# *Digital Humanities* — это что-то новое или мы уже давно этим занимаемся?

## Интервью с Игорем Пильщиковым

Беседовал *Михаил Маяцкий*

**Игорь Пильщиков** — доктор филологических наук, ведущий научный сотрудник Института мировой культуры МГУ им. М. В. Ломоносова, старший научный сотрудник Эстонского гуманитарного института Таллиннского университета, ведущий научный сотрудник Института языкознания РАН, главный редактор Фундаментальной электронной библиотеки «Русская литература и фольклор», научный редактор Русской виртуальной библиотеки, соредактор журналов *Philologica* (1994–2013) и *Studia Metrica et Poetica*.

Начнем с простого вопроса: когда вы узнали про интернет? Сегодня такой вопрос прозвучит странно, но ведь интернет появился всего 20–25 лет назад.

Интернет появился в 1991 году, а через несколько лет началось бурное развитие русскоязычного сегмента интернета. Я об интернете услышал впервые в Англии в 1992–1993 годах и тут же стал смотреть, а что есть в этой новой среде по интересующим меня славистическим вопросам. Например, я вбил в тогдашний поисковик фамилию Батюшкова, которым я как историк литературы много занимался и продолжаю заниматься, и обнаружил, что никакой особой литературы о нем в интернете по его состоянию на 1993 год не было.

Вы с самого начала воспринимали интернет как место, где можно найти такого рода сведения?

Я рассматривал интернет как место, где можно получать, обрабатывать и продуцировать какую-то информацию. Я ви-

дел в нем тогда некое расширение компьютера. В Англии, куда я приехал в 1991 году, все студенты уже работали на персональных компьютерах, хотя это были и смешные на сегодняшний взгляд модели (с 286-м процессором, с 386-м...). Но тем не менее это были компьютеры! Я был единственным человеком во всем университете, который печатал свои статьи на машинке. Люди с других факультетов приходили посмотреть на меня, потому что с машинкой тогда уже мало кто хотел и умел обращаться. А мой первый компьютер был вообще без жесткого диска. Тогда же, в Англии, я начал использовать компьютер в издательской деятельности. Сначала я участвовал в издании журнала Британского нео-формалистического кружка *Essays in Poetics* (как приглашенный редактор одного из разделов). Потом мы вместе с покойным ныне Максимом Шапиром задумали и начали издавать двуязычный русско-английский журнал *Philologica*, взяв курс на редактирование профессиональных текстов в компьютерной среде, на профессиональную редакционно-издательскую обработку. Но тогда, в 1994–1995 годах, еще не было электронных публикаций и электронной текстологии. А уже через три года мой старый друг и соученик по Новосибирскому и Тартускому университетам Евгений Горный, один из пионеров русского интернета, задумался о том, как сделать профессиональную филологическую электронную библиотеку. К тому времени уже существовала Библиотека Мошкова, возникшая практически одновременно с национальной доменной зоной *.RU* (1994). Максим Мошков создал, по сути, русский аналог проекта «Гутенберг» (*Project Gutenberg*). Это была первая — стихийная — стадия дигитализации. Тогда все понимали, что новая среда есть, а содержания в ней никакого нет, и важно ее наполнить какими-то текстами. А уж что мы начнем цифровать раньше — Толстого или Достоевского, было абсолютно все равно. Даже была такая идеологическая установка, манифестированная в проекте «Гутенберг»: все равно, какого качества текст, по какому изданию, давайте сначала наполним информационное пространство, а потом пользователи сами текст до ума доведут, доправят. Это имело смысл на первом этапе наполнения интернета информацией, однако этот этап быстро себя исчерпал (хотя сам по себе процесс стихийной оцифровки бумажных источников не остановился и поныне — и, надеюсь, никогда не остановится). Но в конце 1990-х годов встал вопрос о создании профессиональных ресурсов, планирующих порядок вводимой информации, вид ее представления, цели введения, приоритеты и поэтапную про-

грамму комплектования. Неслучайно ведь на определенной стадии та же *Lib.Ru*, Библиотека Мошкова, тоже выделила у себя большой подраздел *Az.Lib.Ru* (он носит название «*Lib.ru*: Классика»), который устроен уже по принципу профессиональной филологической библиотеки. Там тексты планомерно вычитываются, приводятся в соответствие с печатным изданием, которое они представляют, там появился план публикаций и т.д. Но на первом этапе этого не было, хотя уже на рубеже веков всем стало понятно, что электронная библиотека — наиболее удобный способ представления литературных текстов и разработки инструментария для работы с ними. Поскольку электронная библиотека — это не просто некая коллекция текстов, а коллекция, структурированная и пополняемая, она, как и традиционные библиотеки, имеет планы, которые строятся исходя из нужд целевой аудитории, на которую рассчитана библиотека.

