Летняя школа по русской литературе. 2015. Т. 11. № 4. С. 344-351.

Дмитрий Сичинава
(Москва)

НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА В ЛИТЕРАТУРОВЕДЕНИИ: КРАТКИЙ ОБЗОР

В статье идёт речь о возможных применениях Национального корпуса русского языка в литературоведении и даётся краткий обзор нескольких конкретных примеров его использования (на материале существующих работ других авторов).

Ключевые слова: корпусная лингвистика, поэтика, метрика, литературоведение.

The paper deals with the Russian National Corpus (the RNC) as a research tool in the history and study of literature and poetic. A brief overview of its usage in papers by different authors is presented.

Key words: corpus lingustics, poetics, verse studies, history of literature.

Что в Корпусе есть

Национальный корпус русского языка (http://ruscorpora.ru, далее — Корпус с прописной буквы) существует, доступен для поиска в Интернете и развивается более 10 лет (ему посвящены сборники НКРЯ 2003—2005, НКРЯ 2006—2008 и ряд других публикаций). Это электронное собрание русских текстов (самых разных – от объявлений и уличных разговоров до «Войны и мира» и синодальной Библии), снабженное разнообразной разметкой (то есть информацией, которую можно учитывать при поиске), как на метатекстовом уровне (дата создания, жанр, тип текста, имя автора), так и на уровне слов и их сочетаний (морфологический разбор, лексико-семантическая информация и др.). Объём Корпуса – более полумиллиарда словоупотреблений. Корпус состоит из нескольких подкорпусов с разным составом и типом разметки. В так называемый «основной» корпус входят прозаические тексты от начала XVIII в. до наших дней с морфологической и семантической разметкой; в «синтаксический» — современные тек-

сты с разметкой синтаксических структур; в «газетный» — тексты современных СМИ; в «поэтический» — стихотворения разных веков с разметкой сильных мест строки, метра, типов рифмовки, строфики и т. д., в «мультимедийный» — подтекстовки видео- и аудиозаписей (вместе с самими записями, которые можно просмотреть и прослушать), в основном кинофильмов (здесь в ряде фильмов размечена информация о жестикуляции актеров), в «исторический» — тексты XI—XVII вв. и новые церковнославянские тексты, снабжённые специфическим орфографическим поиском и морфологической информацией; в «параллельный» — выровненные по предложениям оригинальные и переводные тексты на двух или более языках (один из которых русский).

Национальный корпус никогда специально не осмыслялся и не рекламировался как средство работы для литературоведов; основным его потребителем считался лингвист. Кроме того, как возможные пользователи мыслились программисты, а также люди, пишущие, воспринимающие и оценивающие тексты (журналисты, редакторы, переводчики, преподаватели русского языка для носителей и иностранцев и т. п.). Пожалуй, исключением можно было считать разве что поэтический подкорпус, состоящий только из художественных текстов и снабжённый разметкой стиховых параметров, имеющих, разумеется, очевидный лингвистический коррелят, но относимых к поэтике. Тем не менее все подкорпуса (а не только поэтический) довольно активно используются литературоведами (и «филологами» более широкого плана, чем «только лингвисты»), а также представителями иных гуманитарных дисциплин (историками, антропологами и под.).

Чего в корпусе пока нет

Прежде чем перейти к обзору этого использования, необходимо сделать ряд предупреждений «отрицательного» характера, а именно, насчет того, чего в Корпусе *нет* (и для чего исследователю надо наряду с корпусом использовать другие ресурсы). В Корпусе, как общее правило, нет различных вариантов текста и текстологической справки (за всеми подобными комментариями необходимо обращаться к изданиям). Орфографический режим, состав текста, атрибуция и датировки отражают авторитетные издания уровня «Библиотеки

поэта» и академических собраний сочинений (все дальнейшие уточнения датировок, текста или атрибуции, вносимые исследователями, попасть в Корпус могут только постольку поскольку кто-нибудь из составителей или пользователей их заметит; систематической проверки библиографии не проводится).

