Национальный корпус русского языка: перспективы использования в лингвистических исследованиях и в преподавании // *Вестник Азиатско-Тихоокеанской ассоциации преподавателей русского языка и литературы*. № 2-3 Владивосток, 2011. С. 62-67

Савчук С.О.
Институт русского языка
им. В.В. Виноградова РАН
Москва

НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА: ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ В ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ И ПРЕПОДАВАНИИ*

В докладе раскрываются возможности Национального корпуса русского языка как инструмента лингвистических исследований, даются рекомендации по его использованию в практике преподавания русского языка.

Национальный корпус русского языка, лингвистическая аннотация, корпусная лингвистика, преподавание русского языка

Национальный корпус русского языка (НКРЯ) – мощный лингвистический ресурс, предназначенный для научных исследований и преподавания филологических дисциплин. Для того чтобы корпус стал, как того бы хотелось его создателям, необходимым и обязательным инструментом филолога и использовался максимально эффективно, необходимо представлять себе принципы его устройства и функциональные возможности.

В корпусе выделяется три составляющих: 1) коллекция текстов в электронной форме; 2) система лингвистической аннотации; 3) поисковая машина, обеспечивающая выполнение поисковых запросов. Каждая из этих составных частей может существовать либо как самостоятельный ресурс (текстовая коллекция как электронная библиотека, система лингвистической аннотации как особым образом представленное описание языка), либо в составе других информационных систем (поисковая машина как неотъемлемая часть любой информационной системы). Соединение этих компонентов в единое

* Работа выполнена при поддержке: РФФИ (10-06-00151а), Программы ОИФН РАН «Генезис и взаимодействие социальных, культурных и языковых общностей», Программы Президиума РАН «Корпусная лингвистика».

целое создает то, что называется электронным корпусом текстов в современном смысле этого слова. Рассмотрим роль каждого из компонентов в составе корпуса.

1. Коллекция текстов. Размер и состав коллекции зависит от целей и задач составителей корпуса. Национальный корпус живого языка с развитой литературной традицией не может содержать всех текстов, созданных на этом языке, но он должен быть представительным, то есть с максимальной полнотой отражать речевого употребление во всем разнообразии функциональных разновидностей. Строгой «формулы» репрезентативности не существует, и разработчики корпусов обычно руководствуются общими соображениями, такими как статистика производства и потребления письменных текстов, культурная значимость тех или иных жанров и авторов, наконец, практическая доступность тех или иных конкретных текстов. Таким образом, при составлении представительной коллекции текстов нельзя обойтись без знаний о функциональном устройстве современного языка и его истории, о соотношении литературного языка с нелитературными сферами (территориальными, социальными, профессиональными диалектами), об истории литературного языка и ключевых этапах литературного процесса и пр. В свою очередь, исследования, проводимые на основе корпусных данных, имеют обратное влияние на эти знания, расширяя и уточняя их.

Национальный корпус русского языка объединяет самые разные тексты: прежде всего прозаические оригинальные тексты, представляющие русский литературный язык (с начала XVIII века), переводные сочинения (параллельно с оригиналом), поэтические тексты, устную речь (записи публичной и непубличной речи и речи в кино), а также тексты, представляющие нелитературную форму современного русского языка – диалектную.

Основанием для организации текстов в подкорпуса служат особенности разметки и поиска. Если определенный массив текстов имеет единую разметку и тип поиска, он объединяется в самостоятельный модуль и располагается на отдельной вкладке на сайте ruscorpora.ru. Таким образом, RUSCORPORA – это

не один корпус, а несколько.

2. Лингвистическая аннотация, или разметка состоит в том, что каждая словоформа текста сопровождается в корпусе тегами, содержащими лингвистические сведения об этой словоформе. Объем этих сведений зависит от типа аннотации и ее глубины. Большинство корпусов размечены *морфологически* – это означает, что в тегах указана лексема (словарная форма), которой принадлежит данная словоформа, информация о части речи и набор грамматических признаков. Например, в НКРЯ словоформа *книгою* сопровождается следующим набором тегов: {книга=S,f,inan=ins,sg}, что в данном случае совпадает со стандартным морфологическим разбором существительного в курсе русского языка: начальная форма *книга*, существительное среднего рода, неодушевленное, в форме творительного падежа, единственного числа. Подробнее о морфологическом стандарте в НКРЯ см. [5, с. 111-135].

Если в корпусе используется *семантическая* разметка, то теги, сопровождающие словоформу, содержат сведения о семантике слова. В НКРЯ для каждого слова указывается таксономический класс [4, с. 159]: для предметных имен – это лица, животные, растения, вещества и материалы и др.; для прилагательных – размер, форма, цвет, вкус, запах, место, время, свойство человека и др.; для глаголов – движение, эмоция, речь, поведение человека и др.). Помимо этого размечены деривационные классы (уменьшительные, увеличительные, отглагольные, отадъективные имена и пр.).

