradov Russian Language Institute of the Russian Academy of Sciences
akozerenko@mail.ru

Lyashevskaya O. N.
HSE University
olesar@yandex.ru

Morozov D. A.
NSU
morozowdm@gmail.com

Kuznetzova Y. N.
MSU, Institute of Linguistics of the Russian Academy of Sciences
kuznetsova.yn@gmail.com

Makhova A. A.
Vinogradov Russian Language Institute of the Russian Academy of Sciences
discourse@yandex.ru

Piskounova S. V.
saruwatari.lara@gmail.com

Bujlova N. N.
Lopukhin Federal Research and Clinical Center of Physical-chemical Medicine of Federal Medical Biological Agency
bnn@rcpcm.ru

Borodina D. G.
St. Petersburg State University
daria-borodina2001@yandex.ru

Vinogradova I. I.
Prosveshchenie Publishers
irinaivinogradova@yandex.ru

Sizov V. G.
IITP (Kharkevich Institute)
victor.sizov@gmail.com

Dyachenko P. V.
IITP (Kharkevich Institute)
pavelvd@iitp.ru

Kazennikov A. O.
IITP (Kharkevich Institute)
kazennikov@gmail.com

Vlasova N. A.
A.K. Ailamazyan Institute of Program Systems of the Russian Academy of Sciences
nathalie.vlassova@gmail.com

Glazkova A. V.
University of Tyumen
a.v.glazkova@utmn.ru

Stolyarov S. S.
NSU
s.stolyarov@g.nsu.ru

Garipov T. A.
NSU
garipov154@yandex.ru

Smal I. A.
NSU
vanasmal@mail.ru

Gubar'kova Ya. N.
Yandex
karmastina-ya@yandex-team.ru

Abstract

The Russian National Corpus has existed for over 20 years and is a unique linguistic tool. However, the technical limitations of the software platform on which it was implemented significantly narrowed its development prospects. In 2020, work was launched on a comprehensive update of the RNC software platform, as a result of which the National Corpus switched to a new generation 2.0 platform. The implemented deep changes concerned both the development of functionality that meets modern approaches to corpus linguistics, and a fundamental restructuring of the platform architecture as a whole, from data preparation and indexing systems to the user interface. A separate area of development of the capabilities of the RNC was associated with the implementation of neural network models used for metadata tagging, disambiguation, word-formation markup, etc.

This article provides a short description of the new corpus platform as of 2024. The description includes key parameters of changes in the architecture of the RNC platform and its user interface, descriptions of new corpus data analysis services and the specifics of their implementation, as well as a description of the experience of using neural network models for tasks related to corpus data markup.

The purpose of the article is to describe the technological layer of changes implemented in the National Corpus of the Russian Language as part of a large-scale update carried out in recent years.

Keywords: corpus linguistics; Russian National Corpus language; architecture of software platforms; markup of language data

DOI: 10.28995/2075-7182-2025-23-XX-XX

Национальный корпус русского языка 2.0: корпусная платформа, инструменты анализа, нейросетевые модели разметки данных

Бонч-Осмоловская А. А.
ИРЯ им. В.В. Виноградова РАН
abonch@gmail.com

Гладилин С. А.
ИППИ им. А.А. Харкевича РАН,
ФИЦ ИУ РАН
gladilin@iitp.ru

Козеренко А. Д.
ИРЯ им. В.В. Виноградова РАН
akozerenko@mail.ru

Ляшевская О. Н.
НИУ ВШЭ
olesar@yandex.ru

Морозов Д. А.
НГУ
morozowdm@gmail.com

Кузнецова Ю. Н.
МГУ, ИЯз РАН
kuznetsova.yn@gmail.com

Махова А. А.
ИРЯ им. В.В. Виноградова РАН
discourse@yandex.ru

Пискунова С. В.
saruwatari.lara@gmail.com

Буйлова Н. Н.
Федеральный научно-клинический центр физико-химической медицины
bnn@rcpcm.ru

Бородина Д. Г.
СПбГУ
daria-borodina2001@yandex.ru

Виноградова И. И.
Издательство Просвещение
irinaivinogradova@yandex.ru

Сизов В. Г.
ИППИ им. А.А. Харкевича РАН
victor.sizov@gmail.com

Дьяченко П. В.
ИППИ им. А.А. Харкевича РАН
pavelvd@iitp.ru

Казенников А. О.
ИППИ им. А.А. Харкевича РАН
kazennikov@gmail.com

Власова Н. А.
Институт программных систем им. А.К. Айламазяна РАН
nathalie.vlassova@gmail.com

Глазкова А. В.
Тюменский государственный университет
a.v.glazkova@utmn.ru

Столяров С. С.
НГУ
s.stolyarov@g.nsu.ru

Гарипов Т. А.
НГУ
garipov154@yandex.ru

Смаль И. А.
НГУ
vanasmal@mail.ru

Губарькова Я. Н.
Яндекс
karmastina-ya@yandex-team.ru

Аннотация

Национальный корпус русского языка существует уже более 20 лет и представляет собой уникальный лингвистический инструмент. Однако технические ограничения программной платформы, на которой он был реализован, существенно сужали перспективы его развития. В 2020 году были запущены работы по комплексному обновлению программной платформы НКРЯ, в результате которого Национальный корпус перешел на платформу нового поколения 2.0. Реализованные глубинные изменения касались как развития функционала, отвечающего современным подходам корпусной лингвистики, так и фундаментальной перестройки архитектуры платформы в целом, начиная от систем подготовки и индексации данных и заканчивая пользовательским интерфейсом. Отдельное направление развития возможностей НКРЯ было связано с внедрением нейросетевых моделей, использующихся для разметки метаданных, снятия омонимии, словообразовательной разметки и др.

В настоящей статье представлено краткое описание новой корпусной платформы по состоянию на 2024 г. Описание включает в себя ключевые параметры изменений архитектуры платформы НКРЯ и его пользовательского интерфейса, описания новых сервисов анализа корпусных данных и специфики их реализации, а также описание опыта использования нейросетевых моделей для задач, связанных с разметкой корпусных данных.

Цель статьи заключается в описании технологического пласта изменений, реализованных в Национальном корпусе русского языка в рамках масштабного обновления, проведенного в последние годы.

Ключевые слова: корпусная лингвистика; Национальный корпус русского языка; архитектура программных платформ; разметка языковых данных

1 Введение

Национальный корпус русского языка был открыт для публичного доступа 29 апреля 2004 года. В этот момент объем единственного корпуса насчитывал 30 миллионов словоупотреблений. За более чем двадцать лет своего развития Национальный корпус не только заметно увеличился по объему и разнообразию данных, достигнув объема в 2,2 млрд словоупотреблений, но и претерпел концептуальную эволюцию. Изначальная задумка «Русского Стандарта» (Сичинава 2005) состояла в подготовке представительного собрания русских текстов, снабженных морфологической разметкой и предназначенных для удобного поиска при лингвистическом исследовании. НКРЯ в его современном состоянии охватывает тысячелетнюю историю развития русского языка и «представляет как язык предшествующих эпох, так и современный, в разных социолингвистических вариантах — литературном, разговорном, просторечном, диалектном»[1]. Традиционные поисковые инструменты расширяются сервисами статистического анализа и визуализации корпусных данных, принципиально изменились внутренние возможности управления корпусом, начиная от подготовки корпусных данных и кончая гибкими настройками интерфейса. Выход Национального корпуса русского языка на текущий уровень развития потребовал технологической трансформации корпуса как электронного ресурса, происходившей в одном русле с основными процессами современной корпусной лингвистики. В настоящей статье представлено краткое описание новой корпусной платформы по состоянию на 2024 г. Лингвистический аспект обновлений был подробно рассмотрен в статье (Савчук и др., 2024).