В 1998–1999 годах Горный предложил сделать электронную библиотеку профессионального типа, мы назвали ее «Русская виртуальная библиотека» (РВБ). А еще одним активным участником стал Владимир Литвинов, ныне технический директор РВБ. Почти одновременно с нами стартовал другой филологический (но не библиотечный) интернет-проект — *Ruthenia*. Иначе говоря, настало время «филологического» интернета — филологи начали понимать, что им тоже что-то нужно от этой новой среды. Мы открыли библиотеку 1 декабря 1999 года, то есть уже 15 лет назад. Было интервью по радио, и нас услышали коллеги, которые занимались приблизительно тем же самым, но пока вне интернета, на дисках,— это Константин Вигурский и Александр Штольберг, которые тогда работали в Научно-техническом центре «Информрегистр», то есть, так сказать, в «электронной книжной палате» РФ. Согласно существующему законодательству, в России книги регистрируются (по крайней мере, до последнего времени регистрировались) в Книжной палате, а электронные тексты — в «Информрегистре». И вот они, скооперировавшись с Институтом мировой литературы, еще в середине 1990-х начали делать проект «Русская классика на компакт-дисках». Первым проектом было издание Грибоедова, вторым — Пушкина.

Почему первым стал именно Грибоедов?

Он был выбран как, с одной стороны, абсолютный классик, а с другой — автор небольшого корпуса текстов, и на нем обкатывались разные методики (в частности, та поисковая систе-

ма, которая потом стала большим «Яндексом», была обкатана, в частности, на Грибоедове). И когда этот проект перерос в проект ФЭБ, «историческая» связка между ним и «Яндексом» сохранилась. Некоторые виды поиска были впервые опробованы в ФЭБ — например, сделанный «Яндексом» по нашему заказу поиск в старой орфографии (когда пользователь формирует запрос в нынешнем правописании, а результат выдается и в дореформенном, и в пореформенном). Теперь и большой «Яндекс», а с определенного момента и *Google* ведут поиск по русским текстам и в дореволюционной, и в современной орфографии. Но я забегаю вперед. После интервью на радио коллеги предложили встретиться, и в результате я оказался редактором сразу двух электронных библиотек.

Значит, радио? То есть старые медиа играют свою роль в истории новых?

Ну конечно. Культура ведь едина, и все формы коммуникации дополняют друг друга. Коллеги, занимаясь разработкой информационно-поискового инструментария, принципов электронной публикации, принципов подготовки и подачи текстов в электронной среде, услышали, что есть еще такие же чудаки, которые занимаются тем же самым. Таким образом, мы двумя пересекающимися по составу командами стали делать две библиотеки. Новый сайт открылся 1 июня 2002 года и получил название «Фундаментальная электронная библиотека „Русская литература и фольклор“» (ФЭБ). В этом проекте активное участие, особенно на начальных этапах, принимали институты Российской академии наук — Институт мировой литературы и Пушкинский Дом, но в целом он не привязан жестко ни к конкретным институтам, ни к определенным авторам. Мы занимались и Пушкиным, и Ломоносовым, и Маяковским, и фольклором, нас интересуют разные вещи, и в зависимости от новых интересов появляются новые участники. Например, для пока не оконченного словаря русского языка XVIII века соучастником стал Институт лингвистических исследований в Петербурге, где готовят оригинальное, «бумажное» издание этого словаря.

Какая идея, программа стоят за всем этим?

Мы начали разрабатывать сразу несколько стратегий создания электронных библиотек. Одна стратегия — развитие «вглубь» (как в ФЭБ). Выбираются несколько базовых имен — Пушкин,

Грибоедов, Гоголь, Толстой, Маяковский, Есенин и т. д., и по ним аккумулируется все наиболее важное: все авторитетные издания, монографии, статьи, словари… Другой тип развития — экстенсивный (как в РВБ): представить разных авторов если не полным, то репрезентативным корпусом текстов, хотя бы по одному авторитетному изданию для каждого автора. И это тоже дало свои плоды. Ведь когда мы 10 лет назад начали публиковать на РВБ электронные версии академических и научно-популярных изданий русских писателей XVIII века, никаких текстов XVIII века в сети еще не было, а теперь они есть и пользуются читательским спросом (удивительный, кстати, факт!). Когда в ФЭБ мы начали публиковать в цифровой форме пушкинскую периодику (с 1903 года до нынешнего времени) — «Пушкин и его современники» под редакцией Модзалевского, «Временник Пушкинской комиссии» довоенный и послевоенный, «Пушкин: исследования и материалы»,— всех этих материалов в сети не было, а филологам они нужны ежедневно. Собранные в одном месте, представленные в едином формате, проиндексированные одной поисковой системой, эти материалы заиграли по-новому, превратились в качественно новый инструментарий.