В корпусах устной речи используется аннотация, отмечающая место ударения в слове. В мультимедийном корпусе размечены сочетания согласных, позиции гласных и др., что существенно для постановки и решения фонетических и орфоэпических задач, кроме того, используется специально разработанная аннотация жестов и речевых действий [2].

Метатекстовая аннотация – это информация, приписанная каждому тексту в целом. Она характеризует текст как с помощью «внешних» признаков (автор, его пол, возраст, дата создания текста, целевая аудитория, место и время

издания), так и собственно текстовых категорий (сфера функционирования, жанровая и стилистическая принадлежность, тематика) [6]. Метатекстовая разметка отражает архитектуру текстового состава корпуса и служит для формирования пользовательских корпусов.

Следует иметь в виду, что правила (стандарты) аннотации в разных системах могут отличаться друг от друга, что объясняется возможностью множественных интерпретаций одного и того же языкового явления. Например, в грамматической теории не существует единого мнения о характере видовых противопоставлениях глаголов в русском языке, поэтому в одних системах видовая пара может рассматриваться как реализация одной лексемы (*убедить* и *убеждать* как формы *св* и *нсв* глагола *убеждать*), а в других системах – *убедить* и *убеждать* рассматриваются как разные глаголы.

3. *Система поиска*. Система лингвистической разметки превращает библиотеку текстов в подобие гигантской картотеки, в которой каждая словоформа расписана на несколько «карточек» – по числу лингвистических признаков, приписанных словоформе. Для извлечения информации из этой картотеки в корпусе используется *поисковая система*, которая позволяет сортировать «карточки» по любому из признаков. Чем более функциональна система поиска, тем большее число лингвистических задач можно решать с помощью корпуса. «Дружественный» интерфейс, понятная логика и удобство формирования запросов обеспечивают скорость решения поставленных задач. Например, при обычном контекстном поиске результаты выдаются в виде предложений, которые в дальнейшем подлежат ручной обработке. При поиске по ключевым словам контексты обрабатываются автоматически и результаты выдаются в виде конкорданса, что в большинстве случаев значительно облегчает и ускоряет обработку данных.

Система лингвистической аннотации каждого типа отражает систему правил соответствующих разделов науки о языке. Поэтому корпус можно эффективно использовать при изучении лингвистических дисциплин, прежде всего тех, категории которых были использованы в аннотировании текстов.

Покажем это на примере морфологической аннотации.

Изучение курса морфологии русского языка в школе и в вузе начинается с теоретического освоения системы правил (например, состава частей речи, системы категорий имени существительного) и завершается выработкой умения выполнять морфологический анализ текста, при котором каждой словоформе приписываются присущие ей грамматические признаки.

Для пользователей корпуса ситуация выглядит перевернутой. Размеченный (морфологически проанализированный) текст для них – не цель, а данность, так же как и набор морфологических категорий и набор инструментов поиска, с помощью которых можно строить запросы. Запросы могут быть трех типов: 1) относительно словоформ, слов и словосочетаний (напр., употреблялась ли форма *городы* в текстах XVIII в.); 2) относительно грамматических признаков (напр., актуально ли употребление форм 2-го родительного падежа для текстов современного периода); и 3) относительно различных комбинаций лексем и грамматических признаков (напр., проверить, продуктивна ли фразеологическая модель *нужный как* кому/чему кто/что). В ответ на запросы выдаются контексты, в которых представлены языковые единицы, найденные по запрашиваемым признакам.

Корпус можно использовать для решения учебных задач, формулируя простые запросы, например: привести примеры субстантивированных прилагательных; на основе анализа найденных в корпусе примеров определить, какие синтаксические функции могут выполнять слова *холодно, интересно* и установить частеречную принадлежность этих слов и т.д. На подобные запросы корпус выдает десятки, и даже сотни примеров – понятно, насколько это облегчает труд преподавателя, упрощая составление многовариантных заданий для самостоятельных, лабораторных и проверочных работ [3; 7].

Помимо обучающих целей корпус может быть использован для решения исследовательских задач. Благодаря корпусу сбор материала перестает быть рутиной, поглощающей значительную часть времени, отпущенного на исследование, и превращается в увлекательное занятие. Исследователь имеет возможность сосредоточиться на творческих участках работы: оптимизировать

поисковые запросы, чтобы извлечь из корпуса максимум необходимой информации при минимуме ненужной; рассматривать предмет исследования в разных аспектах; обрабатывать большой объем данных для подтверждения выводов статистическими результатами. Скорость, с которой подбираются примеры, приобщает к исследовательскому процессу и студентов, которые получают возможность за короткий срок провести небольшую творческую работу. О практике использования НКРЯ в учебном процессе см. [3, 9].