1.1 НКРЯ 2.0 в свете основных тенденций развития современных корпусов

Ключевыми параметрами развития современных корпусных технологий, задающими тот научный контекст, в котором формируется стратегия технологических изменений Национального корпуса русского языка, на наш взгляд, являются (1) критическое увеличение объемов корпусов (от

[1] https://ruscorpora.ru/page/corpora-about/

сотен тысяч словоупотреблений к миллиардам), (2) переход на стандартизованные системы лингвистической разметки, (3) внедрение инструментов статистического анализа корпусных данных и, как результат технологического развития, (4) расширение областей практического применения корпусов. Подробный обзор работ в этих направлениях можно увидеть в расширенной версии настоящей статьи.

С учетом этих параметров новая платформа по сравнению со старой приобрела следующие принципиальные свойства.

Объемы корпусов

Старая платформа: К 2020 году общий объем всех корпусов НКРЯ составляет около 1 млрд словоупотреблений. При этом грамматическая омонимия снята менее чем в 1% от всех словоупотреблений. Дальнейшее расширение корпуса затруднено в связи с архитектурными ограничениями платформы.

Новая платформа: К 2024 году общий объем всех корпусов НКРЯ составляет более 2,2 млрд словоупотреблений, грамматическая омонимия снята примерно в 65% от всех словоупотреблений. Новая платформа спроектирована для поддержки данных НКРЯ в объеме до 100 млрд словоупотреблений.

Разметка данных

Старая платформа: Корпусные данные снабжаются морфологической разметкой, осуществленной с помощью программы MyStem (Зобнин, Носырев 2015). Алгоритм позволяет строить гипотетические разборы для слов, которых нет в грамматическом словаре Зализняка, но не позволяет с достаточной точностью снимать омонимию.

Новая платформа: Сохраняется морфологическая разметка алгоритмом MyStem, к ней добавлена разметка данных с помощью нейросетевой модели Rubic. Модель размечает не только морфологические, но и синтаксические характеристики словоформ, а также снимает омонимию не только по леммам, но и по словоизменительным признакам.

Развитие инструментов корпусного анализа

Старая платформа: Основным инструментом корпусного анализа является выдача по поисковому запросу в формате конкорданса или KWIC (key word in context). Пользователь имеет возможности сортировки результатов по дате создания текста и другим релевантным параметрам, а также по правому/левому контексту в формате KWIC. Обобщенная информация об изменениях частотностей слова представлена в виде диахронического графика по конкретной словоформе. Пользователь имеет доступ к n-граммам, предпосчитанным по словоформам.

Новая платформа: Инструменты корпусного анализа существенно расширены как на уровне запроса, так и на уровне представления выдачи. На уровне поисковых запросов появился поиск по коллокациям, на уровне выдачи — анализ частотности запроса, допускающий разные способы сортировки и представления данных. Появилась возможность получить предпосчитанную информацию о слове в корпусе в целом — «Портрет слова», куда входят его скетчи, контекстуально близкие слова, выявленные на основе расчета семантических векторов, однокоренные слова в корпусе.

Целевая аудитория НКРЯ

Старая платформа: Платформа ориентирована на подготовленного пользователя-лингвиста, который использует корпус как источник материала для лингвистических исследований.

Новая платформа: Новая платформа ставит своей задачей расширить аудиторию пользователей, в том числе привлекая менее подготовленных пользователей, не работавших ранее с языковыми корпусами. Корпус существует в мобильной версии, имеет богатейшую документацию, логика интерфейса минимизирует усилия пользователя по получению информации.

Ниже будут более подробно рассмотрены три аспекта технической реализации новой корпусной платформы. Во-первых, это концептуально новые подходы к архитектуре корпуса, корпусному ядру и веб-интерфейсу. Во-вторых, разработанные сервисы для корпусного анализа данных. В-третьих, нейросетевые модели, использованные для разметки данных.

2 Корпусная платформа нового поколения: примененные подходы и решения

2.1 Общая архитектура системы

К корпусной платформе нового поколения предъявлялись требования не только соответствия современным стандартам сервисов, предоставляемых крупными лингвистическими корпусами, но и обеспечения гибкости для последующей модификации и развития в соответствии с перспективными подходами, которые могут возникнуть в будущем. На момент разработки имелась существенная неопределенность, в целом характерная для ИТ-проектов: у нас не было представления о полном функционале, который в будущем потребует поддержки в НКРЯ. С развитием корпусных технологий возникают новые потребности, и корпусная платформа должна быть готова к их реализации. Поэтому требовалось организовать программную систему так, чтобы в будущем по возможности облегчить добавление нового функционала. Для этого мы стремились обобщить различные требования (уже имеющиеся или же потенциально возможные) к функционалу в однородные с точки зрения технической реализации группы. Таким образом, перед корпусной платформой ставилась задача поддержки не конкретных видов функционала, а целых функциональных групп; конкретные виды функционала рассматривались как представители этих групп. Такой подход потребовал унификации корпусных данных: атрибуты одинакового типа обрабатываются одними и теми же алгоритмами, а значит должны быть единообразно представлены.

Структурно программная система разделена на три независимые части. Вычислительное ядро реализует универсальный функционал для целой функциональной группы, лингвистическое ядро обеспечивает поддержку конкретных функций, используя реализованный функционал вычислительного ядра, а интерфейсный модуль осуществляет взаимодействие лингвистического ядра с пользователями. Таким образом, например, добавление новой функциональности, касающейся приписанных к токенам атрибутов, выполняется в вычислительном ядре и влияет одновременно на все атрибуты, и таким образом может не затронуть или незначительно затронуть лингвистическое ядро, а поддержка новых атрибутов легко реализовывается в лингвистическом ядре и не требует изменений вычислительных алгоритмов.

Такое разделение позволило использовать в корпусной платформе разные вычислительные ядра в зависимости от размера и структурной сложности корпуса. Так, например, для больших корпусов применяется вычислительное ядро, построенное на базе поисковой системы ElasticSearch, в то время как для сложно структурированной разметки небольшого Синтаксического корпуса лучше подошло вычислительное ядро на базе реляционной базы данных MySQL.

Выделение отдельного интерфейсного модуля важно, поскольку подходы к построению графических интерфейсов пользователя быстро меняются, а отдельный модуль легче заменить.

Было выделено 12 различных видов разметки, поддерживаемой платформой, и охватывающих структурные единицы разного уровня: от морфемной и акцентологической разметки внутри токена до метаразметки целых текстов. Все поддерживаемые НКРЯ виды разметки были преобразованы к этим видам. В дальнейшем при возникновении новых атрибутов, попадающих в один из выделенных видов, не потребуется изменения вычислительного ядра.

