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 (full version)

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 detailed description of the new corpus platform as of 2024. The description includes an overview of the current technological development of corpora and corpus platforms, 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: 10.28995/2075-7182-2025-23-1001-1042

Национальный корпус русского языка 2.0: корпусная платформа, инструменты анализа, нейросетевые модели разметки данных (полная версия)

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 подробное описание новой корпусной платформы по состоянию на 2024 г. Описание включает в себя обзор современного технологического развития корпусов и корпусных платформ, 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[1] существует уже более 20 лет и представляет собой уникальный лингвистический инструмент. Однако технические ограничения программной платформы, на которой он был реализован, существенно ограничивали перспективы его развития, поэтому 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 с 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 подробное описание новой корпусной платформы по состоянию на 2024 г. Описание включает в себя четыре раздела. Первый раздел представляет из себя обзор современного технологического развития корпусов и корпусных платформ, во втором разделе определены ключевые параметры изменений архитектуры платформы НКРЯ и его пользовательского интерфейса, третий раздел содержит описания новых сервисов анализа корпусных данных и специфики их реализации. Наконец, четвертый раздел посвящен описанию опыта использования нейросетевых моделей для задач, связанных с разметкой корпусных данных.

[1] https://ruscorpora.ru/

Цель статьи заключается в описании технологического пласта изменений, реализованных в Национальном корпусе русского языка в рамках масштабного обновления, проведенного в последние годы. Лингвистический аспект обновлений был подробно рассмотрен в статье (Савчук и др., 2024).

1.1 Обзор современных направлений технологического развития лингвистических корпусов

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: количество словоупотреблений, представленных в 22 корпусах НКРЯ, достигло 2 миллиардов, 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»[2]. 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.

Ниже будут вкратце представлены ключевые параметры развития современных корпусных технологий; это позволит очертить тот научный контекст, в котором формируется стратегия технологических изменений Национального корпуса русского языка. Такими параметрами, на наш взгляд, являются критическое увеличение объемов корпусов (от сотен тысяч словоупотреблений к миллиардам) (1.1.1), переход на стандартизованные системы лингвистической разметки (1.1.2), внедрение инструментов статистического анализа корпусных данных (1.1.3) и, как результат технологического развития, расширение областей практического применения корпусов (1.1.4).

1.1.1 Увеличение объемов корпусов

Вопрос о том, какой размер должен иметь идеальный корпус, не имеет однозначного ответа (Reppen 2021). Качественные и широко используемые лингвистами современные корпуса имеют очень разный объем. Одни из самых объемных корпусов сегодня принадлежат к типу TenTen, это целое семейство корпусов на более чем 40 языках, доступных на платформе SketchEngine[3]. По их названию понятно, что целевой размер таких корпусов — 10^{10} слов. Все корпуса этого типа собираются автоматически в Интернете по единому алгоритму, который, в частности, предполагает очистку от повторяющихся текстов и удаление нерелевантных фрагментов, и поэтому могут считаться сопоставимыми. Но вообще для корпусов, собранных автоматически, размер в 10^{10} слов не является пределом, например, почти на порядок больше размеры корпуса NOW (News on the Web)[4].

Однако существуют и корпуса принципиально иного типа, а именно репрезентативные и сбалансированные, при создании которых ставится весьма амбициозная цель подобрать коллекции текстов таким образом, чтобы корпус отражал язык в целом. Определение репрезентативности каждого корпуса — это многоэтапная работа, которая всегда ведется на разработанных специально для этого корпуса теоретических основаниях (Biber 1993). Репрезентативность невозможно измерить никакими известными метриками, поэтому создание репрезентативного корпуса представляет собой отдельную исследовательскую задачу для создателей корпуса. (McEnery, Hardie 2012). Сбалансированность предполагает задание пропорций разных текстов, входящих в корпус, с учетом разнообразных параметров, например, жанров, тем, стилей, иногда также эпохи

[2] https://ruscorpora.ru/page/corpora-about/
[3] https://www.sketchengine.eu/documentation/tenten-corpora/
[4] https://www.english-corpora.org/now/

написания (см. пример корпуса китайского языка Sinica (Chen et al., 1996), корпуса современного турецкого языка (Aksan Y. et al., 2012) или корпуса немецкого языка XX века (Geyken 2007)). Расчет и подготовка корпусных коллекций для таких корпусов — это сложная и кропотливая работа, зачастую требующая дополнительных ресурсов, например, расшифровки устных текстов, распознавания старых печатных или рукописных изданий, поэтому объемы сбалансированных корпусов заведомо существенно меньше, чем собранных автоматически.

Одним из самых больших сбалансированных корпусов является COCA (Corpus of Contemporary American English)[5] с объемом чуть больше миллиарда слов. В этом корпусе баланс понимается следующим образом: создатели корпуса выделили 8 категорий (субтитры, устный, художественная литература, журналы, газеты, научные журналы, блоги, Интернет-страницы) и равномерно распределили тексты между ними.

Также следует упомянуть один из самых известных национальных корпусов — Чешский национальный корпус, который насчитывает примерно 5 миллиардов слов. В целом, корпус является репрезентативным, но хотя бы отчасти сбалансированными можно считать только те его части, которые относятся к пятилетним коллекциям типа SYN20XX[6], а их объем на порядок меньше.

Важно подчеркнуть, что рост объемов корпусов явился ответом на изменившиеся технические возможности: хранение большого объема текстов онлайн, увеличение мощностей памяти и процессоров. Таким образом, безусловным стандартом построения современных лингвистических корпусов, претендующих на общие задачи отображения языковых явлений в синхроническом или диахроническом срезе, стали не только требования к большому объему данных, но, что принципиально, отсутствие технических ограничений для их кратного увеличения. Еще одним фактором, существенно повлиявшим на процесс создания больших корпусов, является расширение применения и улучшение качества инструментов для автоматической разметки. Не в последнюю очередь это стало возможно благодаря процессам стандартизации автоматической разметки, происходившим в последние десятилетия.

1.1.2 Стандартизованная разметка корпусов

Ключевое отличие языковых корпусов от коллекций текстовых файлов (электронных библиотек) состоит в том, что тексты в корпусе снабжены лингвистической разметкой. Как отмечается в (McEnery, Wilson 2001), разметка существенно расширяет возможности использования корпусов и спектр исследовательских вопросов, которые можно решать с их помощью. Разметки языковых корпусов представляют собой лингвистические абстракции, соответствующие базовым уровням языка, начиная с грамматики. Самая «базовая» разметка — морфологическая — обеспечивает возможность поиска по морфологическим признакам слова (например, по частям речи) и соотносит разные формы слова с общей для них «начальной формой» — леммой. Синтаксическая разметка дополняет текст информацией о связи слов в предложении, маркируя главное слово и его зависимые, а также тип синтаксической связи между главным и зависимым. Семантическая разметка корпусов соотносит лексически полнозначные слова с соответствующими семантическими категориями или классами. Дискурсивная разметка выходит за пределы одного предложения, выделяя кореферентные единицы или отношения между предложениями или их частями (элементарными дискурсивными единицами). Могут быть размечены и другие лингвистические единицы — морфемы, интонационные контуры, диалогические роли и т.д. (Newmann Cox 2021).

В 1980–1990-х годах, когда, благодаря резкому увеличению электронных текстов, корпуса стали создаваться очень активно, выбор способа разметки соответствующего набора тегов был делом исключительно рабочей группы, разрабатывающей корпус. Очень характерно в этом смысле звучит один из семи принципов лингвистической разметки, сформулированных в (Leech 1993): ни одна схема разметки не может претендовать на то, чтобы быть корпусным стандартом. Существовала презумпция, что оптимальная схема разметки «вырастает» из конкретных задач, для которых собирается корпус (так называемый bottom-up подход). Однако через некоторое время стало понятно, что отсутствие единого подхода к лингвистической аннотации тормозит в целом «корпусную отрасль», мешает взаимной совместимости ресурсов и возможности их по-

[5] https://www.english-corpora.org/coca/
[6] https://wiki.korpus.cz/doku.php/en:cnk:syn:verze11

вторного использования, затрудняет применение готовых инструментов (например, готовый к использованию синтаксический парсер может опираться на иную схему разметки, чем та, которая используется в корпусе) и мешает использованию и интеграции данных лингвистических корпусов в системы, связанные с автоматическим анализом текста. Как отмечается в (Ide et al., 2017:115), «тысячи часов были потрачены на преобразование данных, представленных в одном формате, в другой, который будет работать для других целей или с другим программным обеспечением, или, что еще хуже, на воссоздание тех же ресурсов с целью решения конкретных задач». Важным фактором, повлиявшим на изменение отношения к стандартизации разметки от позиции «выбор разметки — дело рабочей группы» к использованию унифицированных единых стандартов стал переход компьютерной лингвистики от правиловых алгоритмов к методам машинного обучения и нейросетевому моделированию. Корпуса стали главным источником данных для решения всего спектра задач автоматического анализа и генерации языка, появляются открытые корпусные ресурсы, которые могут быть использованы для машинного обучения, корпуса создаются под конкретные задачи с помощью готовых инструментов и ценность их повторного использования многократно возрастает, особенно тогда, когда речь идет о коммерческих решениях. Таким образом, на сегодняшний день вопрос стандартизации лингвистической разметки является центральным и приоритетным при подготовке корпусных языковых данных. При этом по-прежнему не существует единого стандарта для лингвистического аннотирования; вариативность здесь, с одной стороны, связана с вариативностью возможных форматов разметки, а с другой стороны, с разнообразием языковых данных — лингвистических характеристик языковых структур, специфичных лингвистических паттернов отдельных языков и т.д.

Одним из самых старых стандартов текстовой разметки является Text Encoding Initiative[7] — TEI. Идея стандарта возникла в 1987 году и была направлена на формализацию документации для создания машиночитаемых текстовых объектов гуманитарного знания (литература, лингвистика, история и др.), явившись, таким образом, одним из первых стандартов, используемых для цифровизации культурного наследия. TEI — это, по сути, набор тегов и атрибутов для отображения самых разных параметров текста, которые используются в рамках синтаксиса XML (первоначально SGML). В дальнейшем развиваются близкие TEI корпусные форматы, также основанные на SGML/XML, предлагающие более последовательную и одновременно более простую схему: стандарт кодирования корпусов CES (Corpus Encoding Standard) (Ide 1998) и включающая его инициатива EAGLES (Expert Advisory Group for Language Engineering Standards) (Calzolari, McNaught, Zampolli 1996). Именно в стандарте CES была впервые сформулирована концепция standoff аннотаций — т.е. таких аннотаций, которые не вписываются непосредственно в текст, но связаны с аннотируемым словом через указатель или id. Такой формат позволяет отделять исходный текст от его разметки, а также облегчает дополнение разметки новыми уровнями. Именно этот подход является сегодня мейнстримом. Схемы и форматы, разработанные в рамках CES и EAGLES, можно рассматривать как подготовительную фазу к более масштабному процессу регулирования и стандартизации аннотации корпусов, который был запущен в начале 2000-х годов, благодаря созданию специальной группы «Управление языковыми ресурсами» (“*Languages resource management*”)[8] (Kiyong, Laurent 2010) в рамках ISO — международной организации по стандартизации. Принципиально новым решением стандарта LAF, разработанным рабочей группой, в дальнейшем реализованным в XML-формате GRAF (Graphic annotation format), стал переход от иерархической организации аннотаций к графовой модели представления данных (Ide et al., 2017). LAF основывается на двух фундаментальных принципах: во-первых, создается абстрактная модель данных, которая явным образом разграничивает структуру (физический формат) данных и их содержимое (название лейблов и категорий), во-вторых, вслед за стандартом CES утверждается принцип standoff аннотаций, хранящихся отдельно от текстов. Абстрактная модель данных LAF представляет собой нецикличный направленный граф с параметрами, связывающий с помощью «якорей» сегменты текста с узлами-лейблами. Формат GRAF, который развивает и продолжает идею LAF, предназначен для использования в качестве «стержневого формата» (pivot format), является не самостоятельным форматом, но инструментом совмещения разных форматов разметки для создания многоуровневых аннотаций, которые могут в него переводиться и, наоборот, из него извлекаться. Таким образом, был сделан важный шаг в направлении

[7] https://tei-c.org/
[8] https://www.iso.org/committee/297592/x/catalogue/

совместимости форматов. Разработанная абстрактная графовая модель, лежащая в основе лингвистических аннотаций корпусов, изоморфна многим современным форматам, в том числе и такому общему формату, как формат Linked Open Data (открытых связанных данных) RDF/OWL.

Одновременно с институциональной разработкой форматов лингвистической разметки корпусов происходил процесс развития форматов «де факто», т.е. таких, которые являлись результатом развития индустрии компьютерной лингвистики. Так, оказалось, что формализм синтаксических зависимостей, представляющий предложение как систему связанных элементов «вершина — зависимое», является наиболее удобным для автоматического синтаксического парсинга. Рост популярности этого подхода выразился в распространении трибанков (корпусов синтаксически размеченных предложений). Использование трибанков для обучения парсеров естественным образом привело к необходимости стандартизации синтаксической разметки. Важным этапом для формирования общего формата представления синтаксических отношений стало соревнование (shared task) по зависимостному парсингу, проведенное на конференции CONLL в 2006 году и включавшее в себя задачу установления синтаксических зависимостей для текстов на 13 языках (Buchholz, Marsi 2006). Формат представления данных CONLL-X, предложенный на конференции, стал де факто стандартом. В этом формате каждый токен представляет собой строку из 10 колонок, которые заполняются лингвистической информацией, включая отношения между вершиной и его зависимым. Этот формат задает модель данных, но не устанавливает никаких ограничений собственно на те условные обозначения (лейблы), которыми заполняются колонки, а также на те критерии, по которым устанавливаются синтаксические отношения. В результате даже следующие CONLL-X формату трибанки на разных языках чрезвычайно затруднительно сравнивать. Ответом на эту проблему стала инициатива «универсальных зависимостей» Universal Dependencies (Nivre et al., 2016), (De Marneffe et al., 2021), созданная в 2014 году группой ученых под руководством Иоакима Нивра. Universal Dependencies (UD) предлагает готовый универсальный набор из трех лейблов — часть речи, морфологические параметры и название синтаксического отношения, а также единые правила, определяющие то, как устанавливаются синтаксические зависимости. К моменту запуска инициативы были представлены трибанки на 10 языках, выполненные в этом формате, а в настоящий момент количество представленных трибанков превышает 200, включая данные на древних языках, таких как латынь, древнегреческий, церковнославянский. Безусловному распространению принципов UD способствовало развитие парсера UDPipe, обученного на существующих в коллекции UD трибанках а также приспособленного для обучения на любой коллекции в формате CONLL-X, обогащенного лейблами универсальных зависимостей, этот формат получил название CONLL-U (Straka, Hajic, Straková 2016). На сегодняшний день этот формат наиболее широко используется для подготовки корпусных данных для задач NLP, а также для корпусной лингвистики в целом.

В России распространение CONLL-U было во многом стимулировано соревнованием морфосинтаксических парсеров GRAMEVAL, прошедшим в 2020 году (Lyashevskaya et al., 2020).

