2024. № 2 ВОПРОСЫ ЯЗЫКОЗНАНИЯ С./Pp. 7–34
Voprosy Jazykoznanija

Национальный корпус русского языка 2.0: новые возможности и перспективы развития

Светлана Олеговна Савчук
Институт русского языка им. В. В. Виноградова РАН, Москва, Россия; savsvetlana@mail.ru

Тимофей Александрович Архангельский
Университет Гамбурга, Гамбург, Германия; timarkh@gmail.com

Анастасия Александровна Бонч-Осмоловская
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия;
Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия;
abonch@gmail.com

Ольга Валерьевна Донина
Воронежский государственный университет, Воронеж, Россия; olga-donina@mail.ru

Юлия Николаевна Кузнецова
Московский государственный университет имени М. В. Ломоносова, Москва, Россия;
Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия;
kuznetsova.yn@gmail.com

Ольга Николаевна Ляшевская
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия;
Институт русского языка им. В. В. Виноградова РАН, Москва, Россия; olesar@yandex.ru

Борис Валерьевич Орехов
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия;
nevmenandr@gmail.com

Мария Владимировна Подрядчикова
независимый исследователь; mpodr2015@gmail.com

Аннотация: В статье подводятся итоги проекта фундаментальной реконструкции и модернизации платформы Национального корпуса русского языка, осуществленного в 2020–2023 гг. В фокусе статьи новые возможности, которые открываются для лингвистов и более широкой аудитории, в частности, улучшение репрезентативности имеющихся корпусов, создание новых корпусов, новая разметка, полученная с помощью применения нейросетевых моделей, новые интерфейсные решения. Более детально рассматриваются три ярких новых компонента: ресурсный — новый корпус «Социальные сети», поисковый — Панхронический корпус, объединяющий поиск по корпусам разных периодов, и аналитический — функциональный комплекс статистики и визуализации данных.

Ключевые слова: квантитативная лингвистика, корпусная лингвистика, русский язык

Благодарности: Исследование проводилось в рамках работ, поддержанных грантом Министерства науки и высшего образования № 075-15-2020-793. Авторы выражают свою признательность за ценную помощь и плодотворное сотрудничество Д. В. Сичинаве, А. Н. Дышканту, С. Ю. Толдовой, Н. С. Горбунову, Д. А. Фурсиной, А. А. Маховой, С. В. Пискуновой, Н. Н. Буйловой,

Д. Г. Бородиной, А. Д. Козеренко, И. И. Виноградовой, С. А. Гладилину, Д. А. Морозову, В. Г. Сизову, П. В. Дяченко, А. О. Казенникову, Н. А. Власовой, А. В. Глазковой, С. С. Столярову, Т. А. Гарипову, И. А. Смалю.

Для цитирования: Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Донина О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития. *Вопросы языкознания*, 2024, 2: 7–34.
DOI: 10.31857/0373-658X.2024.2.7-34

Russian National Corpus 2.0: New opportunities and development prospects

Svetlana O. Savchuk
Vinogradov Russian Language Institute, Russian Academy of Sciences, Moscow, Russia; savsvetlana@mail.ru

Timofey Arkhangelskiy
Hamburg University, Hamburg, Germany; timarkh@gmail.com

Anastasiya A. Bonch-Osmolovskaya
HSE University, Moscow, Russia; Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia; abonch@gmail.com

Ol'ga V. Donina
Voronezh State University, Voronezh, Russia; olga-donina@mail.ru

Yuliya N. Kuznetsova
Lomonosov Moscow State University, Moscow, Russia; Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia; kuznetsova.yn@gmail.com

Ol'ga N. Lyashevskaya
HSE University, Moscow, Russia; Vinogradov Russian Language Institute, Russian Academy of Sciences, Moscow, Russia; olesar@yandex.ru

Boris V. Orekhov
HSE University, Moscow, Russia; nevmenandr@gmail.com

Mariya V. Podryadchikova
independent researcher; mpodr2015@gmail.com

Abstract: The paper provides an overview of the results of the fundamental reconstruction and modernization project of the National Corpus of the Russian Language platform, carried out from 2020 to 2023. The focus of the paper is on the new opportunities that are opening up for linguists and a wider audience. This includes improving the representativeness of existing corpora, creating new corpora, new annotation obtained through the application of neural network models, and new interface solutions. Three notable new components are examined in more detail: a resource-related one, which is the new Social Networks corpus, a search-related one, which is the Panchronic corpus that combines searches across corpora from different periods, and an analytical one, which is the functional complex of statistics and data visualization.
Keywords: corpus linguistics, quantitative linguistics, Russian

Acknowledgements: The current research was a part of research work supported by the Ministry of Science and Higher Education grant No. 075-15-2020-793. The authors express their gratitude for the valuable assistance and fruitful cooperation to D. V. Sichinava, A. N. Dyshkant, S. Y. Toldova, N. S. Gorbunov, D. A. Fursina, A. A. Makhova, S. V. Piskunova, N. N. Builova, D. G. Borodina, A. D. Kozerenko, I. I. Vinogradova, S. A. Gladilin, D. A. Morozov, V. G. Sizov, P. V. Dyachenko, A. O. Kazennikov, N. A. Vlasova, A. V. Glazkova, S. S. Stolyarov, T. A. Garipov, I. A. Smal'.

For citation: Savchuk S. O., Arkhangelskiy T., Bonch-Osmolovskaya A. A., Donina O. V., Kuznetsova Yu. N., Lyashevskaya O. N., Orekhov B. V., Podryadchikova M. V. Russian National Corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, 2024, 2: 7–34.

DOI: 10.31857/0373-658X.2024.2.7-34

Введение

Национальный корпус русского языка был открыт для публичного использования в 2004 г. [Сичинава 2005]. За прошедшее время корпус увеличился многократно и по своим объемам, и по разнообразию представленных данных и типов аннотации, однако к 2020 г. стало понятно, что дальнейшее развитие Национального корпуса тормозится из-за накопившихся нерешенных технологических проблем: отсутствия универсальной платформы для развертывания лингвистических корпусов, медленной индексации и нерегулярных обновлений, ограниченности поисковых возможностей, негибкого, плохо адаптируемого интерфейса взаимодействия с корпусной базой. Кроме технологических сложностей, была еще и концептуальная сложность, заключавшаяся в изолированном существовании разных корпусов в составе НКРЯ, из-за этого процессы подготовки и загрузки данных требовали дополнительных усилий программистов. Накопилось и «содержательное» отставание, проявлявшееся в недостаточной репрезентативности диахронических данных основного корпуса, и явная недостаточность статистического функционала относительно современных стандартов корпусной лингвистики. Иными словами, в результате более чем 15 лет своего плодотворного развития, Национальный корпус русского языка оказался в той точке, когда критическим образом стали необходимы глубокие системные и концептуальные решения о том, как должна быть в дальнейшем организована экосистема корпуса, которая бы обеспечивала возможности по его поддержке, обновлению и внедрению современных корпусных технологий. В 2020 г. началась работа научного коллектива, созданного на базе консорциума пяти научных институтов и вузов и поддержанная субсидией Министерства науки и образования. Глобальной целью проекта стало создание компьютерно-лингвистической платформы нового поколения, ядром которой является Национальный корпус русского языка (НКРЯ). Работы по проекту были сгруппированы в три взаимосвязанных направления:

— инфраструктурное направление, в рамках которого была заложена программно-технологическая основа для создаваемой компьютерно-лингвистической платформы;
— ресурсное направление, обеспечивавшее развитие и пополнение всех корпусов платформы;
— научно-исследовательское направление, в котором созданная платформа была апробирована в конкретных лингвистических исследованиях.

Цель настоящей статьи — очертить наиболее значимые изменения, связанные с новым этапом развития Национального корпуса русского языка. В фокусе статьи будут в первую очередь новые возможности, которые открываются для лингвистов и для более широкой аудитории. В части 1 представлен общий краткий обзор модернизации НКРЯ, части 2, 3 и 4 посвящены наиболее интересным и многообещающим обновлениям: новому корпусу «Социальные сети» (2), Панхроническому корпусу (3) и инструментам статистического анализа и визуализации данных (4).

1. Общий обзор изменений Национального корпуса русского языка

Работа над так называемым «Корпусом 2.0» — технологически и концептуально обновленным (а в части архитектуры даже пересобранным) Национальным корпусом русского языка — велась в нескольких направлениях: технической модернизации программной архитектуры, ресурсного пополнения и обновления данных и разметки. Эти направления тесно взаимосвязаны: без технологической перестройки корпусной платформы невозможно было выстроить процесс оперативного пополнения корпусов (каждый корпус раньше существовал изолированно и требовал сложной настройки), устаревшая модель интерфейса ограничивала возможности по развитию разметки корпусов, добавлению новых типов метаданных и новых поисковых возможностей. Принципиально важна и обратная связь — данные, подготовленные с помощью нового инструмента автоматической разметки и снятия омонимии, позволили внедрить инструменты статистического анализа. Таким образом, Корпус 2.0 стал мощным аналитическим инструментом и на уровне отдельных лемм (портрет слова), и на уровне их сочетаний (простых частотностей и коллокаций), и на уровне отбираемых пользователем подкорпусов (портрет подкорпуса). Ниже мы кратко перечислим наиболее значимые изменения, произошедшие в Корпусе на этапе 2.0.

1.1. Пополнение корпусов, новые корпуса

К концу 2022 г. общий объем коллекций текстов НКРЯ превысил 2 млрд словоупотреблений. Благодаря увеличению базы текстов корпуса на сотни миллионов слов Корпус 2.0 теперь представляет собой наиболее полную и разнообразную коллекцию текстов на русском языке, собранную в одном месте и объединенную удобным интерфейсом.

Особое внимание было уделено улучшению показателей репрезентативности в области жанров и периодов, ранее мало представленных в Корпусе. Так, коллекции нехудожественной прозы и публицистики **Основного корпуса** были пополнены текстами научных трудов XVIII — начала XIX в., были добавлены документы, научные работы, описания путешествий, подборки мемуаров, публицистика (в том числе из журналов «Сын отечества», «Московский наблюдатель», «Вестник Европы»), тексты дневников и воспоминаний XIX–XXI вв. из проекта «Прожито» и др. С другой стороны, увеличена представленность образцов современного языка, в частности была добавлена проза XX — начала XXI в., коллекция современных путеводителей, собрание современных научных текстов разных жанров (тезисы, программы, учебные пособия, задачи, конспекты), коллекция производственно-технических инструкций и пособий. Коллекции **Газетного корпуса** расширили временной охват: теперь они включают хронологический диапазон с 1983 по 2021 г. Газетный корпус достиг более 800 млн словоупотреблений.

Существенно были расширены коллекции корпусов со специальной разметкой, причем принципиально важным является не только увеличение объема, но и качественные характеристики источников, расширение репрезентативности по параметрам, определяемым спецификой корпуса. Так, в **Диалектный корпус** были добавлены тексты из различных регионов и диалектных зон: северные говоры (Архангельская область), среднерусские (Тверская область), южнорусские (Смоленская, Тамбовская области), говоры позднего формирования (Поволжье, Урал, Сибирь). Расширение **Параллельных корпусов** коснулось не только существенного увеличения имеющихся коллекций и жанрового разнообразия представленных в них текстов, но и создания новых языковых пар, таких, например, как сербско-русская, румынско-русская, корейско-русская, хинди-русская. В **Древнерусский корпус** были добавлены знаменитые памятники русской литературы, такие как «Сказание

о Борисе и Глебе», «Поучение Владимира Мономаха», «Слово о полку Игореве» и др., а также коллекции пергаменных и бумажных деловых документов. В **Мультимедийный корпус** были добавлены большие коллекции текстов устной научной речи (доклады на конференциях, учебные и популярные лекции, теле- и радиопередачи), устной политической речи (интервью, пресс-конференции, выступления на митингах, собраниях и съездах, ток-шоу на радио и ТВ и мн. др.). Пополнение **Поэтического корпуса** включало в себя тексты поэтов второй половины XX в., а также большую коллекцию русских переводов античной поэзии: «Илиаду» Гомера в переводе Н. И. Гнедича, «Энеиду» Вергилия в переводах В. Я. Брюсова и С. М. Соловьева и сатиры Горация в переводе А. А. Фета.

