ВЕСТНИК РОССИЙСКОЙ АКАДЕМИИ НАУК, 2024, том 94, № 9, с. 795–803

С КАФЕДРЫ ПРЕЗИДИУМА РАН

О НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА

© 2024 г. Е.В. Рахилина[a,b,*]

[a]*Национальный исследовательский университет "Высшая школа экономики", Москва, Россия*
[b]*Институт русского языка им. В.В. Виноградова РАН, Москва, Россия*

**E-mail: rakhilina@gmail.com*

Поступила в редакцию 29.07.2024 г.
После доработки 05.08.2024 г.
Принята к публикации 10.08.2024 г.

Статья посвящена проекту создания Национального корпуса русского языка (НКРЯ) – мощной справочно-информационной системы по русскому языку, которая была разработана консорциумом организаций РАН с участием компании "Яндекс". Описаны история создания Корпуса, основной его функционал и пути совершенствования, а также наиболее технологичные подкорпуса – поэтический, параллельный, мультимедийный; приведены примеры их работы. Особое внимание уделено последним разработкам, которые финансируются в том числе из средств гранта Министерства образования и науки РФ, что позволило внедрить в НКРЯ технологии искусственного интеллекта. Впечатляющим результатом стал так называемый панхронический корпус, который отражает всю тысячелетнюю историю русского языка и делает её доступной для пользователей. Автор отмечает роль НКРЯ как инструмента научных исследований в области лингвистики и филологии, IT-специалистов, а также для методики преподавания русского языка.

Ключевые слова: Национальный корпус русского языка, русский язык, технологии искусственного интеллекта, корпусная лингвистика.

DOI: 10.31857/S0869587324090025, **EDN:** FCHBQK

ИЗ ИСТОРИИ СОЗДАНИЯ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА

Национальный корпус русского языка (НКРЯ) – это большое и специальным образом обработанное электронное собрание русских текстов, которое служит справочно-информационной системой по русскому языку. В 2024 г. этот проект отмечает своё двадцатилетие, в апреле поисковый ресурс под названием "Национальный корпус русского языка" был впервые размещён в Интернете для свободного доступа. Этому предшествовала большая подготовительная работа специалистов разных научных направлений.

РАХИЛИНА Екатерина Владимировна – доктор филологических наук, профессор, руководитель Школы лингвистики НИУ ВШЭ, главный научный сотрудник ИРЯ РАН.

У истоков проекта стоит Институт русского языка им. В.В. Виноградова РАН, который стал тем центром, где под руководством академика РАН В.А. Плунгяна трудилась команда основных разработчиков НКРЯ и определялись стратегии развития корпуса и корпусных технологий. Важно отметить, что с самого первого дня корпус создавался в сотрудничестве с компанией "Яндекс". Дело в том, что И.В. Сегалович, один из основателей компании, был и одним из главных инициаторов проекта Национального корпуса русского языка. Примерно 25 лет назад он организовал в Независимом математическом университете своего рода конкурс лингвокомпьютерных проектов. Победил в этом конкурсе проект корпуса русского языка со снятой вручную грамматической омонимией[1], когда одинаковые на вид формы слов вручную размечаются как разные, что обеспечивает высокую точность при поиске. Победив, проект получил небольшую фи-

[1] Грамматическая омонимия – совпадение форм слов в разных грамматических позициях, например, формы винительного и именительного падежа существительных мужского рода (*стол*).

нансовую поддержку, но дело было даже не в сумме (хотя в то время любая поддержка исследований была важной), а в понимании того, что с научной точки зрения проект оказался чрезвычайно привлекательным: лингвистических корпусов в мире тогда было немного, и замысел выглядел новым и амбициозным. В результате на базе небольшого пилотного массива русских текстов был создан первый вариант корпуса, тогда работавший на поисковом движке “Яндекса”, а И.В. Сегалович стал постоянным заинтересованным участником совещаний по его расширению и совершенствованию. Но даже после того, как Сегаловича не стало, “Яндекс” остался партнёром НКРЯ[2].

