О ПЕРСПЕКТИВАХ ФИЛОЛОГИЧЕСКОГО КОРПУСА

Б.В. Орехов

Лингвистическое сообщество в целом осознало, что создание корпуса не является простым количественным фактором, ускоряющим поиск примеров. Это особая идеология, формирующая исследовательский взгляд на предмет, задающая контуры и даже тематику научных сюжетов.

Дж. Лич еще в 1992 г. настаивал на том, что словосочетание «корпусная лингвистика» отсылает не к области знания, но в большей степени к методологическому базису лингвистических исследований: «В принципе (и часто на практике) корпусная лингвистика с легкостью соединяется с другими направлениями лингвистики: мы можем изучать фонетику, синтаксис, социолингвистику и другие аспекты лингвистики с помощью корпуса, и когда мы так делаем, это можно назвать соединением корпусных методик и предмета фонетики, синтаксиса, социолингвистики и т. д.». Но за этим не следует упускать главного: корпусная лингвистика «не просто появившаяся методология изучения языка, это новая исследовательская инициатива, и на самом деле новый философский подход ⟨...⟩ Так что технология в данном случае (как это веками было в естественных науках) играет более важную роль, чем просто поддержка и упрощение процесса исследования: я вижу в этом сущность нового типа знания и путь к новому способу познания языка» [*Leech*: 105–106].

Во многом сходные мысли высказаны в 2008 г. в статье В.А. Плунгяна «Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики»[1]. В.А. Плунгян формулирует сложившуюся ситуацию со всей возможной определенностью: «Корпус в каком-то смысле вернул лингвистам их подлинный объект – тексты на естественном языке в максимально полном объеме».

Эти обеспеченные появлением корпусов очевидные успехи лингвистики заставляют поставить вопрос о создании специализированного филологического корпуса, который мог бы послужить платформой и импульсом для новых исследовательских шагов в старейшей гуманитарной науке.

Определение предмета филологии не является очевидным и допускает известную вариативность (см. [*Орехов*: 74–82]). Если

[1] Русский язык в научном освещении. 2008. № 2 (16). С. 7–20.

сосредоточиться на понимании филологии как науки и прикладной дисциплины, ориентированной на работу с культурно значимыми художественными текстами, то у нее есть два очевидных аспекта: один связан с текстологией рукописей (и вообще источников, например, позволяющих уточнить датировку), изучением редакций и вариантов произведений, второй включает в себя разного рода комментирование, толкование, т.е. создание специального научного инструментария для диалога между текстом и читателем. Первый аспект восходит к позитивистской филологии Вильгельма Шерера, предполагавшего, что научный метод в литературоведении должен ограничиваться документированием фактов (часто внешних по отношению к собственно реальности текста). Второй имеет более сложное происхождение, включающее достижения филологической науки с конца XIX по середину XX века (компаративистика, формализм, структурализм).

Если попытаться сформулировать, как эти два направления должны быть отражены в гипотетическом филологическом корпусе, то в центре внимания окажутся две принципиально различные вещи: филологически корректное представление текстов и соединение в рамках корпуса самого текста-источника и комментаторской надстройки. Первое обращает нас к уровню принципов формирования текстовой коллекции, особенностей поисковых механизмов и интерфейса, второе – к уровню разметки корпуса.

Говоря о представлении художественных текстов в корпусе приходится констатировать, что между лингвистическим корпусным подходом и традиционным филологическим существует известный антагонизм[2]. Наличные лингвистические корпуса синтагматичны, это линейным образом устроенные коллекции текстов, в которых появление дублирующих экземпляров, незначительно отличающихся вариантов, видится однозначным недостатком. Дело в том, что корпуса используются в лингвистике для подсчета частотности (слов, грамматических форм и пр.); как особенное достоинство корпусной лингвистики это отмечают и Дж. Лич, и В.А. Плунгян. В случае, если в корпусе один и тот же текст будет повторяться (пусть даже частично), это создаст неприятный перекос в количественных данных, например, некоторое слово в результатах поиска будет отображаться чаще, чем оно реально встречается в языке. Поэтому варианты одного и того же текста

[2] См. его экспликацию, напр.: *Бодрова А.С., Пильщиков И.А*. Проблемы корпусного подхода к задачам авторской лексикографии // Авторская лексикография и история слов: К 50-летию выхода в свет «Словаря языка Пушкина». М., 2013. С. 59–61.

из корпуса обычно изымаются, чтобы не создавать путаницы при количественных исследованиях; обычно в текстовой коллекции корпуса остается только один вариант, при этом создателям корпуса не всегда важно, насколько этот вариант аутентичен и отражает авторскую волю: такими филологическими «деталями» в системе больших чисел корпуса можно пренебречь.