Наиболее радикально обновлен **Обучающий корпус**, предназначенный для использования в школьном преподавании русского языка и литературы, в него было добавлено более 1000 новых текстов. Теперь в Обучающем корпусе есть все основные произведения из российской школьной программы по русской литературе, включая те, которые рекомендуются для внеклассного чтения. В дополнение к собственно пополнению коллекции Обучающего корпуса был разработан раздел «Упражнения на основе Корпуса», в котором представлены упражнения, составленные на материале Обучающего корпуса и других корпусов НКРЯ. Упражнения относятся к разным разделам школьного курса русского языка и предназначены для самостоятельной работы на уроке и дома, а также для контроля знаний.

Кроме пополнения уже имеющихся корпусов были подготовлены и опубликованы новые корпуса. Это, во-первых, **корпус «Русская классика»**. Корпус включает художественные, публицистические и эпистолярные произведения из собраний сочинений русских классических писателей. Произведения русских классических писателей имеют особый статус для истории русского литературного языка. Если считать, что литературный язык — такой, который «обработан мастерами», тексты этих мастеров и составляют ядро корпуса русского литературного языка. С таким корпусом можно сверяться как с нормативным, а не узусным источником, из него можно извлекать авторитетные примеры для академических грамматик, словарей и учебных пособий.

Во-вторых, это **корпус «От 2 до 15»**. Корпус включает в себя литературу на русском языке, которую читают современные дети и подростки. Тексты подобраны по результатам масштабных опросов детей, подростков, учителей и родителей. Каждый текст размечен в соответствии с возрастом, в котором его обычно наиболее интересно читать. Для автоматической разметки фрагментов текстов по минимальному возрасту, в котором они предположительно будут понятны читателям, была создана нейросетевая модель [Morozov et al. 2022].

В-третьих, был подготовлен **корпус «Восточнославянская эпиграфика»**. Основу коллекции эпиграфических текстов в составе НКРЯ составляют средневековые надписи, собранные в базе www.epigraphica.ru. Для корпуса отобраны только ранее опубликованные тексты XI–XV вв., написанные на славянских языках, а также содержащие азбуки или цифры. В состав первого релиза корпуса входят 663 надписи общим объемом более 5200 слов. Корпус снабжен пословной морфологической разметкой со снятой омонимией.

Наконец, еще одним новым корпусом в семействе корпусов НКРЯ стал **корпус «Социальные сети»**. Этот корпус, основанный на данных VK и Telegram, а также ряда других источников, отражает живые языковые изменения за пределами литературного языка. Такие языковые явления почти не фиксируются в текстах Основного и Газетного корпусов. Более подробно особенности и структурные характеристики корпуса социальных сетей будут представлены в разделе 2.

1.2. Использование нейросетевых моделей для разметки текстов корпуса

Еще одним важным шагом в развитии НКРЯ стал переход на использование нейросетевых моделей для подготовки и анализа данных. Наиболее значимым результатом стало

развитие модели Rubic [Lyashevskaya et al. 2023], которая использовалась для автоматического снятия омонимии и разметки новых данных Основного и газетных корпусов. Модель использует стандарт разметки CONLL-U [Straka et al. 2016], совмещающий морфологическую и синтаксическую разметку. 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 г. [Lyashevskaya et al. 2020]. Ранее в Основном корпусе противопоставлялись тексты со снятой (вручную) морфологической омонимией (около 6 млн словоупотреблений) и основная масса текстов, где были представлены все возможные разборы; в большинстве корпусов со специальной разметкой присутствовала только неснятая омонимия. В результате использования нейросетевой модели для Основного и газетных корпусов это противопоставление фактически снято: действительно, при нынешних объемах Корпуса морфологическая омонимия в нем может быть разрешена только программными методами. Автоматическое снятие омонимии открывает возможности для развития новых статистических инструментов, таких как поиск по коллокациям, поиск семантически близких слов (подробнее об этих инструментах будет говориться в разделе 4). Совмещение морфологической и синтаксической разметки, характеризующее используемый стандарт CONLL-U, расширяет поисковый функционал, позволяя включить на больших объемах текстов инструменты синтаксического поиска (типы и направления синтаксических связей, синтаксические роли слов), а также выявить для каждого слова так называемые «скетчи» — устойчивые словосочетания с заданными синтаксическими отношениями (см. раздел 4). Эксперименты с нейросетевыми моделями проводились и для абсолютно новых задач, так, например, с помощью модели rutermextract[1] были размечены ключевые слова в текстах корпуса региональных СМИ. Одно ключевое слово (например, для газетной заметки о росте травматизма в зимние праздники) может состоять из однословного ключа (*праздник*, *переломы*) либо из двусловного сочетания (*таяние снега*). Разметка по жанрам корпуса социальных сетей также была сделана автоматически с помощью обучения нейросети (см. подробнее раздел 2). Наконец, еще одно новое экспериментальное направление приложения нейросетевых моделей связано со словообразовательным разбором. Автоматические разборы были сгенерированы нейросетевым алгоритмом (модель НейроКРЯ), ядром которого является свёрточная нейросеть, архитектура которой была предложена в работе [Sorokin, Kravtsova 2018]. Нейросеть была обучена на двух источниках данных. Разметка морфем в Обучающем корпусе опирается на [Тихонов 2002], содержащий около 100 тыс. лексем. Для каждого слова 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, где по состоянию на май 2023 г. даны разборы для 75 тыс. лексем (310 тыс. неуникальных морфем) [Ляшевская и др. 2009]. Словарь создан на основе словаря морфем [Кузнецова, Ефремова 1986]. Применение модели НейроКРЯ в Основном корпусе позволяет объединять в словообразовательные гнезда даже те слова, которых нет в словаре, т. е. в исходных обучающих данных.

В целом экспериментальное направление использования нейросетевых моделей для подготовки данных и их анализа является сегодня наиболее перспективным для развития всей экосистемы Национального корпуса русского языка, особенно с учетом объема и масштаба его пополнений. Ожидается, что с дальнейшим развитием возможностей нейросетевого моделирования процесс подготовки новых данных будет значительно упрощен, а аналитический функционал корпуса получит дополнительные возможности и для новых, и для старых данных.

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

1.3. Новые интерфейсные решения

Внешний облик Национального корпуса русского языка претерпел существенные изменения, причем эти изменения коснулись не только дизайна страницы в интернете, но и общего подхода к представлению информации о Корпусе. Обновление интерфейса началось с концептуально новой главной страницы, дающей сводную информацию о всех корпусах и их объеме, а также подробного пользовательского руководства. Поисковая строка на главной странице — это так называемый «обзор возможностей»: простой запрос, подобный обычному запросу в поиске, ведет пользователя на страницу агрегированных результатов, которые кроме собственно выдачи-конкорданса включают в себя график Панхронического корпуса (см. подробнее раздел 3), случайное стихотворение из Поэтического корпуса со словом из запроса, карточку свойств слова с его грамматическими и семантическими характеристиками и список коллокаций из Основного корпуса. Таким образом, это набор иллюстраций, «на лету» демонстрирующий, какую информацию можно найти в разных корпусах НКРЯ по простому запросу. Были усовершенствованы формы поиска и отбора подкорпуса, в частности, в форму поиска добавлена возможность поиска по словоформе (в ряде корпусов — также с использованием регулярных выражений), а при отборе подкорпуса появилась возможность выбрать диапазон дат обновления версий корпуса. Эта функция важна при воспроизведении результатов корпусного исследования, поскольку при пополнении корпуса показатели частотности могли измениться. Цель многочисленных изменений интерфейса — сделать работу с корпусом более понятной и удобной. Стоит отметить, что за этими изменениями стоит еще и глобальная перестройка всей внутренней архитектуры корпуса, в результате которой появилась принципиальная возможность гибкой настройки и адаптации интерфейса под специфические особенности корпусов [Гладилин, Козеренко 2022]. Постепенно все корпуса Национального корпуса русского языка были переведены на новый интерфейс.

Основные направления развития НКРЯ, очерченные выше, объединила общая задача — создание новой корпусной платформы, в которой, с одной стороны, интегрированы и оптимизированы уже имеющиеся корпуса и концептуальные решения, а с другой стороны, реализован модульный подход, обеспечивающий необходимую гибкость и доступность для изменения и дальнейшего развития. Для рассказа обо всех важных результатах, достигнутых в рамках проекта, объема настоящей статьи недостаточно. Среди наиболее интересных и значимых результатов проекта, которые мы бы хотели осветить подробнее, выделим три: новые корпусные данные, введенные в научный оборот с помощью корпуса «Социальные сети», объединение текстов исторических и современных корпусов в формате Панхронического корпуса, а также разнообразные инструменты статистики и визуализации, обеспечивающие возможности применения современных квантитативных методов для анализа корпусных данных.

2. Корпус «Социальных сетей»

2.1. Электронная коммуникация в НКРЯ

Сфера электронной коммуникации (или компьютерно-опосредованного общения, язык интернета) стала предметом лингвистического внимания с момента массового распространения интернета, см. [Иванов 2000; Бергельсон 2002; Кузьмина 2003; Трофимова 2004; Капанадзе 2005; Горошко 2007; Какорина 2008] и др. В исследованиях начала XXI в. поднимались вопросы о специфике этого модуса существования языка: гипертекстовой природе

текста в среде сети Интернет, его гибридном характере, о тенденциях в орфографии, языковой игре и др. При создании Национального корпуса русского языка сфера электронной коммуникации была указана среди основных сфер функционирования, и тексты электронной коммуникации были включены в состав Основного корпуса письменных текстов (самые ранние датированы 2002 г.). В корпусе представлены образцы существовавших в тот период типов компьютерно-опосредованной коммуникации: форумы, чаты в ICQ, записи в «Живом журнале», СМС-сообщения, комментарии в СМИ. В дальнейшем, по мере развития этой сферы общения, в корпус включались новые типы текстов — ленты новостей и групповые чаты, форумы по интересам, блоги на других платформах.

Учитывая, что тексты электронной коммуникации представляют собой особое подмножество в составе Основного корпуса письменных текстов, в среднем существенно сильнее нормированном орфографически, при подготовке текстов проводилась предварительная нормализация орфографии, при которой всем формам с отклонениями от стандартных написаний приписывалась правильная стандартная форма, которая в дальнейшем получала морфологическую аннотацию и участвовала в лексико-морфологическом поиске. Благодаря этой предварительной работе в подкорпусе электронной коммуникации по запросу какой-либо лексемы мы можем получить разные варианты ее написания, например: по запросу леммы *ничто* будут выдаваться контексты, среди которых встретятся формы *ничего*, *ничо*, *ничё*, *ниче*; по запросу *красавчик* среди контекстов встретим написания *красавчег*, *кросавчег.*

Помимо нормализации орфографии стандарт подготовки текстов предполагал ручное редактирование и метатекстовую разметку: удаление повторов, разметку ников и реплик в форумах и др. Все это делает поиск в корпусе более точным, однако существенный минус такого подхода — трудозатратность при подготовке текстов, которая по своей сложности приближается к оформлению транскриптов устной речи. Как следствие медленный процесс пополнения корпуса перестал соответствовать темпу развития электронной коммуникации, которая с распространением социальных сетей (в русскоязычном сегменте интернета примерно начиная с 2006 г.) испытала взрывной рост.

Для изучения сферы электронной коммуникации в ее современном состоянии создан корпус социальных сетей, опирающийся на иные принципы отбора, подготовки и организации материала.

2.2. Принципы создания корпуса «Социальные сети»

В разработке концепции корпуса принимали участие Б. В. Орехов, С. О. Савчук, Д. В. Сичинава. В отличие от небольшого подкорпуса электронной коммуникации (текущий объем не превышает 3,4 млн словоупотреблений), которому отведена определенная доля в общем массиве текстов Основного корпуса, корпус «Социальные сети» должен иметь **большой объем** (от 100 млн словоупотреблений) и регулярно пополняться, что позволит получать при его использовании статистически значимые результаты. Объем корпуса на конец 2023 г. составлял 157 млн словоупотреблений.

Представительность (репрезентативность) корпуса социальных сетей обеспечивается тем, что в его состав включены тексты разных типов интернет-коммуникации: интернет-форумы, записи в блогах, сообщения в мессенджерах, при этом ставится задача как можно более широкого охвата социальных сетей, популярных форумов и каналов известных блогеров. **Региональная репрезентативность** корпуса состоит в том, что наряду с общероссийскими социальными сетями в него включены текстовые коллекции локальных сетей — местные форумы, группы VK и Telegram, популярные в регионе блогеры. В настоящее время в корпусе присутствуют коллекции текстов локальных соцсетей Воронежа и Воронежской области (Большой Воронежский форум, Воронежский рыболовный

форум, группы VK «Типичный Воронеж», «Регион-36» и др.), подготовленные в Воронежском государственном университете Н. С. Горбуновым, Д. А. Фурсиной, П. Д. Есиповой, А. Ю. Луценко, А. С. Шудриковой и др. [Донина и др. 2024 (в печати)]. В следующем году планируется расширить региональную коллекцию материалами популярных блогеров и форумов, собранную воронежскими коллегами в соцсетях соседних Курской, Ростовской и Тамбовской областей.