В работе над НКРЯ принимали участие и специалисты других академических институтов, прежде всего Института проблем передачи информации им. А.А. Харкевича РАН (ИППИ РАН). В своё время с этим институтом сотрудничал академик, известный математик А.П. Ершов из Новосибирска, который в конце 1970-х годов, то есть ещё до широкого распространения персональных компьютеров и Интернета, сформулировал задачу создания Машинного фонда русского языка (подробнее см. [1]), а затем способствовал организации в прикладном математическом институте, каким был ИППИ РАН, лингвистической лаборатории под руководством академика РАН Ю.Д. Апресяна. В этой лаборатории удалось создать специализированный синтаксический подкорпус НКРЯ. Недавно, в 2020 г., ИППИ РАН (во главе консорциума научных и учебных институций) получил мегагрант Минобрнауки России на тотальную технологическую реконструкцию корпуса на базе новейших технологий. Общими усилиями проект был блестяще выполнен. Благодаря целевому финансированию консорциуму удалось добиться существенных успехов в совершенствовании корпуса. Как коллективных участников этого проекта отметим здесь Институт лингвистических исследований РАН, специалистов российских вузов, в первую очередь Школы лингвистики НИУ ВШЭ и филологического факультета Воронежского университета. В результате слаженной работы консорциума сегодня НКРЯ содержит более 2 млрд словоупотреблений, множество специализированных подкорпусов и поисковых возможностей.

ОБ ОСНОВНОМ ФУНКЦИОНАЛЕ НКРЯ: ПОИСК И СТАТИСТИКА

Национальный корпус русского языка – это не гомогенный ресурс, а собрание почти 20 разных подкорпусов (не считая параллельных), по которым может производиться поиск. Важнейшими являются основной корпус, включающий тексты с XVIII в. по настоящее время, и газетный, представляющий большое количество текстов нескольких центральных СМИ с начала XXI в. Каждый подкорпус – это результат сочетания научных лингвистических достижений и компьютерных технологий: и те, и другие участвуют в обеспечении сложного поиска по корпусу. Действительно, поскольку в корпусе можно найти не только слово в определённой форме или все формы слов (как в любом поисковике), но и *все возможные* слова, выступающие в определённой форме (скажем, *все* глаголы в повелительном наклонении 2 лица множественного числа, как: *садитесь*, *проходите, устраивайтесь*, *(не) стесняйтесь* и проч. и проч. – заранее их список исследователю не может быть известен), в него встроена русская грамматика, на которую может опираться такой поиск, а также разметка текстов, к которой он апеллирует.

Правила (и разметка) должны быть технологичны – для удобства поисковика, но одновременно просты и наглядны, чтобы быть доступны для пользователя, в том числе не обладающего профессиональной подготовкой теоретического лингвиста. На рисунке 1 представлен фрагмент расширенной грамматической таблицы с чек-боксами для выбора нужных значений (и цепочек таких значений), в которой достаточно легко ориентироваться, задавая поисковые параметры. Внутри основного подкорпуса можно выбирать параметры текста: автора, тематику, жанр и проч.; для этого разработчики предварительно проводят разметку текстов по этим параметрам (метаразметка). Таким образом, можно выявлять особенности употребления лексики, например её частотность, в разных жанрах и типах текстов, а также у разных авторов.

Скажем, по основному корпусу буквально за несколько минут удаётся определить, насколько часто те или иные авторы используют частицы в своих прозаических произведениях. Для этого надо выбрать тексты автора в качестве подкорпуса, задать по этому подкорпусу поиск слов, принадлежащих к классу частиц, а потом в графе “Статистика” посмотреть результат – долю частиц из расчёта на миллион словоупотреблений. Задав такой поиск для основных писателей XIX в., мы почти мгновенно построили условную шкалу от минимального значения к максимальному. Она выглядит так: Пушкин – Герцен – Лермонтов – Толстой – Чехов – Гончаров – Тургенев – Салтыков-Щедрин – Достоевский. Литературоведы или специалисты по языку писателей могут проинтерпретировать эти результаты, принимая во внимание тенденцию роста употребления частиц в русском языке в целом и в связи с формированием сложной прагматической системы русского языка, а также оценить эмоциональность языка Достоевского, прозрачность прозы Пушкина и т.д. Сам корпус, конечно, никаких интерпретаций не даёт, но, как видим, он обеспечивает исследователям доступ к очень ценным и нетривиальным данным.

[2] В частности, неоценимую помощь корпусу оказали сотрудники “Яндекса” А.И. Зобнин и И.И. Виноградова.