Автоматическая разметка корпусов текста является ключевым решением для обработки больших объемов данных, поскольку ручная разметка невозможна в разумные сроки. Принципиально важным является не только приписывание потенциально возможных тегов словоформе, но выбор единственно правильных тегов и соответствующей леммы, т.е. снятие морфологической омонимии. Эта задача решается с помощью современных методов, основанных на глубинном обучении. Модель Rubic, дополнившая традиционно используемый в НКРЯ алгоритм MyStem, которая была применена для разметки корпусов НКРЯ, будет подробно описана в 4.1. Важным результатом применения модели Rubic стало внедрение синтаксической разметки и автоматическое снятие омонимии для больших объемов данных Основного и Газетных корпусов (более 1,24 миллиарда словоупотреблений). Этот результат позволил не только существенно расширить возможности поисковых запросов, но разработать сервисы комплексного анализа выдачи, отвечающие современным требованиям корпусных инструментов и методов лингвистических исследований.

1.1.3 Инструменты статистического анализа

Наличие встроенных статистических инструментов существенно упрощает лингвисту работу с корпусными данными, потому что, во-первых, позволяет отказаться от статистических подсчетов

вручную, а во-вторых, помогает увидеть важные статистические особенности и распределения языковых явлений.

Большинство крупных корпусов мирового уровня уже немыслимы без статистических инструментов. В целом, в корпусах существует всего три основных типа выдачи: конкорданс (чаще всего в формате KWIC), коллокации и частотные списки (Stefanowitsch 2020). Конкорданс — исторически первый и наиболее просто реализуемый способ выдачи, без которого корпус не может считаться корпусом. Под коллокациями в корпусной лингвистике понимаются привычные и повторяющиеся сочетания слов (Firth 1957). Все статистические метрики, применяемые для извлечения коллокаций, идейно основаны на том, что в текстах естественного языка слова встречаются друг с другом не случайно: их сочетаемость ограничена, по крайней мере, грамматикой и семантикой, поэтому она подчиняется определенным статистическим распределениям, а значит, коллокации могут извлекаться автоматически (Evert 2008). Коллокации позволяют исследователю получить в концентрированном виде информацию о сочетаемости того или иного слова, найти слова, которые связаны с ним более тесными связями. Некоторые современные корпуса ориентированы именно на поиск коллокаций, наиболее яркий пример — корпуса, работающие на платформе Sketch Engine[9]. Главная функция, реализованная на этой платформе, а именно построение скетчей слова, основана на обнаружении коллокаций с учетом синтаксических связей. Еще один корпус, нацеленный в том числе на поиск коллокаций, это COCOCO[10] (Kopotev et al., 2015). Встроенные инструменты для поиска коллокаций также существуют, например, в таких корпусах, как Британский национальный корпус[11], Чешский национальный корпус[12], Корпуса Университета Лидса[13].

Частотные списки могут иметь разнообразные применения, в частности, для составления пособий для изучения некоторого языка как иностранного, но если говорить о чисто исследовательских целях, то частотные списки особенно полезны для выявления отличий корпусов друг от друга или корпуса и его подкорпуса. Помимо лемм, упорядоченных по частоте встречаемости, частотный список также может включать грамматические формы слов некоторого языка. Впервые такая задача в неавтоматическом режиме была решена еще в 1982 году на данных Брауновского корпуса (Francis, Kučera 1982). Было разработано несколько приложений для построения частотных списков, среди известных следует упомянуть, по крайней мере, WordSmith Tools[14] и MonoConc[15] (оба приложения коммерческие). Однако открытых корпусов со встроенными инструментами для построения частотных списков (в том числе для пользовательских подкорпусов) мало. Впервые эта возможность была реализована на платформе Sketch Engine. В Британском национальном корпусе функционал построения частотных списков существует только в очень ограниченном виде.

Корпусом с наиболее разнообразными статистическими инструментами можно без сомнения назвать Чешский национальный корпус. Помимо коллокаций, корпус позволяет получить данные о частотных распределениях лемм и любых сочетаний грамматических признаков. Еще одним чрезвычайно наглядным инструментом является Word at a Glance (Machálek 2020a). Он позволяет увидеть в одном окне самую разнообразную информацию о слове: его частотность в жанровых подкорпусах, сочетаемость в рамках коллокаций, похожие слова, изменение частотности во времени, а также несколько примеров с заданным словом из корпуса.

Развитие инфраструктуры для многомерных корпусных исследований — унификация разметки, развитие статистических инструментов и корпусных приложений, наконец, значимое увеличение объемов самих корпусов способствовали тому, что корпусные методы стали применяться не только для изучения лингвистических явлений, но для решения более широких прикладных лингвистических задач, таких как преподавание, переводоведение, автоматический анализ языка, а также в разных областях социальных наук.

[9] https://www.sketchengine.eu/
[10] https://cococo.cosyco.ru/about.html
[11] https://www.english-corpora.org/bnc/
[12] https://www.korpus.cz/
[13] http://corpus.leeds.ac.uk/internet.html
[14] https://www.lexically.net/wordsmith/
[15] https://monoconc.com/

1.1.4 Расширение аудитории корпусных исследований

Первоначально корпуса создавались для лингвистических исследований и сбора естественных примеров для описания разнообразных языковых явлений. Однако по мере развития корпусной инфраструктуры и методов корпусного анализа, корпуса стали применяться и для других задач. В первую очередь корпуса стали использоваться в преподавании языка как иностранного, начало этой методике было положено еще несколько десятилетий назад (Wray 2013). Можно сказать, что существует отдельный подход data-driven learning, который подразумевает, что изучающий язык использует корпус как конкорданс для обнаружения характерного для некоторого слова языкового поведения (Johns, King 1991), (Boulton 2011). На основе корпусов создаются учебные материалы, упражнения и словари (McCarthy 2008). Параллельные корпуса находят наиболее широкое применение в переводоведении и обучении переводчиков (Doval, Sánchez Nieto 2019), (Beeby, Rodríguez, Sánchez-Gijón 2009), (Zanettin 2013).

Безусловно, корпуса имеют огромное значение для решения разных задач, связанных с обработкой естественного языка, в том числе и как обучающие коллекции для машинного обучения. Среди классических задач можно назвать обучение чат-ботов (Shawar, Atwell 2005), разрешение омонимии (Roll, Correia, Berger-Tal 2018), определение тональности (Yang, Lin, Chen 2007), (Schrauwen 2010), классификации текстов (Curtotti, Mccreath 2010) и многоге другое.

Корпусные методы широко применяются в судебной лингвистике, например, для установления авторства текста или для целей семантического анализа и определения значения слова (Coulthard 1994), (Heffer 2005), (Coulthard, Johnson, Wright 2017), (Баранов 2023).

Использование корпусного подхода в социальных науках определяет специфику и тематику собираемых корпусов (Wiedemann 2013). Корпуса используются для решения таких задач, как обнаружение «языка ненависти» (Poletto et al., 2021) или для анализа сообщений из соцсетей при разного рода катастрофах (Imran, Mitra, Castillo 2016). При этом узкая тематическая направленность задачи не обязательно влияет на размер корпуса. Так, в период с января 2020 года до декабря 2022 года был собран мультиязыковой корпус сообщений прессы, посвященный коронавирусу, который достиг объема 1,2 млрд словоупотреблений (Davies 2021). В целом, можно сказать, что на сегодняшний день мы видим примеры использования корпусных методов практически в любой области знаний, так или иначе связанной с естественным языком, от библиотечного дела (Bowker 2018) до экспериментальной психологии (Chatrand 2022).

1.2 НКРЯ 2.0 в свете основных тенденций развития современных корпусов

Обзор современного состояния развития корпусной лингвистики позволил выдвинуть принципиальные требования к новой платформе по сравнению со старой. Ниже обобщены ключевые результаты развития платформы 2.0.

Объемы корпусов

Старая платформа: К 2020 году общий объем всех корпусов НКРЯ составляет около 1 миллиарда словоупотреблений. При этом грамматическая омонимия снята лишь менее чем в 1% от всех словоупотреблений (только вручную снятая омонимия в Основном, Обучающем и Устном корпусах, а также снятая при помощи процессора Этап-3 и проверенная вручную омонимия в СинТагРусе). Дальнейшее расширение корпуса затруднено в связи с архитектурными ограничениями платформы.

Новая платформа: К 2024 году общий объем всех корпусов НКРЯ составляет около 2 миллиардов словоупотреблений, грамматическая омонимия снята примерно в 65% от всех словоупотреблений (добавилась автоматически снятая омонимия в Основном корпусе и обоих корпусах СМИ). Новая платформа позволяет увеличить объем данных НКРЯ до 100 миллиардов словоупотреблений.

Разметка данных

Старая платформа: Корпусные данные снабжаются морфологической разметкой, осуществленной с помощью программы MyStem (Зобнин, Носырев 2015), представляющей собой контаминацию недетерминированного конечного автомата и наивного байесовского классификатора. Алгоритм позволяет строить гипотетические разборы для слов, которых нет в грамматическом словаре Зализняка, и ранжирует леммы по вероятности в случае омонимии. Используется собственный тип разметки НКРЯ.

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. Используется разметка формата CONLL-U, разработаны принципы взаимной трансформации разметок НКРЯ и CONLL-U. Подробнее о том, как работает модель Rubic, будет изложено в разделе 4.1.

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, который может быть построен только 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 (например, анализ частотности не только встретившихся словоформ, но и обобщение результатов на уровне грамматических параметров). Появилась возможность получить предпосчитанную информацию о слове в корпусе в целом — «Портрет слова», куда входят его скетчи (коллокации на основе базовых синтаксических связей), контекстуально близкие слова, выявленные на основе расчета семантических векторов, однокоренные слова в корпусе. Описания методов, использованных для подготовки «Портрета слова», приводятся в разделах 3.2 и 3.3. Подробное описание пользовательских возможностей, реализованных в новых корпусных инструментах, было представлено в статье (Савчук и др., 2024).

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 языковым корпусами. Корпус существует в мобильной версии, имеет богатейшую документацию, логика интерфейса минимизирует усилия пользователя по получению информации. Подробно идеология обновления интерфейса новой корпусной платформы представлена в разделе 2.2.

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

Разработка новой платформы НКРЯ включала в себя перестройку на уровне «бэкенда» — архитектуры индексации, поиска и статистической обработки корпусных данных, а также концептуальное обновление на уровне «фронтенда», т.е. пользовательского интерфейса, с помощью которого пользователь взаимодействует с корпусами НКРЯ. Для взаимодействия между корпусной архитектурой и веб-интерфейсом был создан новый инструмент для работы с корпусами, позволяющий взаимодействовать через скрипты, использующие API, без необходимости пользовательского интерфейса. Ниже мы последовательно рассмотрим подходы и решения для каждого из этих направлений.

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. Например, вместо поддержки конкретного списка возможных атрибутов, приписываемых каждому токену в тексте, вводилось понятие «атрибута, приписанного к токену», и программировались универсальные алгоритмы, рассматривающие список атрибутов заданного типа как параметр.

Такой подход потребовал унификации корпусных данных: атрибуты одинакового типа обрабатываются одними и теми же алгоритмами, а значит должны быть единообразно представлены. Унификация была достигнута либо путем изменения разметки исходных текстов, либо за счет подключения небольших модулей-конвертеров, что позволяло избежать переобучения лингвистов-разметчиков, готовящих тексты корпуса.

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.

Были выделены следующие виды разметки, поддерживаемой платформой:

Структурная единица, реализуемая вычислительным ядром	**Структурная единица, реализуемая лингвистическим ядром**	**Приписываемые к структурной единице атрибуты**
токен	слово	лемма, грамматические параметры, морфемное членение, семантические, орфоэпические и др. параметры (в случае снятой омонимии) словоформа, а также вспомогательные параметры: повтор лексемы, знаки препинания до и после слова, начало/конец предложения, слово с заглавной буквы (в случаях как снятой, так и не снятой омонимии)
разбор	вариант снятия омонимии	лемма, грамматические параметры, морфемное членение, семантические, орфоэпические и др. параметры (в случае неснятой омонимии)
сегмент	предложение / фраза в устных корпусах / грамота в корпусе «Берестяные грамоты»	в настоящее время нет атрибутов, но предусмотрена их поддержка при необходимости
текст	письменный текст / устный текст / выровненные между собой тексты в параллельных корпусах (=текст+перевод) / текст+последовательность аудио- или видеофрагментов в мультимедийных корпусах	мета-атрибуты: тип текста, жанр, тематика и т.д.
фрагмент (последовательность подряд идущих сегментов)	абзац, строфа, реплика, клипотекст[16], зона выравнивания (в параллельных корпусах)	говорящий (для реплик и клипотекстов), вычисленный возраст аудитории (в детском корпусе), речевые акты (приписываются к клипотексту)
подмножество токенов внутри сегмента	клауза, группа, микросинтаксическая конструкция	тип клаузы, вид микросинтаксической конструкции, лемма микросинтаксической конструкции (заложено в платформу, но в настоящее время поддерживается ограниченно)

[16] В мультимедийный корпусах клипотекст — минимальная единица, состоящая из отрывка видео или аудио и соответствующего ему текста, а также набора жестов и речевых действий.

Структурная единица, реализуемая вычислительным ядром	**Структурная единица, реализуемая лингвистическим ядром**	**Приписываемые к структурной единице атрибуты**
последовательность токенов, пересекающая границы сегментов	стихотворная строка	метр, количество стоп/иктов/слогов, клаузула, схема
ребро графа токенов внутри сегмента	синтаксическая связь, лексико-функциональная связь	направление связи, тип связи, вспомогательный предлог (в лексико-функциональной связи)
меж-токены (фантомы)	эллидированное слово[17]	как у токенов, но отсутствует словоформа
выравнивание фрагментов	выравнивание текста с переводом в параллельных корпусах, выравнивание текста с видео/аудио в мультимедийных корпусах	нет атрибутов
соответствие токенов внутри выровненных фрагментов	«пословное выравнивание» (спроектировано, но в данный момент не используется)	нет атрибутов
внутритокенная разметка	ударение, морфема, шрифтовое выделение	атрибуты морфологического разбора, тип шрифтового выделения

Таблица 1: Виды разметки, поддерживаемые НКРЯ

В дальнейшем при возникновении новых атрибутов, попадающих в один из выделенных классов, не потребуется изменения вычислительного ядра. Вся необходимая функциональность будет в этом случае обеспечена лишь не очень значительными изменениями лингвистического ядра. Ниже мы покажем на трех примерах, каким образом вычислительный модуль позволяет гибкую настройку нового лингвистического функционала.

Пример 1. Организация поиска в мультимедийном корпусе

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 тем же способом, что и два параллельных текста, выровненных между собой. Это позволяет переносить функциональные возможности, разработанные для параллельных корпусов, и на мультимедийные. Так, для параллельных корпусов программная платформа нового поколения поддерживает поиск по запросу, условия которого накладываются на тексты на обоих языках в выровненной паре. Это позволяет, например, искать такие английские предложения, содержащие слово *cat*, перевод которых содержит слово *кошка* (а не *кот*). При этом условия на каждом языке могут накладываться не только на отдельные слова, но и на их сочетания (например, на расстояния между словами) (Сичинава 2022). Новый запрос обеспечения возможности поиска по комбинации жестов в мультимедийном корпусе (с возможностью комбинации условий

[17] Эллидированному слову в тексте приписываются все атрибуты, которые есть у обычного слова, за исключением собственно словоформы — чтобы по ним можно было осуществлять поиск.