Тематическое разнообразие достигается за счет включения в корпус текстов, относящихся к разным тематическим областям, что в случае социальных сетей не представляет трудности, поскольку в настоящее время свои страницы в интернете имеют органы власти и управления, предприятия и организации, группы по интересам, профессиональные сообщества, а современные блогеры вышли далеко за рамки личных дневников — сегодня люди ведут не только персональные блоги, но и рассчитанные на широкую публичность: корпоративные, экспертные, тематические, новостные, кулинарные, туристические, музыкальные, игровые, автомобильные. Многое из этого тематического разнообразия представлено в текущей версии корпуса, например, «Чат для художников», «Ворон и Ёжка. Почти серьезный канал о жизни Воронежа» в Telegram, «kolokolschool. Гончарная школа», «Благотворительный фонд Милосердие», «Привет, Воронеж. Новости» и др. в VK.

2.3. Материал и способ представления данных

Планируемый большой объем корпуса социальных сетей может быть получен только при автоматическом сборе и автоматической обработке материала. Все тексты взяты из открытых источников: VK, Telegram, Livejournal, Liveinternet, Blogspot, Большой Воронежский форум и др. В сборе материала принимали участие Б. В. Орехов (сбор и обработка), Е. И. Пискунова, А. Б. Хазова, группой студентов и сотрудников ВГУ под руководством О. В. Дониной подготовлена воронежская коллекция.

Подготовка электронных версий текстов включает очистку их от html-разметки, элементов верстки веб-страниц и снабжение текстов xml-разметкой, используемой в корпусе; поиск и удаление дублей — повторяющихся текстов, которые неизбежно присутствуют в социальных сетях; поиск и удаление текстов на иностранных языках — вся эта работа выполняется с помощью программных средств. Подобные проблемы решались при создании газетных корпусов в составе НКРЯ, с ними сталкиваются все составители веб-корпусов.

В базе данных корпуса тексты организованы по-разному. Для основной массы текстов один документ включает один текст — пост за определенную дату. Для части текстов сохранена их диалоговая природа: один документ включает пару сообщений — исходный пост и комментарий к нему, они также имеют точную датировку. Часть текстов, в основном это многостраничные форумы из воронежской коллекции, представлена в одном документе без разделения на отдельные сообщения — они объединены общей темой, но точная датировка таких документов невозможна, она представляет собой интервал в несколько дней, месяцев и даже лет, если общение на форуме продолжается. Некоторый разнобой в датировке может отражаться на точности статистической обработки, например, при отображении распределения результатов поиска по датам, поэтому в дальнейшем предполагается унифицировать способ подачи дат текстов.

Ручная орфографическая нормализация текстов не проводилась, так что все они представлены в оригинальной орфографии, см. примеры (1) и (2).

(1) *Ребятки, 2.08. 15 в 18: 00 я жду всех-всех-всех на афтепати к fdwg и препати к моему дэрэ-))) буду рада видеть всех желающих сделать чин-чин во славу фитнеса и за мое здоровье😘😘регистрация не обязательно, количество мест не ограничено-)))) просто приходим и…. место: pub daddy, ул. Солдатская, 6-А* [vk (28.07.2015)]

(2) *Кароче, все ж уже понятно. Давайте так. Вы как настоящий школьничег еще раз провозгласите, что я слился, и на этом мы прекратим душный разговор ни о чем. Вторй вариант, вы возьмете себя в руки, перечитаете чат, и будете общаться по существу. Вариант со школьничком, мне в вашем сдучае представляется наиболее вероятным*😄🙂 [Rozetked Discuss. telegram Rozetked Discuss (10.07.2021)]

Таким образом, при лемматизации формы слова в нормативном и ненормативном (нестандартном) написании не объединяются в одну единицу и не опознаются как ее варианты. Поэтому для того, чтобы получить полное представление обо всех вхождениях слова, нужно учитывать все варианты его написания и осуществлять поиск по каждому из них. Например, с помощью двух запросов по словоформе со звездочкой *короч** и *кароч** получаем набор вариантов написания вводного слова, с которого начинается второй пример: *короче*, *короч*, *кароче*, *карочи*, *кароч*, *карочь*, *карочии*, *карочееее*, *карочеееееееее*. Как для автоматически определенных лемм, так и для словоформ при вводе с клавиатуры в форме запроса появляется всплывающая подсказка (suggest) — список присутствующих в корпусе последовательностей с таким началом; это позволяет пользователю ориентироваться в богатстве лексики и орфографии корпуса социальных сетей.

В дальнейшем предполагается опробовать на текстах социальных сетей механизм отождествления различных орфографических вариантов слова, который использовался при разработке Панхронического корпуса (см. раздел 3 настоящей статьи). В результате все возможные варианты написания слова можно будет получить по одному запросу на любой из вариантов.

2.4. Лингвистическая разметка и организация поиска

В корпусе используются два вида разметки — морфологическая разметка словоформ и метаразметка текстов. Морфологическая разметка выполнена на основе Mystem (https://yandex.ru/dev/mystem/) с неснятой морфологической омонимией. Это дает повышенную долю шума при поиске в сравнении с подкорпусом электронной коммуникации в составе основного корпуса, особенно в случае несловарных слов и нестандартных написаний. Однако, как уже говорилось, отказ от нормализации орфографии дает большой выигрыш во времени при подготовке текстов, а качество автоматической лемматизации и снятия морфологической неоднозначности в текстах с нестандартной орфографией и пунктуацией еще требует дополнительных исследований и оценки.

Метаразметка текстов включает минимум параметров: дата создания текста, интернет-платформа, на которой опубликован текст, регион охвата соцсети, тип текста, жанр текста, автор (или его условное имя) и название текста (последние два параметра определены для части текстов). Отбор подкорпуса возможен по дате, в том числе можно задать диапазон дат, по названию социальной сети или блог-платформы, по региону, по типу и жанру текста. Отбор подкорпуса по типу текста позволяет искать отдельно в исходных постах, то есть текстах автора блога, и в комментариях пользователей и подписчиков. В том случае, если выбран комментарий, текст исходного поста будет виден, но поиск по нему не ведется. Поиск по жанрам возможен благодаря экспериментальной разметке, выполненной автоматически для основного массива текстов корпуса, на чем необходимо остановиться подробнее.

2.5. Автоматическая разметка жанров

Социальные сети представляют большой интерес для исследователей речевых жанров, что отражено в многочисленных работах. Жанры интернета (или жанры 2.0, как их еще

называют) рассматриваются в сопоставлении с жанрами в других сферах коммуникации (прежде всего разговорной, публицистической, художественной); отмечается подвижность системы интернет-жанров, их зависимость от технологических платформ, вариативность, изменчивость; предлагаются различные классификации жанров, обсуждается роль корпусных методов в изучении речевых жанров, ср. [Щипицина 2009; Шмелева 2012; Дементьев 2016; Литвиненко 2016; Горошко, Землякова 2017; Кириллов 2017; Шилихина 2018; Карасик 2019; Егорова 2021] и др. Накоплен богатый материал, однако до общепринятой типологии, как признают исследователи, еще далеко.

Большинство исследователей сходятся в том, что форум, блог, канал в соцсетях представляют собой сложные образования — гипер- или мегажанры. Отдельные же сообщения, которыми обмениваются участники общения в форуме, авторские записи в блогах и комментарии к ним могут иметь разную жанровую природу: сообщать новость, привлекать внимание к какому-то событию, объявлять о каком-то мероприятии, рекомендовать какую-то покупку, давать полезные советы, и наоборот, содержать просьбу о помощи. В системе метатекстовой разметки НКРЯ определение жанровой принадлежности текста, как и его тематики, относится к базовым текстовым характеристикам и производится вручную. На материале корпуса социальных сетей было решено провести эксперимент по автоматическому определению жанровой принадлежности текстов с использованием нейросетевой модели. Для разметки использована модель RuRoBERTa, дообученная на текстах корпуса. Работы проводились группой нейроразметки (А. В. Глазкова, Д. А. Морозов, Н. А. Власова, Т. А. Гарипов, С. С. Столяров, И. А. Смаль).

Эксперимент проводился в два этапа. На первом этапе на основе анализа текстов корпуса социальных сетей и изучения исследовательской литературы была составлена номенклатура жанров, наиболее типичных для основного массива корпуса (всего девять жанров). Из основного и регионального газетного корпусов отобраны коллекции текстов, относящихся к этим жанрам (объемом не менее 200 документов каждая), составивших датасет для обучения модели. С помощью обученной модели был размечен основной массив текстов корпуса социальных сетей, а результаты разметки были протестированы экспертами. Эксперты оценивали корректность определения жанра, отмечая случаи ошибочных решений и предлагая альтернативный вариант. Особое внимание было уделено категории «неопределенный жанр», в которую попадали тексты, не относящиеся ни к одному из заданных жанров, либо не имеющие четких жанровых признаков (например, фрагменты диалога). Последующий анализ этих текстов позволил выявить еще целый ряд типичных для социальных сетей жанров, например, *инструкция*/*рекомендация*/*совет*, *цитаты*, *афоризмы*, *история*, *интернет-рейтинг*, *гороскоп*. Они были включены в итоговый список, используемый для разметки корпуса. Обучающий датасет также был расширен за счет текстов, относящихся к этим дополнительным жанрам.

На втором этапе модель была обучена на расширенном датасете и использована для повторной разметки массива корпуса. Полученные результаты в настоящее время представлены на сайте НКРЯ. Собираются данные об ошибках в определении жанровой принадлежности, которые в дальнейшем будут учтены при коррекции модели. В целом жанровое распределение текстов в текущей версии корпуса социальных сетей выглядит следующим образом. Наиболее велика доля текстов, относящихся к жанрам объявления/анонса (объединены в одну группу) — 20,6 % и информационного сообщения (новости) — 19 %. Большая группа жанров занимает нишу от 2 % до 10 %: инструкция/рекомендация/совет (6,1 %), отзыв/рецензия (5,1 %), рецепт (4 %), цитаты/афоризмы (3 %), поэзия (3 %), история (2,5 %), анекдот (2,3 %). К жанрам с долей менее 2 % в общей структуре относятся оценка (1,3 %), поздравление (1,1 %), интернет-рейтинг (0,4 %), гороскоп (0,4 %), вопрос (0,1 %), подпись к фото (0,06 %). К текстам блогов и форумов из воронежской коллекции, которые представлены большими файлами, не разделенными на отдельные посты, методику автоматического определения жанров применить оказалось невозможно, на их долю приходится, соответственно, 6,4 % и 6,1 % словоупотреблений.

Наконец тексты, жанровую принадлежность которых определить не удалось, отнесены к неопределенной категории и составляют 13 % от общего количества словоупотреблений. В дальнейшем предстоит повторный анализ текстов этой категории с последующим дообучением модели; результатом этого этапа будет уменьшение доли неопределенных в жанровом отношении текстов.

В целом эксперимент по автоматическому определению жанров электронной коммуникации следует признать успешным: накопленный богатый опыт ручной метатекстовой разметки и достаточно большие коллекции текстов разных жанров удалось продуктивно использовать для обучения нейросетевых моделей. На основе моделей в дальнейшем планируется проводить автоматическую разметку текстов по разным метатекстовым и текстовым признакам. В ближайшие планы развития корпуса входит разметка тематики текстов, их тональности, поиск ключевых слов, а также разметка невербальных компонентов (эмотиконов).

3. Панхронический корпус

Панхронический корпус — собрание текстов, охватывающее несколько общепризнанных диахронических периодов языка (в том числе и языка-предка вместе с языком-потомком; разумеется, решение вопроса о том, в каких случаях перед нами «древне-L язык», а в каких «язык M, являющийся предком языка L», зависит от традиции и формализовано быть не может). Такой корпус возможен только для языка (группы языков) с очень длинной более или менее непрерывной письменной историей. Среди таких ресурсов, например, французская база данных Frantext (https://www.frantext.fr/) или латинский Corpus Corporum Цюрихского университета [Roelli 2014]. Ведется работа над панхроническим корпусом чешского языка HiČKoK, который объединяет коллекции Чешского национального корпуса и корпусов, собиравшихся другими историками языка, на базе единой разметки в формате Universal Dependencies (https://korpus.cz/hickok). Он охватывает «исторический континуум» с XIII по XXI в.