Нет установки и на включение в Корпус всех текстов некоторого периода вообще или всех текстов данного автора; при составлении корпусов приоритет отдаётся репрезентативности выборки, а не полному включению. И снова исключением может считаться поэтический подкорпус, где некоторые авторы представлены полными собраниями своих стихотворений (но, опять-таки, обычно без вариантов). Таким образом, если некоторое слово не встречается в Корпусе ранее 1850 и после 1950 г., это не даёт никакой гарантии того, что до и после этих дат оно в русских опубликованных (и даже оцифрованных) текстах неизвестно. Для исследований такого рода необходимо использовать вместе с Корпусом также ресурсы типа feb-web или Google Books (где, в свою очередь, как и в обычном поисковике, существенно ограничен морфологический поиск, — в частности, нельзя задавать грамматические параметры, — ограничен поиск по метапризнакам, многие тексты многократно повторяются, что сбивает статистику и пр.), а также, разумеется, в нужных случаях сверять цитаты, их контекст и все другие релевантные признаки, обращаясь к печатным изданиям.

Проблема «пожеланий» пользователя к поэтическому корпусу НКРЯ поднимается также в работе Р. Г. Лейбова [8], одним из лейтмотивов которой являются реплики персонажа, обозначенного как «Неблагодарный Пайщик». Это замечания к текущему состоянию Корпуса: соответствующие недостатки не являются для Корпуса принципиальными и могут быть более или менее легко устранены (например, несовершенство жанровой классификации, отсутствие поиска по длине текста или по заглавиям).

При всех этих «недостатках» Корпус, безусловно, многократно облегчает первичный поиск информации и оценку масштабов явления по сравнению с фронтальным просмотром текста, обращением к памяти, составлением картотеки и т. д.

Лексика

Перейдем к конкретным примерам. Например, лингвист и антрополог Н. Б. Вахтин [3] исследовал по корпусу совместную сочетаемость устойчивого эпитета *дикий* с топонимом *Сибирь*, выработку русского литературного клише, закрепляющегося за образом Сибири. «Сочетаний со словами *дикий*, *дикость* — тоже довольно много: в Сибири *дикая природа*, она населена *диким народом*, живущим на ее *диких пространствах*, исполненных *дикой красоты*» [3, с. 207]. Приводится множество находимых в Корпусе примеров этого стереотипа, от «Воспоминаний» Ф. В. Булгарина до «Интересной жизни» П. Ф. Нилина. Здесь же обсуждается и эволюция стереотипа *сибиряка* — могучего, предприимчивого, солидного, молчаливого, свободолюбивого, идеального солдата.

Корпус позволяет исследовать диахронию использования той или иной лексики. Например, Ф. М. Достоевский приписывал себе изобретение глагола *стушеваться* (в Корпус вошла и повесть «Двойник», где это слово якобы было им введено в русский язык, и «Записные книжки» Достоевского, где содержится утверждение о приоритете автора). Тем не менее Корпус находит датированный 5 февраля 1826 г. фрагмент «Дневника» А. В. Никитенко, из которого видно, что слово существовало и не требовало комментария за поколение до литературного дебюта Достоевского: *Но я знаю его, знаю, что он честолюбив, а честолюбие, сопровождаемое успехом, с каждым шагом вперед умаляет в глазах честолюбца предметы, остающиеся у него позади, и так до тех пор, пока они совсем стушуются, и он уже не видит больше ничего, кроме самого себя.* В. А. Плунгяну [12] принадлежит исследование употребления лексемы *жуть* в русском языке: встретившаяся у Лескова как модный неологизм, экзальтированное «чужое слово» (еще в кавычках), *жуть* затем становится одним из «фирменных» слов русского модернизма в стихах и прозе: *бездонная жуть мировой пустоты раскрывает себя в необъятном, всепоглощающем, черном величии* (И. А. Новиков, 1907), *сладкая волна неизъяснимой жути ожгла ему грудь* (А. Белый, 1909), *Полнится мутями Все бытие: Полнится жутями Сердце мое* (он же, 1901—1921) и др. Затем это слово становится просторечно-сниженным: *«Ой, как я плохо играю!» — думала Таня, глядя на экран. — «Жуть!»* (Аксенов, 1963).