на жесты с условиями на текст) оказался фактически аналогичным поиску по условиям, накладываемым на слова на двух языках (с точностью до замены слов одного из языков на жесты). Таким образом, реализация поиска по комбинации жестов потребовала лишь небольших изменений в лингвистическом ядре.

Пример 2. Поиск по совпадению и различию атрибутов токена

В настоящее время система поддерживает только поиск по совпадению некоторых атрибутов одного слова в тексте с соответствующими атрибутами соседнего слова. Это делается при помощи дополнительных приписанных к токену атрибутов «повтор лексемы», «повтор числа» и т.д. Однако для многих лингвистических исследований принципиальна важна возможность поиска по совпадению атрибутов слов, не идущих в тексте подряд.

Наш анализ показал, что если интерпретировать совпадение двух атрибутов графовой связью между двумя токенами, то условие на то, что у слова А некий атрибут совпадает с соответствующим атрибутом слова Б, может быть представлено как условие на графовую связь аналогично поисковому запросу по расстоянию или по заданной синтаксической связи. Таким образом, для выполнения поставленной задачи потребуется дополнить поисковый индекс разметкой графовых связей отдельного типа, отражающей совпадение атрибутов между токенами. После этого потребуются лишь небольшие изменения лингвистического ядра, чтобы возможность такого поиска был включена в корпусную систему. Такие изменения уже включены в план работ на будущее.

Пример 3. Разметка сложности текста для читателей разного возраста

При создании корпуса «От 2 до 15» для отдельных частей произведения размечался показатель сложности текста в терминах предполагаемого возраста читателя (Morozov, Glazkova, Iomdin 2022). Такое приближение часто используется в прикладных исследованиях, посвященных созданию алгоритмов автоматической оценки сложности текста. В этом случае читательский опыт респондентов одного возраста считается схожим, а верхняя граница сложности устанавливается на уровне последних классов школы или младших курсов университета. В ситуации с корпусом «От 2 до 15» возрастная разметка была экспериментально собрана таким образом, чтобы определить наиболее популярные художественные произведения у носителей языка различного (школьного) возраста. Считался порог, при котором 50% респондентов к этому возрасту прочли книгу.

Авторами исследования была разработана нейросетевая модель, позволяющая предсказывать сложность текста. На вход модель получает фрагмент текста (один или несколько абзацев), на выходе возвращает предполагаемый возраст читателя.

Перед разработчиками платформы встала задача организации поиска по атрибуту уровня читательского опыта. Анализ показал, что атрибут необходимо приписывать к фрагментам текста (а нарезку текста на фрагменты организовать в соответствии с разметкой). Таким образом, путем несложных изменений в лингвистическом ядре может быть реализован поиск[18], аналогичный поиску по условиям на говорящего, реализованному в устном корпусе.

Приведенные примеры показывают, каким образом базовый список единиц, поддерживаемых вычислительным ядром (см. выше Таблицу 1), не только обеспечивают уже существующую функциональность поиска, но и открывают гибкие возможности для ее расширения по мере развития корпусной специальной разметки и запросов на новые исследовательские возможности.

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. Их создатель берет на себя задачи по подготовке и хранению корпусных данных, снабжению их инструментами поиска и лингвистического анализа, предоставлению удобного интерфейса для доступа к этим инструментам, поддержания и своевременного обновления программного и аппаратного обеспечения, необходимого

[18] В настоящий момент разметка фрагментов моделью недоступна на корпусной платформе. Фрагментам приписана сложность всего текста, что является временным и упрощенным решением. Полноценная разметка фрагментов планируется к внедрению на корпусную платформу в первой половине 2025 г.

для функционирования системы. Все это существенно снижает порог вхождения для пользователей, позволяя им приступить к работе без развертывания корпуса и инструментов к нему на собственном персональном компьютере. Однако такой подход сужает спектр возможных исследований, поскольку 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. Благодаря такой возможности пользователь может строить сложные комбинации из многочисленных однотипных запросов, решая задачи, недоступные ему через интерфейс (например, повторить один и тот же запрос для тысячи различных лемм). Автоматизация обеспечивает меньше возможностей, чем развертывание корпуса и инструментов к нему на собственном компьютере, однако не требует существенных ресурсов на стороне пользователя и позволяет избежать проблемы с авторскими правами, возникающей при открытой публикации всех текстов корпуса.

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 (поисковый запрос полностью формируется на стороне бэкенда и передается на сторону фронтенда при помощи API).

Таким образом, в настоящее время нельзя утверждать, что API решает задачу автоматизации. Однако такое использование API 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, а также разработать библиотеку на языке Python для реализации наиболее популярных сценариев использования корпуса через API. Выбор языка Python обусловлен его популярностью среди специалистов по компьютерной лингвистике.

В качестве базового протокола для передачи данных в рамках API мы использовали протокол сериализации структурированных данных Protocol Buffers[19] (ProtoBuf).

Можно выделить следующие виды запросов к API, соответствующие основным видам использования корпусной платформы:

1. запрос основных статистических данных о корпусах НКРЯ;
2. запрос данных о конфигурации конкретного корпуса: доступных в корпусе видах выдачи и их параметрах, возможных настройках отображения, сортировки и группировки выдачи;
3. запрос набора доступных в конкретном корпусе поисковых форм (как, например, поиск точных форм или поиск коллокаций), а также их состава — набора доступных для задания полей и их иерархии;
4. запрос типов атрибутов, имеющихся в корпусе, и возможных значений для тех атрибутов, которые выбираются из списка;
5. поисковый или аналитический запрос с указанием желаемого типа представления результата (например, конкорданс, KWIC или частотность) из числа доступных в корпусе;
6. запрос «Портрета» конкретного слова с указанием списка типов выдачи из числа доступных в корпусе;
7. запрос «Портрета» конкретного корпуса.

Для каждого вида запроса разработаны форматы сообщений, которыми обмениваются клиент и сервер. Использование протокола ProtoBuf позволяет автоматически проверять соответствие формата сообщения требуемому, что снижает вероятность ошибки.

[19] https://protobuf.dev/

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

Как было показано во вступительной обзорной части, метод корпусного анализа вышел далеко за пределы собственно академических лингвистических исследований. НКРЯ активно используется исследователями из смежных гуманитарных дисциплин, преподавателями, писателями и переводчиками. Кроме того, поиск в Корпусе не всегда связан с профессиональной деятельностью пользователя; он может служить инструментом для удовлетворения общего интереса к русскому языку и культуре. Одной из ключевых задач при разработке новой корпусной платформы стало изменение дизайна и функциональности интерфейса таким образом, чтобы доступ к данным корпуса не требовал специальной лингвистической подготовки, устранялись существующие барьеры и Национальный корпус становился доступным для всех, кто проявляет к нему интерес.

Доступность корпусной платформы для широкого круга пользователей предполагает выполнение следующих требований:

- возможность использовать платформу с разнообразных устройств: настольного компьютера, планшета, смартфона;
- поддержка иноязычных пользователей, использующих Корпус для изучения русского языка;
- возможность получения всего имеющегося спектра информации по минимальному запросу;
- визуальная наглядность отображаемых результатов;
- уменьшение количества лишних действий пользователя;
- развитая система контекстных подсказок и руководств и анонсов, помогающая пользователю ориентироваться в программной системе.

Рассмотрим, как новый интерфейс корпусной платформы решает перечисленные задачи.

2.3.1 Адаптация интерфейса под пользовательские устройства

По данным SimilarWeb[20], на январь 2025 около 70% пользователей интернета пользуются им с мобильных устройств: смартфонов или планшетов. Доля мобильных пользователей постоянно растет. Не являются исключением и пользователи НКРЯ.

Перед разработчиками платформы нового поколения стояла задача, с одной стороны, обеспечить аналогичность интерфейса при доступе с различных устройств, а с другой — учитывать особенности каждого вида устройств при построении интерфейса для них. Для этого новая версия сайта поддерживает набор стилевых таблиц (в соответствии с технологией CSS) для различных размеров экрана, соответствующих наиболее распространенным классам мобильных и настольных устройств — от самого миниатюрного смартфона до настольного компьютера с хорошим разрешением экрана. Разработанное веб-приложение реализует 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). Это 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).

[20] https://www.similarweb.com/ru/platforms/

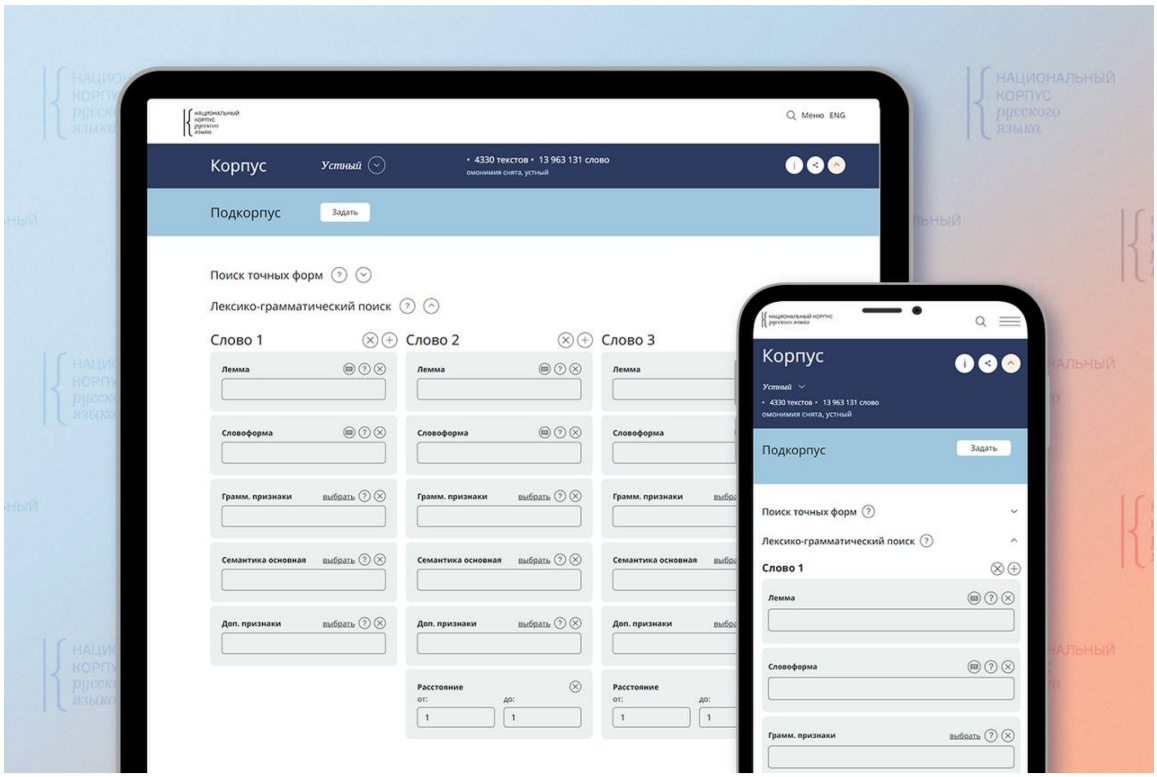


Рис. 1: Слева — версия сайта для ПК и планшета, справа — оптимизированная для мобильных устройств.

2.3.2 Поддержка иноязычных пользователей

Иностранные пользователи, использующие Корпус для изучения и исследования русского языка, заинтересованы в получении примеров и контекстов на русском языке. В то же время им может быть сложно выбрать настройки и другие параметры поиска, пользуясь русскоязычной терминологией. Поэтому для пользователей сайта, основным языком которых не является русский, реализована возможность переключить язык интерфейса на английский.

2.3.3 Получения всего имеющегося спектра информации по минимальному запросу

В то время как для опытных пользователей, хорошо знакомых с функционалом системы, естественно максимально конкретизировать запрос, заранее понимая, в каком виде ожидаются результаты, новые пользователи Корпуса, а также пользователи, недостаточно понимающие весь функционал системы, предпочитают наглядно увидеть все возможные результаты по своему запросу, чтобы выбрать, какие из них им интересны. Это позволяет снять барьер на вхождение и дать широкой аудитории инструменты для решения собственных задач. Для профессиональной аудитории привычно и ценно самостоятельно строить сложные поисковые запросы, анализировать большое количество данных, иметь возможность дополнительно обработать выдачу и сделать собственные научные выводы. В то же время у новой аудитории задача часто другая — быстро получить простой ответ на свой вопрос. Благодаря модульности и расширяемости платформы, нам удается на основе одного и того же внутреннего инструментария строить разный интерфейс для профессионалов и широкой аудитории.

Поскольку все содержательные лингвистические операции (поиск и анализ результатов) выполняются интерфейсом не напрямую, а через API, открывается возможность использовать одни и те же вызовы API в различных местах интерфейса. Приведем несколько примеров.

В 2022 году в НКРЯ появился сервис «Обзор возможностей», который дает пользователям представление о ключевых возможностях, доступных в НКРЯ, знакомит с общими принципами

устройства интерфейса, показывает, какие виды результатов можно получить, информирует о типичных ошибках при конструировании поисковых запросов. При этом используются те же самые вызовы API, что и при соответствующих запросах в основном интерфейсе этих корпусов.

Сервис «Портрет слова» также позволяет не конструировать несколько запросов к поисковому и другому функционалу и затем самостоятельно комбинировать их результаты, а делает это автоматически. Пользователю нужно лишь ввести начальную форму слова, после чего в визуально компактной и понятной форме сервис представит разнообразную информацию для всех имеющихся разборов заданной леммы. Пользователь увидит скетчи слова (как список коллокаций для основных синтаксических отношений), все грамматические формы слова (без необходимости искать и сравнивать разборы в разных примерах) и так далее.

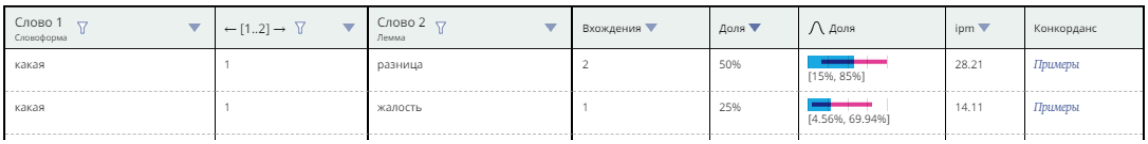
Из «Портрета слова» налажен переход в полный функционал поиска и наоборот, из результатов поиска, кликнув на разбор любого слова, можно перейти к его «Портрету». Такие перекрестные ссылки, позволяющие пользователю переходить из сервиса в сервис, не теряя контекст своего запроса, — это еще один пример того, как новый интерфейс помогает пользователям осваивать возможности корпуса.

2.3.4 Визуальная наглядность отображаемых результатов

Графическое представление результатов дает возможность быстро и эффективно донести до пользователя сложную для восприятия информацию. Так, например, барометр частотности в «Портрете слова» позволяет одним взглядом оценить, насколько частотной является лемма.