Панхронические корпуса применяются как для исследования многовековых тенденций в развитии языка, проявляющихся непрерывно на протяжении долгого периода (таких как грамматикализация вида или одушевленности, изменение предложного управления и т. п.), так и для феноменов, фиксируемых в письменных источниках со значительными временны́ми интервалами — так называемых «скрытых» (submerged) явлений (ср. [Adams, Vincent (eds.) 2016] для латинского языка). Как известно, значительная часть лексем, фразеологизмов, конструкций эпизодически фиксируется письменными источниками древнерусской эпохи, а затем проявляется лишь в текстах Нового времени, диалектных словарях и т. п. (ср. [Зализняк 2024: 159 и сл.]), что связано с жанровой и стилистической неполнотой дошедшего до нас закрытого корпуса древних текстов.

Вполне естественно, что панхронический корпус строится путем объединения уже существующих. Важнейший стимул при его создании — это именно стремление получить поиск «в одном окне», не повторяя запрос много раз в разных поисковых формах с разными орфографическими принципами, форматами запроса и т. п. Что касается корпусов разных исторических периодов, то над ними обычно работают разные команды специалистов, затем получающие возможность привести свои результаты к «общему знаменателю»: и латинский, и чешский панхронический корпус возникли именно после появления такого единого формата разметки, сделавшего возможным единый интерфейс поиска. Заметим, что команда НКРЯ приняла решение объединить разные корпуса на общей платформе независимо от проектов, посвященных латыни и чешскому (о последнем нам стало известно уже после публикации первого релиза нашего Панхронического корпуса), — оно вполне логично вытекает из общей задачи.

Важным моментом стал количественный и качественный скачок в методиках обработки большого объема данных при помощи механизмов машинного обучения — будь то морфологический анализ или распознавание древних рукописей. Если ранее подготовка такого анализа данных предполагала полностью ручной набор, разметку и нормализацию (особенно это касалось текстов на нестандартизированном языке с неустойчивой орфографией и т. п.), сейчас обучение на «золотом стандарте» размеченных вручную текстов с последующей экспертной посткоррекцией обеспечивает статистику и быстрый поиск для куда большего количества текстов — и наличие в разметке, рассчитанной на пользователя-специалиста, того или иного числа ошибок искупается самим фактом наличия значительного массива текстов, доступного для поиска, разметка которого постоянно улучшается.

О перспективе Панхронического корпуса на базе НКРЯ Д. В. Сичинава упоминает в статье о развитии Старорусского корпуса [Сичинава 2016: 209–210], в то время еще не имевшего морфологической разметки и представлявшего собой «слабое звено» между полностью размеченными вручную Древнерусским корпусом и Корпусом берестяных грамот (оба представляют тексты до XV в.), с одной стороны, и также морфологически аннотированным (тогда в основном с неснятой омонимией) Основным корпусом (куда входят тексты начиная с 1700 г.), с другой стороны. Тогда же О. Н. Ляшевская возглавила работу по автоматической лемматизации и снятию омонимии не только в Старорусском, но и в Основном корпусе НКРЯ; изначально оба корпуса планировалось разметить именно при помощи совместимого набора признаков, допускающих сквозной поиск. Появилась также идея синтаксической разметки всего временно́го диапазона на базе Universal Dependencies (тоже независимо от чешского проекта), ее интеграция в Панхронический корпус — дело будущего, но версии Старорусского и Основного корпусов в соответствующем формате уже сформированы. Работы над созданием Панхронического корпуса начались в 2020 г. Первая версия опубликована в ноябре 2022 г.; второй релиз, включивший также новый корпус «Восточнославянская эпиграфика», вышел в конце 2023 г.

Панхронический корпус в составе НКРЯ объединяет четыре разработанных в составе НКРЯ исторических корпуса. Это Древнерусский корпус [Мишина, Пичхадзе 2015], Старорусский корпус [Гаврилова и др. 2016], корпус «Берестяные грамоты» [Сичинава 2022] и появившийся недавно корпус «Восточнославянская эпиграфика» [Sitchinava, Dyshkant 2021; Sitchinava 2023]. Пятой составной частью Панхронического корпуса стал Основной корпус, самые ранние тексты которого относятся к рубежу XVII и XVIII вв.; с этой же временно́й отметки начинается и Поэтический корпус, который также планируется включить в Панхронический. Нами принято решение не делать объем текстов в пределах каждого временно́го промежутка одинаковым (иногда такая количественная нормализация осуществляется в исторических корпусах, например, в американском корпусе COHA [Davies 2010]). Основной корпус включен полностью, несмотря на то что его объем в 30 раз больше, чем пяти исторических вместе взятых. При этом 70 % хронологического диапазона приходится на исторические корпуса.

Перечисленные выше пять корпусов-компонентов продолжают независимое существование в рамках НКРЯ для исследователей, занимающихся именно этими массивами текстов, а не макродиахроническими сюжетами. Привычный пользователям язык запросов, глубина разметки, орфография и т. п. в них сохраняется; впрочем, в части из них теперь доступны для поиска поздне- и раннедревнерусские леммы (см. ниже), интерфейс между которыми разработан специально для Панхронического корпуса. Таким образом, Панхронический корпус и его корпуса-компоненты способствуют взаимному усовершенствованию и обогащению информацией.

Панхронический корпус планируется обновлять раз в год, после пополнения входящих в него корпусов-компонентов. Его разметка также обновляется с учетом встретившихся в новых текстах лексем.

3.1. Представление лемм

Корпуса-компоненты Панхронического корпуса размечались на базе разных стандартов разметки и разного морфологического описания. Стандарт подачи лемм зависел от словаря соответствующей эпохи (Словаря древнерусского языка XI-XIV вв., Словаря русского языка XI-XVII вв., Грамматического словаря русского языка А. А. Зализняка); особую проблему представляла лемматизация имен собственных и образованных от них прилагательных, названий жителей, этнонимов, которые в большинство исторических словарей не включаются. Корпус берестяных грамот был размечен на основании словоуказателя ко второму изданию «Древненовгородского диалекта» А. А. Зализняка [2004], стандарты которого несколько отличаются от стандартов главных древнерусских словарей. Этот словоуказатель ориентирован на позднедревнерусский фонетический состав.

Кроме того, и внутри Древнерусского корпуса разные тексты, восходившие к изначально независимым базам данных, были размечены в соответствии с разными принципами (например, краткие прилагательные могли даваться как отдельные леммы или как словоизменительные формы). Внутри Основного корпуса, который также представлял собой с самого своего появления в 2004 г. объединение текстов, размеченных по двум разным морфологическим стандартам, похожие различия были к началу 2020-х гг. сглажены (усилиями О. Н. Ляшевской и Д. В. Сичинавы), но не в полной мере. Пятый корпус-компонент Панхронического корпуса — «Восточнославянская эпиграфика» — был добавлен позже, и нового размножения форматов удалось избежать. Он размечался по тем же стандартам, что и Древнерусский корпус.

Ключевая задача объединения этих ресурсов — связать соответствиями (не обязательно одно-однозначными, но позволяющими сквозной поиск интересующих пользователя слов и контекстов) представлений слов на следующих трех этапах:

— Раннедревнерусский (до падения и прояснения редуцированных): *загърдѣтисѧ*,
— Позднедревнерусский и старорусский (после падения и прояснения редуцированных): *загордѣтися*,
— Современный русский язык (русский язык Нового времени): *загордеться* (ныне устаревшее и просторечное, в отличие от более стандартного *загордиться*, ср.: *Загорделась, Санечка, загорделась, хоть бы разок приехала* [Чехов, «Иванов»]).

Кроме этих фонетических соответствий, возможны дополнительные, как регулярные (например, отвердение *ц*), так и нерегулярные, включающие в себя и морфологические замены (древнему *ужина* соответствует современное *ужинъ*, а наряду с устаревшим *загордеться* представлено более стандартное и более частотное в корпусе *загордиться*). Следует учитывать также явление расщепления лемм, связанного с фонетическими процессами, влиянием церковнославянской нормы и/или семантической дифференциацией (*съборъ* > *сбор*, *собор*; *грьчьскыи* > *греческий*, *грецкий*), а также появления новых омонимов (*мочи* [глагол], *мочь* [существительное] > *мочь*).

Чтобы «связать» три исторических корпуса — Древнерусский, Старорусский и корпус «Берестяные грамоты» (а затем и добавившийся позже четвертый, «Древнерусская эпиграфика») — генерировался полный список лемм в первых двух корпусах, а затем строилось их соответствие в представлении другой эпохи. Допустим, в Корпусе берестяных грамот встретилась лемма *сгонити*, для нее порождалась раннедревнерусская лемма *съгонити*; в Древнерусском корпусе для формы *звьнигородьць* была построена поздняя лемма *звенигородецъ*. Для этого используется механизм регулярных выражений со сложными условиями и полным последующим ручным контролем.

Что касается аналогичной карты соответствия позднедревнерусских лемм (полученных из всех трех исторических корпусов) и лемм современного русского корпуса, она строилась при помощи алгоритма, разработанного Т. А. Архангельским. Этот алгоритм учитывал как

регулярные правила, так и расстояния символов между леммами-кандидатами. Т. А. Архангельский предложил модификацию классического алгоритма расстояния между строками символов, известного как расстояние Левенштейна, где стоимость любой операции (замена, удаление или вставка символа) зависит от контекста, в котором находится символ в строке. Эти функции определяются с использованием нейронной модели, обученной на 50 тыс. пар слов, предварительно выровненных с учетом расстояния Левенштейна и выборочно проверенных. Получившаяся карта соответствий была усовершенствована Т. А. Архангельским и Д. В. Сичинавой вручную и в дальнейшем корректируется по мере работы с корпусом, для включения исправлений в новые релизы.

На последнем этапе проведения алгоритмического объединения карт А. Н. Дышкант собрал единые списки панхронических лемм, охватывающие все четыре интегрируемых корпуса. В итоговой версии панхронического поиска лемм в любом тексте каждая лемма может быть приведена к нормализованной форме раннедревнерусского периода (*възбранити*), позднедревнерусского / старорусского (*возбранити*) и современного русского (*возбранить*). Поиск осуществляется одновременно для всех вариантов лемм: таким образом, пользователь, вводя один и тот же запрос *возбранить*, сможет найти упоминания слова *возбранить* из XVIII в., *възбранити* из XI в. и *возбранити* из XV в. Эти формы могут быть найдены при выборе любого из фонетически-орфографических обликов. Важно отметить, что пользователь не обязан знать заранее, в какой именно форме он ищет интересующее его слово. В Панхроническом корпусе по мере ввода леммы появляется всплывающая подсказка (suggest; см. также выше о корпусе «Социальные сети»), в которой представлены начинающиеся с тех же букв леммы, и пользователь может выбрать нужный ему вариант.

Начиная с версии 2023 г. в Панхроническом корпусе введено также различие между межчастеречными омонимами/омографами, которым соответствуют отдельные строки соответствия лемм. Например, древнерусское *и* — местоимение (соответствующее совр. *он*) и союз; древнерусское и старорусское *дати* — глагол (современное *дать*) и союз ‘чтобы’ (утрачен современным языком), современный омограф *пропасть* — глагол *пропáсть* (древнее *пропасти*) и существительное *прóпасть* (не изменившее внешнего облика).

Полученные соответствия не являются однозначными: нашей целью было предоставить возможность найти все варианты данного слова. Например, при запросе *ивангородский* будут найдены словообразовательные варианты *ивангородский*, *иванегородский*, *иваниго-родский*, *иваногородский, иванягородецкий*. Отметим, что имена собственные и дериваты от них весьма избирательно охвачены историческими словарями, так что новая информация для исторической лексикографии очевидна уже на уровне разметки.

3.2. Грамматика

Грамматический набор признаков в Панхроническом корпусе отличается от разметки корпусов-компонентов. Обычно в этом наборе отсутствуют признаки, которые размечены только в части корпусов (например, управление предлогов, вид глагола, счетная форма), либо их интерпретация унифицирована. Например, информация об одушевленности представлена только в форме винительного падежа для слов *о*-склонения, форм множественного числа, местоимений и адъективов. Грамматические признаки, утраченные в современном языке (такие как аорист и имперфект), сохранены в исторических корпусах, поскольку они позволяют формировать информативные запросы, охватывающие большой временной диапазон из нескольких корпусов-компонентов.

Грамматические параметры влияют также и на выделение лемм в разных корпусах. Например, в Древнерусском корпусе компаратив *выше*, краткая форма *высокъ* и полная форма *высокыи* — разные лексемы; в других корпусах подача этой леммы унифицирована (хотя конкретная орфография может различаться).