В то же время для филолога художественный текст представляет собой не единый объект, а парадигму вариантов, часть которых осталась в рукописи, а часть выходила в разных прижизненных изданиях. Например, существует изрядное число только печатных версий пушкинского романа в стихах: «В течение 12 лет перед российским читателем одиннадцатикратно появлялись книги, на обложке и титульном листе которых было название “Евгеній Онѣгинъ” ⟨...⟩ Некоторые его фрагменты представали перед читателем в разных вариантах – либо автор вносил изменения в ранее напечатанный текст, либо в текст вкрадывались разного рода искажения, ошибки, опечатки» [*Перцов*: 649–650]. Кроме того, важная часть парадигмы – это рукописные варианты, отвергнутые автором на этапе работы над замыслом. Для восстановления последовательности этой работы, которая отражается в творческой истории текста, такого рода варианты исключительно важны, но по оговоренным причинам они не включаются в корпуса даже специализированного типа (например, варианты отсутствуют в поэтическом корпусе в составе Национального корпуса русского языка[3]). Путь упрощения, по которому приходится идти лингвистам, оказывается неприемлемым для филологов, воспринимающих исключение вариантов как насилие над текстом и научной традицией его изучения.

Разрешить эти затруднения непросто. По крайней мере, сложившаяся ситуация означает, что существующие и апробированные технологии поисковых механизмов и интерфейсов корпуса не подходят для филологически ориентированной системы. Более того, чтобы не растерять достоинств корпусных технологий при соединении их с потребностями филологической традиции, необходимо провести серьезную проектную работу, которая определила бы следующее:

а) каким образом следует считать частотность при поиске? Простое количество некоторого слова, посчитанного по всем

[3] Одно из немногих исключений, скорее, напоминающее случайность: две редакции стихотворения Н.С. Гумилева «Баллада» («Пять коней подарил мне мой друг Люцифер...», 1908), в хронологически предшествующем варианте это 4-й текст из цикла «Сказка о королях» («Пять могучих коней мне дарил Люцифер...», 1905).

вариантам всех текстов, не будет означать осмысленного числа, которое мы могли бы положить в основу квантитативного исследования. В то же время отказ от частотности означал бы отказ от одного из важнейших достижений корпусной лингвистики;

б) как располагать варианты одного текста в выдаче? В традиционном корпусе, вследствие его синтагматичности, все тексты находятся на одном уровне иерархии, в результатах поиска они также отображаются последовательно, без графического оформления разницы в статусе. Однако варианты рукописные и печатные, автографы и списки, черновики и беловики образуют более сложную систему, отражение которой в поиске требует соответствующих интерфейсных решений;

в) как технически организовать сопоставление вариантов? С композиционной точки зрения редакции одного текста зачастую находятся в весьма сложных отношений друг с другом: одна и та же строфа поэтического произведения может оказаться в начале одного и в конце другого варианта, вычеркнутое в одной строке слово может возникнуть в другой. Какие-то отрезки одних вариантов совсем не находят параллелей в других. Запутанные отношения между редакциями лермонтовского «Демона» хорошо иллюстрируют трудности, возникающие при соположении текстов, которые в принципе стоило бы признать вариантами одного и того же произведения. Очевидно, такие отношения в рамках электронной системы следует хранить в сложно организованных структурах данных (вероятно, в чем-то они могут напоминать стандарты кодирования видео), а при сопоставлении каждый раз принимать филологически корректное решение относительно обоснованности той или иной параллели.

При этом приходится быть готовым к тому, что ни одно решение не будет принято сообществом безоговорочно. Так, некоторые вынужденные, технически обусловленные, особенности представления текстов в «Параллельном корпусе переводов “Слова о полку Игореве”»[4] неоднократно критиковались филологами, несмотря на то, что воспринимались создателями ресурса как малосущественные. Среди таких особенностей разбиение текста «Слова...» на отрывки в соответствии со схемой, принятой в издании Р.О. Якобсона, и обратная перестановка фрагмента текста, описывающего начало похода Игоря и солнечное затмение, в то место, где этот фрагмент находился в варианте, которым располагали первые

[4] Один из первых филологически ориентированных корпусов действует с февраля 2007 г.; располагается в Интернете по адресу: http://nevmenandr.net/slovo/

издатели (в некоторых изданиях и переводах этот фрагмент переставлен дальше от начала в место, которое для него определяет не художественная, а аристотелева логика).