С помощью круговых, столбчатых диаграмм, географических карт и графиков в интерфейсе «Портретов корпусов» НКРЯ подается информация о структуре и составе корпусов. В «Портретах подкорпусов» можно с помощью сравнительных диаграмм проанализировать, насколько пользовательский подкорпус отличается от корпуса в целом.

Для того чтобы экспертная аудитория могла делать более глубокие и обоснованные научные выводы, важно учитывать особенности механизмов, с помощью которых проводились расчеты. Принципиально новый подход, реализованный в новом интерфейсе, состоит в том, что ограничения примененных методов анализа данных не только описываются в руководстве пользователя, но и визуализируются сразу при выдаче. Так, в виде выдачи «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, в разные периоды времени, 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»

Еще одним примером доступной визуализации является облако похожих слов, реализованное в функционале «Портрета слова». «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). Такие визуализации стали возможны благодаря внедрению инструментов статистического анализа и портретирования (см. разделы 3.1 и 3.2).

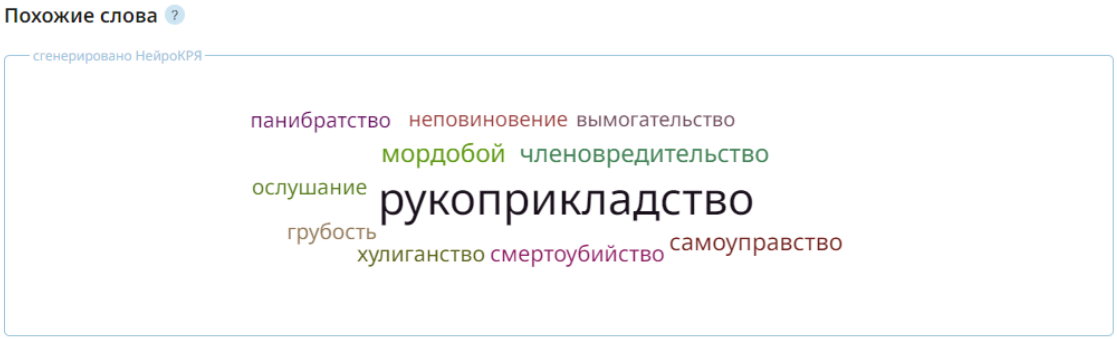


Рис. 4: Виджет «Похожие слова»

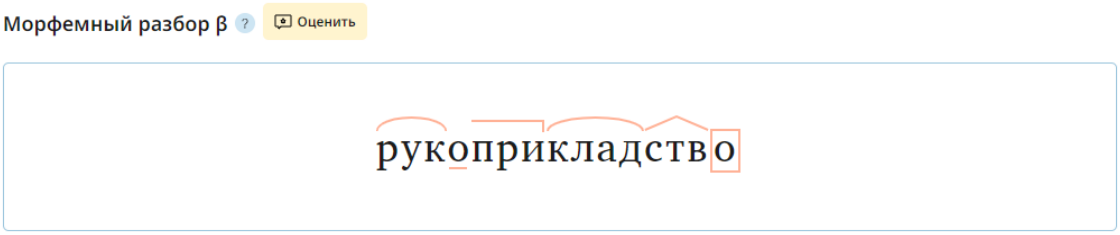


Рис. 5: Виджет «Морфемный разбор»

В нескольких сценариях поиска по Корпусу элементы интерфейса были намеренно перегруппированы по сравнению со старой версией корпусной платформы. Особенно заметна перегруппировка 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, поскольку в тексте слова также обычно располагаются на одной строке друг за другом.

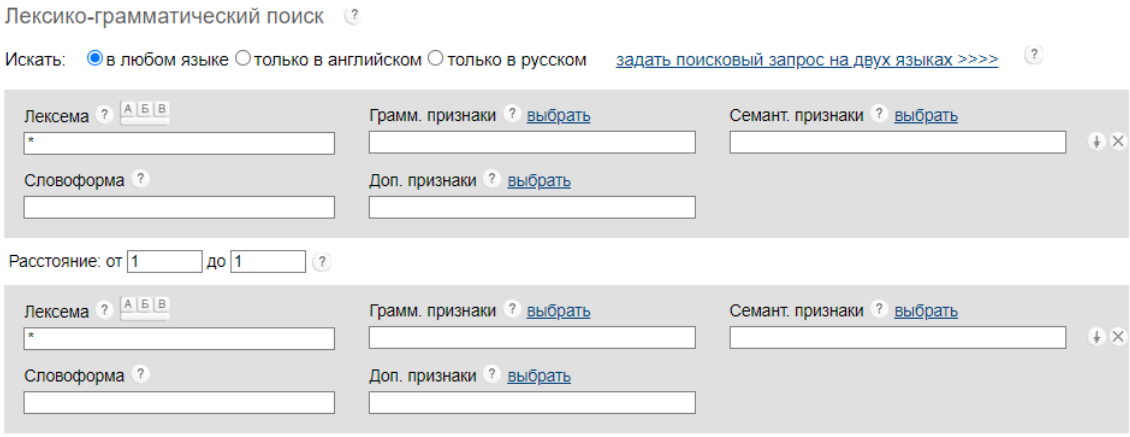


Рис. 6: Старое расположение условий лексико-грамматического поиска

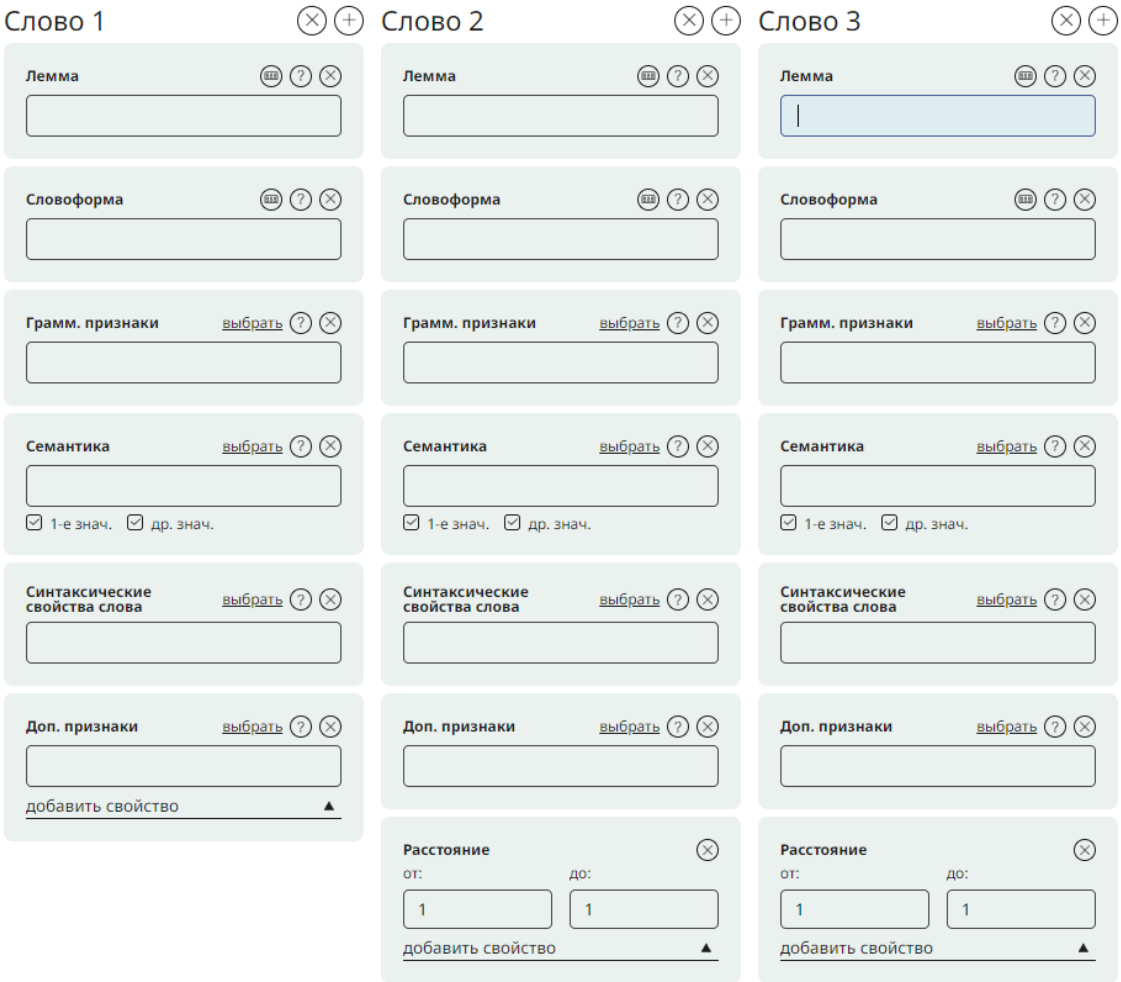


Рис. 7: Новое расположение условий лексико-грамматического поиска

Перегруппировка затронула и отображение результатов поиска в параллельных корпусах (Рис. 8 и 9). Теперь оригинальный фрагмент располагается слева, а переводы справа (можно переключаться между разными переводами). Это позволяет разместить на одном экране ПК больше примеров. Для мобильных устройств реализовано переключение с помощью слайдера, что более привычно для пользователей смартфонов.

Рис. 8: Старое расположение результатов поиска в параллельных корпусах

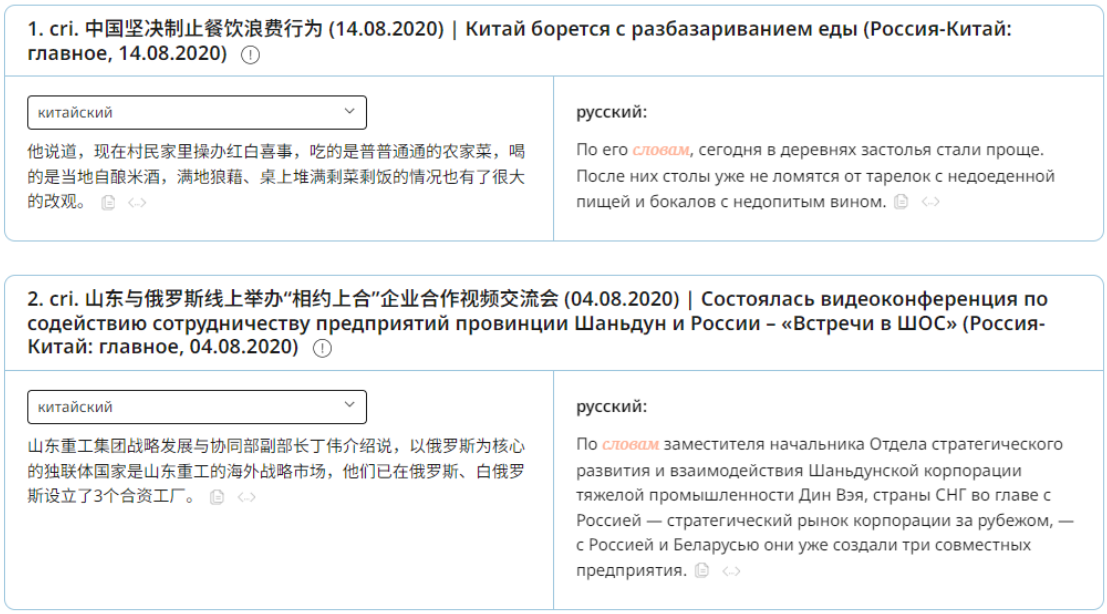


Рис. 9: Новое расположение результатов поиска в параллельных корпусах

2.3.5 Уменьшение количества лишних действий пользователя

Любой пользователь, вне зависимости от квалификации, совершает ряд действий в интерфейсе системы при каждом обращении к корпусу. Такие действия не должны отнимать много усилий, а напротив, должны быть максимально быстрыми и очевидными.

В новом интерфейсе непосредственно с главной страницы организован доступ к поиску по любому из корпусов, а также выведены ссылки для доступа к другому функционалу, который часто используется.

Существенная экономия достигается за счет отказа от перезагрузки веб-страницы при каждой операции в интерфейсе. Определенные достижения в этом направлении были уже в старом интерфейсе: некоторые сложные операции в нем выполнялись без перезагрузки страницы. Новый интерфейс изначально спроектирован так, чтобы уменьшить количество перезагрузок страниц.

Частичное изменение содержимого страницы в нем реализуется через асинхронный запрос к серверу с последующим переписыванием части страницы после получения ответа. Обработчики таких запросов на стороне сервера легковесны благодаря использованию таких же асинхронных запросов к API (см. раздел 2.2).

Все возможные на сегодняшний день пользовательские настройки поисковой выдачи собраны в едином меню «Настройки», что позволяет пользователю быстро их найти.

Выбор настроек запоминается в браузере пользователя и применяется для следующих поисковых запросов. Аналогично запоминаются предпочтительный вид поиска в каждом корпусе (соответствующая форма поиска будет всегда показываться открытой), вид выдачи, который будет открываться по умолчанию, а также режим отображения (или скрытия) подробной информации о запросе в шапке корпуса.

При постоянной работе с поиском по корпусу для удержания контекста важно всегда иметь перед глазами запрос, с которым работаешь. Полезным нововведением является отображение в шапке корпуса не только информации о параметрах искомого слова, но и о параметрах заданного пользователем подкорпуса. Возможность в любой момент вернуться к форме и откорректировать любые параметры сокращает усилия в сравнении с заданием параметров заново.

Для обмена результатами исследований, в том числе при публикации в научных журналах, теперь можно воспользоваться короткими ссылками на запрос и кнопкой «Скопировать пример», с помощью которой в буфер обмена помещается информация о примере и его выходных данных.

2.3.6 Система контекстных подсказок и анонсов, руководство пользователя

Для того чтобы менее подготовленные пользователи могли быстрее научиться пользоваться новым функционалом, в новом интерфейсе поддерживается регулярно обновляемое руководство пользователя, доступен поиск по руководству.

В разделе «Совет дня» «Обзора возможностей» регулярно размещается подробная информация о наиболее интересных нововведениях.

Актуальность вспомогательной информации, размещенной на сайте Корпуса, поддерживается с помощью системы управления контентом, которая позволяет в онлайн режиме структурировать, тегировать и редактировать онлайн анонсы, статьи, контекстные подсказки и руководство пользователя.

Корпуса НКРЯ различаются по данным, типам разметки и функциональным возможностям запросов и анализа. Тем не менее, общая концепция интерфейса, ориентированная на удобство, доступность и прозрачность взаимодействия пользователя с сервисами, сохраняется для всех корпусов. Важным является единый стандарт интерфейса, применяемый во всех корпусах, который обеспечивает для пользователя интуитивную легкость перехода между ними и расширяет доступность специализированных ресурсов. Этот аспект особенно ярко проявляется в новых сервисах анализа данных корпуса, например, таких как «Портрет слова». Хотя часть инструментов еще доступна не для всех корпусов, унифицированный шаблон помогает пользователю ориентироваться в информации и одновременно включает в себя перспективы дальнейшего развития сервисов в специализированных корпусах. Описание технологических аспектов, лежащих в основе развития новых сервисов, будет представлено в следующем разделе.