Например, в корпусной платформе нового поколения тексты, состоящие из слов, и аннотации жестикуляции в видеозаписи, состоящие из отдельных жестов, представляются при помощи одного и того же программного механизма. Таким образом, выровненный с текстом видеоряд внутренне представляется и обрабатывается тем же способом, что и два параллельных текста, выровненных между собой.

2.2 Автоматизированное взаимодействие корпусной платформы с другими лингвистическими системами

В настоящее время корпуса, доступные для пользователей через сеть Интернет, являются одним из важнейших инструментов корпусной лингвистики. Однако такой подход позволяет применять к корпусным данным только набор инструментов, реализованный в интерфейсе веб-сайта. Частично преодолеть это ограничение можно за счет предоставления пользователю возможности автоматизировать выполнение запросов к корпусу.

Автоматизация запросов к корпусной платформе осуществляется при помощи программного интерфейса приложений (англ. Application Programming Interface, API). Разработанный для нужд НКРЯ API обеспечивает возможность выполнения произвольных запросов, доступных через интерфейс. Однако в настоящее время API не доступен стороннему пользователю, а используется только самим графическим интерфейсом системы. Такой подход позволил нам отделить реализацию интерфейса пользователя от непосредственно поискового сервера. Хотя в настоящее время нельзя утверждать, что API решает задачу автоматизации, успешное применение в архитектуре системы подтверждает его универсальность: поскольку любая операция с корпусом преобразуется интерфейсом в запрос к API, можно сделать вывод, что API обеспечивает все необходимые возможности. После определенной доработки планируется сделать API общедоступным.

API поддерживает запросы, соответствующие основным видам использования корпусной платформы:

- запрос основных статистических данных о корпусах НКРЯ;
- запрос данных о конфигурации конкретного корпуса;
- запрос набора доступных в конкретном корпусе поисковых форм;
- запрос типов атрибутов, имеющихся в корпусе, и их допустимых значений;
- поисковый/аналитический запрос;
- запрос «Портрета» конкретного слова или конкретного корпуса.

2.3 Новая концепция интерфейса корпусной платформы, ориентированная на широкий круг пользователей

Интерфейс новой корпусной платформы обеспечивает возможность её использования с самых различных устройств: от смартфонов до настольных компьютеров с большим размером экрана. Для этого применяется подход «первичности мобильной версии» (англ. mobile first), в соответствии с которым самые первыми в списке вариантов стилевых таблиц находятся таблицы для самых миниатюрных мобильных устройств (обладающих не только самыми маленькими размерами экрана, но и наименьшими вычислительными ресурсами), что позволяет им прекратить ресурсоемкий для них дальнейший перебор вариантов. Пользователю автоматически открывается версия, наиболее подходящая для размера устройства, с которого он выходит в интернет (Рис. 1).

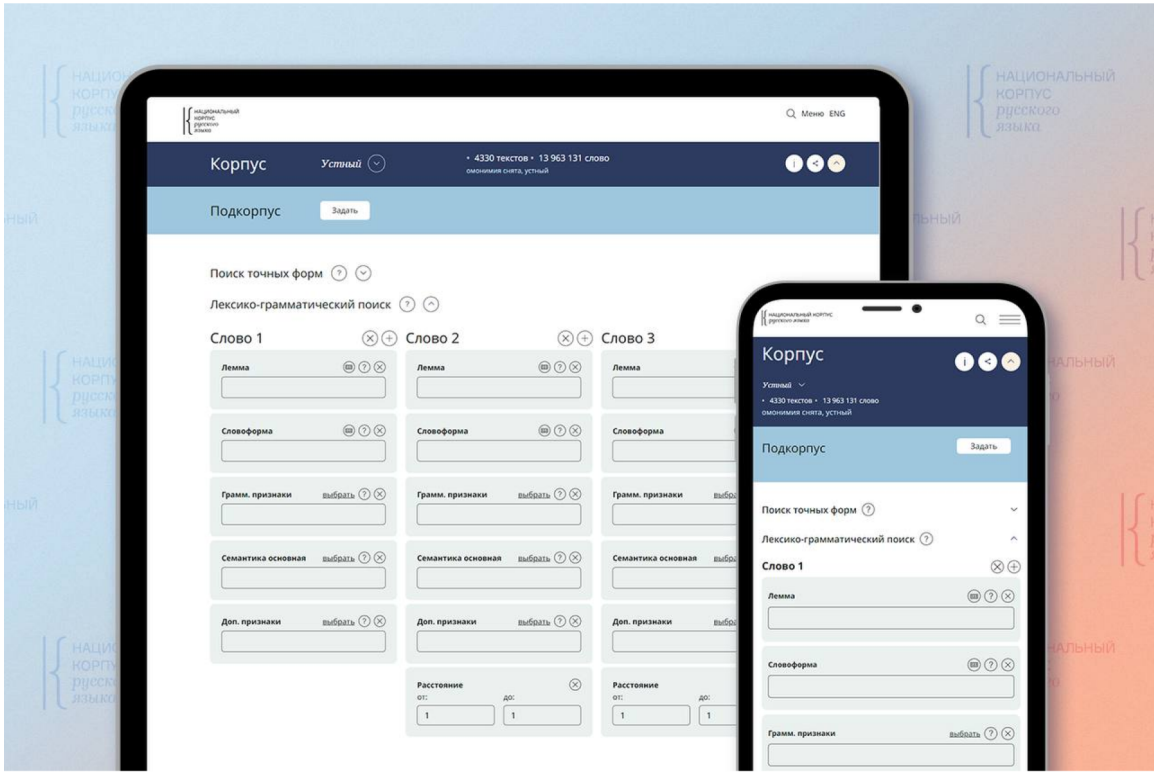

Рис. 1: Слева — версия сайта для ПК и планшета, справа — оптимизированная для мобильных устройств.

Реализован ряд комплексных лингвистических инструментов, позволяющих на одном экране получать разностороннюю лингвистическую информацию. Так, сервис «Портрет слова» в визуально компактной и понятной форме предоставляет разнообразную информацию для всех имеющихся разборов заданной леммы.

Там, где это возможно, результаты представляются с помощью круговых, столбчатых диаграмм, географических карт и графиков. В инструменте «Частотность» показываются доверительные интервалы для рассчитанной частотности (Рис. 2). При отображении графиков указываются временные границы, за пределами которых данных слишком мало для достоверных выводов. Под графиком отображается тепловая шкала, описывающая количество текстов, в которых найдены результаты, позволяющая оценить, насколько рост частотности слова является случайным выбросом в конкретном тексте или же объективно наблюдаемым явлением (Рис. 3).

Слово 1 Словоформа	← [1..2] →	Слово 2 Лемма	Вхождения	Доля	Доля	ipm	Конкорданс
какая	1	разница	2	50%	[15%, 85%]	28.21	Примеры
какая	1	жалость	1	25%	[4.56%, 69.94%]	14.11	Примеры

Рис. 2: Вид выдачи «Частотность»

Рис. 3: Вид выдачи «График»

«Похожие слова» отображаются в виде облака тегов, в котором размер букв и удаленность слов друг от друга характеризуют степень близости контекстов употребления слов (Рис. 4). Для морфемного разбора использована наглядная нотация, принятая в школьном преподавании русского языка (Рис. 5).