В текстах, входящих в Панхронический корпус, снята лексическая и грамматическая омонимия. В Старорусском корпусе и большинстве текстов Основного корпуса приписывание лемм и грамматических характеристик сделано автоматически при помощи нейросетевых механизмов, причем в Старорусском корпусе предусмотрена ручная посткоррекция ошибок нейросетевого разбора.

3.3. Семантика

Для построения запроса по семантической разметке в Панхроническом корпусе используется привязанное к каждому входящему в запрос слову поле «семантика», сопровождаемое формой с отдельными страницами по знаменательным частям речи и местоимениям. Эта форма устроена полностью аналогично форме лексико-семантического поиска в тех корпусах НКРЯ, в которых выступает современный русский язык. Соответственно, можно задавать, например, для предметных имен — названия частей тела или этнонимы, для абстрактных имен — названия физических свойств или эмоций, для прилагательных — цветообозначения, для глаголов — глаголы движения или физического воздействия на объект и т. п. Фасетная классификация, используемая в семантической разметке современных текстов НКРЯ, описана, в частности, в работе [Рахилина и др. 2009].

Как указывает Д. В. Сичинава [2024: 347], «ограничения принятого решения очевидны. Разумеется, семантически размечены те и только те древне- и старорусские слова, которые имеют вошедшие в современный семантический словарь НКРЯ когнаты. Поскольку лексическая семантика подвержена историческим изменениям, а ряд слов, включая частотные, вообще утрачен современным языком — или же маргинализован в нем — и в семантическом словаре НКРЯ отсутствует, в семантической разметке исторических текстов заведомо есть неточности и неполнота, на данном этапе сознательно заложенные в проекте, и к ней надо относиться с осторожностью». Например, помету «отрицательная оценка» получают слова *тварь* или *челядь*, совершенно нейтральные в древнерусском или церковнославянском. Случаи полисемии и омонимии (опять же, являющиеся таковыми в современном русском, например, *вода* в значении ‘бессодержательный текст’ во всех контекстах ищется на «отрицательную оценку»), не снятые в семантической разметке «современного» НКРЯ, унаследованы и семантической разметкой Панхронического корпуса.

«Тем не менее полученное при таком подходе высокое покрытие исторических текстов семантической разметкой и диахроническая стабильность семантических классов большинства лексем в значительной мере компенсируют неизбежные недостатки такой разметки и делают оправданным само это предприятие» [Сичинава 2024: 347]. Например, запрос «аорист от глаголов движения» дает на исторических текстах большую и корректную выборку (*поиде*, *придохъ*, *понесохъ*, *повезоша*, *поѣха, слѣзе* и мн. др.).

3.4. Применение

В Панхроническом корпусе можно строить релевантные для нескольких веков истории русского языка запросы типа «предлог *по* с предложным падежом», «история существительного *забава*», «одушевленность названий животных», «сочетаемость глаголов движения с абстрактным субъектом», «имена собственные на *-славъ*» на всем этом массиве текстов, не вводя каждый раз пять запросов в интерфейс всех корпусов поочередно.

На материале Панхронического корпуса могут быть построены нормализованные частотные графики (количество вхождений на миллион) на всем хронологическом диапазоне запроса. Разумеется, выводы на основании частотного графика (а в будущем — также полученных на корпусе коллокаций и частотных словарей, когда такой функционал будет доступен) должны делаться осторожно и с учетом специфики текстов, включенных в корпус в разные периоды. Указанная проблема относится к азбучным основам квантитативной лингвистики. О схожей опасности предостерегает А. А. Зализняк [2024: 297], анализируя работы исследователя, который «подсчитывает количество имперфектов по отношению не к общему числу словоформ прошедших времен, а к общему объему текста»: «Единственно, чем хороша такая статистика, — это тем, что так легче считать. В остальном она никуда не годится: какой смысл может иметь подсчет среднего числа имперфектов на страницу текста в условиях, когда в одном тексте описания событий в прошлом могут быть представлены, скажем, вдвое чаще, чем в другом?» В ряде случаев заметные пики частот — статистические выбросы — связаны с конкретными текстами, для которых характерна высокая частота той или иной конструкции, а также с общим объемом текстов, доступных для того или иного интервала.

Решению этих проблем могут поспособствовать такие подходы, как использование в Панхроническом корпусе обобщенных жанровых признаков для отбора подкорпуса, альтернативные алгоритмы подсчета частот и другие. В релизе 2023 г. уже задействован метатекстовый признак «категория текста» (можно отдельно рассматривать литературные, богословские, бытовые тексты и т. п.), а кроме того, построенный средствами визуализации НКРЯ график можно корректировать как по оси абсцисс, так и по оси ординат, выводя «за кадр» отдельные выбросы.

4. Статистика и визуализация

Современная корпусная лингвистика очень во многом опирается на квантитативные методы анализа данных: сравнения частотностей лексико-грамматических конструкций, анализ коллокационных метрик, сравнение частотных словарей корпусов. В отличие от традиционного конкорданса эти методы предполагают работу с обобщенными данными, а не с изолированными примерами, поэтому результаты анализа представляются в виде таблицы или в виде графиков и диаграмм, визуализирующих дистрибуцию лексико-грамматических единиц или метаданных корпусных текстов. Корпус 2.0 предлагает целый пакет таких возможностей, включающий в себя поиск по коллокациям, подсчет частотностей, представление данных в виде графиков и автоматическое построение диаграмм. Кроме этого, в корпусе появился новый функционал, связанный с исследованием свойств отдельных лексем — «портрет слова». Ниже будет представлен краткий обзор возможностей основных инструментов статистики и визуализации, внедренных в корпусную платформу 2.0.

4.1. Коллокации

Поиск по коллокациям является, пожалуй, наиболее значительным (и долгожданным) нововведением из всего блока преобразований, относящихся к статистике и визуализации. В корпусной лингвистике понятие коллокаций несколько отличается от понятия, принятого в лексикографии: к коллокациям не предъявляется требование некомпозициональности, а коллокат понимается просто как единица, встречающаяся с заданным словом (ключом) чаще, чем случайно. Для того чтобы определить меру такой неслучайности, в НКРЯ используются хорошо известные в корпусной лингвистике метрики: t-score, MI^3, LogDice

и Loglikelihood[2]. Все они основаны на частотности каждого из элементов коллокации и размере корпуса, однако дают несколько разное ранжирование. Например, такая классическая метрика, как t-score, дает ранжирование, близкое простой частоте, а MI[3] выводит на более высокие позиции более редкие коллокаты. Пользователь может выбрать метрику, наиболее подходящую под его конкретные исследовательские задачи, и ранжировать результаты по ней.

Как уже говорилось выше, внедрение поиска по коллокациям потребовало доработки поискового интерфейса: функция поиска по коллокациям теперь доступна в нем наравне с поиском точных форм и лексико-грамматическим поиском. Интерфейс дает пользователю широкие возможности, поскольку как ключ, так и коллокаты могут быть заданы не просто как леммы, но и определены с помощью набора грамматических и/или семантических признаков; кроме того, можно задать ограничения на тип и направление синтаксической связи между ключом и коллокатом. Синтаксические параметры поискового запроса по коллокациям позволяют искать не просто наиболее частотные ассоциаты слова, но и наиболее типичные заполнители его активных и пассивных валентностей.

4.2. Частотность

Принципиально новые возможности дает такой новый формат выдачи результатов лексико-грамматического поиска, который называется в интерфейсе НКРЯ «частотность». Благодаря его внедрению исследователь избавлен от необходимости составлять частотный список лексем и их сочетаний, встретившихся в той или иной конструкции, вручную: теперь все данные о частотном распределении результатов поиска выводятся автоматически в одном окне. Например, нас интересует сочетание «предлог *с* + форма родительного падежа на *-у*»: в современных текстах лидируют сочетания *с виду*, *с ходу*, *<сбить> с толку*, *с голоду*, *с размаху*, *с глазу <на глаз>*, *с краю*. Как мы можем видеть на рис. 1, выдача по частотности показывает нам распределение лексем в рассматриваемой конструкции, а именно долю вхождений каждой лексемы в полученной выдаче.

Слово 1 Лемма	←[1..2]→	Слово 2 Лемма	Вхождения	Доля	Доля	ipm	Конкорданс
с	1	голод	2857	13,49%	[13,03%, 13,95%]	7,63	Примеры
с	1	вид	2828	13,35%	[12,9%, 13,81%]	7,55	Примеры
с	1	толк	2296	10,84%	[10,43%, 11,26%]	6,13	Примеры
с	1	ход	1589	7,5%	[7,15%, 7,86%]	4,24	Примеры
с	1	глаз	1490	7,03%	[6,7%, 7,39%]	3,98	Примеры
с	1	размах	1253	5,91%	[5,6%, 6,24%]	3,35	Примеры
с	1	бок	1010	4,77%	[4,49%, 5,06%]	2,7	Примеры
с	1	край	885	4,18%	[3,92%, 4,46%]	2,36	Примеры
с	1	пол	586	2,77%	[2,55%, 3%]	1,56	Примеры
с	1	час	513	2,42%	[2,22%, 2,64%]	1,37	Примеры
с	1	разбег	475	2,24%	[2,05%, 2,45%]	1,27	Примеры
с	1	перепуг	462	2,18%	[1,99%, 2,39%]	1,23	Примеры
с	1	мах	310	1,46%	[1,31%, 1,63%]	0,83	Примеры
с	1	бой	276	1,3%	[1,16%, 1,46%]	0,74	Примеры
с	1	жир	253	1,19%	[1,06%, 1,35%]	0,68	Примеры
с	1	испуг	248	1,17%	[1,03%, 1,32%]	0,66	Примеры
с	1	верх	189	0,89%	[0,77%, 1,03%]	0,5	Примеры

Рис. 1. Распределение частотностей существительных,
которые встречаются в конструкции с предлогом *с* в родительном падеже на *-у*

[2] Для подсчета коллокаций были использованные формулы, приведенные в работе [Evert, Krenn 2003].

Потенциал этого статистического инструмента состоит в возможности разных группировок результата поиска: мы можем сгруппировать результаты (т. е. вывести общее количество найденных единиц) не только по более привычным для пользователей корпуса леммам и словоформам, но и по грамматическим признакам, обозначенным набором грамматических тэгов, причем при настройке выдачи возможна и группировка по расстояниям между элементами запроса. Таким образом, частотность как режим отображения результатов поиска обеспечивает пользователя данными не только о лексическом распределении слов запроса, но и о моделях управления, паттернах синтаксической упорядоченности, а также помогает выявлять лексико-грамматические конструкции.

4.3. Графики

Графики, показывающие хронологическое распределение слова по текстам корпуса, существовали и в предыдущей версии НКРЯ. Но в текущей версии корпуса внедрены два новшества, которые делают графики еще более полезным и достоверным инструментом.

Во-первых, появилась возможность сравнить на графике результаты нескольких сложных запросов (до пяти), в том числе и заданных на разных подкорпусах, причем сравнивать можно не только результаты поиска по точным формам, но и результаты лексико-грамматического поиска. Хронологическое распределение результатов нескольких запросов визуализировано на одном графике, где каждому запросу соответствует отдельная линия. Второе нововведение заключается в тепловой шкале, дополняющей собственно график и отражающей сравнительное количество найденных документов по запросу на каждый год. Количество документов визуализируется с помощью цветовой насыщенности. Благодаря тепловой шкале пользователь имеет возможность сразу соотнести частотность словоупотреблений с количеством найденных документов и скорректировать интерпретацию, см. рис. 2.