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, как показано в разделе 1.1.3, 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. Инструменты квантитативного корпусного анализа могут быть применены уже на этапе поиска, как, например, при поиске по коллокациям или же использоваться для дополнительного исследования результатов поискового запроса. Кроме того, квантитативный анализ может быть

проведен для всего корпуса в целом или для выбранного подкорпуса. В данном разделе мы рассмотрим категории аналитических инструментов, которые вошли в лингвистическое ядро программной платформы НКРЯ нового поколения.

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. А значит, статистические характеристики подкорпусов не могут быть предварительно рассчитаны на этапе индексации. Все расчеты выполняются вычислительным ядром непосредственно при обработке запроса. Это возможно, поскольку для осуществления расчетов не требуются сами тексты, достаточно лишь их заголовков и метаданных, что существенно снижает объем обрабатываемой информации.

В то же время, статистическая информация о корпусе в целом может быть для ускорения вычислена и сохранена в момент индексации. Если в будущем будут выявлены фиксированные подкорпусы, статистика по которым востребована существенно чаще остальных, то в целях ускорения вычисление статистик по ним также может быть перенесено на этап 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. Также этот инструмент используется для вывода информации о частотности и формах в «Портрете слова».

Это наиболее ресурсоемкие вычисления, выполняемые непосредственно в момент запроса, поскольку количество результатов поискового запроса может быть очень велико, а в экстремальных случаях — даже превышать размер самого корпуса (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. Так, 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. При этом абсолютные показатели получаются из относительных нормированием на полное количество результатов поиска. Как использование для вычислений случайной подвыборки, так и малый размер всей выборки могут вести к существенному падению точности получаемых результатов, поэтому система рассчитывает статистические доверительные интервалы, которые отображаются напротив вычисленных значений.

Для ускорения вычислений в оперативной памяти поддерживается полный индекс отдельных атрибутов слов. Это существенно повышает требования к аппаратному обеспечению (на каждый атрибут требуется несколько гигабайт оперативной памяти), но резко ускоряет расчеты.

Также в будущем возможно применение кэширования результатов расчетов. Это может значительно увеличить скорость в ситуации, когда пользователи, переключаясь, делают одинаковый запрос несколько раз за небольшой промежуток времени (например, переключаются туда-обратно между разными экранами в интерфейсе).

Особой подкатегорией этой категории является статистика по метаатрибутам примеров, встретившихся в выборке. Фактически, это статистика по подкорпусу текстов, отобранных по запросу, при этом каждый текст в ней участвует с весом, равным числу найденных в нем вхождений.

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 (см. об этом подробнее в разделе 3.3). 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.

Таким образом, 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.

В целом, можно заключить, что 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 рамках рассмотрения аналитических инструментов новой корпусной платформы НКРЯ мы уделим особое внимание двум нейросетевым моделям, которые были разработаны специально для сервиса «Портрет слова». Первая — модель словообразовательного разбора в «Портрете слова». Информация о внутренней структуре слова, получаемая в результате работы модели позволяет не только отобразить его морфемный состав, но связать слово с однокоренными словами. Пользователь может, таким образом, с помощью одного клика перейти с портрета исходного слова на портрет его однокоренного. Семантическая информация о слове обеспечивается с помощью векторной модели дистрибутивной семантики, подсчитанной для разных корпусов. Виджет «Похожих слов» таким образом обеспечивает связность исходного слова и его квазисинонимов и ассоциатов. Такой переход также реализован в интерфейсе. Ниже будет рассмотрена каждая из моделей.

3.2 Модель словообразовательного разбора

Помимо широко используемых во многих корпусах видов разметки, таких как, например, разметка морфологических свойств слов, в НКРЯ встречается и специализированная разметка. Одним из видов такой разметки является словообразовательная, то есть разметка морфем, из которых состоит слово, и их типов. Словообразовательная разметка востребована как для лингвистических исследований (Гришина и др., 2009), так и для обучения русскому языку. Многие из орфографических правил, вызывающих наибольшие затруднения, связаны именно с морфемной структурой слова — например, правописание безударных гласных в корне или правописание приставок[21]. В НКРЯ словообразовательная разметка присутствует в двух корпусах: Основном и Обучающем. Важно отметить, что в НКРЯ разметка применяется к лемме слова без учета формы

[21] https://yandex.ru/company/researches/2016/ya_spelling

или контекста конкретного словоупотребления. С точки зрения архитектуры системы морфемный разбор является атрибутом, приписанным структурной единице «разбор». Для каждого разбора указан список морфем, их тип (приставка, корень, интерфикс, суффикс, окончание или постфикс) и линейная позиция в слове.

В основе разметки словообразовательной структуры 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 (310 тысяч неуникальных морфем). Этот словарь составлялся на основании идеологии «Словаря морфем русского языка» А. И. Кузнецовой и Т. Ф. Ефремовой (Кузнецова, Ефремова 1986). Принципы этой идеологии — значительная (хотя и не максимальная) дробность выделения морфем и соотносимость с другими лексемами аналогичного строения. Поэтому морфемное деление в разметке корпуса не совпадает с принятым, например, в школе. В исконных словах могут выделяться морфемы, даже если слово без них употребляется маргинально (*у-лыб-а-ть-ся*, ср. *у-смех-а-ть-ся*) или если мотивированность этимологии слова для современного носителя неочевидна (*на-сек-ом-ое*, *вос-точ-н-ый*). В иностранных словах заимствованные основы членятся (например, *ре-волюц-и-я*, *квит-анци-я*), если усматривается семантическое и структурное соответствие между ними и лексемами похожего строения (ср. *э-волюц-и-я*, *рас-квит-а-ть-ся*). Разбираются в том числе и служебные части речи, а также имена собственные и производные от них.

Разметка морфем в Обучающем корпусе опирается на разработанный на основе «Морфемно-орфографического словаря» А. Н. Тихонова (Тихонов 2002), словарь морфемного анализа **Morphodict-T**. Этот словарь содержит около 100 тыс. лексем. Морфемный состав слова в Morphodict-T 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 Основном корпусе: например, указанные выше слова анализируются как *улыб-а-ть-ся*, *насеком-ое*, *восточ-н-ый*, *революци-я*, *квитанци-я*. В «Портрете слова», представленном в Обучающем корпусе, дается морфемное строение только слов, относящихся к знаменательным частям речи, — нарицательным существительным, прилагательным, глаголам и наречиям.

На Рис. 10 и Рис. 11 видна разница между морфемным разбором в Основном корпусе, построенном на основе словаря Тихонова, и разбором в Обучающем корпусе, построенном на основе словаря Кузнецовой и Ефремовой, в виджетах «Портрета слова» этих корпусов.

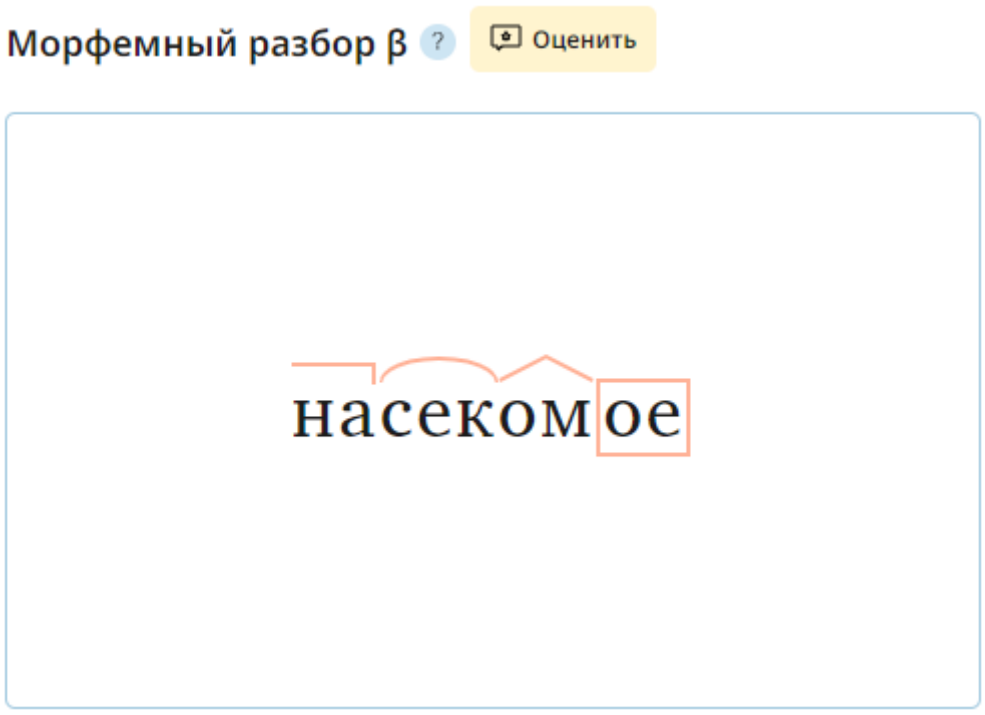


Рис. 10: Основной корпус. Морфемный разбор на основе словаря Кузнецовой и Ефремовой

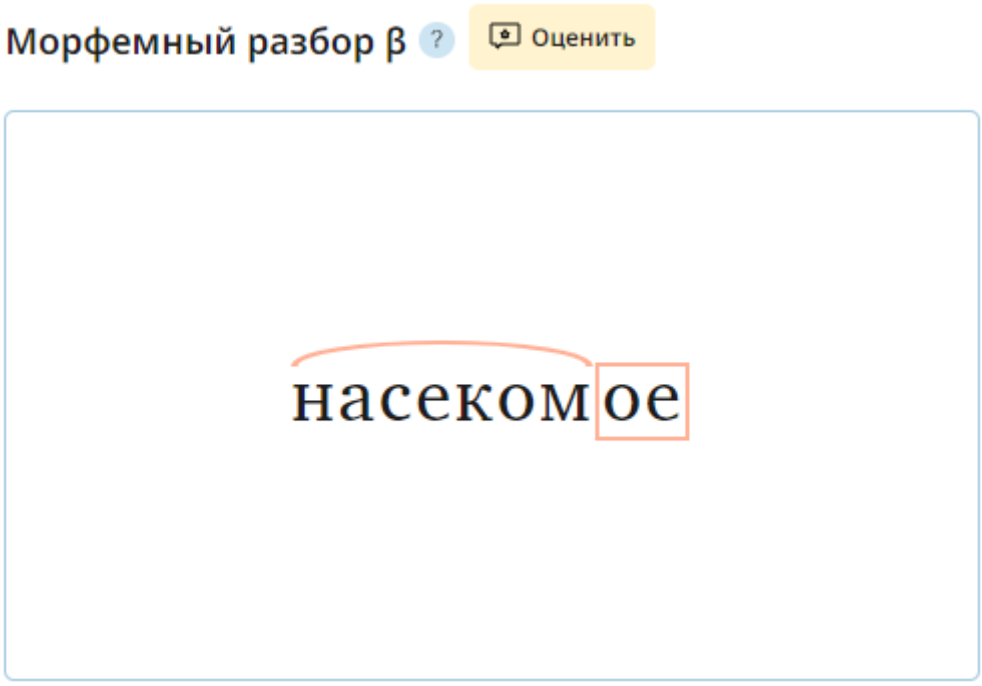


Рис. 11: Обучающий корпус. Морфемный разбор на основе словаря Тихонова

Из описанного выше следуют две основные проблемы словообразовательной разметки в НКРЯ. Во-первых, 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 (с порогом встречаемости хотя бы 3 раза в 3 различных текстах), то есть имеющаяся ручная разметка далека от полноты. Во-вторых, существование различных, противоречащих друг другу подходов к морфемному членению заметно осложняет автоматическое пополнение словарей. При этом нельзя утверждать, что авторы конкретного словаря строго придерживаются единого для этого словаря алгоритма морфемного членения; в частных случаях применяются локальные решения, не удовлетворяющие описанному алгоритму (Иомдин 2019).

Несмотря на описанные проблемы, выборку достаточного размера можно использовать для обучения алгоритмов автоматического морфемного анализа. Как и во многих других областях обработки естественного языка, использование методов машинного обучения может обеспечивать высокое качество результатов. Так, в 2018 году был представлен алгоритм генерации морфемных разборов на базе ансамбля сверточных нейронных сетей (Sorokin, Kravtsova 2018), который показал значительный прирост качества по сравнению с ранее существовавшими алгоритмами построения морфемных разборов. Мы проанализировали качество работы предложенного алгоритма при его обучении на словарях Morphodict-T и Morphodict-K при помощи кросс-валидации по пяти выборкам и пришли к выводу, что модель, обученная на Morphodict-K, показывает значительно более высокие результаты. Особенно различается качество автоматической разметки по доле полностью верных разборов, что может свидетельствовать о более высокой внутренней согласованности данных в словаре Morphodict-K. Все полученные результаты приведены в Таблице 2.

Метрика	Morphodict-T	Morphodict-K
F-мера для границ морфем	98,09	98,66
Точность (precision) для границ морфем	97,79	98,58
Полнота (recall) для границ морфем	98,38	98,74

Метрика	Morphodict-T	Morphodict-K
Доля слов с верно определенными границами морфем (без учета их типа)	96,61	97,40
Доля полностью верных разборов	88,49	90,82

Таблица 2: Сравнение доли верных разборов в Morphodict-T и Morphodict-K

В то же время использованная нами модель не лишена недостатков. Исследование (Garipov, Morozov, Glazkova 2023) показало значительное снижение качества при тестировании на словах, содержащих корни, не встретившиеся в обучающей выборке, что может свидетельствовать о недостаточной обобщающей способности алгоритма. В дальнейшем мы планируем подробнее изучить возможные способы устранения этого недостатка.

В настоящий момент модель, полученная в результате обучения 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 (в интерфейсе корпуса это отображается пометой «сгенерировано НейроКРЯ» на рамке виджета). На сегодняшний день мы собираем отклики пользователей о качестве генерации и готовимся включить доработанную модель в поисковые возможности Основного корпуса. Параллельно с этим мы изучаем возможность добавления автоматических разборов и в Обучающий корпус.

Обе модели размещены в открытом доступе и доступны для выгрузки в разделе «Нейросетевые модели» на сайте НКРЯ.[22]

На Рис. 12 продемонстрирован разбор слова *эстетика* в Основном корпусе, порожденный моделью НейроКРЯ (происхождение разбора указано в рамке виджета).