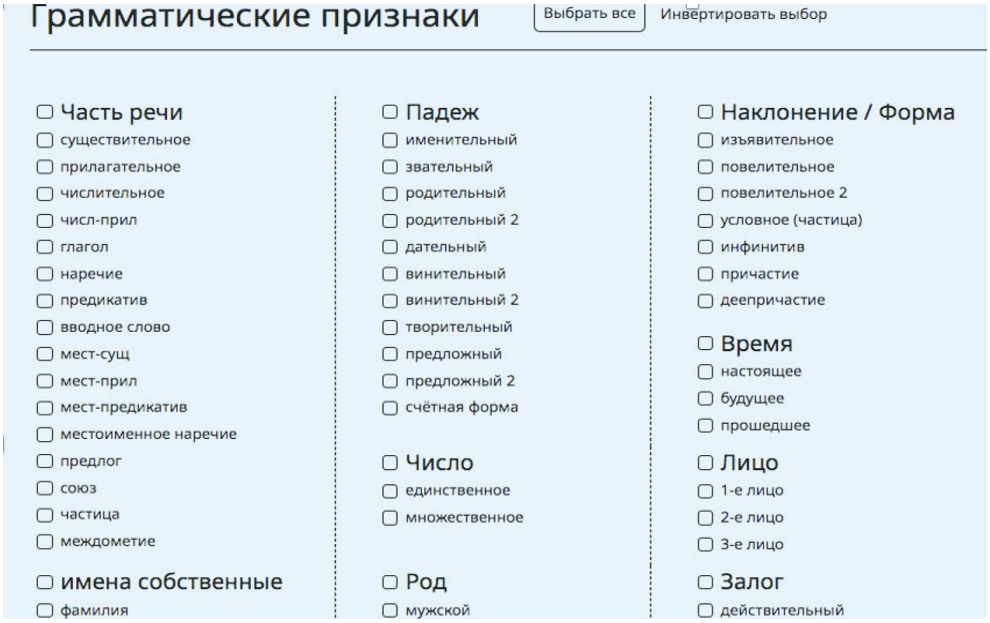

Рис. 1. Таблица параметров грамматического поиска

Ключ	Коллокат	Совместная частота	Частота ключа	Частота коллоката	LogDice	Loglikelihood	MI3	t-score	Агр. мера	Конкорданс
точный	диагноз	58	2302	3762	10.04	794.37	15.95	7.61	11.02	Примеры
точный	адрес	178	2302	27599	9.57	2135.81	17.32	13.33	15.19	Примеры
точный	слепок	12	2302	624	9.20	169.46	13.02	3.46	6.71	Примеры
точный	математически	11	2302	640	9.10	152.84	12.73	3.32	6.50	Примеры
точный	прогноз	33	2302	7224	9.03	371.02	13.60	5.74	8.83	Примеры
точный	индикатор	12	2302	1767	8.87	144.34	11.98	3.46	6.44	Примеры
точный	подсчет	17	2302	3515	8.86	192.96	12.33	4.12	7.13	Примеры
точный	перевод	94	2302	31260	8.82	980.76	15.28	9.68	12.16	Примеры

Рис. 2. Статистика коллокаций с прилагательным *точный*

Статистика – сильная сторона НКРЯ, особенно она важна для оценки частотности коллокаций[3]. Сейчас для коллокаций в графе “Статистика” исследователь имеет доступ к нескольким метрикам, которые позволяют уточнять полученные количественные оценки в разных техниках (рис. 2).

Статистику можно визуализировать. В новом интерфейсе НКРЯ появилась страница “Портрет слова”, где собирана вся информация, касающаяся интересующей пользователя лексической единицы, включая визуализации. Там можно увидеть не только подробную картину частотности сочетаемости, например, слова *точный* – в разных типах сочетаний (это так называемые скетчи – ключ поиска, например, с определяемым существительным – *точная копия*, с наречием – *математически точный* и т.п.), но и похожие на него слова (*безошибочный, чёткий, правильный, подробный* и др.), а также график исторического изменения его частотности (постепенный рост употреблений начиная с середины XIX в. с небольшим снижением к современному периоду).

ТЕХНОЛОГИЧНЫЕ ПОДКОРПУСА

Отдельно следует остановиться на высокотехнологичных подкорпусах НКРЯ, требующих особенно сложной разметки. Начнём с параллельного под-

[3] Коллокация (фразеологическое сочетание) – словосочетание, имеющее признаки синтаксически и семантически целостной единицы, в котором выбор одного из компонентов осуществляется по смыслу, а выбор второго зависит от выбора первого (например, *ставить условия* – выбор глагола *ставить* определяется традицией и зависит от существительного *условия*, при слове *предложение* используется другой глагол – *вносить*).