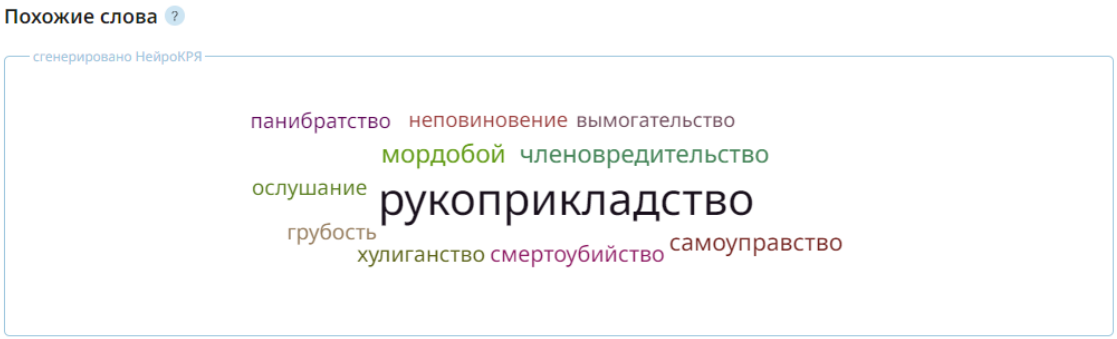

Рис. 4: Виджет «Похожие слова»

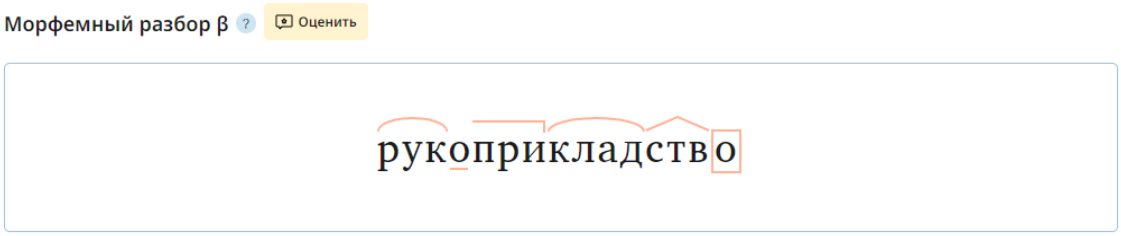

Рис. 5: Виджет «Морфемный разбор»

В нескольких сценариях поиска по Корпусу элементы интерфейса были намеренно перегруппированы по сравнению со старой версией корпусной платформы. Так, в интерфейсе задания условий лексико-грамматического поиска группы условий на искомые слова в словосочетании теперь расположены в одну строку слева направо (Рис. 6 и 7). Такой подход визуально более интуитивен для пользователей и соответствует расположению слов в тексте.