Возможно корпусное изучение лексики и как предмета идиостиля. А. А. Белов [1, с. 57-58] изучает слово *зной* у Тютчева, и на основании статистических данных приходит к выводу о том, что это слово представляет собой поэтизм, причем особенно характерный для лирики Тютчева по сравнению с прочими поэтическими текстами:

«Судя по всему, причина, по которой частотность *зноя* в поэтическом корпусе оказывается выше, чем частотность *тепла* – практически та же самая по которой частотность *тепла* оказывается выше в общеязыковом корпусе, а именно: большая семантическая нейтральность *тепла* по сравнению со *зноем* (отсутствие у тепла семы «высокой степени проявления признака»). По этой причине *тепло* воспринимается как менее экспрессивная и оттого «менее поэтичная» лексическая единица. Как видно, частотность *зноя* в тютчевских стихотворениях превышает частотность *жара* в два раза, при том, что в поэтическом подкорпусе, напротив, совместная частотность *жара* и *жары* превышает частотность *зноя* более, чем в четыре раза (слова *жарá* у Тютчева, как уже было отмечено, нет вообще); в отношении же *тепла* наблюдается полное тождество – его частотность как в текстах Тютчева, так и в поэтическом подкорпусе предельно мала» [1; с. 57-58].

Семантическая разметка Корпуса позволяет искать лексику определенного семантического поля. Недавняя диссертация Е. С. Фоминой [13] посвящена, в частности, образам различных этносов у И. С. Тургенева. Для предварительного сбора информации по этой теме, в частности, можно выстроить подкорпус из произведений Тургенева и ищется лексика с пометой «этноним». При помощи единственного запроса обнаруживаются и *бережливые немцы,* и *чопорные англичане*, и русский герой, который *избегал русских за границей.*

Изучение стиха

Поэтический подкорпус уже стал объектом целой серии исследований, объединенных, в частности, в сборники «Корпусные исследования русского стиха» [9], [10]. Их предметом стала, например, ритмика свободного стиха [5], анализ кратких атрибутивных форм типа *красна девица* или *пластмассовы цветочки* в стихе, в том числе современном, и связь их употребления со стилем поэтического течения [6], лексика

текстов Батюшкова, созданных в период душевной болезни, на фоне поэтической лексики его времени [11] и мн. др.

Особый масштабный проект посвящен исследованию частотности лексики в «высокой», «элитарной» поэзии (включенной в Корпус) и так называемой «наивной», «непрофессиональной» поэзии (публикуемой, в частности, в Интернете на сайтах типа Стихи.Ру; этот массив ценен своим огромным объёмом и сравнительной однородностью приёмов). Изучавшие его А. А. Бонч-Осмоловская и Б. В. Орехов [2] приходят к следующим выводам:

«Мы ранжировали леммы (лексемы – *Д. С.*) по разнице в позициях в частотнике (частотном словаре. – *Д. С.*) наивной поэзии и поэтического корпуса. Выяснилось, что такие слова, которые находятся существенно выше в "высоком" списке, чем в списке наивных поэтов, интуитивно ощущаются как поэтизмы: *око, взор, меж, уста, единый, мгла, дух, бездна*. И напротив, слова, которые демонстрируют наибольший рост в списке наивных поэтов по сравнению со списком поэтического корпуса, — это характерные "прозаизмы" из сферы вещного мира (*фото, сигарета*) и быта (*проблема*). Первый факт нуждается в отдельном осмыслении. Несмотря на ожидаемое восприятие наивными поэтами традиционно-поэтического инструментария лексики, его часть оказывается, по-видимому, совершенно невостребованной. Притом, что наивные авторы явно ориентируются на классические образцы, наиболее архаичные особенности этих образцов отсеиваются и в конструируемый концепт поэтического не включаются. Механизм такого отбора чрезвычайно любопытен и должен быть исследован специально» [2].

Ряд исследований на материале поэтического подкорпуса провёл Р. Г. Лейбов, в том числе статистически исследовал «привязанность» той или иной лексики к выделенной позиции – рифме [7]. Степень зарифмованности слова связана с концептуальной нагрузкой, которую оно несёт, влиянием «конечного числа лексем, образующих вокруг имени облако потенциальных тематико-сюжетных линий» [7, с. 196]; так, слово *Полтава* в 72% случаев попадает в патриотически окрашенный рифменный ряд *слава – держава – кровавый…*, и не случайно после покорения Варшавы Суворовым в 1794 г. название столицы Польши тоже постепенно втягивается в этот ряд.