корпуса, точнее целого семейства (около 30) подкорпусов текстов-переводов с русского и на русский общим объёмом около 180 млн слов. Понятно, что в первую очередь речь идёт о переводах на и с крупных европейских языков: английского, немецкого, французского. Это достаточно большие массивы текстов, выровненных по предложениям: каждому предложению с помощью особой программы сопоставлен его перевод. При поиске лексемы, грамматического значения или сочетания мы получаем доступ к переводным эквивалентам, причём, как правило, от лучших переводчиков. Легко выяснить, что трудное русское слово *вообще-то* переводится на английский (если переводится – часто оно опускается) как *actually, indeed, in fact, really, to tell the truth* и многими другими способами. В НКРЯ есть параллельные подкорпуса практически для всех славянских языков (белорусского, болгарского, польского, сербского, словенского, украинского, чешского). Помимо больших языков представлены средние и малые (бурятский или хакасский), а также языки со сложной морфологией (финский, эстонский) и сложной графикой (армянский, хинди, японский, корейский, китайский): разметка, выравнивание и морфологический анализ этих корпусов требуют дополнительных трудоёмких исследований[4].

Другой подкорпус, который потребовал вложения больших ресурсов – это мультимедийный корпус русского языка (МУРКО), созданный специально для изучения жестикуляции, сопровождающей речь. Такие исследования в последние годы стали популярны во всём мире: оказалось, что автоматическое порождение речи, поразительные успехи в котором были достигнуты усилиями специалистов по автоматическому анализу звучащей речи, неполно, если оно не воспроизводит естественную интонацию и жестикуляцию. Работы по документации русской жестикуляции в НКРЯ начались давно, около 15 лет назад. В итоге был создан корпус видеоресурсов (прежде всего кинофильмов и публичных выступлений), разделённых на последовательность клипов, каждому из которых придан соответствующий фрагмент сопровождающего его устного текста. Текст расшифрован, так что по нему можно найти те жесты, которые сопровождают, например, частицу *вот* или конструкцию *иди сюда!* Более того, несколько фильмов и выступлений были размечены по жестам, так что возможен и обратный поиск, то есть поиск языковых выражений, которые соответствуют, например, поднятому вверх указательному пальцу правой руки или движению головы снизу вверх. Этот корпус не имеет аналогов в мире; он послужил основой для уникальной монографии "Русская жестикуляция", написанной его создателем Е.А. Гришиной на материале интереснейших новейших данных [3, 4].

Наконец, несколько слов о подкорпусе поэтических текстов, в котором собрана вся русская поэзия – от Кантемира и Ломоносова до Бродского и Гандлевского, причём работа по его пополнению современными поэтическими текстами продолжается. Однако не сам по себе объём корпуса (хотя 13 млн слов для коротких стихотворных строчек – это действительно очень много) является в этом проекте поразительным: помимо обычной грамматической, корпус снабжён специальной стиховедческой разметкой (по строфике, метрике, рифме и т.п.), соответственно, в нём можно осуществлять поиск по этим параметрам, прослеживая пути развития русской поэзии и сравнивая их с общемировыми. Например, можно задать поиск по слову в позиции рифмы, скажем, чтобы узнать, какая рифма, когда и кем из поэтов (и в каком стихотворении) была придумана к чрезвычайно трудному для рифмовки слову *вечером* (настолько трудном, что вплоть до начала XX в. поэты ставили в позицию рифмы только его близкий синоним *ввечеру*, ныне устаревший). Ответ на такого рода загадки корпус выдаёт мгновенно: первым зарифмовал эту форму Андрей Белый ("Из бисерных высот", 1902): *вечером – глетчеров*. Маяковский (1928) использует менее точную, но тоже нетривиальную глагольную рифму: *вечером – увековечили*. Корпус показывает, что было довольно много попыток зарифмовать *вечером* с прилагательным или причастием: *вечером – клетчатыми*, как у И. Эренбурга (1915), *вечером – встреченными*, как у С. Парнок (1915), *вечером – незамеченным*, как у А. Кусикова (1920) и, наконец, с местоимением *вечером – нечего*, как у А. Введенского (1920)[5].

Другой интересный сюжет – связь строфики с метрикой. Хорошо известно, что сонеты (которым присуща особая строфика) пишутся пятистопным ямбом. Здесь связь строфики с метрикой прямая и непосредственная. Однако корпус позволяет обнаружить и сделать предметом изучения нестандартные сонеты (существуют сонеты, написанные, например, хореем). Без корпуса такого рода исследования невозможны или требуют огромных затрат времени, он стал для стиховедов общедоступным уникальным инструментом, который существенно продвигает и *формализует* эту область, превращая её в полноценную науку.

У поэтического корпуса есть и другая важная роль – организующая. Как и в случае с нестандартными параллельными корпусами, в особенности мультимедийным, поэтическая разметка сама по себе является результатом большой теоретической работы по стиховедению, а терминологический указатель к корпусу поэтических текстов, вывешенный на сайте корпуса, – компактным онлайн учебником по этому предмету, созданным ведущими стиховедами страны [6].