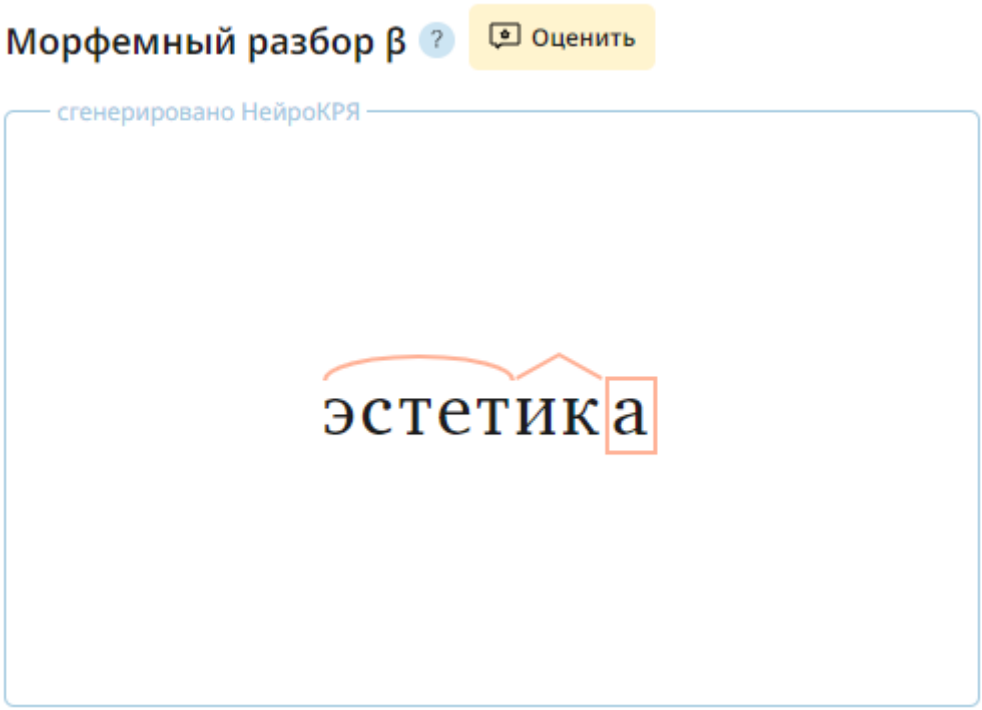


Рис. 12: Основной корпус: морфемный разбор, сгенерированный НейроКРЯ

[22] 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.

Использование для представления слов и текстов многомерных векторов повсеместно встречается в обработке естественного языка. В таких задачах, как классификация текстов, внутритекстовая разметка, генерация текста, конвертация слов в наборы чисел (многомерные вектора) является первым этапом работы с текстом. Наиболее простые алгоритмы получения таких представлений, например, one-hot кодирование, фактически никак не учитывают семантику кодируемого, тогда как современные языковые модели, например, BERT (Devlin et al., 2019), опираются не только на семантику кодируемого слова, но и на конкретный контекст его употребления. Промежуточным звеном между этими подходами являются различные алгоритмы построения статических эмбеддингов, например, CBoW и Skip-gram (Rehurek, Sojka 2011). Эти алгоритмы опираются на дистрибутивную гипотезу: предполагается, что слова, регулярно употребляющиеся в похожих контекстах, имеют схожую семантику. Полученные таким образом векторные представления позволяют, в том числе, оценивать семантическую схожесть между словами через косинусное расстояние между их векторами, что, в свою очередь, может быть использовано для поиска слов-ассоциатов.

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». При построении модели использовался алгоритм CBOW. Статические векторные модели на базе Основного корпуса строились и раньше, например, в работе (Kutuzov, Kuzmenko 2017). Однако в ходе нашей работы для Основного корпуса была обучена модель с использованием лемм, сгенерированных моделью RuBic (подробнее о модели см. раздел 4.1). Это позволило существенно уменьшить количество ошибочно сгенерированных или неправильно токенизированных лемм среди похожих слов. Сравнивая похожие слова для Основного и Старорусского корпусов, можно отслеживать семантические дрейфы (изменения значений слов со временем), а сравнивая Обучающий, корпус «Русская классика» и корпус Центральных СМИ, можно исследовать, например, журналистские штампы или употребление слов в разных типах дискурса. Приведем примеры, иллюстрирующие разницу ассоциатов для слов *игра*, *нива* и *трубить* в разных корпусах (Рис. 13–18):

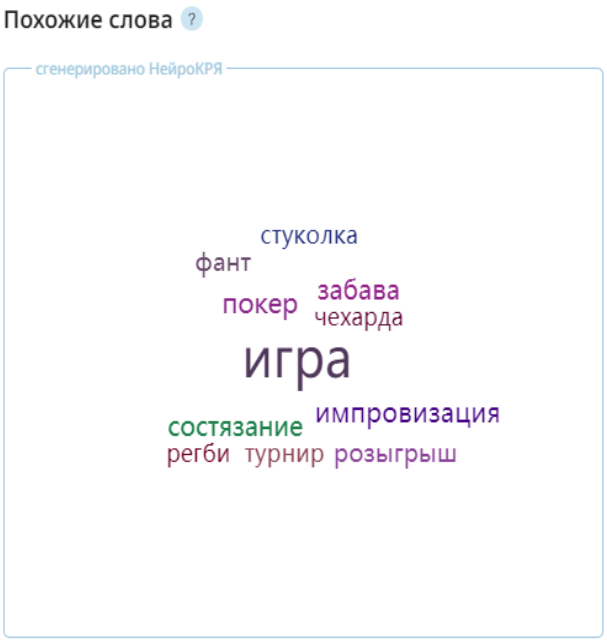


Рис. 13: Похожие слова для слова *игра* в Основном корпусе

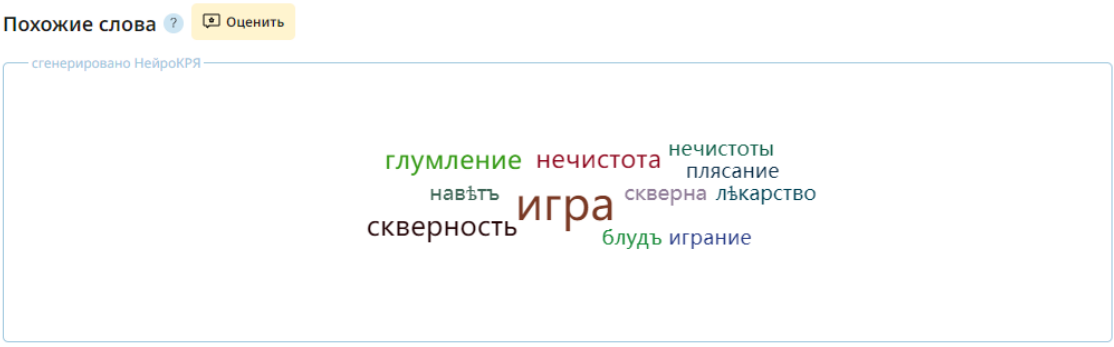


Рис. 14: Похожие слова для слова *игра* в Старорусском корпусе

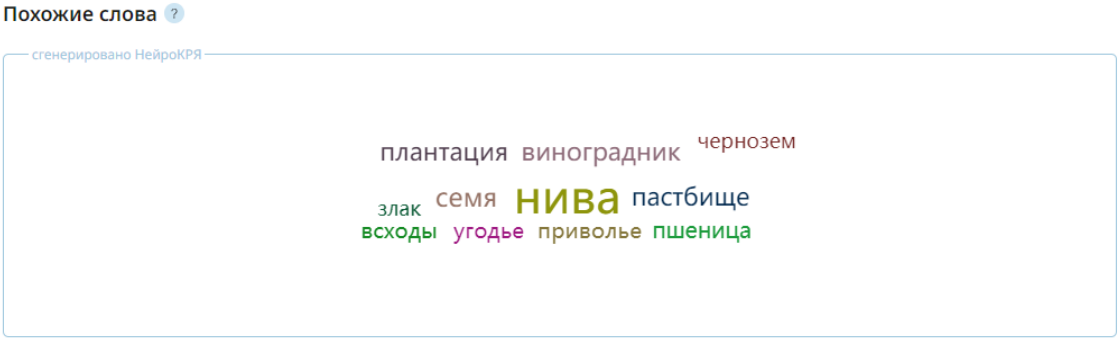


Рис. 15: Похожие слова для слова *нива* в Обучающем корпусе

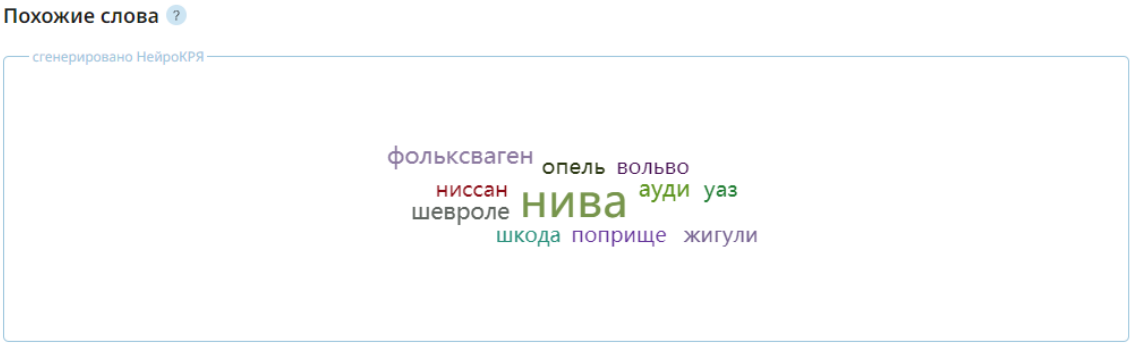


Рис. 16: Похожие слова для слова *нива* в корпусе Центральных СМИ

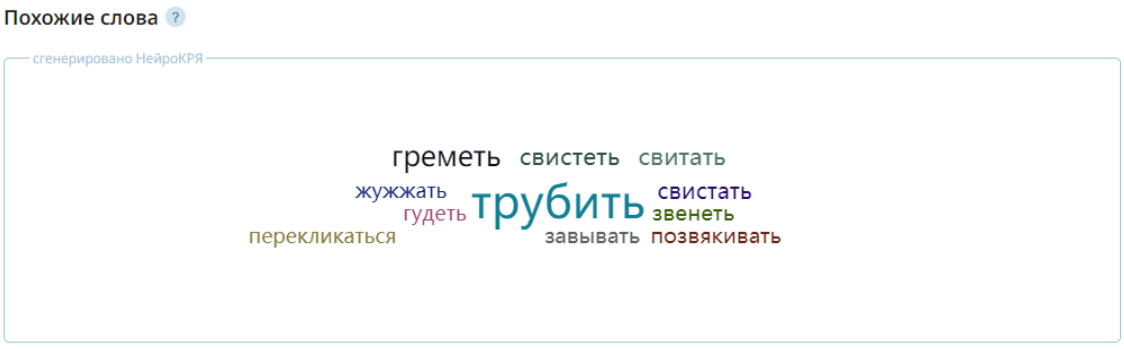


Рис. 17: Похожие слова для слова *трубить* в корпусе «Русская классика»

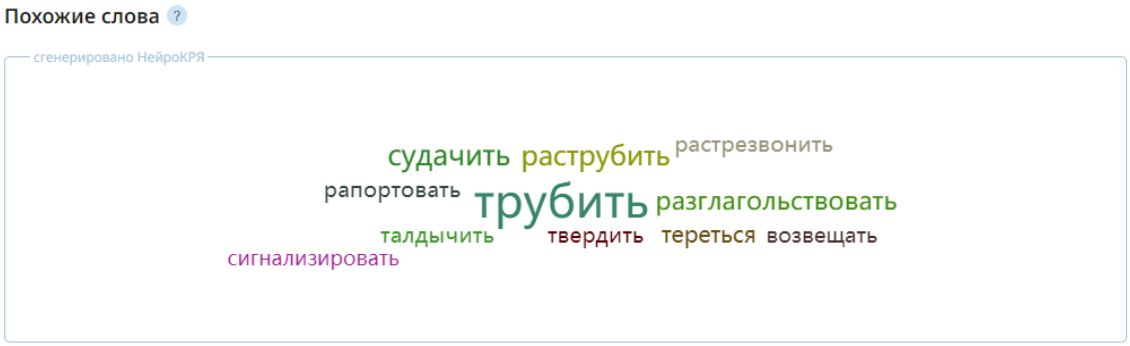


Рис. 18: Похожие слова для слова *трубить* в корпусе Центральных СМИ

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. Модель, обученная на Основном корпусе, показала корреляцию 0.367 (по Спирмену) на выборке RuSimLex365 (Kutuzov, Kunilovskaya 2018) на Корпусе региональных СМИ — 0.227.

Все семь моделей размещены в открытом доступе на странице «Нейросетевый модели» сайта НКРЯ[23] и могут быть использованы для научных и прикладных исследований.

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. 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. Как уже было сказано выше в (1.1.3), такая неопределенность блокировала развития статистических и аналитических сервисов, поскольку омонимичные формы или разборы существенно зашумляют любые подсчеты. Применение интеллектуальных моделей для разметки данных позволяет значительно ускорить включение текстов в корпус при сохранении высокого качества лингвистической разметки. Ниже будут последовательно 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

Задача, которую решает нейросетевая модель Rubic, — это автоматическая разметка текста, а именно: лемматизация, морфологическая характеристика для всех токенов, включая определение части речи, построение дерева синтаксической зависимости предложения. Каждая из этих операций предполагает разрешение омонимии. Таким образом, Rubic представляет собой альтернативу морфологическому анализатору MyStem (Зобнин, Носырев, 2015), ранее применявшемуся для обработки текстов НКРЯ и основанному на грамматическом словаре, и, кроме того, выполняет синтаксическую разметку. При разработке модели Rubic'a 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.

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

4.1.1 Принципы работы Rubic

Архитектура Rubic'a (Lyashevskaya et al., 2023) основана на архитектуре модели qbic, победившей в соревновании по обработке русского языка GramEval2020 (Anastasyev 2020). Используется однослойный LSTM-энкодер, комбинирующий векторизованное представление слов, получаемые из BERT-подобной модели (в текущей реализации, sberbank-ai/ruBert, предобученный на 30 Гб данных) 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. Rubic обучается в мультизадачном режиме, то есть веса классификаторов (а-в) определяются независимо друг от друга.

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 (проза, газетные тексты, поэзия, социальные сети, википедия) общим объемом свыше 2,4 миллиона токенов. Разметка обучающего корпуса текстов проверена вручную. Все данные приводятся в расширенном формате морфологической и синтаксической разметки UD-ext (Lyashevskaya 2019). Такой подход, как уже говорилось выше в (1.1.2), следует глобальной тенденции стандартизации корпусной разметки, и в частности, унификации морфо-синтаксических тегов, обеспечивающую возможность кросс-языковых исследований. Выбранный формат UD-ext ставит в соответствие наборы частеречных тегов и морфологических признаков, используемые при разметке текстов в НКРЯ и в русских корпусах Universal dependencies, см. Таблицу 3.

UPOS	ADJ, ADP, ADV, ADVPRO, ANUM, AUX, CCONJ, COM, DET, INIT, INTJ, NOUN, NUM, PARENTH, PART, PRED, PREDPRO, PRON, PROPN, PUNCT, SCONJ, SYM, VERB, X
FEAT (основные)	Abbr, Animacy, Aspect, Case, Clitic, Degree, Gender, Mood, Number, Person, Tense, Transit, Variant, VerbForm, Voice
FEAT (лексические признаки)	NameType, NumForm, NumType, Poss, Polarity, PronType, Reflex
FEAT (другие дополнительные)	Anom, Hyph, InflClass, Foreign, Typo

Таблица 3: Формат UD-ext

Чтобы улучшить обработку предложений, написанных заглавными буквами, а также с использованием буквы «Е» и разного рода кавычек и отточий, применяется аугментация данных объемом 3200 предложений (40 тысяч словоформ).