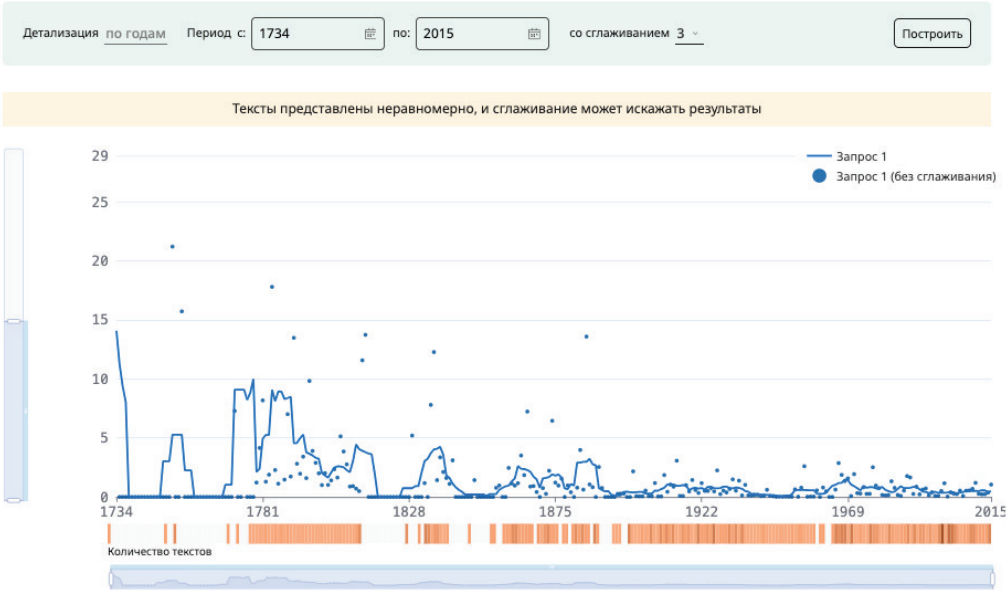

Рис. 2. Хронологический нормализованный график частотности леммы *кофий*. Тепловая шкала отражает количество найденных текстов по запросу за каждый год

Можно заметить, что некоторые колебания графика могут отражать не столько изменение частотности слова, сколько неравномерность хронологического распределения документов, таким образом, повышается надежность и достоверность выводов исследования.

4.4. Сравнение подкорпусов

К новым инструментам, построенным на подсчете распределения частотности, относятся и частотные словари. Первый частотный словарь на материале Основного и устного корпусов НКРЯ был создан еще 15 лет назад [Ляшевская, Шаров 2009], однако за последние годы принципиально изменился состав и объем корпуса, поэтому насущной стала задача обновления его частотного словаря. Реализация решения этой задачи стала скорее необходимым первым шагом для задачи более сложной и амбициозной: сравнения частотных словарей корпуса и подкорпуса. Простая, на первый взгляд, идея потребовала разработки быстродействующих алгоритмов подсчета частотностей «на лету»: поскольку возможности пользователя по созданию собственного подкорпуса практически не ограничены, соответственно, хранить частотные словари всех подкорпусов технически невозможно и требуется считать частотности лемм только после запроса пользователя. Теперь же благодаря новому алгоритму у нас в руках оказался простой инструмент, позволяющий на базовом уровне определить схожесть корпуса и подкорпуса и строить свои выводы с учетом этого знания. Например, одного взгляда на сопоставление частотного словаря Основного корпуса с подкорпусом, состоящим из документов, относящихся к официально-деловой сфере функционирования (см. рис. 3), достаточно, чтобы увидеть существенные отличия на уровне лексики.

Часть речи: Глагол

Скачать

Подкорпус					Корпус		
№	Разница	Лемма	ipm	Вхождения	№	ipm	Вхождения
1	=	быть	8801.01	44818	1	12930.39	4841784
2	=	мочь	2933.61	14939	2	3036.61	1137059
3	↑ 6	иметь	2664.38	13568	9	1052.63	394157
4	↑ 181	установить	1161.93	5917	185	127.97	47920
5	=	бы	1019.37	5191	5	2020.08	756419
6	↑ 37	находиться	1018.58	5187	43	369.96	138533
7	↑ 33	принять	864.82	4404	40	389.85	145980
8	↑ 18	получить	825.75	4205	26	556.75	208476
9	нов	предусмотреть	815.34	4152	>500	—	—
10	↑ 35	являться	753.68	3838	45	364.71	136566
11	↑ 182	указать	720.88	3671	193	122.57	45896
12	↑ 177	производить	699.09	3560	189	126.06	47203
13	↑ 4	дать	615.82	3136	17	731.6	273947
14	↑ 234	определять	605.22	3082	248	98.81	36998
15	нов	осуществлять	594.22	3026	>500	—	—

Рис. 3. Сравнение словарей Основного корпуса и подкорпуса, ограниченного официально-деловой сферой функционирования

Сравниваться могут абсолютные и относительные величины, а также ранги позиций слов в списках, при этом при сравнении подкорпуса и корпуса мы можем сразу увидеть «передвижения» рангов слов подкорпуса относительно списка корпуса и, таким образом, получить информацию о лексической специфике подкорпуса. Объем частотного словаря, доступного пользователю, составляет 500 лексем, при этом пользователь может посмотреть частотный список определенной части речи — существительных, глаголов, прилагательных, наречий. При помощи этого инструмента можно сравнивать с генеральной совокупностью лексикон определенной эпохи, определенного автора и т. п. Сравнение между несколькими корпусами пока не доступно, но со временем может также принести интересные результаты.

Еще одним способом «посмотреть сверху» на состав корпуса (или его подкорпуса) является сравнительный анализ метаатрибутов текстов, реализованный в виде интерактивных диаграмм (в том числе диахронических). Эти внешние параметры, их равномерность или неравномерность распределения пользователь всегда должен принимать во внимание, чтобы учитывать их возможное воздействие на результаты исследования. Так, диаграмма распределения текстов по тематике в Основном корпусе (рис. 4) наглядно показывает, что документы, входящие в корпус, очень разнообразны по этому параметру. Важно учитывать, что на диаграммах выводится только 10 первых значений, а остальные объединяются в категорию «Прочее».

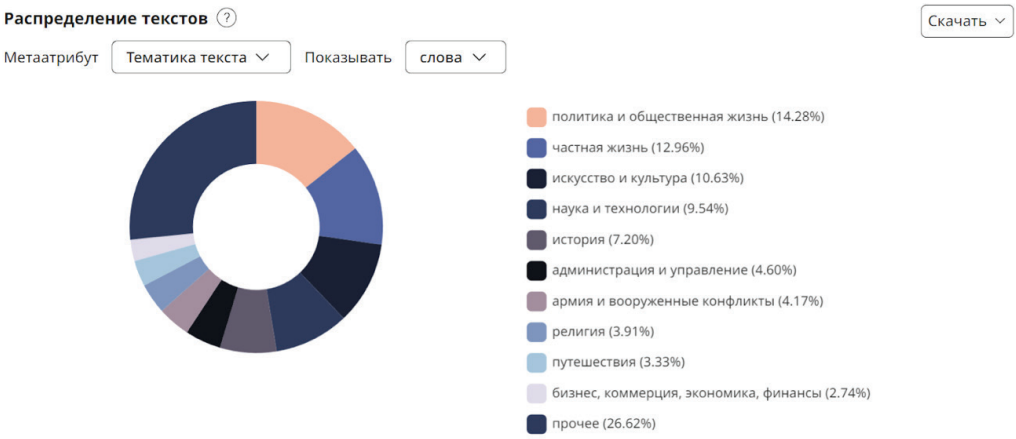

Рис. 4. Распределение текстов по тематике в Основном корпусе

4.5. Портрет слова

Портрет слова представляет собой интерфейс для комплексного анализа свойств отдельной лексемы на материале конкретного корпуса. При разработке концепции портрета слова мы вдохновлялись функционалом сервиса «Word at a glance», представленным в Национальном чешском корпусе (https://www.korpus.cz/slovo-v-kostce/), учитывался также опыт Цифрового словаря немецкого языка (https://www.dwds.de/).

На странице портрета слова собраны общие параметры слова, соответствующие его грамматической и семантической разметке в корпусе, а также в компактном виде представлена самая разная информация о функционировании слова в заданном корпусе. В своем максимальном разнообразии портрет слова реализован в Основном корпусе.

Кроме пяти случайных примеров из конкорданса, диахронического графика и диаграммы частотности слова по метаатрибутам текстов корпуса, в портрет слова включены и другие блоки, которые в комплексе дают пользователю разнообразную информацию о том, как слово существует и функционирует в языке. Во-первых, в портрет слова входят скетчи — предподсчитанные коллокации, распределенные по типу синтаксической связи. Каждый скетч представляет собой список из 10 самых частотных коллокаций в заданных синтаксических отношениях. Для каждой части речи был определен свой список таких отношений. Например, для существительных скетчи включают в себя прилагательные-определения к заданному существительному, глаголы, для которых это существительное является подлежащим, глаголы, для которых оно же является прямым дополнением или косвенным дополнением, другие существительные, связанные с заданным сочинительной связью. Скетчи позволяют пользователю быстро оценить сочетаемость выбранной лексемы, прежде чем приступать к более детальному исследованию (см. рис. 5).

Скетчи

душа Существительное

Определения		*Сказуемые*		*Глаголы с прямым дополнением*		*Глаголы с косвен*
1. мертвый	10,41	1. болеть	7,83	1. чаять	9,2	1. кривить
2. человеческий	9,81	2. рваться	6,39	2. отвести	9,02	2. любить
3. живой	8,93	3. жаждать	6,3	3. наполнять	8,89	3. болеть
4. бессмертный	8,42	4. гореть	5,95	4. раздирать	8,62	4. отдохнуть
5. чистый	8,05	5. искать	5,92	5. отдать	8,37	5. отдыхать
6. грешный	7,89	6. просить	5,57	6. отводить	8,36	6. покривить
7. добрый	7,89	7. стремиться	5,52	7. упокоить	8,04	7. ненавидеть
8. детский	7,81	8. петь	5,51	8. погубить	7,95	8. чувствовать
9. чужой	7,43	9. лежать	5,51	9. спасти	7,84	9. стремиться
10. христианский	7,41	10. радоваться	5,44	10. продать	7,79	10. привязаться

Показать все коллокации

Рис. 5. Скетчи для слова *душа* по основным синтаксическим связям существительных: определения к *душе*, глаголы-сказуемые, глаголы, которые управляют существительным *душа* как прямым дополнением

В портрет слова добавлен еще один совершенно новый блок информации о слове — морфемный разбор словообразовательной структуры слова (подробнее о разметке словообразовательной структуры слова говорилось выше в разделе 1.2, посвященном развитию нейросетевых моделей в проекте). Морфемный разбор строится не только для тех слов, которые присутствуют в базовом словаре морфемного анализа [Ляшевская и др. 2009], но и для тех слов, которые не находятся в этом словаре, в таком случае их морфемная структура строится автоматически с помощью нейросетевой модели НейроКРЯ. Например, слово *кроссовки* отсутствует в словаре морфемного анализа, так что его членение (*кросс-ов-к-и*) предсказано алгоритмом (см. рис. 6). Такие разборы снабжены специальным признаком «сгенерировано НейроКРЯ».

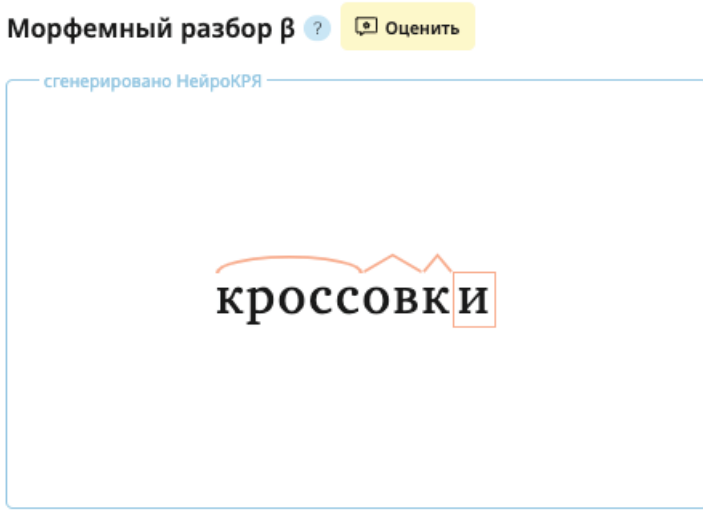

Рис. 6. Автоматически предсказанный морфемный разбор слова *кроссовки*

Еще один сервис портрета слова, который обеспечивается моделью НейроКРЯ, — это автоматический подбор однокоренных слов. Пользователю выводятся 10 наиболее частотных однокоренных слов с заданным словом, встречающихся в корпусе (см. рис. 7).

Рис. 7. Однокоренные слова к слову *гарантировать*, сгенерированные с помощью алгоритма НейроКРЯ

Дистрибуция форм слова в корпусе отображена с помощью таблицы «формы слова» (см. рис. 8). Важно подчеркнуть, что сравниваются именно орфографические варианты слова, так, например, мы можем увидеть формы слова в старой и новой орфографии, а также получить информацию о том, какие формы слова не встречаются в корпусе вовсе. Формы слова *плен* показывают не только наличие в корпусе этого слова в старой орфографии, но и конкуренцию форм предложного падежа и почти полное отсутствие парадигмы

множественного числа. Насыщенность цветового фона показывает частотность формы, относительная частотность ipm формы выдается во всплывающем окне.