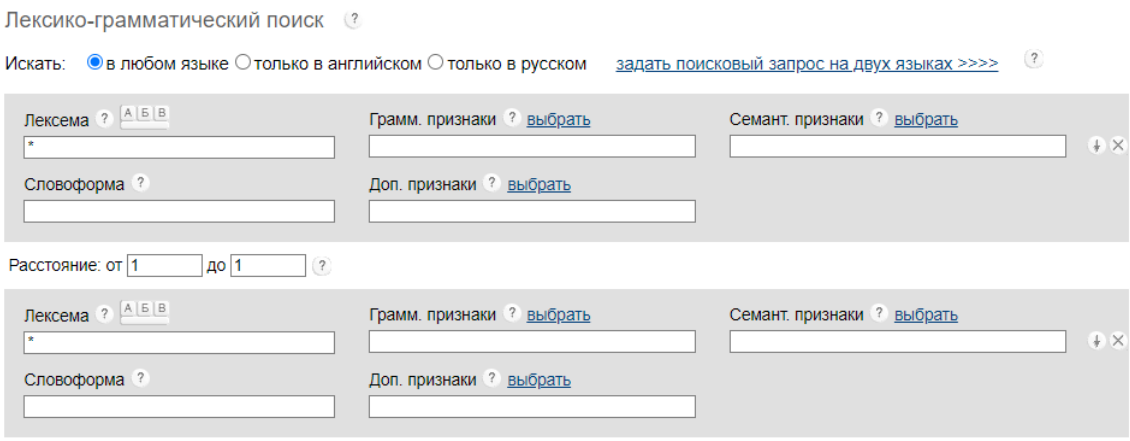

Рис. 6: Старое расположение условий лексико-грамматического поиска

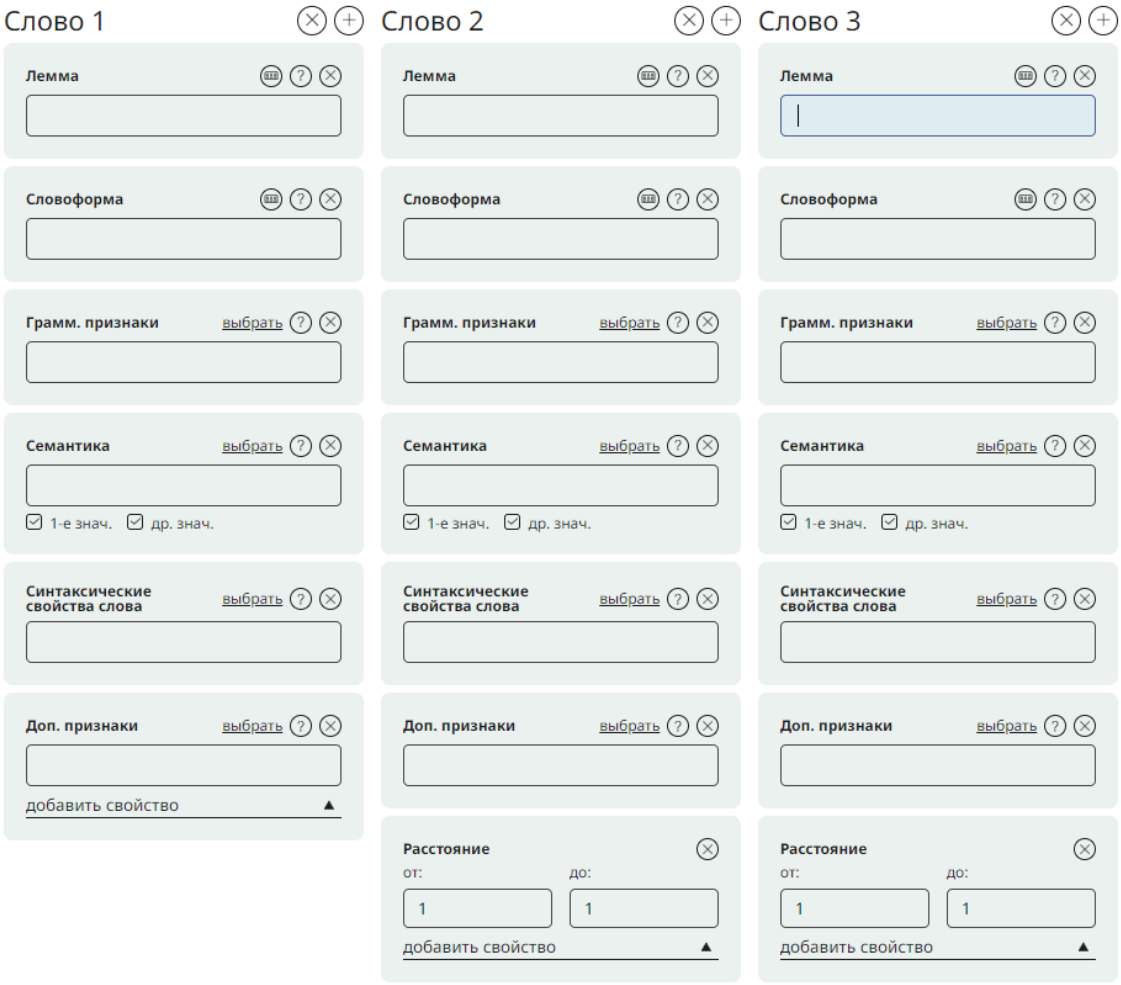

Рис. 7: Новое расположение условий лексико-грамматического поиска

Новый интерфейс ориентирован на уменьшение количества лишних действий пользователя. Так, выбор настроек, предпочтительный вид поиска в каждом корпусе, вид выдачи, открывающийся по умолчанию, а также режим отображения информации о запросе в шапке корпуса запоминаются в браузере пользователя и применяются для следующих поисковых запросов. В шапке

корпуса отображаются параметры не только искомого слова, но и заданного пользователем подкорпуса. Возможность в любой момент вернуться к форме и откорректировать любые параметры запроса сокращает усилия в сравнении с заданием параметров заново.

В новом интерфейсе поддерживается регулярно обновляемое руководство пользователя и поиск по нему. Система управления контентом позволяет в онлайн режиме структурировать, тегировать и редактировать все сопроводительные тексты.

Для иностранных пользователей весь интерфейс сайта переведен на английский язык.

3 Инструменты анализа корпусных данных

3.1 Основные направления развития статистико-аналитической компоненты НКРЯ

Современные методы корпусной лингвистики в наибольшей степени ориентированы на использование количественного анализа распределения языковых единиц. Именно поэтому современные корпусные платформы не ограничиваются лишь конкордансами для отображения словоупотреблений, но включают дополнительные сервисы, которые позволяют систематизировать, обобщать и статистически оценивать результаты анализа корпусных данных.

Программная платформа НКРЯ нового поколения реализует широкий спектр аналитических инструментов обработки корпусных данных. Эффективность их реализации обеспечена за счет предварительных вычислений на этапе индексации текстов, использования (при необходимости) приближенных вычислений по рандомизированной подвыборке и эффективных по времени доступа, но затратных по памяти механизмов кэширования. В вычислительном ядре алгоритмы реализации указанных статистических инструментов подразделяются на:

- алгоритмы, выполняемые разово в процессе индексации: результат работы такого алгоритма сохраняется в базе данных и готов к использованию при обработке запроса пользователя;
- алгоритмы, выполняемые в процессе пользовательского запроса на основе текстов корпуса и/или показателей, вычисленных и сохраненных в процессе индексации;
- алгоритмы, выполняемые в процессе пользовательского запроса на основании случайного подмножества результатов поиска. Результаты работы такого алгоритма являются приблизительными, поэтому они применяются в случае, когда точное вычисление в процессе запроса невозможно из-за ограничений на время ожидания.

3.1.1 Инструменты статистической характеризации корпусов и подкорпусов

Статистические инструменты этого типа позволяют строить портрет корпуса и подкорпуса, получать статистическое распределение текстов по значениям мета-атрибутов и строить диахронические графики. Поскольку корпусная платформа позволяет рассматривать произвольный набор текстов, заданный пользовательскими условиями, как подкорпус, количество гипотетически возможных подкорпусов очень велико. Таким образом, статистические характеристики подкорпусов не могут быть предварительно рассчитаны на этапе индексации.

3.1.2 Инструменты статистической характеризации результатов поиска словосочетаний

Статистические характеристики этого типа позволяют строить распределение, удовлетворяющее поисковым условиям словоформ и лемм и получать наиболее распространенные в результатах поиска n-граммы. Это наиболее ресурсоемкие вычисления, выполняемые непосредственно в момент запроса, поскольку количество результатов поискового запроса может превышать размер всего корпуса целиком (в случае, если одно и то же слово входит в несколько разных словосочетаний, удовлетворяющих условию поиска). При отображении результатов поиска вычисления могут быть прекращены, как только нужное количество примеров сформировано, но при подсчете статистики должны быть учтены все результаты или их репрезентативная подвыборка. В случае если количество результатов поиска превышает миллион, в качестве такой выборки рассматривается случайное подмножество в миллион результатов и производится расчет относительных показателей только на основе них.

3.1.3 Инструменты статистической характеризации лемм

Инструменты этой категории позволяют, например, находить похожие слова, то есть слова, встречающиеся в одинаковых контекстах. В виджете «Похожие слова» отображаются ближайшие семантические ассоциаты слова; коэффициент близости слов подсчитывается с помощью моделей дистрибутивной семантики, построенных на актуальных материалах основного корпуса НКРЯ. Вычисления характеристик этого типа требует предрасчетов на этапе индексации текстов с сохранением информации, привязанной к каждой лемме. В момент пользовательского запроса происходит статистический расчет на основе сохраненной информации, а не самих текстов корпуса, что критично снижает вычислительную сложность.

3.1.4 Статистические коллокации

Статистические коллокации для произвольного запроса могут быть вычислены только в момент пользовательского запроса, но скетчи (заранее фиксированные для каждой части речи наборы коллокаций с учетом синтаксических связей) эффективнее вычислять на этапе индексации и сохранять для каждой леммы, поскольку количество лемм в корпусе ограничено и составляет не более нескольких сотен тысяч (с порогом встречаемости хотя бы 3 раза в 3 различных текстах). Для остальных лемм такая информация не сохраняется.

3.2 Аналитические инструменты для сервиса «Портрет слова»

В рамках рассмотрения аналитических инструментов новой корпусной платформы НКРЯ мы уделим особое внимание двум нейросетевым моделям, используемым в сервисе «Портрет слова»: модели словообразовательного разбора и семейству векторных word2vec-моделей, используемых в виджете «Похожие слова».

3.2.1 Модель словообразовательного разбора

В НКРЯ словообразовательная разметка лемм присутствует в двух корпусах: Основном и Обучающем. С точки зрения архитектуры системы морфемный разбор является атрибутом, приписанным структурной единице «разбор». Для каждого разбора указан список морфем и тип каждой из них.

В основе разметки в Основном корпусе лежит специально разработанный для корпуса словарь морфемного анализа **Morphodict-K**, где по состоянию на май 2023 года даны разборы для 75 тыс. лексем. Этот словарь составлялся на основании идеологии «Словаря морфем русского языка» (Кузнецова, Ефремова 1986). Принципы этой идеологии — значительная дробность выделения морфем и соотносимость с другими лексемами аналогичного строения.