[4] Подробнее о параллельных корпуса в составе НКРЯ на современном этапе см., например [2].

[5] Подробнее об этой нестандартной рифме см. [5].

НКРЯ И ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

В 2023 г. в НКРЯ началась разметка данных при помощи искусственного интеллекта. Это колоссальный прорыв, который потребовал перестройки всего корпуса (недаром проект назывался НКРЯ 2.0), перехода на новую платформу и новый интерфейс. Стали доступны новые опции, например, синтаксическая разметка текстов и поиск по коллокациям, разнообразная статистика, визуализации и прочее. Фактически создана компьютерно-лингвистическая платформа нового поколения как основа национальной справочно-информационной системы по русскому языку. Возможности этой поисковой системы чрезвычайно разнообразны, но здесь хотелось бы выделить один выдающийся результат, о котором разработчики мечтали еще 20–25 лет назад. Тогда в осуществление этой мечты поверить было невозможно.

Речь идёт о создании так называемого панхронического поиска – ресурса, который технологически объединил основной и исторические корпуса. Основной корпус до сих пор существует как самостоятельный подкорпус и представляет собой, как мы уже говорили, собрание текстов разных типов и жанров с начала XVIII в. и до наших дней – то есть за три с лишним столетия. Конечно, за это время язык существенно изменился, мы говорим иначе, чем Ломоносов, Пушкин и даже Толстой. Тем не менее заинтересованный читатель может понять (и перевести на современный язык, упрощая и не обращая внимание на детали) подавляющее большинство текстов трёхвековой давности. Более старые тексты – старорусские (XV–XVII вв.) и древнерусские (XI–XVI вв.) были выделены в отдельные, исторические, подкорпуса: они доступны в основном специалистам, обычному человеку их читать трудно. Ввиду произошедших языковых изменений – в грамматике, лексике и орфографии – без специальной подготовки сложно не только отождествить форму слова, но иногда и определить саму лексему. Поэтому у русистов, специалистов по современному русскому и русистов-историков, корпусный инструментарий был разным: синхронные данные обычно не становились предметом интереса историков, а синхронистам была неизвестна глубокая история.

С возникновением панхронического корпуса произошла в подлинном смысле революция: теперь можно проследить (причём начиная с самых первых письменных памятников) развитие любого слова, конструкции, приставки, грамматической формы и т.д. вопреки тому, что изменились орфография, грамматика и язык в целом. Это может сделать любой пользователь, не обязательно профессиональный историк языка (правда, потом всё равно придётся разбирать полученные примеры из летописей и грамот). Мы задаём поиск в современной орфографии, а получаем *все* примеры, включая древние. Иллюстрацией может быть рисунок 3, где показана выдача по запросу к панхроническому корпусу на глагол *падать*. Показаны начальные примеры из 37 365 найденных.

Таким образом, панхронический корпус представляет всю тысячелетнюю историю развития

Рис. 3. Примеры на глагол *падать* (первые по времени) в панхроническом корпусе

русского языка и, как универсальный инструмент, до определённой степени стирает границы между синхронными и диахроническими исследованиями, создаёт основу для сотрудничества и совместных проектов разных лингвистических направлений русистики. В частности, исторические данные русского языка становятся доступным и ценным материалом для типологических обобщений по языковым изменениям и сдвигам значений слов, особенно если учесть значимую историческую дистанцию в тысячу лет. Развитие этого ресурса обещает многие научные открытия.

ПОЛЬЗОВАТЕЛЬСКИЙ ПОТЕНЦИАЛ НКРЯ

Мы уже достаточно обосновали интерес, который проявляют к корпусу исследователи языка (русисты, типологи, историки, литературоведы и стиховеды) благодаря его научному потенциалу – справочному и статистическому. Другой важный класс пользователей корпуса – это, конечно, преподаватели русского языка. Для них и их учеников, школьников и студентов, разработан специальный *обучающий подкорпус*, в который вошли тексты школьной программы по словесности, специально размеченные в строгом соответствии со школьными учебными стандартами (которые, по понятным причинам, в некоторых случаях отступают от академических, существенно их упрощая).