Некоторые достижения корпусной лингвистики позволяют оценивать часть этих вопросов как принципиально имеющие ответ. Так, особенное место в типологии корпусов занимают параллельные корпуса, которые призваны отразить межъязыковое взаимодействие и варьирование, представляя интуитивно воспринимаемые как идентичные тексты на разных языках, т.е. тексты и их переводы на другой язык. Хорошим примером могут служить Параллельный корпус в составе Национального корпуса русского языка (http://ruscorpora.ru/search-para.html) или корпус ParaSol (http://www.slavist.de/). Не случайно, что параллельные корпуса устроены сложнее, чем обычные поисковые лингвистические системы: они требуют особой настройки или даже создания специальных поисковых механизмов. Однако именно параллельные корпуса работают с текстами как с парадигмой вариантов, хотя вариантами являются не отличающиеся друг от друга редакции одного текста на одном языке (как это должно быть в составе филологического корпуса), а переводные эквиваленты одного текста на разных языках. Тем не менее, результаты поиска в параллельном корпусе представляются в приемлемом виде (решение проблемы под литерой *б*): текст оригинала визуально отделен от контекстов переводов, а сопоставление текстов на разных языках (называемое *выравниванием*, *alignment*) перед помещением в корпус выполняется в полуавтоматическом режиме с помощью специальных программ (решение проблемы под литерой *в*).

Действительно, уже сейчас можно предсказать, что важной проблемой при создании филологического корпуса станет автоматическое сравнение вариантов, которое фиксировало бы разницу между этапами работы над текстом (взятая в целокупности она называется «творческой историей»). Скорее всего, имеющиеся средства выравнивания текстов, используемые для параллельных корпусов, без настройки и введения дополнительной функциональности, ориентированной на специфику поэтической речи, нельзя будет использовать в построении текстовой базы. В то же время компьютерная лингвистика успела разработать полезные алгоритмы (см. алгоритм *шинлов*) нахождения повторяющихся фрагментов текста. Их адаптация к нуждам филологического корпуса позволит во многом автоматизировать трудоемкий текстологический процесс сравнения вариантов. В связи с этим можно обратить внимание на прототип электронной системы представления вариантов художественных текстов на материале стихотво-

рения «Анчар»[5]. В среде цифровых гуманитарных технологий[6] на Западе также идут работы в этом направлении, получившем общее обозначение «text collation» (см. [*Gilbert*: 139–147]).

Несомненно, что помимо текстовой расшифровки (транскрипции) рукописного варианта в филологическом корпусе должны присутствовать и изображения оригинальных автографов[7], по которым исследователь смог бы уточнить чтение того или иного фрагмента (особенно это актуально в случае, когда почерк не позволяет прочитать все с полной ясностью). Изображения и текст должны быть связаны («залинкованы»), удачные примеры этой визуальной стратегии можно найти в представлении Codex Sinaiticus (http://codexsinaiticus.org/en/manuscript.aspx) и рукописей М. Пруста (http://research.cch.kcl.ac.uk/proust_prototype/index.html).

Реализация связи текста и изображения составляет существенную проблему перспективного филологического ресурса. Автоматизация этой задачи вызывает сомнения, потому что современные системы распознавания изображений (Optical character recognition) хорошо справляются с печатным текстом, но плохо умеют понимать текст рукописный. В этом направлении работа ведется и коммерческими компаниями, и членами академического сообщества, но в целом рассчитывать на то, что в этой части подготовки корпуса компьютер сможет качественно облегчить усилия человека, не приходится. В то же время для человека может быть создано специализированное рабочее место, текстовый редактор, учитывающий потребности текстолога.

Наконец, в филологическом корпусе должен найти отражение и комментаторско-интерпретативный вектор науки о художественных текстах. Надо признать, что подобного рода задачи до сих пор были далеки от тех, которые решались в рамках построения корпусов. С комментированием, введением в оборот важных для понимания культурно значимого произведения научных работ лучше справлялись электронные научные издания (ЭНИ), своего рода стандарт которых был предложен Фундаментальной электронной библиотекой «Русская литература и фольклор» (http://feb-web.ru/). Каждое ЭНИ предоставляет пользователю возможность ознакомиться с интересующим его произведением, с основной литерату-

---

[5] Ресурс доступен в Интернете по адресу http://lcph.bashedu.ru/cgi-bin/anchar-sample.pl

[6] Digital Humanities, Н.В. Перцов предлагает называть эту область «компьютерной гуманитаристикой» (см. [*Перцов*: 647], 2015. С. 647.

[7] Сходные идеи изложены в статье Перцова «Об аутентичном факсимильно-транскрипционном представлении рукописей русских писателей» [Перцов: 627–648].

рой о нем и библиографическими справочниками, формирующими дальний горизонт относящегося к предмету чтения.