Разметка морфем в Обучающем корпусе опирается на разработанный на основе «Морфемно-орфографического словаря» (Тихонов 2002) словарь морфемного анализа **Morphodict-T**. Этот словарь содержит около 100 тыс. лексем, морфемный состав которых определяется в соответствии с практикой морфемного анализа в средней школе. При этом используется более жесткий подход к определению того, какие смысловые связи являются прозрачными в современном языке, и, как правило, выделяется меньшее число морфем, чем в Morphodict-K.

Словари Morphodict-K и Morphodict-T сравнительно малы по отношению к многообразию всех лемм Основного и Обучающего корпусов: в совокупности в этих корпусах содержится более 300 тыс. уникальных лексем. Для слов, не входящих в словари, морфемные разборы строятся автоматически при помощи алгоритма на базе ансамбля сверточных нейронных сетей (Sorokin, Kravtsova 2018). В ходе тестирования доля полностью верных разборов составила 88.5% для словаря Morphodict-T и 90.8% для словаря Morphodict-K (Garipov, Morozov, Glazkova 2023). В настоящий момент модель, обученная на Morphodict-K, интегрирована в сервис «Портрет слова» в Основном корпусе НКРЯ: из 314 935 различных лемм, представленных в сервисе, для 255 821 леммы разбор сгенерирован моделью. Обе модели размещены в открытом доступе и доступны для выгрузки в разделе «Нейросетевые модели» на сайте НКРЯ.[2]

[2] https://ruscorpora.ru/license-content/neuromodels

3.3 Векторные модели в «Портрете слова» (сервис «Похожие слова»)

Так как НКРЯ содержит весьма разнообразные по домену (типу, тематике, жанру и т.д.) и времени создания корпусы, одни и те же слова могут употребляться в них в несовпадающих значениях и наборах контекстов. Для того чтобы обнаружить и визуализировать особенности использования слов в различных корпусах, могут быть использованы модели векторного представления слов. Статические векторные модели на базе Основного корпуса строились и раньше (Kutuzov, Kuzmenko 2017), однако в ходе нашей работы модели обучались с использованием лемм, сгенерированных моделью Rubic или, при их наличии, указанных вручную. Это позволило существенно уменьшить количество ошибочно сгенерированных или неправильно токенизированных лемм в словаре модели.

Мы построили модели семантических векторов для существительных, глаголов, прилагательных и наречий для Основного, Обучающего, Газетных корпусов, Древнерусского, Старорусского, корпуса «От 2 до 15» и корпуса «Русская классика», ср. разницу ассоциатов для слова *трубить* в разных корпусах (Рис. 8–9):

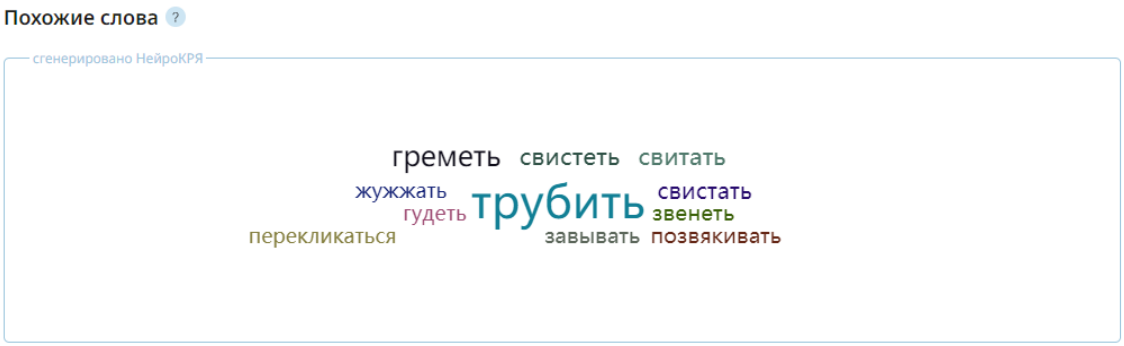

Рис. 8: Похожие слова для слова *трубить* в корпусе «Русская классика»

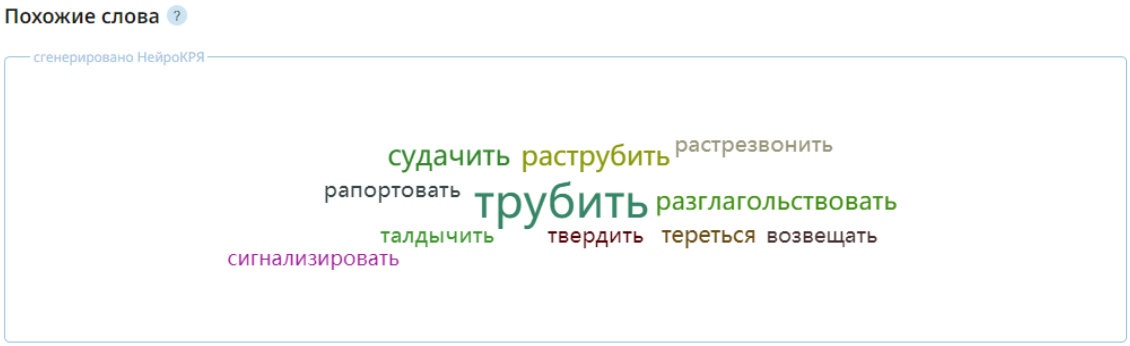

Рис. 9: Похожие слова для слова *трубить* в корпусе Центральных СМИ

Все используемые на сегодняшний день модели имеют одинаковую архитектуру (CBoW из библиотеки (Rehurek, Sojka 2011)) и схожие параметры обучения: окно размером 5, порог встречаемости 5-10 в зависимости от корпуса. Все семь моделей размещены в открытом доступе на странице «Нейросетевый модели» сайта НКРЯ[3] и могут быть использованы для научных и прикладных исследований.

[3] https://ruscorpora.ru/license-content/neuromodels

4 Нейросетевые модели в разметке данных и метаданных НКРЯ

Наиболее ресурсно затратным этапом при подготовке корпуса является этап корпусной разметки. Это касается как собственно внутритекстовой лингвистической разметки, так и разметки текстовых метаданных. В прошлом алгоритмы автоматической разметки морфологии показывали недостаточно высокое качество, поэтому центральным решением старой платформы НКРЯ был отказ от снятия омонимии — словоформе приписывались все возможные разборы, выбор релевантного разбора оставался на стороне пользователя. Такая неопределенность блокировала развитие статистических и аналитических сервисов, поскольку омонимичные разборы существенно зашумляют любые подсчеты. Применение интеллектуальных моделей позволяет значительно повысить качество автоматической разметки и исключить это ограничение. Ниже будут кратко представлены модели, которые сейчас используются при подготовке данных НКРЯ. Это, во-первых, нейросетевая модель морфосинтаксической разметки Rubic, а во-вторых, комплекс моделей для разметки метаданных: жанров в корпусе «Социальные сети» и ключевых слов в текстах Газетного корпуса. Использование этих моделей уже изменило экосистему НКРЯ, открыв новые возможности для значительного облегчения наиболее трудоемкого этапа подготовки корпусных данных.