Практика показала, что корпус можно использовать не только при изучении морфологии, лексикологии, семантики, синтаксиса, фонетики и орфоэпии, орфографии, но и таких разделов, как фразеология, словообразование, стилистика и культура речи, язык художественной литературы и его история. Покажем это на нескольких примерах.

Проверить продуктивность фразеологической модели можно следующим образом. Строим запрос многословного сочетания: *глагол* с семантическим признаком ‘движение и изменение положения в пространстве’ + *как* + существительное в форме им-вин. п. + *в* + существительное в форме предложного падежа. Рассматриваем контактное расположение элементов (задаем расстояние 1) и учитываем знаки препинания (в дополнительных признаках при слове *как* указываем «слово после запятой»). В результате получаем более 200 контекстов с разными глаголами указанной семантической группы: ***крутиться/кружиться/вертеться*** как *белка в колесе*, как *ось/спица в колесе*, как *колесики в часовом механизме*, как *вода/щепка в водовороте*, как *чаинки в чае*, как *белье в центрифуге*, как *мусор в вихре*; ***метаться*** как *зверь/лев/львица/тигр/тигрица/птица в клетке*, как *птица в силке*, как *заяц в западне*, как *мыши в ловушке*, как *кошка в доме;* ***тонуть*** как *пень в болоте,* как *мыши в ведре,* как *камни в воде,* как *утюг в океане* и т.д. Наряду с оборотами, зафиксированными в словарях: *ворочаться, как медведь в берлоге, крутиться, как белка в колесе, плавать, как рыба в воде, кататься как сыр в масле*, - в текстах, как видно из приведенных выше примеров, встречаются многочисленные обороты, построенные по данной модели, что свидетельствует

о ее продуктивности.

Недостаток изучения фонетики и орфоэпии в вузе состоит в том, что студенты, как правило, знакомятся с образцами произношения в виде транскрипции, а не в виде звучащей речи. Мультимедийный корпус в составе НКРЯ, снабженный орфоэпической разметкой, поможет провести наблюдения над звучащими образцами и познакомиться с такими особенностями «старшей нормы» литературного языка, как «эканье» *([в'э]нок, [б'эл'э]соватое, [м'э]тался*); произношение твердых согласных перед [э]: *бу[ɣ]а'л[тэ]р, эф[фэ]кт*; произношение сочетания *щн* как [шн] словах *беспомощн[шн]о, в сущн[шн]ости, всено'щн[шн]ого*; произношение j в формах местоимений (*им, их [jим], [jих]*) и др.

Словесное ударение традиционно считается одним из самых сложных аспектов русской грамматики. Это связано не только с его свойствами – разноместностью и подвижностью, но и с тем, что акцентная система русского языка на протяжении последних трех веков находится в процессе перестройки, причем изменения здесь происходят стремительно, порой их можно наблюдать в течение жизни одного поколения. Быстрое изменение системы служит источником появления словоизменительных и словообразовательных вариантов и вынуждает постоянно пересматривать отношение к ним с точки зрения норм литературного языка. Поэтому рекомендации грамматических описаний, словарей и справочников могут существенно отличаться друг от друга и от устной речевой практики, что создает дополнительные трудности при изучении русского языка, в особенности для тех, кто изучает его за пределами России. Акцентологический корпус, созданный в составе Национального корпуса русского языка, предоставляет возможность изучать словесное ударение не на основе словарей, а наблюдая реальные тексты. С учетом особенностей русской акцентной системы он с самого начала был задуман как диахронический. Историческая часть представлена поэзией XVIII–XX вв. – эти тексты служат в русистике традиционным источником для исследования норм словесного ударения предшествующих эпох. Современное

состояние ударения отражено в записях устной речи, в которых ударение расставлено в соответствии с реальным произношением.

Акцентологический корпус предназначен не только для специалистов по русской акцентологии. Он позволяет решать и учебно-методические задачи: его можно использовать как справочный ресурс при изучении русской грамматики, как материал для составления упражнений, учебных пособий для изучающих русский язык. Корпус погружает в стихию живой речи, наглядно представляет ее вариативность и изменчивость, позволяет получить ответ на вопрос, почему предписания нормативных пособий отличаются друг от друга, помогает более осмысленно и творчески подойти к выбору авторитетных рекомендаций. Так, например, с помощью корпуса можно выяснить, как распределяются в текстах допустимые нормой вариантные формы им.-вин. п. существительного *ветер*. Для формы *ве'тры* отмечено 305 вхождений, первое – в 1735 г. (В.К. Тредиаковский, М.В. Ломоносов). Форма *ветра'* встретилась 15 раз, первая фиксация относится к 1827 г. (Е.А. Баратынский), остальные контексты – к концу XIX в. и к XX в., при этом 10 вхождений отмечено в поэзии, 5 – в транскриптах кино. Для других слов группы, например *парус, якорь, месяц,* соотношение иное: *па'русы/паруса'* – 24/196; *я'кори/якоря'* – 5/23. Формы на *-и/-ы* сосредоточены в поэзии XVIII и начала XIX в. (в архаизованной речи), формы на *-а* появляются в первой трети XIX в. Для слова *месяц* форма *ме'сяцы* встретилась 14 раз, форма *месяца'* в корпусе на зафиксирована.