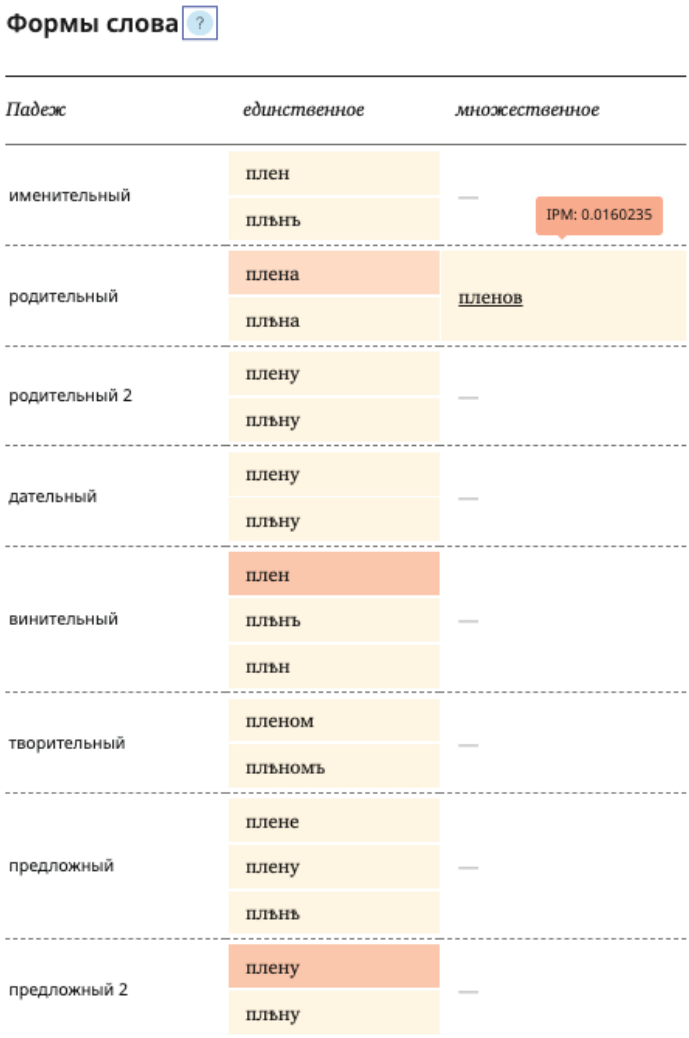

Рис. 8. Таблица распределения форм слова *плен* в корпусе

Кроме морфологически и синтаксически связанных слов портрет слова позволяет увидеть так называемые ассоциаты слова, или «похожие слова» — слова, которые наиболее часто встречаются в корпусе в одном и том же контексте. В основе представляемого ассоциативного ряда лежат предподсчитанные модели дистрибутивной семантики — матрицы взаимной встречаемости слов. Характерно, что разные корпуса дают разные списки «похожих слов» к одной и той же лексеме, потому что наиболее частотные контексты одного и того же слова в них будут разными (см. рис. 9).

В настоящий момент портрет слова имеется у всех корпусов НКРЯ, однако его наполненность различна в разных корпусах. Дальнейшее наполнение виджетами портретов слова связано с развитием нейросетевых технологий корпуса, таких как автоматическое снятие омонимии, модели дистрибутивной семантики, технологии морфемного анализа.

Рис. 9. «Похожие слова» для слова *звезда*
в Основном корпусе (слева) и в корпусе Региональных СМИ (справа)

5. Дальнейшие направления развития корпуса

Выше был дан общий обзор основных изменений в составе корпусов и пользовательском функционале НКРЯ, которые стали результатом работы проектной группы в 2020–2023 гг. Следует подчеркнуть, что в основе всех этих изменений лежит фундаментальная и масштабная перестройка архитектуры корпуса. Результатом преобразований стало создание новой корпусной платформы, которая должна обеспечить технологические возможности для устойчивого развития корпусов НКРЯ, их текстового пополнения, подготовки данных для индексации, гибкой настройки интерфейсов.

Создание платформы стало необходимым условием развития технологий нейросетевого моделирования для разметки данных корпуса и создания новых аналитических инструментов: ведь сложность автоматических алгоритмов состоит не только в их экспериментальной настройке и оптимизации на не самых обычных (а часто весьма экзотических) данных, но и во внедрении в собственно технологическую базу корпуса. Очевидно, что решение проблемы интеграции экспериментальных автоматических методов открывает огромный горизонт для дальнейшего развития — экстенсивного, связанного с «выравниванием» функциональных возможностей разных корпусов, с автоматическим снятием омонимии, разметкой метаданных, разработкой статистических аналитических экспериментов, и интенсивного, углубленного, связанного с созданием совершенно новых типов разметок и аналитических инструментов.

Наконец, еще одним направлением развития, обеспеченным технологическими преимуществами новой платформы, является движение в сторону персонализации пользовательской активности. Уже сейчас регистрация на сайте корпуса дает пользователю расширение доступного функционала, такого, как, например, возможность сравнения хронологической дистрибуции нескольких запросов. В дальнейшем пользователи смогут иметь возможность выстроить в своем личном кабинете свой собственный персонифицированный профиль НКРЯ, отражающий личные настройки, историю запросов, сохраненные выдачи и отбор подкорпусов и т. д. Таким образом, перспективное развитие Национального корпуса русского языка 2.0 включает не только совершенствование инструментов и ресурсов для изучения русского языка в его историческом и современном многообразии, но и развитие пользовательской инфраструктуры лингвистических — и шире — общегуманитарных исследований.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Бергельсон 2002 — Бергельсон М. Б. Языковые аспекты виртуальной коммуникации (Языковое поведение в сети Интернет). *Вестник МГУ. Сер. 19. Лингвистика и межкультурная коммуникация*, 2002, 1: 55–67. [Bergel'son M. B. Linguistic aspects of virtual communication (Linguistic behavior on the Internet). *Vestnik MGU. Ser. 19. Lingvistika i mezhkul'turnaya kommunikatsiya*, 2002, 1: 55–67.]

Гаврилова и др. 2016 — Гаврилова Т. С., Шалганова Т. А., Ляшевская О. Н. К задаче автоматической лексико-грамматической разметки старорусского корпуса XV–XVII вв. *Вестник ПСТГУ. Серия III: Филология*, 2016, 2(47): 7–25. [Gavrilova T. S., Shalganova T. A., Lyashevskaya O. N. On the problem of automatic lexical and grammatical markup in the Old Russian corpus of the XV–XVII centuries. *Vestnik PSTGU. Seriya III: Filologiya*, 2016, 2(47): 7–25.]

Гладилин, Козеренко 2022 — Гладилин С., Козеренко А. Новый интерфейс поиска для НКРЯ: системное описание и реализация. *Информационные технологии и системы 2022 (ИТС 2022): материалы конференций.* Шилин Л. Ю. и др. (ред.). Минск: БГУИР, 2022, 113–121. [Gladilin S., Kozerenko A. The new search interface for the RNC: System description and implementation. *Informatsionnye tekhnologii i sistemy 2022 (ITS 2022).* Conf. proc. Shilin L. Yu. et al. (eds.). Minsk: Belarusian State Univ. of Informatics and Radioelectronics, 2022, 113–121.]

Горошко 2007 — Горошко Е. И. Теоретический анализ Интернет-жанров: к описанию проблемной области. *Жанры речи: Сб. науч. ст.* Вып. 5. *Жанр и культура.* Дементьев В. В. (ред.). Саратов: Наука, 2007, 119–127. [Goroshko E. I. Theoretical analysis of Internet genres: Towards a description of the problem area. *Zhanry rechi*. Coll. of papers. No. 5. *Zhanr i kul'tura*. Dement'ev V. V. (ed.). Saratov: Nauka, 2007, 119–127.]

Горошко, Землякова 2017 — Горошко Е. И., Землякова Е. А. Полиформатный мессенджер как жанр 2.0 (на примере мессенджера мгновенных сообщений Telegram). *Жанры речи*, 2017, 1(15): 92–100. [Goroshko E. I., Zemlyakova E. A. A multi-format messenger as a genre 2.0 (on the example of the Telegram instant messenger). *Zhanry rechi*, 2017, 1(15): 92–100.]

Дементьев 2016 — Дементьев В. В., Степанова Н. Б. Корпусная генристика: проблема ключевых фраз. *Жанры речи*, 2016, 1(13): 24–41. [Dement'ev V. V., Stepanova N. B. Corpus genristics: The problem of key phrases. *Zhanry rechi*, 2016, 1(13): 24–41.]

Донина и др. 2024 (в печати) — Донина О. В., Фурсина Д. А., Горбунов Н. С. Создание регионального подкорпуса: от идеи до воплощения. *Труды международной конференции «Корпусная лингвистика-2023»* (в печати). [Donina O. V., Fursina D. A., Gorbunov N. S. Creation of a regional subcorpus: From idea to implementation. *Trudy mezhdunarodnoi konferentsii «Korpusnaya lingvistika-2023»* (in print).]

Егорова 2021 — Егорова В. И. Социальные сети и их речевые жанры. *Russian Linguistic Bulletin*, 2021, 3(27): 123–128. [Egorova V. I. Social networks and their speech genres. *Russian Linguistic Bulletin*, 2021, 3(27): 123–128.]

Зализняк 2004 — Зализняк А. А. *Древненовгородский диалект.* М.: Языки славянской культуры, 2004. [Zaliznyak A. A. *Drevnenovgorodskii dialekt* [Old Novgorod dialect]. Moscow: Yazyki slavyanskoi kul'tury, 2004.]

Зализняк 2024 — Зализняк А. А. *Слово о полку Игореве: взгляд лингвиста.* 4-е изд. М.: Альпина, 2024. [Zaliznyak A. A. *Slovo o polku Igoreve: vzglyad lingvista* [The Tale of Igor's Campaign: A linguist's view]. Moscow: Al'pina, 2024.]

Иванов 2000 — Иванов Л. Ю. Язык интернета: заметки лингвиста. *Словарь и культура русской речи.* М.: Азбуковник, 2000, 131–147. [Ivanov L. Yu. The language of the Internet: Notes of a linguist. *Slovar' i kul'tura russkoi rechi.* Moscow: Azbukovnik, 2000, 131–147.] http://faq-www.ru/lingv.htm.

Какорина 2008 — Какорина Е. В. СМИ и интернет-коммуникация (интернет-форум как новый коммуникативно-речевой жанр). *Современный русский язык: активные процессы на рубеже XX–XXI веков*. Крысин Л. П. (отв. ред.). М.: Языки славянских культур, 2008, 549–578. [Kakorina E. V. Mass media and Internet communication (Internet forum as a new communicative and speech genre). *Sovremennyi russkii yazyk: aktivnye protsessy na rubezhe XX–XXI vekov.* Krysin L. P. (ed.). Moscow: Yazyki slavyanskikh kul'tur, 2008, 549–578.]

Капанадзе 2005 — Капанадзе Л. А. На границе письменного и устного текста: структура и тенденции развития электронных жанров. *Голоса и смыслы. Избранные работы по русскому языку.* М.: ИРЯ РАН, 2005, 305–320. [Kapanadze L. A. On the border of written and oral text: The structure and

trends in the development of electronic genres. *Golosa i smysly. Izbrannye raboty po russkomu yazyku.* Moscow: Vinogradov Russian Language Institute, 2005, 305–320.]

Карасик 2019 — Карасик В. И. Жанры сетевого дискурса. *Жанры речи*, 2019, 1(21): 49–55. [Karasik V. I. Genres of online discourse. *Zhanry rechi*, 2019, 1(21): 49–55.]

Кириллов 2017 — Кириллов А. Г. Трансформация жанра блога в программах обмена мгновенными сообщениями. *Жанры речи*, 2017, 2(16): 260–267. [Kirillov A. G. The transformation of the blog genre in instant messaging programs. *Zhanry rechi*, 2017, 2(16): 260–267.]

Кузнецова, Ефремова 1986 — Кузнецова А. И., Ефремова Т. Ф. *Словарь морфем русского язык*а. М.: Русский язык, 1986. [Kuznetsova A. I., Efremova T. F. *Slovar' morfem russkogo yazyka* [Dictionary of morphemes of the Russian language]. Moscow: Russkii yazyk, 1986.]

Кузьмина 2003 — Кузьмина М. В. Компьютерный вид общения «чат» как жанр естественной письменной речи: основные характеристики. *Естественная письменная русская речь: исследовательский и образовательный аспекты: материалы конф.* Ч. II: *Теория и практика современной письменной речи.* Голев Н. Д. (ред.). Барнаул: Изд-во Алтайского ун-та, 2003, 86–91. [Kuz'mina M. V. Computer speech style "chat" as a genre of natural written speech: Basic features. *Estestvennaya pis'mennaya russkaya rech': issledovatel'skii i obrazovatel'nyi aspekty.* Conf. proc. P. II*: Teoriya i praktika sovremennoi pis'mennoi rechi.* Golev N. D. (ed.). Barnaul: Altai State Univ. Press, 2003, 86–91.]