4.1 Нейросетевая модель морфосинтаксической разметки для русского языка Rubic

Задачи, которые решает нейросетевая модель Rubic, — это автоматическая лемматизация, морфологическая характеристика для всех токенов, включая определение части речи, и построение дерева синтаксической зависимости предложения. Таким образом, Rubic представляет собой альтернативу морфологическому анализатору MyStem (Зобнин, Носырев, 2015), ранее применявшемуся для обработки текстов НКРЯ и основанному на грамматическом словаре. При разработке модели ставились задачи улучшения обработки «несловарных» разборов (например, *ажник, летось, сподтишка*), просторечных и грамматически аномальных форм и конструкций (например, *силов*, *хоцца, подумашь, ефту*), словоформ, записанных в нестандартной орфографии (в том числе петровской эпохи, в дореволюционной орфографии, а также в советской орфографии до реформы 1956 года). Модель также должна корректно обрабатывать архаичные формы из церковнославянского языка (например, *бысть, быша, многая лета*) согласно соглашениям, принятым для исторических корпусов НКРЯ.

4.1.1 Принципы работы Rubic

Архитектура Rubic (Lyashevskaya et al., 2023) основана на архитектуре модели qbic, победившей в соревновании по обработке русского языка GramEval2020 (Anastasyev 2020): однослойном LSTM-энкодере, комбинирующем векторизованные представления слов, получаемые из BERT-подобной модели (в текущей реализации, sberbank-ai/ruBert), и морфологические пометы, приписываемые анализатором PyMorphy2. Полученное представление анализируется тремя декодерами, выполняющими задачи классификации для выбора: а) части речи и грамматических признаков, б) леммы и в) дерева зависимостей. Для обучения модели использовались специально подготовленные обучающие данные на основе корпусов СинТагРус, UD-Taiga (Droganova, Zeman 2018; Droganova, Lyashevskaya, 2018) и НКРЯ (Lyashevskaya et al., 2020). Они охватывают тексты различных временных эпох и жанров общим объемом свыше 2,4 миллиона токенов. Все данные приводятся в расширенном формате морфологической и синтаксической разметки UD-ext (Lyashevskaya 2019). Целевая синтаксическая разметка представляется в формате CONLL-U, который затем сохраняется в синтаксисе XML, принятом в НКРЯ. Это позволяет, с одной стороны, сохранить подход к синтаксической разметке, принятый в Universal Dependencies, а с другой — использовать морфологическую разметку в стандарте НКРЯ.

4.1.1.1 Токенизатор

Практически любой анализ текста начинается с его разбиения на фрагменты (токенизации). Анализ существующих алгоритмов токенизации (udpipe (Straka, Hajic, Straková 2016), razdel[4], spacy[5],

[4] https://github.com/natasha/razdel
[5] https://github.com/explosion/spaCy

nltk (Bird, Loper, Klein 2009), PyMorphy2 (Korobov 2015), MyStem[6], rusenttokenize[7]) показал, что качество сегментации текста на предложения у таких подходов достаточно невелико (F1 на тестовой выборке сложных предложений GOLD не превысило 0.55). В связи с этим было принято решение обучить на имеющихся данных собственную модель для токенизации. В качестве архитектуры была выбрана модель Stanza (Qi P. et al. 2020). Мы обучили модель со стандартными параметрами обучения на открытых датасетах Тайга (Shavrina, Shapovalova 2017) и СинТагРус[8], а также на внутренних данных из корпусов прозы XX–XXI веков, поэзии, корпусов со старой орфографией XVIII века, а также текстах новостей XXI века. Полученная модель на выборке GOLD продемонстрировала высокое качество по отдельным словам (F1=0.94) и значительно лучшее качество по предложениям (F1=0.63). Для дополнительного сравнения новой модели с предыдущим решением была подготовлена расширенная тестовая выборка, схожая по составу с реальными данными (выборка TEST). На материале этой выборки обученная модель значительно превзошла лучший из рассмотренных алгоритмов, достигнув метрики F1=0.95 по предложениям и 0.99 по отдельным словам. Модель токенизатора доступна для скачивания на странице «Нейросетевые модели» сайта НКРЯ.[9]

4.1.1.2 Классификатор морфологических признаков

После токенизации текста происходит приписывание каждому выделенному слову морфологических признаков. Классификатор морфологических признаков работает на принципе полного морфологического тега, иными словами, входом и выходом модели служит набор, состоящий из частеречного и грамматических тегов вида «NOUN|Animacy=Anim|Case=Nom|Gender=Fem|Number=Sing». В отличие от классификаторов, в которых каждая грамматическая категория определяется независимо, данное решение позволяет избежать потери части разбора (например, одушевленности, в случае если вероятность ее определения в контексте низкая) и свести требуемый набор грамматических помет для частеречных классов и подклассов к грамматическому стандарту корпуса.

4.1.1.3 Лемматизатор

Размеченное морфологическими пометами слово попадает в лемматизатор. В основе лемматизации лежат правила преобразования словоформы в лемму вида «удалить последовательность символов в конце строки длины N, удалить последовательность символов в начале строки длины M > добавить последовательность D в конце строки > применить маску капитализации / декапитализации». Используются преобразования, встретившиеся в обучающих данных более 3 раз, чтобы исключить влияние несистемных опечаток и другого шума. В зависимости от объема обучающих данных, такой подход дает от 1000 до 2000 правил. Для улучшения качества лемматизации предусмотрена проверка наиболее вероятных гипотез лемм с данными словаря, составленного вручную.

4.1.1.4 Разметка синтаксического дерева

Наконец, на последнем этапе разметки происходит анализ синтаксического дерева каждого предложения. При построении гипотез синтаксических деревьев используется подход Т. Дозата и К. Маннинга (Dozat, Manning, 2016) на основе глубокого биаффинного внимания для определения пар связанных словоформ и метки синтаксического отношения.

4.1.2 Результаты

Качество работы модели Rubic'a оценивалось на коллекции тестовых данных, представляющих разные сферы употребления языка, см. Таблицу 1.

[6] https://yandex.ru/dev/mystem/
[7] https://github.com/deeppavlov/ru_sentence_tokenizer
[8] https://ruscorpora.ru/page/corpora-datasets/
[9] https://ruscorpora.ru/license-content/neuromodels

	fiction	news	poetry	social	wiki
Часть речи	0.9922	0.9893	0.9923	0.9777	0.9808
Леммы	0.9930	0.9923	0.9846	0.9848	0.9780
Морфологические признаки	0.9591	0.9517	0.9654	0.9528	0.9423
Неименованные синтаксические связи	0.9599	0.9563	0.9106	0.9296	0.9457
Именованные синтаксические связи	0.9530	0.9425	0.8942	0.9153	0.9231

Таблица 1: Результаты работы Rubic'а на тестовом множестве

Модель хорошо справляется с определением частей речи, морфологическим разбором некоторых грамматических категорий, таких как сравнительная степень, переходность, вид. При лемматизации модель демонстрирует высокое качество обработки для слов продуктивных парадигм. Результаты синтаксического парсинга показывают, что модель чаще всего правильно анализирует большинство часто употребляемых конструкций — вводные и сочиненные конструкции, предложные и атрибутивные группы и т.п. Также хорошо определяются дальние связи (например, субъект на расстоянии 5-10 слов от предиката). Более подробно особенности генерируемой разметки рассмотрены, например, в работе (Lyashevskaya et al., 2023).