Обучающий подкорпус значительно меньше НКРЯ как такового и имеет более скромный функционал, но школьные задания (вплоть до олимпиадных) можно с его помощью выполнять. Ученики, будучи продвинутыми пользователями самых разных обучающих ресурсов, легко справятся с этой работой. Русский язык как объект изучения (а в определённой степени и русская литература) получает своего рода микроскоп, под которым можно с интересом рассматривать грамматику, лексику, сочетаемость слов, стилистику языка. Со своей стороны учитель может быстро, буквально одним нажатием кнопки, подбирать нужные примеры для упражнений, контрольных и домашних заданий – при необходимости из произведений того автора, с творчеством которого ученики знакомятся на уроках литературы. Так что для учителя обучающий корпус – полезный помощник. Чтобы усилить эту помощь, разработчиками в сотрудничестве с опытными педагогами и преподавателями педагогических вузов создаются специальные методические материалы для школ и вузов. Развивается специальный портал STUDIORUM (https://studiorum.ruscorpora.ru), где эти материалы собираются.

Отдельная задача – разработка корпуса для преподавателей русского языка как иностранного. Помимо источника примеров, которые можно получать не только из текстов, но и из видеофрагментов мультимедийного корпуса с короткими характерными диалогами, НКРЯ обеспечивает доступ к новым онлайнресурсам по русскому языку – небольшим словарным и фразеологическим базам, включая мобильную версию, которые так необходимы иностранцам, изучающим русский язык. Совсем недавно сотрудники Школы лингвистики НИУ ВШЭ разработали два таких ресурса,

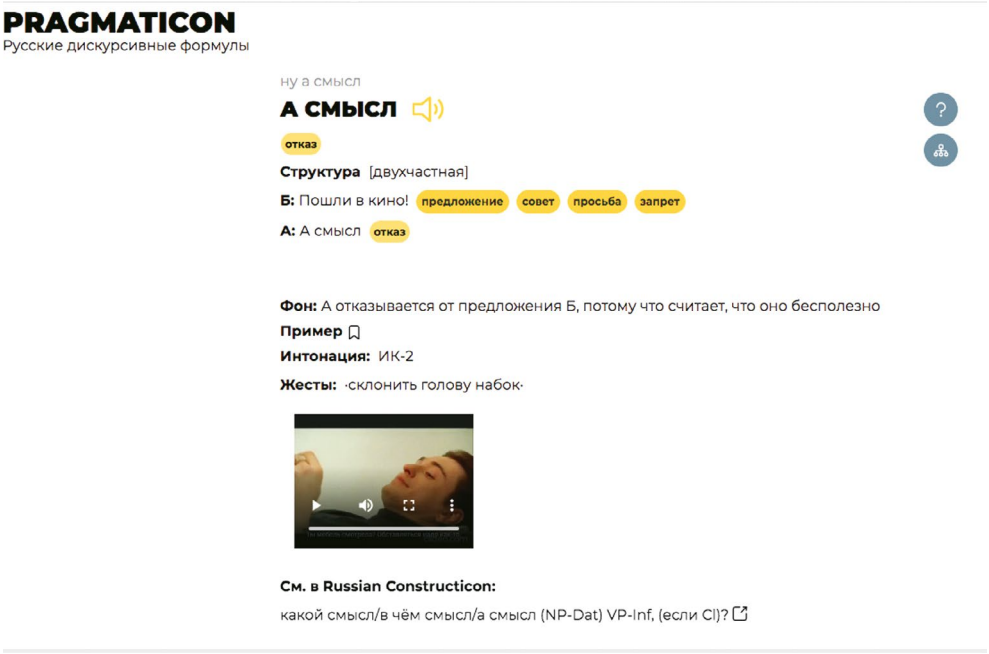

Рис. 4. Описание дискурсивной формулы в учебном ресурсе “Прагматикон”

сопряжённых с НКРЯ, – Русский конструктикон[6] и Прагматикон[7]. В первом собраны, описаны и проиллюстрированы характерными примерами на базе НКРЯ самые простые грамматические конструкции русского языка (4 тыс. единиц), например, со значениями большого количества (ср. *куча денег*, *потоки писем, реки людей*), сравнения (ср. *молодец, не то что ты*), многократного действия (ср. *вспоминал на каждом шагу*) и др. Второй ресурс включает так называемые дискурсивные формулы [7], а именно, неоднословные реплики, которые в русском языке используются в значении *да* (согласие, подтверждение) или нет (отрицание, отказ), ср.: *вот это да! ну и ну! а ты как думал? без проблем, ни в коем случае, ни под каким видом, да ладно (тебе)*! и многие другие – более 600 единиц. С помощью НКРЯ описана их интонация, сопутствующая жестикуляция, условия употребления и приведены примеры, в том числе в виде фрагментов из мультимедийного корпуса, где видна жестикуляция и слышна интонация, с которой произносится формула.