Rubic работает с текстовыми данными, представленными в формате CoNLL-U. При подготовке текстов НКРЯ используется отдельная модель токенизации (см. ниже). На этапе предобработки для улучшения качества работы модели токены, состоящие из смеси кириллических букв и других символов (знаков ударения, диакритик и разного «шума»), очищаются специальным модулем. Кроме того, набор правил premodern2modern приводит тексты, представленные в старой орфографии разных периодов, к современной орфографии. Поскольку размеченные данные поступают в формате xml и на индексацию также уходят данные в формате xml, предусмотрены инструменты конвертации форматов xml -> CoNLL-U и обратной конвертации CoNLL-U -> xml.

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 и семантическую разметку в стандарте НКРЯ.

4.1.1.1 Токенизатор

Практически любой анализ текста начинается с его разбиения на фрагменты (токенизации). В рамках задачи снятия морфологической неоднозначности наиболее подходящими размерами фрагментов являются предложения и отдельные слова. В простейшем случае для разбиения текста на предложения можно воспользоваться разбиением по набору знаков препинания (например, точка, вопросительный и восклицательный знаки), а для разбиения на слова — по пробелам. Однако в реальности такой подход не может дать высокого качества как из-за использования знаков препинания в других целях (например, в инициалах), так и из-за опечаток (например, пропущенные пробелы). Так как качество разбиения критически важно для последующего анализа, мы провели ряд экспериментов по выбору оптимальной модели токенизации. Нами были протестированы предобученные алгоритмы udpipe (Straka, Hajic, Straková 2016), razdel[24], spacy[25], nltk (Bird, Loper, Klein 2009), pymorphy2 (Korobov 2015), MyStem[26], rusenttokenize[27], а также их комбинации. Тестирование проводилось на специально подготовленной выборке сложных случаев (выборка GOLD). Лучший результат продемонстрировало сочетание алгоритма razdel для сегментации на предложения и spacy для сегментации на слова. Полные результаты тестирования приведены в Таблице 4.

		Сегментация на предложения					
		razdel	*spacy*	*nltk*	*rusenttokenize*	*udpipe*	*mystem*
Токенизация	*razdel*	F1-s: 0.5562 F1-w: 0.8946	F1-s: 0.4577 F1-w: 0.8946	F1-s: 0.3299 F1-w: 0.8946	F1-s: 0.4545 F1-w: 0.8946	F1-s: 0.4532 F1-w: 0.8946	F1-s: 0.3441 F1-w: 0.8946
	spacy	**F1-s: 0.5562** **F1-w: 0.9272**	F1-s: 0.4577 F1-w: 0.9273	F1-s: 0.3299 F1-w: 0.9271	F1-s: 0.4545 F1-w: 0.9272	F1-s: 0.4132 F1-w: 0.9272	F1-s: 0.3441 F1-w: 0.9272
	nltk	F1-s: 0.5562 F1-w: 0.9060	F1-s: 0.4577 F1-w: 0.9032	F1-s: 0.3299 F1-w: 0.9050	F1-s: 0.4545 F1-w: 0.9041	F1-s: 0.4132 F1-w: 0.9041	F1-s: 0.3441 F1-w: 0.9041
	pymorphy	F1-s: 0.5562 F1-w: 0.8734	F1-s: 0.4577 F1-w: 0.8734	F1-s: 0.3299 F1-w: 0.8734	F1-s: 0.4545 F1-w: 0.8734	F1-s: 0.4132 F1-w: 0.8735	F1-s: 0.3441 F1-w: 0.8735
	udpipe	F1-s: 0.5562 F1-w: 0.8387	F1-s: 0.4577 F1-w: 0.8387	F1-s: 0.3299 F1-w: 0.8387	F1-s: 0.4545 F1-w: 0.8387	F1-s: 0.4132 F1-w: 0.8387	F1-s: 0.3441 F1-w: 0.8387
	MyStem	F1-s: 0.5562 F1-w: 0.8720	F1-s: 0.4577 F1-w: 0.8720	F1-s: 0.3299 F1-w: 0.8720	F1-s: 0.4545 F1-w: 0.8720	F1-s: 0.4132 F1-w: 0.8720	F1-s: 0.3441 F1-w: 0.8720

Таблица 4. Результаты тестирования различных алгоритмов сегментации и токенизации по двум метрикам: F1-s (по предложениям) и F1-w (по отдельным словам), аналогичным использованным на соревновании CoNLL 2018[28]

[24] https://github.com/natasha/razdel
[25] https://github.com/explosion/spaCy
[26] https://yandex.ru/dev/mystem/
[27] https://github.com/deeppavlov/ru_sentence_tokenizer
[28] https://universaldependencies.org/conll18/

Анализ допускаемых алгоритмом ошибок и создание специальных эвристических алгоритмов постобработки позволило достичь качества 0.95 по метрике F1. При этом не удалось обнаружить эвристик, которые бы позволили значительно улучшить результат сегментации на предложения (F1=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), разработанная Stanford NLP Group на базе рекуррентной нейронной сети LSTM, показавшая хорошие результаты в аналогичной задаче для английского языка. Эта модель является двухслойной. Первый слой состоит из одномерной сверточной нейронной сети и слоя BiLSTM. Выходные данные CNN объединяются со скрытыми состояниями BiLSTM и передаются на второй слой BiLSTM. В качестве функции потерь используется кросс-энтропия. Мы обучили модель со стандартными параметрами обучения на открытых датасетах Тайга (Shavrina, Shapovalova 2017) и СинТагРус[29], 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 тестовой выборке продемонстрировала схожее качество по отдельным словам 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 предобученных алгоритмов, достигнув метрики F1=0.95 по предложениям и 0.99 по отдельным словам (Таблица 5).

	razdel + spacy + эвристики	stanza
Выборка GOLD	F1-s: 0.556 F1-w: 0.952	F1-s: 0.637 F1-w: 0.944
Выборка TEST	F1-s: 0.943 F1-w: 0.992	F1-s: 0.956 F1-w: 0.996

Таблица 5. Сравнение работы комбинированного подхода (сегментация на предложения при помощи razdel, токенизация при помощи spacy и дополнительные эвристики постобработки) и обученной на данных корпуса модели Stanza. Сравнение проводится по двум метрикам: F1-s (по предложениям) и F1-w (по отдельным словам).

Модель токенизатора доступна для скачивания и размещена на странице «Нейросетевые модели» сайта НКРЯ.[30]

4.1.1.2 Классификатор морфологических признаков

После сегментации текста на предложения, а предложений — на слова, начинается этап приписывания 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» (ср. разбор формы *кошка*). 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 далее 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

[29] https://ruscorpora.ru/page/corpora-datasets/
[30] https://ruscorpora.ru/license-content/neuromodels

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 (ср. наблюдение «о менее 1000 классов лемматизации» в работе Michurina et al., 2021). Модуль лемматизации получает на вход словоформу с меткой ее части речи, определенной морфологическим модулем. Для улучшения качества лемматизации предусмотрена возможность сравнения наиболее вероятных гипотез лемм с данными словаря, составленного вручную по данным существующих корпусов с ручной разметкой и словарей.

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 6.

	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 6: Результаты работы Rubic’a на тестовом множестве

Модель хорошо справляется с определением частей речи, морфологическим разбором некоторых грамматических категорий, таких как сравнительная степень, переходность, вид. При лемматизации высокое качество обработки модель демонстрирует 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).

Критическим образом на качество анализа влияют ошибки в токенизации исходных данных, попадающих на вход Rubic’a.

Как правило, ошибки при автоматической разметке одновременно наблюдаются и в лемматизации, и в морфологических признаках, и в дереве синтаксических зависимостей. Это может быть связано со структурой предложения, например, из-за недостаточности контекста для одиночных слов в клаузе (ср., например, предложение прямой речи «- *Бушую».*), при наличии оборванных словоформ или перед многоточиями. Анализатор сталкивается с трудностями при определении части речи, морфологических и синтаксических признаков на границе предложений. Например, в начале предложения правильно определяется категория существительного, но возникает

ошибка в недостаточном/избыточном определении классов имен собственных (*Надежда, Воробьиха*). В той же позиции начала предложения может неправильно определиться категория одушевленности, что влечет за собой также ошибку и в синтаксическом разборе:

Белены объелись или выпили лишнее!
белена NOUN _ Animacy=Anim|Case=Nom|Gender=Masc|Number=Plur 2 nsubj

При морфологическом анализе ошибки возникают из-за наличия в текстах авторских искажений и скандирования (*рр-рр-а-а-аз, стооооой* и т.п.), аномальных, но при этом частотных форм (*испужамшись*) слов с нестандартной капитализацией (*варЮ*), а также содержащих кавычки и букву «Е».

Синтаксический анализатор достаточно уверенно определяет зависимости в пределах клаузы (в именных группах, группах глагола, предикативов и наречий), между прямой речью и клаузой, сопровождающей прямую речь, между клаузами в сложносочиненных предложениях. Ошибки возникают при определении вершины в безглагольных клаузах, в клаузах с эллипсисом вершины, в определении обращений и дискурсивных элементов, не выделенных пунктуационно. Наблюдаются ошибки в том случае, когда две клаузы связаны отношением сочинения VS. паратаксиса («*Упал, упал человек, тонет!*» — *упал, упал* (бессоюзное сочинение глаголов, имеющих общий субъект)) должно быть связано как conj, но Rubic размечает как parataxis (при этом *тонет* правильно размечается как conj)).

С точки зрения зависимости качества работы модели от жанра текста можно сказать, что несколько большую трудность вызывают тексты соцсетей, энциклопедические, математические тексты, поэзия. Очевидной причиной ошибок являются нестандартные синтаксические конструкции и пунктуация этих текстов, эллипсис и оборванные высказывания, редкие имена собственные, использование в данных жанрах предложений большой длины. Кроме того, можно заметить, что при обработке таблиц и списков литературы, транслируемых токенизатором в одно предложение, синтаксический модуль пытается трактовать отношения между словоформами как предикативные и именные зависимости, а не как «плоские» отношения типа списков.

В целом, анализ качества работы модели показывает, что самым слабым звеном среди модулей является лемматизатор. Часто ошибочно лемматизируются короткие слова с выпадающим гласным (*мох, пес* и т.п.), существительные с чередованием в конце слова (*котенок — котята* и т.п.), глаголы с чередованиями в корне (*жать, плыть, выть, петь, скрести* и т.п.), аббревиатуры (г. — *город/господин/грамм*), степени сравнения прилагательных и наречий на «по-» (*потише — тихий, тио, тико)*. Для улучшения качества лемматизации корпуса на этапе постобработки используются эвристики для замены ошибочных форм по спискам, составленным вручную для форм разных частей речи (около 50 тысяч правил).

4.1.3 Перспективы

Rubic планируется улучшить модулем, позволяющим делать альтернативные (правильные) морфологические разборы, включая часть речи и лемму. Для этого прежде всего необходимо разработать формат добавления альтернативных разборов в обучающие данные, а именно — изменения и дополнения в формат CoNLL-U, который используется Rubic'ом. Отметим, что некоторые возможности представления альтернативных разборов уже есть в текущей версии обучающих данных для Rubic'a. Так, уже разработан формат представления альтернативных разборов, таких как разбор формы *расположенный* как причастия и прилагательного. Кроме того, леммы, не соответствующие литературному варианту русского языка (*удилом* (ед. ч. от sg. tt.), *шеколад* — дневники А.К. Гладкова) могут быть связаны с литературным вариантом леммы через вспомогательные таблицы, аналогичные тем, которые используются в панхроническом поиске (подробнее о принципах работы панхронического поиска см. в статье (Савчук и др., 2024).

Как уже говорилось выше в (4.1.2), качество работы лемматизатора, используемого в Rubic'e, требует дальнейших улучшений. Для повышения эффективности работы нейросети можно было бы предложить несколько другой алгоритм лемматизации. В настоящий момент распределение ответов лемматизатора следует, условно, закону Парето: доля правильных вариантов составляет

от 50 до 95%, остальное — неправильные ответы разного рода, убывающие по частоте. Улучшенная модель лемматизации могла бы быть более чувствительна к словоформам непродуктивных парадигм, а также могла бы учитывать не только длину, но и буквенный состав «псевдоокончаний», лучше разрешать омонимию слов с пересечением парадигм (*белка*/*белок*, *стрелка*/*стрелок*). Можно также предложить выделить в отдельный модуль обработку аббревиатур. Это связано с тем, что у аббревиатур другое распределение операций лемматизации, но при этом их определение все же чувствительно к контексту. Кроме того, использование аббревиатур при обучении ухудшает качество лемматизации обычных слов.

Возможно дополнение структуры Rubic'a алгоритмом обработки искаженных и просторечных форм (*гавагите, г'азог'ву, бегат, стоооой, по-до-жди-и-и-и* и т.п.), которые часто встречаются в корпусных данных, подлежащих разметке.

Качество работы Rubic'a может быть повышено за счет целевого пополнения всех обучающих текстовых множеств для охвата языковых явлений и жанрового разнообразия в объеме, достаточном для обучения. Кроме пополнения обучающих данных, необходима и чистка обучающего корпуса, исправление опечаток, ошибок разметки, расстановки знаков препинания. В частности, это актуально для разрешения омонимии «*причастие-прилагательное*», «*предикатив-наречие-прилагательное*». Также при работе по улучшению обучающих данных необходимо уделить внимание редким синтаксическим конструкциям и длинным предложениям, при автоматическом разборе которых Rubic часто допускает ошибки.

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.

Особенность разметки жанров в случае корпуса «Социальные сети» заключается в том, что на начальном этапе в корпусе отсутствовала ручная разметка жанров. Поэтому первым этапом работы стала экспертная оценка выборки текстов и формирование списка широко представленных жанров. Для этого была сформирована выборка, содержащая около трех тысяч текстов, выбранных случайным образом. В результате экспертной оценки были выделены 13 жанров, наиболее широко представленных в выборке. К ним были добавлены жанры «Биография», «Гороскоп» и «Интернет-рейтинг» как содержащие довольно характерные тексты. Тексты прочих жанров были объединены в класс «Неопределенная категория». Список жанров и распределение текстов в выборке по результатам оценки экспертов представлены в Таблице 7 в столбцах «Жанр» и «Количество текстов, размеченных экспертами» соответственно.

Распределение текстов по жанрам по результатам экспертной оценки случайной выборки текстов получилось неравномерным. К наиболее широко представленным классам (жанрам) — «Анонс | объявление», «Неопределенная категория» и «Информационное сообщение» — относятся 1019, 671 и 343 текста соответственно, то есть больше двух третей от объема рассмотренной выборки. Неравномерное распределение жанров затрудняет обучение моделей для автоматической разметки категорий текстов, поэтому с целью выравнивания количества примеров в классах в набор обучающих данных для ряда жанров были добавлены дополнительные тексты. Дополнительные тексты были преимущественно получены из текстов Основного и Регионального корпусов. Для жанра «Интернет-рейтинг» были использованы тексты корпуса «Социальные сети», содержащие характерные для данного жанра фразы. Итоговое распределение текстов в наборе данных для обучения модели автоматической разметки жанров представлено в столбце «Всего текстов» в Таблице 7.