Таким образом, тексты корпуса дают динамическую картину процесса перестройки парадигмы множественного числа в группе существительных с исконным неподвижным ударением на основе, в результате которой на смену старой схеме приходит схема с ударением на окончании. Начиная с XIX в. у этих слов появляется вариант именительного падежа множественного числа с новым ударным окончанием –а', который в течение XIX-XX в. получает интенсивное распространение. В современном литературном языке лексемы *парус* и *якорь* характеризуются полностью перестроенными акцентными парадигмами, для лексем *ветер* и *месяц* процесс перестройки еще не

закончился, причем находится на разных стадиях: если оба варианта *ве’тры* и *ветра’* признаются нормативными, то для слова *месяц* в значении ‘период времени’ при допустимости форм косвенных падежей с наконечным ударением *месяца’м*, *месяца’ми*, *месяца’х* форма им. мн. *месяца’* отвергается литературной нормой.

Возможности корпуса в изучении языка художественной литературы еще совсем не исследованы. Между тем в НКРЯ содержится самая большая электронная коллекция текстов XVIII в., основной фонд классики XIX и начала XX в., богатая коллекция литературы советского периода и современной прозы. Разнообразный в жанровом отношении состав текстов в соединении с поисковыми средствами корпуса дает возможность исследовать особенности использования художественных средств в индивидуальном языке писателя, в произведениях определенного периода, литературного направления, в языке поэзии и прозы и т.д. Покажем это на примере сравнений. Сравнение считается источником фигур изобразительности; по образному выражению А. Белого, сравнения – «раскрытие глаза», они «вводят нас в мир глаза» писателя [1, с. 288]. Например, на материале корпуса можно показать, как глаз разных писателей видит один и тот же признак. Поиск сочетания «широкий + как» в подкорпусах текстов разных авторов дает нам целый ряд сравнений: *волна, широкая, как волны моря*; *луна, красная и широкая, как медный щит* (И.С. Тургенев); *рот широк, как у налима*; *мысль моя широка, как море*; *гладкая юбка, широкая, как колокол; диван широк, как двухспальная кровать* (А.П. Чехов); *море, широкое, как страсть, и страсть, широкая, как море*; *улицы, ставшие широкими, как черные блестящие моря*; *широкая, как шкап, грудная клетка* (В.А. Набоков); *борода широка, как у старого козла*; *широкая, как лошадиное копыто, ладонь*; *съел широкого, как печной заслон, чебака*; *сидел за широким, как двухспальная кровать, письменным столом* (М.А. Шолохов). Анализ подобного материала, поиск сходства и объяснение отличий – увлекательное занятие для исследователей стилистики художественной речи.

Список использованной литературы

1. Белый Андрей. Мастерство Гоголя. М.: МАЛП, 1996. 351 с.

2. Гришина Е.А. Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: Новые результаты и перспективы. СПб: Нестор-История, 2009. С. 175-214

3. Добрушина Н.Р. Корпусные методики обучения русскому языку // Там же. С. 335-351

4. Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. М.: Индрик, 2005. С. 155-174.

5. Ляшевская О.Н., Плунгян В.А., Сичинава Д.В. О морфологическом стандарте Национального корпуса русского языка // Там же. С. 111-135.

6. Савчук С.О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции. // Там же. С. 62-88.

7. Савчук С.О., Сичинава Д.В. Обучающий корпус русского языка // Национальный корпус русского языка: Новые результаты и перспективы. М.: Нестор-История, 2009. С. 317-334

8. Национальный корпус русского языка www.ruscorpora.ru.

9. Портал "Национальный корпус русского языка и преподавание" http://studiorum.ruscorpora.ru/index.php?option=com_tag&Itemid=75

Savchuk, S.O.

Institute for Russian Language Russian Academy of Sciences

Moscow

The Russian National Corpus: perspectives of its use as a tool for research and language teaching

The paper presents the main characteristics of the Russian national corpus, available on open access on the site http://ruscorpora.ru. Corpora composition, types of linguistic annotation and ways of their effective use according to their purposes are being described.

Russian National Corpus, annotated corpora, corpus linguistics, teaching Russian