Третий класс пользователей НКРЯ – это специалисты в области IT. Проект традиционно сотрудничает с компанией “Яндекс”, однако в рамках работы над корпусом создаются открытые коллекции выверенных датасетов для машинного обучения, и эта работа будет продолжена.

ПЕРСПЕКТИВЫ НКРЯ

Современный онлайн ресурс не может не развиваться – иначе он умирает. Во-первых, программное обеспечение в современной реальности достаточно быстро устаревает, нуждается в оптимизации и постоянной поддержке. Во-вторых, обновляются передовые технологии, и то, что вчера было современным, безнадёжно устаревает. Кроме того, необходимо соответствовать мировому уровню таких ресурсов, как НКРЯ, то есть национальных корпусов разных языков, а ещё лучше их превосходить.

В первую очередь нуждается в развитии глубокая синтаксическая нейроразметка, внедрение которой только началось в НКРЯ и сопряжено с постоянной работой над системными ошибками и оптимизацией нейросетей. Другая задача искусственного интеллекта применительно к корпусу – семантическая нейроразметка, оставшаяся в планах разработчиков: она позволила бы существенно улучшить поиск и открыла бы новые технологические возможности для русской лексикографии, пока ещё серьёзно отстающей от мирового уровня. Как видим, работа над корпусом далеко не завершена: русский язык настолько разнообразен и богат, что охват его современным инструментарием требует неустанных усилий.

Если говорить о пополнении и расширении корпуса, то в первую очередь необходимо существенно увеличить объём исторических данных и пополнять исторические корпуса. Это сложная работа, масштабная и часто ручная. Хотелось бы, чтобы корпус охватывал все значимые тексты древнерусского периода – истоки русской культуры и литературы, но также и по возможности всё разнообразие текстов XV–XVII вв., когда наблюдалась наибольшая вариативность норм русского языка. Однако и XVIII–XIX вв. представлены в корпусе ещё недостаточно: тексты газет, в том числе провинциальных (очень трудных для обработки ввиду плохой сохранности), журналов, писем, записок, сочинений писателей-любителей очень важны для полноты картины русского языка: опыт показывает, что часто именно они фиксируют разговорные конструкции своего времени, позже утраченные. Очень важно и пополнение подкорпуса устной, то есть современной разговорной речи, и банка жестикуляций. Мы уже говорили о корпусе русской поэзии, который ждёт и пополнения современными текстами, и дальнейшей работы по совершенствованию разметки как основы для автоматического определения метра; об этом в своё время мечтали академик-математик А.Н. Колмогоров и член-корреспондент РАН, выдающийся филолог М.Л. Гаспаров. Имеется задел по параллельным корпусам малых языков и языков с нетривиальной графикой или морфологией. Его необходимо реализовывать, одновременно пополняя уже имеющиеся большие переводные коллекции.

Пополнение и развитие корпуса русского языка нуждается в новых подкорпусах. Это корпуса русского языка так называемых нестандартных носителей, например, русскоговорящих детей разного возраста, включая письменную речь старших дошкольников и младших школьников. Устные детские тексты, начиная с первых слов и конструкций, представляют онтогенез русской речи (этот процесс – становления речи конкретного ребёнка – обычно сопоставляют с филогенезом, то есть со становлением языка в исторической перспективе). Помимо прочего, письменные тексты позволяют отслеживать типичные ошибки, что важно для обучения грамотному письму. Для решения этих задач необходимы коллекции больших данных, и НКРЯ может их аккумулировать.

Ещё один тип нестандартного русского языка связан с географией. Русский язык развивается на разных территориях, как английский в Индии или Австралии, США, Шотландии, Ирландии или испанский в Южной Америке. Это особые территориальные варианты языка, тесно взаимодействующие с каким-то другим языком и нуждающиеся в изучении. Существование таких вариантов говорит о силе языка, который получает новые стимулы для развития. Статус этих вариантов близок к статусу диалектов (диалектный подкорпус в рамках НКРЯ имеется), средства сбора и документации подобных вариантов до определённой степени уже разработаны. Хорошо

[6] https://constructicon.ruscorpora.ru
[7] https://pragmaticon.ruscorpora.ru

известны казахский русский, армянский русский, дагестанский русский и т.п. Локальные работы по сбору и анализу таких данных ведутся в рамках небольших проектов (в частности, в Высшей школе экономики), однако нужна концентрация этих усилий и новые Владимиры Ивановичи Дали, которые, пользуясь новыми технологиями НКРЯ, обнародовали бы особенности русского языка в этих его ипостасях.