Для корпуса схемы связи текста и метатекста, по всей видимости, нужно будет изобретать его разработчикам, так как сформированные в рамках ЭНИ и в рамках корпусной лингвистики подходы плохо монтируются между собой. Применяемая в корпусах *разметка* обычно включает заранее заданный набор параметров, значения которых приписываются словам или сочетаниям слов. Круг тем, которые могут возникнуть в связи с художественным текстом, принципиально не ограничен, поэтому даже с опорой на существующие исследования и с учетом того, что уже разработаны компьютерные программы семантического анализа, представляется невозможным автоматизировать и унифицировать смысловую (или даже хотя бы тематическую) разметку помещенных в филологический корпус произведений. Иными словами, добиться той же легкости филологически значимой разметки текста, что и морфологическая разметка, для традиционного лингвистического корпуса, по всей видимости, не удастся. При этом важно помнить, что корпусная разметка не должна быть научным исследованием сама по себе, она лишь достаточно простой инструмент, исследование облегчающий. Это добавляет трудностей в задачу наложения смыслового слоя на художественный текст.

Очевидное (хотя и паллиативное) решение обозначенной проблемы – это привязка к художественному тексту уже существующих филологических комментариев. В распоряжении современного исследователя имеются претендующие на полноту комментаторские труды Ю.М. Лотмана, В.В. Набокова («Евгений Онегин»), Л.И. Соболева («Война и мир») и др. Очевидный плюс такого подхода – простота решения. В минусах будут числиться отсутствие унифицированной подачи комментария и вытекающая из этого невозможность превращения комментария в корпусную разметку. Кроме того, несмотря на высокий профессионализм и добросовестность авторов, комментарий в силу естественных причин не может быть полон (можно вспомнить хотя бы о работах, вышедших в печати уже после появления комментария).

Таким образом, создатели филологического корпуса не могут ограничиться привязкой существующих комментариев к тексту. Им следует также предусмотреть обращение к имеющейся посвященной исходному тексту научной литературе. В идеальном случае следовало бы построить систему автоматического извлечения из литературоведческих трудов релевантной для интерпретации художественного произведения информации (соответствующая область компьютерной лингвистики называется information extraction).

Вторым базовым элементом разметки филологического корпуса должны стать эксплицитно обозначенные и доступные для поиска случаи интертекстуальных схождений как внутри корпуса одного автора (самоцитирования), так и за его пределами. Отчасти эту проблему решают уже упомянутые технологии нахождения дубликатов и повторов в текстах (алгоритм шинглов), но не следует забывать и о более сложных случаях интертекста, указания на которые можно будет найти только в специальной литературе.

К сожалению, далеко не все необходимые для создания филологического корпуса технологии разработаны, хотя во многом современная наука вплотную подошла к решению многих технических задач. Все это означает, что по крайней мере в первое время придется использовать ручной труд экспертов. Как кажется, удобным инструментом ручной текстовой разметки может стать программная платформа «Annotation Studio», разрабатываемая в Массачусетском технологическом институте (http://www.annotationstudio.org/). Она предоставляет несколько удобных возможностей привязывания к произведению филологически важной информации как текстового, так и мультимедийного характера. Кроме того, эта платформа предполагает низкий порог вхождения для экспертов-филологов, которые при этом могут не являться экспертами в области компьютерных технологий.

Итак, очевидно, что текстологическая часть филологического корпуса должна включать в себя, по меньшей мере, все доступные редакции и варианты произведения, которые исследователь имел бы возможность читать, сравнивать между собой (используя при этом технические средства поиска и визуализации различий) и с изображением рукописи или аутентичного печатного издания. При этом требующими проработки оказываются разнообразные детали реализации этой идеи: от устройства поисковой статистики до технологии связывания расшифровки и изображения.

Кроме текстологической части филологический корпус должен иметь комментаторско-интерпретационную. В ее рамках текст должен быть связан со сделанными в литературоведческих исследованиях выводами о структуре, смысловых акцентах, текстуальных источниках произведения. Конкретное воплощение связей на уровне поиска и интерфейса пока может быть прорисовано только в самых общих чертах, поскольку аналогичной функциональности пользователям лингвистических корпусов до сих пор не предоставлялось.

ЛИТЕРАТУРА

Бодрова А.С., Пильщиков И.А. Проблемы корпусного подхода к задачам авторской лексикографии // Авторская лексикография и история слов: К 50-летию выхода в свет «Словаря языка Пушкина». М., 2013. С. 59–61.

Орехов Б.В. Что такое филология? // Вестник Башкирского государственного педагогического университета им. М. Акмуллы. 2010. № 3. С. 74–82.

Перцов Н.В. Лингвистика, поэтика, текстология. Избранные статьи. М., 2015. С. 649–650.

Плунгян В.А. Корпус как инструмент и как идеология: О некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. 2008. № 2 (16). С. 7–20.

Gilbert P. Automatic collation: A technique for medieval texts // Computers and the Humanities. Jan. 1973. Vol. 7. Issue 3. P. 139–147.

Leech G. Corpora and theories of linguistic performance // J. Svartvik (ed). Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82. Stockholm. 4–8 August 1991. Berlin: Mouton de Gruyter, 1992. P. 105–122.