Жанр	Количество текстов, размеченных экспертами	Количество дополнительных текстов	Всего текстов	Источник дополнительных текстов
Анекдот	78	153	231	Основной корпус
Анонс | объявление	1019	-	1019	-
Биография	2	454	456	Основной корпус, Региональный корпус
Вопрос	36	-	36	-
Гороскоп	8	204	212	Основной корпус, Региональный корпус
Инструкция | совет | рекомендация	81	-	81	-
Интернет-рейтинг	8	253	261	Тексты корпуса «Социальные сети», подобранные по ключевым словам
Информационное сообщение	343	-	343	-
История	44	192	236	Тексты, вручную подобранные экспертами
Неопределенная категория	671	-	671	-
Отзыв | рецензия	215	-	215	-
Оценка	34	150	184	
Поздравление	45	674	719	Основной корпус, Региональный корпус
Поэзия	145	-	145	-
Прецедентный текст	119	-	119	-
Рецепт	93	223	316	Основной корпус, Региональный корпус
Итого	2941	2303	5244	

Таблица 7: Набор данных для разметки жанров в корпусе «Социальные сети»

Для разметки жанров была выбрана предварительно обученная модель RuRoBERTa [31] (Zmitrovich et al., 2023), повторяющая архитектуру модели RoBERTa для англоязычных текстов (Liu et al., 2019). Модель использует токенизацию по принципу Byte-level BPE (Gage 1994). Для предварительного обучения были использованы тексты Википедии, а также коллекции новостных и художественных текстов, текстов веб-ресурсов и русскоязычных субтитров. RuRoBERTa показывает высокие результаты в задачах классификации текстов на русском языке (в частности, в рамках последних соревнований Dialogue Evaluation (Golubev, Rusnachenko, Loukachevitch 2023). Для автоматической разметки жанров модель была дообучена на собранном наборе данных 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 десятикратной перекрестной проверки (кроссвалидации для десяти фолдов). Перекрестная проверка выполнялась следующим образом. Выборка текстов корпуса «Социальные сети», размеченная экспертами, была десять раз разбита на обучающую и тестовую подвыборки по принципу скользящего окна. К обучающей подвыборке были добавлены дополнительные тексты, после чего на дополненной обучающей подвыборке выполнялось дообучение модели с использованием параметров, указанных выше. Тестирование модели выполнялось на тестовой выборке. Таким образом, в результате перекрестной проверки были получены десять значений показателей качества модели на контрольных подмножествах данных. Итоговая оценка качества представляет собой среднее арифметическое этих значений. 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, дообученных на наборе данных, состоящем из текстов, размеченных экспертами, и дополнительных текстов. Объединение предсказаний моделей в ансамбле осуществлялось по принципу усреднения предсказанных вероятностей жанров (soft voting). На этапе постобработки предсказаний модели все сверхкороткие тексты были перенесены в класс «Неопределенная категория». Пороговое значение, являющееся критерием для определения коротких текстов, составляет 40 токенов. Такое значение было получено в результате эмпирического анализа текстов корпуса. Количество токенов определяется с помощью токенизатора модели RuBERT (Kuratov 2019).

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*).

Извлечение ключевых слов из текстов Региональных СМИ выполнено с помощью библиотеки RuTermExtract[32]. Алгоритм, лежащий в основе RuTermExtract, представляет собой адаптированную для русского языка версию алгоритма TermExtract[33]. 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).

На этапе предобработки текстов биграммы, заключенные в кавычки, были объединены символом «_», чтобы алгоритм рассматривал их как униграмму (например, *«Комсомольская правда»* -> *«Комсомольская_правда»*) и не разделял на слова. Предобработанные тексты подавались на вход алгоритму RuTermExtract со следующими параметрами: максимальное количество извлеченных ключевых слов — 20; параметр nested, позволяющий извлекать ключевые слова, лежащие

[31] https://huggingface.co/ai-forever/ruRoberta-large
[32] https://github.com/igor-shevchenko/rutermextract
[33] https://pypi.org/project/topia.termextract

внутри других ключевых слов, — True. С помощью алгоритма для текстов были получены первичные списки ключевых слов в нижнем регистре. К полученным ключевым словам применялась несколько шагов постобработки:

1. замена символа «_», добавленного в биграммы на этапе предобработки, пробелом;
2. удаление ключевых слов, состоящих из трех и более слов;
3. удаление полных и кратких имен в соответствии со списком личных имен;
4. удаление однокоренных униграмм с помощью модели Morphodict-K (см. Раздел (3.2.));
5. проведение нормализации словосочетаний на основе списка правил с помощью библиотеки PyMorphy2 (Korobov 2015);
6. обработка некоторых распространенных ошибок (например, такой ошибкой является постановка второго слова в форму генетива в некоторых именованных сущностях: «*юрий лужкова*» -> «*юрий лужков*»);
7. удаление ключевых слов таким образом, чтобы длина списка составляла не более 15 ключевых слов.

5 Заключение

В статье представлено описание обновленной платформы НКРЯ с технологической точки зрения. Это обновление является важнейшим этапом 20-летнего развития Национального корпуса русского языка. Следует подчеркнуть, что речь идет не об отдельных нововведениях, а о внедрении комплексного подхода, основанного на идеологических принципах, соответствующих современным практикам и стандартам развития корпусных ресурсов, а также общим тенденциям цифрового развития общества. Эти принципы находят свое выражение в переходе к модульной и гибкой архитектуре корпусного ядра и веб-интерфейса, открытой для дальнейших изменений и масштабного пополнения корпусов; в разработке сервисов для анализа данных, которые позволяют исследователям переходить от ручного анализа примеров к количественному анализу, основанному на статистическом обобщении распределения лексических единиц; в интеграции технологий искусственного интеллекта в процесс подготовки корпусных данных; в создании собственных нейросетевых моделей и их размещении в открытом доступе; и, наконец, в ориентации на привлечение более широкой аудитории к работе с Национальным корпусом русского языка, расширение возможностей применения корпусных данных не только в лингвистических исследованиях, но и в педагогике, а также в качестве источника языковых данных для самых разных областей гуманитарного знания.

Литература

[1] Баранов А. Н. (2023). Инструментарий лингвистики в лингвистической экспертизе: корпусные технологии // Язык. Право. Общество. С. 54-58.
[2] Гришина Е. А. и др. (2009). О задачах и методах словообразовательной разметки в корпусе текстов // Полярный вестник (Тромсё), 2009, № 12. С. 5–25.
[3] 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.
[4] Иомдин Б. Л. Как определять однокоренные слова? // Русская речь. 2019. No 1. С. 109–115.
[5] Кузнецова А. И., Ефремова Т. Ф. (1986). Словарь морфем русского языка. Москва: Рус. яз., 1986.
[6] Савчук С. О. (2015). Корпус региональных газет России и зарубежья // Труды Института русского языка им. В. В. Виноградова. 2015. № 6. С. 163—193.
[7] Савчук С. О. и др. (2024). Национальный корпус русского языка 2.0: новые возможности и перспективы развития //Вопросы языкознания. Т. 2. С. 7-34.
[8] Сичинава Д. В. (2005). Национальный корпус русского языка: очерк предыстории // Национальный корпус русского языка: 2003—2005. М.: Индрик. С. 21—30.
[9] Сичинава Д. В. (2022). Корпус берестяных грамот как параллельный // Труды Института русского языка им. В. В. Виноградова. 2022. № 2 (32), 92-106.
[10] Тихонов А. Н. (2002). Морфемно-орфографический словарь. Москва: Астрель: АСТ, 2002.
[11] Aksan Y. et al. (2012). Construction of the Turkish National Corpus (TNC) // LREC. 2012. P. 3223-3227.
[12] 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.

[13] Beeby A., Rodríguez Inés P. and Sánchez-Gijón P. (eds). (2009). Corpus Use and Translating. Corpus Use for Learning to Translate and Learning Corpus Use to Translate, Amsterdam: Benjamins.
[14] Biber D. (1993). Representativeness in corpus design // Literary and linguistic computing. Vol. 8. № 4. P. 243–257.
[15] Bird, St., Loper E., Klein E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.
[16] Boulton A. (2011). Data-driven learning: the perpetual enigma // S. Goźdź-Roszkowski. Explorations across Languages and Corpora, Peter Lang. P. 563-580.
[17] Bowker L. (2018). Corpus linguistics is not just for linguists: Considering the potential of computer-based corpus methods for library and information science research // Library Hi Tech, 36(2). P. 358-371.
[18] Buchholz, S., Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. // Proceedings of the tenth conference on computational natural language learning (CoNLL-X). P. 149-164.
[19] Calzolari, N., McNaught, J., & Zampolli, A. (1996). EAGLES Final Report: EAGLES Editors' Introduction. Pisa, Italy, EAG-EB-EI.
[20] Chartrand, L. (2022). Modeling and corpus methods in experimental philosophy. Philosophy Compass, 17(6).
[21] Chen K. J. et al. (1996). Sinica corpus: Design methodology for balanced corpora //Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation. P. 167-176.
[22] Coulthard, M. (1994). On the Use of Corpora in the Analysis of Forensic Texts', Forensic Linguistics: International Journal of Speech, Language and the Law 1(1). P. 27–43.
[23] Coulthard, M., Johnson, A. and Wright, D. (2017) An Introduction to Forensic Linguistics: Language in Evidence, London: Routledge.
[24] Curtotti M., Mccreath E. (2010). Corpus based classification of text in Australian contracts // Proceedings of the Australasian Language Technology Association Workshop.
[25] Davies, M. (2021). The coronavirus corpus: Design, construction, and use // International journal of corpus linguistics, 26(4). P. 583-598.
[26] De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. Computational linguistics, 47(2). P. 255-308.
[27] Devlin J. et al. (2019). Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019. P. 4171—4186.
[28] Doval, I., Sánchez Nieto, M. T. (2019) Parallel Corpora for Contrastive and Translation Studies. New Resources and Applications. Amsterdam: John Benjamins
[29] Dozat T., Manning Ch. D. (2016). Deep Biaffine Attention for Neural Dependency Parsing https://arxiv.org/abs/1611.01734
[30] 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.
[31] 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.
[32] Evert, S. (2008). Corpora and collocations // A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58, Mouton de Gruyter, Berlin. P. 1212-1248.
[33] Firth, J. R. (1951/1957): Modes of meaning. In: Papers in Linguistics, 1934-1951. Oxford: Oxford University Press.
[34] Francis, W. N., Kučera H. (1982). Frequency Analysis of English Usage: Lexicon and Grammar. Boston: Houghton Mifflin.
[35] Gage F. (1994). A new algorithm for data compression. C Users J. 12, 2. P. 23–38.
[36] 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.
[37] Geyken A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century // Collocations and idioms: Linguistic, lexicographic, and computational aspects. T. 23. P. 41.
[38] Golubev, A., Rusnachenko, N., Loukachevitch, N.V. (2023). RuSentNE-2023: Evaluating Entity-Oriented Sentiment Analysis on Russian News Texts. ArXiv, abs/2305.17679.
[39] Heffer, C. (2005) The Language of Jury Trial: A Corpus-Aided Analysis of Legal-Lay Discourse, Basingstoke: Palgrave.
[40] Ide, N. (1998). Corpus enconding standard: SGML guidelines for encoding linguistic corpora // LREC. P. 463-470.

[41] Ide, N. et al. (2017). Community standards for linguistically-annotated resources // Handbook of linguistic annotation. P. 113-165.
[42] Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. arXiv preprint arXiv:1605.05894.
[43] Johns, T. & P. King (Eds.), (1991), Classroom Concordancing // English Language Research Journal, 4.
[44] Kiyong L., Laurent R. (2010). Towards Interoperability of ISO Standards for Language Resource Management // ICGL 2010. Hong Kong, Hong Kong SAR China. 9p.
[45] Kopotev, M., et al. (2015). Online extraction of Russian multiword expressions // The 5th workshop on balto-slavic natural language processing. P. 43-45.
[46] Korobov M. (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. P. 320-332.
[47] Kuratov, Y. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language / Y. Kuratov, M. Arkhipov // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, Moscow. Vol. 18. Moscow, 2019. P. 333-339.
[48] 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.
[49] 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.
[50] Leech, G. (1993). Corpus annotation schemes // Literary and linguistic computing, 8(4), P. 275-281.
[51] Liu, Y., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.
[52] 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.
[53] 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.
[54] Lyashevskaya O. N. et al. (2020). GRAMEVAL 2020 Shared Task: Russian Full Morphology and Universal Dependencies Parsing. Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference "Dialogue" (2020) Issue 19, P. 553-569.
[55] Machálek, T (2020a). Word at a Glance: Modular Word Profile Aggregator // Proceedings of LREC 2020. P. 7011–7016.
[56] McCarthy, M. (2008). Accessing and interpreting corpus information in the teacher education context // Language Teaching, 41 (4). P. 563-574.
[57] McEnery, T., Hardie, A. (2012) Corpus Linguistics: Method, theory and practice. Cambridge: Cambridge University Press.
[58] McEnery, T., Wilson, A. (2001) Corpus Linguistics. An Introduction. Edinburgh: Edinburgh University Press.
[59] Morozov, D. A., Glazkova A. V., Iomdin B. L. (2022). Text complexity and linguistic features: Their correlation in English and Russian // Russian Journal of Linguistics 26 (2). P. 426–448.
[60] Newman, J., & Cox, C. (2021). Corpus annotation //A practical handbook of corpus linguistics. — Cham : Springer International Publishing, 2021. — C. 25-48.
[61] Nivre, J. et. Al. (2016). Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). P. 1659-1666.
[62] Poletto, F. et al. (2021). Resources and benchmark corpora for hate speech detection: a systematic review // Lang Resources & Evaluation 55. P. 477–523
[63] Rehurek, R., Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2)).
[64] Reppen R. (2021). Building a corpus: what are key considerations? // O'Keeffe A., McCarthy M. (ed.). The Routledge handbook of corpus linguistics. Routledge, 2021. P. 13-20
[65] Roll U., Correia R. A., Berger-Tal O. (2018). Using machine learning to disentangle homonyms in large text corpora //Conservation Biology. V. 32. №. 3. P. 716-724.
[66] Schrauwen S. (2010). Machine learning approaches to sentiment analysis using the Dutch Netlog Corpus // Computational Linguistics and Psycholinguistics Research Center. P. 30-34.
[67] 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.
[68] Shawar B. A., Atwell E. S. (2005). Using corpora in machine-learning chatbot systems //International journal of corpus linguistics. V. 10. №. 4. P. 489-516.

[69] 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.
[70] Stefanowitsch, A. (2020). Corpus linguistics: A guide to the methodology. Berlin: Language Science Press, 2020.
[71] Straka, M., Hajic, J., & Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). P. 4290-4297.
[72] Wiedemann, G. (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences // Historical Social Research/Historische Sozialforschung. P. 332-357.
[73] Wray, A. (2013). Formulaic language // Language Teaching, 46. P. 316–334.
[74] Yang C., Lin K. H. Y., Chen H. H. (2007). Emotion classification using web blog corpora //IEEE/WIC/ACM International Conference on Web Intelligence (WI'07). IEEE, 2007. P. 275-278.
[75] Zanettin, F. (2013). Corpus methods for descriptive translation studies // Procedia-Social and Behavioral Sciences, 95. P. 20-32.
[76] Zmitrovich, D., et al. (2023). A Family of Pretrained Transformer Language Models for Russian. ArXiv, abs/2309.10931.