* * *

• Национальный корпус русского языка имеет *стратегическое значение* для сохранения и документирования русского языка.

• НКРЯ спроектирован как универсальный справочный инструмент по изучению русского языка для всего мирового сообщества лингвистов, гуманитариев и педагогов, специалистов в области искусственного интеллекта.

• Чрезвычайно важно, что этот центральный ресурс мировой русистики находится в России, а его технологическая основа современна и рассчитана на перспективное развитие.

• Проект такого уровня, масштаба и значимости нуждается в постоянной поддержке в рамках госзадания: он требует совершенствования и развития технологий, а также пополнения новыми сложно организованными текстами и их лингвистической обработки.

ЛИТЕРАТУРА

1. *Андрющенко В.М.* Концепция и архитектура Машинного фонда русского языка. М.: Наука, 1989.

   Andryushchenko V.M. Concept and architecture of the Machine Fund of the Russian language. Moscow: Nauka, 1989.

2. *Сичинава Д.В.* Параллельные тексты в составе Национального корпуса русского языка: новые языки и новые задачи // Труды Института русского языка им. В.В. Виноградова. 2019. № 3 (21). С. 41–61.

   Sichinava D.V. Parallel texts in the National Corpus of the Russian Language: new languages and new tasks // Proceedings of the V.V. Vinogradov Russian Language Institute. 2019, no. 3 (21), pp. 41–61.

3. *Гришина Е.А.* Русская жестикуляция с лингвистической точки зрения / Под ред. С.О. Савчук. М.: ЯСК, 2017.

   Grishina E.A. Russian gestures from a linguistic point of view / Ed. S.O. Savchuk. Moscow: YASK, 2017.

4. *Rakhilina E., Cienki A.* 2024. Creation and analysis of the multimedia Russian corpus for gesture research // A. Cienki (ed.). The Cambridge Handbook of Gesture Studies. Cambridge: Cambridge University Press, 2024. Pp. 249–272.

5. *Плунгян В.А. Ввечеру*: о слове, которому поэзия продлила жизнь // Ред. В.А. Плунгян, Л.Л. Шестакова. Корпусный анализ русского стиха. М.: Азбуковник, 2013. С. 68–80.

   Plungyan V.A. In the evening: about the word whose life poetry extended // Ed. V.A. Plungyan, L.L. Shestakova. Corpus analysis of Russian verse. Moscow: Azbukovnik, 2013. Pp. 68–80.

6. *Гришина Е.А., Корчагин К.М., Плунгян В.А., Сичинава Д.В.* Поэтический корпус в рамках Национального корпуса русского языка: общая структура и перспективы использования // Ред. В.А. Плунгян. Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 71–113.

   Grishina E.A., Korchagin K.M., Plungyan V.A., Sichinava D.V. Poetic Corpus within the Framework of the National Corpus of the Russian Language: General Structure and Prospects of Use // Ed. V.A. Plungyan. National Corpus of the Russian Language: 2006–2008. New Results and Prospects. St. Petersburg: Nestor-History, 2009. Pp. 71–113.

7. *Bychkova P., Rakhilina E.* 2023. Towards pragmatic construction typology: The case of discourse formulae // A. Barotto, S. Mattiola (eds.). Discourse Phenomena in Typological Perspective. Amsterdam: John Benjamins, 2023. Pp. 35–63.

ON RUSSIAN NATIONAL CORPUS

E.V. Rakhilina[a,b,*]

[a]*National Research University Higher School of Economics, Moscow, Russia*
[b]*Vinogradov Institute of the Russian Language of the Russian Academy of Sciences, Moscow, Russia*
**E-mail: rakhilina@gmail.com*

The article describes the project of Russian National Corpus (RNC) – a powerful reference and information system in Russian language, created by a consortium of institutions belonging to the Russian Academy of Sciences and with the active participation of Russian IT-company Yandex. The history of the Corpus is presented in great detail: the author comments upon its main functionality and the most technologically advanced subcorpora – poetic, parallel, multimedian, providing examples of their use. Special attention is paid to the latest developments which allow us to introduce modern AI technologies in the RNC; this work was supported by a grant from the Ministry of Education and Science of the Russian Federation. One of the most impressive results is the so-called "panchronic corpus", which encompasses the thousand-year history of the Russian language and provides searching tools within this data array. As of now, RNC is a crucial support for scientific research both in the field of linguistics and philology, as well as for the methodology of teaching Russian as first and second language and in the domain of IT technologies.

Keywords: Russian National Corpus, Russian language, IT technologies, corpus linguistics.