На материале снабженного грамматической разметкой поэтического корпуса возможно задавать запросы, вообще не содержащие конкретных лексем, например, ритмико-синтаксические клише (в терминологии М. Л. Гаспарова) в разных размерах. Последовательность «прилагательное-прилагательное-существительное» в 3-ст. амфибрахии выдаёт ряды похожих строк разных поэтов [4]: *И вечный, напрасный упрек…* (Ростопчина), *И серый походный сюртук* (Лермонтов), *И странные, дикие звуки* (Лермонтов), *И тихие, тихие звуки* (Мей), *И ясно речное стекло* (Бунин), *Да мило кривое окно* (он же).

Параллельные тексты

Параллельный подкорпус открывает большие перспективы для изучения истории художественного перевода, например, с точки зрения перевода культурно значимой лексики. Так, известно, что Р. Райт-Ковалева переводит слово *hamburger* у Дж. Сэлинджера как *котлета*; параллельный подкорпус показывает, что за 20 лет до нее Н. Волжина в переводе «Гроздьев гнева» Дж. Стейнбека это слово старалась либо пропускать вовсе, либо заменять другим известным читателю английским заимствованием – *сэндвич*. Слово *гамбургер* в русском языке утверждается лишь позже, с 1990-х гг.

Таким образом, Корпус представляет собой успешно применяемое подспорье в работе исследователя литературы; вероятно, как это часто бывает, появление нового инструмента позволяет ставить и новые задачи, неосознаваемые или чересчур трудоемкие при применении старых.

Список литературы и источников

1. *Белов А. А.* Лексико-семантическое поле «зной» в поэтических текстах Ф. И. Тютчева. Дисс. … канд. филол. наук. Череповец: ЧГУ, 2008.

2. *Бонч-Осмоловская А. А., Орехов Б. В.* Некоторые применения корпусных методов к наивной поэзии // Статьи на случай: сборник к 50-летию Р. Г. Лейбова. См.: http://www.ruthenia.ru/leibov_50/article_b-osm_orexov.html.

3. *Вахтин Н. Б.* От «дикости» к «другому»: эволюция образов Сибири и Севера в русском языке // Studia Russica Helsingiensia et Tartuensia XII: Мифология культурного пространства: К 80-летию Сергея Геннадиевича Исакова. Тарту, 2011. С. 203—216.

4. *Гришина Е. А., Корчагин К. М., Плунгян В. А., Сичинава Д. В.* Поэтический корпус в рамках НКРЯ: общая структура и перспективы использования // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб., 2009. С. 71—113

5. *Корчагин К. М.* Еще раз о раннем русском свободном стихе // Корпусный анализ русского стиха. М., 2013. С. 219—242.

6. *Кулева А. С.* Языковые особенности поэтического направления: усечённые прилагательные // Корпусный анализ русского стиха. М., 2013. С. 128—141.

7. *Лейбов Р. Г.* Русская слава и польская столица: к истории одного рифменного клише // История литературы. Поэтика. Кино: сборник в честь Мариэтты Омаровны Чудаковой. М., 2012. С. 187—198.

8. *Лейбов Р. Г.* Неблагодарный пайщик: опыты корпусного анализа текста // Корпусный анализ русского стиха. Выпуск 2. М., 2014. С. 48—74.

9. НКРЯ 2003—2005 — Национальный корпус русского языка: 2003—2005. Сб. статей. М., 2005.

10. НКРЯ 2006—2008 — Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб., 2009.

11. *Орехов Б. В.* «Не Аполлон, но я кую сей цепи звенья…»: поздние стихи Батюшкова в свете корпусных данных // Корпусный анализ русского стиха, М., 2013. С. 157—171.

12. *Плунгян В. А.* «Жуть» и «жуткий»: от мистицизма к просторечию // Авторская лексикография и история слов. Материалы Международной научной конференции к 50-летию выхода в свет «Словаря языка Пушкина». М., 2013. С. 145–153.

13. Фомина Е. Национальная характерология в прозе И. С. Тургенева. Диссертация на соискание ученой степени доктора философии. Тарту, 2014.