Литвиненко 2016 — Литвиненко Ж. М. Современная русистика о жанрах интернет-коммуникации: форум, блог, чат. *Вестник ТГПУ*, 2016, 3(168): 48–52. [Litvinenko Zh. M. Modern Russian studies on the genres of Internet communication: forum, blog, chat. *TSPU Bulletin*, 2016, 3 (168): 48–52.]

Ляшевская и др. 2009 — Ляшевская О., Гришина Е., Тагабилева М., Иткин И. О задачах и методах словообразовательной разметки в корпусе текста. *Полярный вестник*, 2009, 12: 5–25. [Lyashevskaya O., Grishina E., Tagabileva M., Itkin I. On the tasks and methods of word-formation markup in a text corpus. *Polyarnyi vestnik*, 2009, 12: 5–25.]

Ляшевская, Шаров 2009 — Ляшевская О. Н., Шаров С. А. *Частотный словарь современного русского языка (на материалах Национального корпуса русского языка).* М.: Азбуковник, 2009. [Lyashevskaya O. N., Sharov S. A. *Chastotnyi slovar' sovremennogo russkogo yazyka (na materialakh Natsional'nogo korpusa russkogo yazyka)* [Frequency dictionary of contemporary Russian based on the Russian National Corpus data]. Moscow: Azbukovnik, 2009.]

Мишина, Пичхадзе 2015 — Мишина Е. А., Пичхадзе А. А. Древнерусский подкорпус Национального корпуса русского языка. *Труды Института русского языка им. В. В. Виноградова*, 2015, 6: 99–115. [Mishina E. A., Pichkhadze A. A. The Old Russian subcorpus of the Russian National Corpus. *Proceedings of the V. V. Vinogradov Russian Language Institute*, 2015, 6: 99–115.]

Рахилина и др. 2009 — Рахилина Е. В., Кустова Г. И., Ляшевская О. Н., Резникова Т. И., Шеманаева О. Ю. Задачи и принципы семантической разметки лексики в НКРЯ. *Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы.* Плунгян В. А. (отв. ред.). СПб.: Нестор-История, 2009, 215–239. [Rakhilina E. V., Kustova G. I., Lyashevskaya O. N., Reznikova T. I., Shemanaeva O. Yu. Tasks and principles of semantic markup of lexicon in the RNC. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy.* Plungian V. A. (ed.). St. Petersburg: Nestor-Istoriya, 2009, 215–239.]

Сичинава 2005 — Сичинава Д. В. Национальный корпус русского языка: очерк предыстории. *Национальный корпус русского языка: 2003–2005.* Плунгян В. А. (отв. ред.). М.: Индрик, 2005, 21–30. [Sitchinava D. V. Russian National Corpus: An outline of the prehistory. *Natsional'nyi korpus russkogo yazyka: 2003–2005*. Plungian V. A. (ed.). Moscow: Indrik, 2005, 21–30.]

Сичинава 2016 — Сичинава Д. В. Старорусские/среднерусские тексты в НКРЯ: от экстенсивной коллекции к корпусу. *Rašytinis palikimas ir skaitmeninė technologijos: VI tarptautinė mokslinė konferencija*, Vilnius, 2016 m. rugpjūčio 22–28 d. Vilnius: Lietuvos mokslo taryba, 2016, 208–210. [Sitchinava D. V. Old/Middle Russian texts in the RNC: from an extensive collection to a corpus. *Rašytinis palikimas ir skaitmeninė technologijos: VI tarptautinė mokslinė konferencija,* Vilnius, 2016 m. rugpjūčio 22–28 d. Vilnius: Lietuvos mokslo taryba, 2016, 208–210.]

Сичинава 2022 — Сичинава Д. В. Корпус берестяных грамот как параллельный. *Труды Института русского языка им. В. В. Виноградова*, 2022, 2: 92–106. [Sitchinava D. V. The corpus of birch bark letters as a parallel corpus. *Proceedings of the V. V. Vinogradov Russian Language Institute*, 2022, 2: 92–106.]

Сичинава 2024 — Сичинава Д. В. Панхронический корпус: интеграция исторических и современных корпусных ресурсов. *Труды Института русского языка им. В. В. Виноградова*, 2: 336–353. [Sitchinava D. V. A panchronic corpus: Integration of historical and contemporary corpus resources. *Proceedings of the V. V. Vinogradov Russian Language Institute*, 2: 336–353.]

Тихонов 2002 — Тихонов А. Н. *Морфемно-орфографический словарь: около 100 000 слов.* М.: АСТ, 2002. [Tikhonov A. N. *Morfemno-orfograficheskii slovar': okolo 100 000 slov* [Morphemic and spelling dictionary: about 100,000 words]. Moscow: AST, 2002.]

Трофимова 2004 — Трофимова Г. Н. *Функционирование русского языка в Интернете: концептуально-сущностные доминанты.* Автореф. дис. … докт. филол. наук. М.: РУДН, 2004. [Trofimova G. N. *Funktsionirovanie russkogo yazyka v Internete: kontseptual'no-sushchnostnye dominanty* [The functioning of the Russian language on the Internet: conceptual and essential dominants]. Abstract of cand. diss. Moscow: RUDN Univ., 2004.]

Шилихина 2018 — Шилихина К. М. Лексические маркеры жанров интернет-коммуникации. *Жанры речи*, 2018, 3(19): 218–225. [Shilikhina K. M. Lexical markers of Internet communication genres. *Zhanry rechi*, 2018, 3(19): 218–225.]

Шмелева 2012 — Шмелева Т. В. Жанр в современной медиасфере. *Жанры речи: сб. науч. ст.* Вып. 8. *Жанр и творчество.* Дементьев В. В. (ред.). Саратов; М.: Лабиринт, 2012, 26–37. [Shmeleva T. V. Genre in the modern media sphere. *Zhanry rechi.* Coll. of papers. No. 8. *Zhanr i tvorchestvo.* Dement'ev V. V. (ed.). Saratov; Moscow: Labirint, 2012, 26–37.]

Щипицина 2009 — Щипицина Л. Ю. *Жанры компьютерно-опосредованной коммуникации*. Архангельск: Поморский ун-т, 2009. [Shchipitsina L. Yu. *Zhanry komp'yuterno-oposredovannoi kommunikatsii* [Genres of computer-mediated communication]. Arkhangelsk: Pomor State Univ., 2009.]

Adams, Vincent (eds.) 2016 — Adams J. N., Vincent N. (eds.). *Early and Late Latin continuity or change?* Cambridge: Cambridge Univ. Press, 2016.

Davies 2010 — Davies M. *The Corpus of Historical American English (COHA).* Electronic resource, 2010. https://www.english-corpora.org/coha/.

Evert, Krenn 2003 — Evert S., Krenn B. *Computational approaches to collocations.* Introductory course at the European Summer School on Logic, Language, and Information (ESSLLI 2003), Vienna. 2003. www.collocations.de.

Lyashevskaya et al. 2020 — Lyashevskaya O. N., Shavrina T. O., Trofimov I. V., Vlasova N. A. GRAMEVAL 2020 shared task: Russian full morphology and universal dependencies parsing. *Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог»*, 2020, 19: 553–569. [Lyashevskaya O. N., Shavrina T. O., Trofimov I. V., Vlasova N. A. GRAMEVAL 2020 shared task: Russian full morphology and universal dependencies parsing. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conf. "Dialogue"*, 2020, 19: 553–569.]

Lyashevskaya et al. 2023 — Lyashevskaya O., Afanasev I., Rebrikov S, Shishkina Y., Suleymanova E., Trofimov I., Vlasova N. Disambiguation in context in the Russian National Corpus: 20 years later. *Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог»*, 2023, 2: 307–318. [Lyashevskaya O., Afanasev I., Rebrikov S, Shishkina Y., Suleymanova E., Trofimov I., Vlasova N. Disambiguation in context in the Russian National Corpus: 20 years later. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conf. "Dialogue"*, 2023, 22: 307–318.]

Morozov et al. 2022 — Morozov D. A., Glazkova A. V., Iomdin B. L. Text complexity and linguistic features: their correlation in English and Russian. *Russian Journal of Linguistics*, 2022, 2(26): 425–447.

Roelli 2014 — Roelli Ph. The Corpus Corporum, a new open Latin text repository and tool. *Archivum Latinitatis Medii Aevi: Bulletin Du Cange*, 2014, 72: 289–304.

Sitchinava, Dyshkant 2021 — Sitchinava D., Dyshkant A. Integration of the Old East Slavic epigraphical databases, corpora and indices. *Scripta & e-Scripta: The Journal of Interdisciplinary Medieval Studies*, 2021, 21: 93–106.

Sitchinava 2023 — Sitchinava D. Multiple interpretation and fragmented texts within a historical corpus: the case of Old East Slavic vernacular writing. *Jazykovedný časopis*, 2023, 74(1): 266–274.

Sorokin, Kravtsova 2018 — Sorokin A., Kravtsova A. Deep convolutional networks for supervised morpheme segmentation of Russian language. *Artificial Intelligence and Natural Language. AINL 2018. Communications in Computer and Information Science*. Ustalov D., Filchenkov A., Pivovarova L., Žižka J. (eds.). Springer: Cham, 2018, 3–10.

Straka et al. 2016 — Straka M., Hajič J., Straková J. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. 4290–4297.

Получено / received 21.11.2023 *Принято / accepted 30.01.2024*

Содержание

Contents

Overviews

Reviews

РОССИЙСКАЯ АКАДЕМИЯ НАУК
Отделение историко-филологических наук

ВОПРОСЫ ЯЗЫКОЗНАНИЯ

Журнал основан в январе 1952 года
Выходит 6 раз в год

2
МАРТ — АПРЕЛЬ

Москва
2024

Статьи отбираются редколлегией журнала на основе анонимного независимого рецензирования.

Индексируется в: Российский индекс научного цитирования (РИНЦ); Brill Linguistic Bibliography (Online); Cambridge University Press Language Teaching (Online); De Gruyter Saur Dietrich's Index Philosophicus; EBSCOhost MLA International Bibliography (Modern Language Association); Elsevier BV Scopus; European Reference Index for the Humanities and Social Sciences (ERIH PLUS); Gale MLA International Bibliography (Modern Language Association); ProQuest Linguistics and Language Behavior Abstracts (Online), Core; ProQuest MLA International Bibliography (Modern Language Association); Russian Science Citation Index (RSCI); Web of Science Core Collection's Emerging Sources Citation Index (ESCI); Wiley-Blackwell Publishing Ltd. Linguistics Abstracts (Online).

Адрес редакции: 119019, Москва, ул. Волхонка, 18/2, Институт русского языка им. В. В. Виноградова РАН, редакция журнала «Вопросы языкознания»

Телефон: +7 495 637-25-16

E-mail: voprosy@mail.ru

Сайт: https://vja.ruslang.ru

ISSN 0373-658X

RUSSIAN ACADEMY OF SCIENCES
Department of History and Philology

VOPROSY JAZYKOZNANIJA

(TOPICS IN THE STUDY OF LANGUAGE)

Founded in January 1952
6 issues per year

2
MARCH — APRIL

Moscow
2024

Articles are selected by the editorial board on the basis of anonymous double-blind independent peer review process.

Abstracting/Indexing: Brill Linguistic Bibliography (Online); Cambridge University Press Language Teaching (Online); De Gruyter Saur Dietrich's Index Philosophicus; EBSCOhost MLA International Bibliography (Modern Language Association); Elsevier BV Scopus; European Reference Index for the Humanities and Social Sciences (ERIH PLUS); Gale MLA International Bibliography (Modern Language Association); ProQuest Linguistics and Language Behavior Abstracts (Online), Core; ProQuest MLA International Bibliography (Modern Language Association); Rossiiskii indeks nauchnogo tsitirovaniya (RINTs); Russian Science Citation Index (RSCI); Web of Science Core Collection's Emerging Sources Citation Index (ESCI); Wiley-Blackwell Publishing Ltd. Linguistics Abstracts (Online).

Address: "Voprosy Jazykoznanija", editorial office, Vinogradov Russian Language Institute (RAS), Volkhonka street, 18/2, Moscow, 119019, Russia

Telephone: +7 495 637-25-16

E-mail: voprosy@mail.ru

Website: https://vja.ruslang.ru

ISSN 0373-658X