4.2 Разметка жанров в корпусе «Социальные сети»

Корпус «Социальные сети» содержит тексты из открытых интернет-источников и включает в себя записи в блогах и сообщения в мессенджерах (подробнее о балансе источников корпуса «Социальные сети» см. в статье (Савчук и др., 2024)). Поскольку понятие «социальные сети» в этом случае трактуется максимально широко, а также в связи с большим объемом корпуса (почти 160 млн словоупотреблений), появилась необходимость в автоматической разметке жанров для систематизации текстов корпуса.

Для разметки жанров была выбрана предварительно обученная модель RuRoBERTa [10] (Zmitrovich et al., 2023). Модель была дообучена на подготовленном вручную наборе данных объёмом около 3 тысяч текстов, размеченном по 16 жанрам, с использованием следующих параметров: скорость обучения — 5e-6, количество эпох обучения — 3, максимальная длина входной последовательности — 256 токенов. При обучении жанрам назначались веса в зависимости от их доли в наборе данных. Качество модели было проверено с помощью кросс-валидации для десяти фолдов. Значение F-меры с макроусреднением составило 54,42% для 16 жанров, доля правильных ответов (accuracy) составила 71,16%.

Для итоговой разметки жанров в корпусе «Социальные сети» был использован ансамбль из трех моделей с усреднением предсказанных вероятностей жанров (soft voting).

4.3 Разметка ключевых слов в Корпусе региональных СМИ

Корпус Региональных СМИ в основном состоит из коротких информационных текстов, опубликованных в газетах различного уровня (Савчук 2015). Каждый текст, как правило, посвящен единственной теме. Для описания тематики и упрощения поиска текстов в корпусе выполнена автоматическая разметка ключевых слов. Одно ключевое слово может состоять из одного существительного в именительном падеже в единственном или множественном числе (*праздник, переломы*) либо из двусловного сочетания (биграммы) с главным словом-существительным (*таяние снега, обычные дни*).

[10] https://huggingface.co/ai-forever/ruRoberta-large

Извлечение ключевых слов из текстов Региональных СМИ выполнено с помощью библиотеки RuTermExtract[11] и набора правил постобработки. Он построен на анализе морфологических характеристик слов и словосочетаний и набора правил для извлечения ключевых слов. Для морфологического анализа в русскоязычной версии используется библиотека PyMorphy2 (Korobov 2015).

5 ЗАКЛЮЧЕНИЕ

В статье представлено краткое описание обновленной платформы НКРЯ с технологической точки зрения. Это обновление является важнейшим этапом 20-летнего развития Национального корпуса русского языка и является результатом внедрения комплексного подхода, соответствующего современным практикам и стандартам развития корпусных ресурсов, а также общим тенденциям цифрового развития общества. Более подробное описание произошедших изменений можно увидеть в расширенной версии настоящей статьи.

Литература

[1] Зобнин А. И., Носырев Г. В. (2015). Морфологический анализатор MyStem 3.0 // Труды Института русского языка им. В. В. Виноградова. 2015. № 6. С. 300–310.

[2] Кузнецова А. И., Ефремова Т. Ф. (1986). Словарь морфем русского языка. Москва: Рус. яз., 1986.

[3] Савчук С. О. (2015). Корпус региональных газет России и зарубежья // Труды Института русского языка им. В. В. Виноградова. 2015. № 6. С. 163—193.

[4] Савчук С. О. и др. (2024). Национальный корпус русского языка 2.0: новые возможности и перспективы развития //Вопросы языкознания. Т. 2. С. 7-34.

[5] Сичинава Д. В. (2005). Национальный корпус русского языка: очерк предыстории // Национальный корпус русского языка: 2003—2005. М.: Индрик. С. 21—30.

[6] Сичинава Д. В. (2022). Корпус берестяных грамот как параллельный // Труды Института русского языка им. В. В. Виноградова. 2022. № 2 (32), 92-106.

[7] Тихонов А. Н. (2002). Морфемно-орфографический словарь. Москва: Астрель: АСТ, 2002.

[8] Anastasyev D., (2020). Exploring pretrained models for joint morphosyntactic parsing of Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference “Dialogue”, volume 19. P. 1–12.

[9] Bird, St., Loper E., Klein E. (2009). Natural Language Processing with Python. O’Reilly Media Inc.

[10] Dozat T., Manning Ch. D. (2016). Deep Biaffine Attention for Neural Dependency Parsing https://arxiv.org/abs/1611.01734

[11] Droganova, K, Lyashevskaya O. (2018). Cross-tagset parsing evaluation for Russian // Digital Transformation and Global Society Third International Conference, DTGS 2018, St. Petersburg, Russia, May 30 – June 2, 2018, Revised Selected Papers, Part I / Ed. by Daniel A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov, O. Koltsova. Issue 858. P. 380-390.

[12] Droganova, K, Lyashevskaya O., Zeman D. (2018). Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks // Proceedings of TLT 2018 International Workshop on Treebanks and Linguistic Theories, 13-14 November 2018, Oslo, Norway. NEALT Proceedings Series. Linköping University Electronic Press, 2018. P. 52-65.

[13] Garipov T., Morozov D. Glazkova A. (2023). Generalization Ability of CNN-Based Morpheme Segmentation // 2023 Ivannikov Ispras Open Conference (ISPRAS), Moscow, Russian Federation. P. 58-62.

[14] Korobov M. (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. P. 320-332.

[15] Kutuzov, A., Kunilovskaya, M. (2018). Size vs. Structure in Training Corpora for Word Embedding Models: Araneum Russicum Maximum and Russian National Corpus. In: van der Aalst, W., et al. Analysis of Images, Social Networks and Texts. AIST 2017. Lecture Notes in Computer Science, vol 10716. Springer, Cham.

[16] Kutuzov A., Kuzmenko E. (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661.

[17] Lyashevskaya O. (2019). A reusable tagset for the morphologically rich language in change: A case of Middle Russian // Komp’juternaja Lingvistika i Intellektual’nye Tehnologii. P 422–434.

[18] Lyashevskaya O. et al., (2023). Disambiguation in context in the Russian National Corpus: 20 years later // Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference “Dialogue” (2023). Issue 22. P. 307-318.

[11] https://github.com/igor-shevchenko/rutermextract

[19] Rehurek, R., Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2)).
[20] Shavrina T., Shapovalova O. (2017) TO THE METHODOLOGY OF CORPUS CONSTRUCTION FOR MACHINE LEARNING: «TAIGA» SYNTAX TREE CORPUS AND PARSER. in proc. of "CORPORA2017", international conference, Saint-Petersbourg.
[21] Sorokin, A., Kravtsova, A. Deep Convolutional Networks for Supervised Morpheme Segmentation of Russian Language // Ustalov, D., Filchenkov, A., Pivovarova, L., Žižka, J. (eds) Artificial Intelligence and Natural Language. AINL 2018. Communications in Computer and Information Science, vol 930. Springer, Cham.
[22] Zmitrovich, D., et al. (2023). A Family of Pretrained Transformer Language Models for Russian. ArXiv, abs/2309.10931.