Какие задачи вы ставили перед собой? Какие препятствия возникали?

Нужно было думать и объяснять себе и другим, что́ делать. Даже то, что может показаться очевидным, было вовсе не очевидно. Возьмем такую простую вещь: нужно ли при представлении текста в электронной форме воспроизводить пагинацию исходного печатного издания? С нашей точки зрения, конечно, нужно! Можно делать новые, оригинальные электронные публикации, а можно воспроизводить какой-то определенный печатный источник, и эти вещи нужно разделять. На стадии проекта «Гутенберг» и раннего Мошкова дело обстояло так: мы не знаем, что это за текст и по какому изданию он приводится. Издание не идентифицировано. До сих пор главная филологическая болезнь интернета — неидентифицированные тексты. Когда каждый может быть редактором и распространителем, что ж удивляться, что подчас распространяется неизвестно что. Когда вы берете с произвольного сайта текст «Анны Карениной», уверены ли вы, что там нет пропусков, что текст правильный? Да и какой текст «Анны Карениной» или «Войны и мира» считать правильным — вопрос далеко не пустой, и ответ на него не очевиден. И вот опубликуете вы научную работу по этой теме,

а вас на смех поднимут: «Знаете, а ведь Толстой такого не писал ни в одной из редакций романа!»

Принципы филологически корректного воспроизведения издания — это воспроизведение структуры и пагинации издания, это обязательная идентификация, библиографическое описание исходного печатного издания, позволяющее соотнести с ним и таким образом идентифицировать его электронную версию. Все это приходилось объяснять и другим, и самим себе! Я помню, как десять лет назад тот же Максим Мошков спрашивал коллег на семинаре в РГБ: зачем нужна страница издания, когда в электронной библиотеке есть полный текст, и если мне нужно найти цитату, то я ее тут же нахожу? Ответ, в общем-то, простой: а если у вас ссылка не на цитату «Мой дядя самых честных правил…», а «Пушкин, ПСС, т. 6, с. 5», то как вы будете искать?

Вопрос об идентификации — принципиальный. Почему *Google Books* является не полноценной библиотекой, а «всего лишь» полигоном для новых технологий? Об этом много писали американские филологи и библиотековеды. Да потому что, с одной стороны, *Google* оцифровывает огромное количество текстов, а с другой — уничтожает «поляну»: вместо библиотеки (некоей электронной версии Библиотеки Конгресса или даже Всемирной электронной библиотеки) получается гигантская сельская лавка, в которой все перемешано, а книги — неизвестно какие. С самого начала не была подключена команда библиографов. Сначала оцифровщики предполагали, что можно будет идентифицировать книги по титульным листам. Это крайне наивное представление, потому что в разные эпохи титульные листы организовывались по-разному, да и взять всю информацию с титульного листа невозможно — иногда она находится в других местах в той же книге, а подчас и в других книгах. Началась путаница и мешанина.

На втором этапе в *Google Books* стали использовать для идентификации библиотечные карточки. Но они делались и делаются в разные периоды и в разных странах по-разному, и сведéние этих разных описаний воедино — задача небанальная. Результат: найти конкретное издание или идентифицировать найденное издание в *Google Books* непросто, а иногда невозможно. Поэтому огромное количество собственно научных задач на *Google Books* невыполнимо. И конечно, автоматическое распознавание книг, содержащих тексты на разных языках, крайне затруднено. Если книга содержит и кириллицу, и латиницу, вполне возможно, она будет распознана вся как кириллица, а части на ла-

тинице не будут корректно распознаны, или наоборот. Русская дореформенная орфография на разных этапах создания *Google Books* распознавалась по-разному. До определенного момента старые книги по ошибке распознавались с помощью анализатора, настроенного на современное правописание и современный словарь, поэтому робот, скажем, распознавал неизвестное слово «сынъ», но не «знал», что оно идентично известному в другом правописании слову «сын». Поэтому, набрав в запросе «сын», слово «сынъ» вы не найдете. На втором этапе тексты в старой орфографии стали распознавать отдельно. А потом распознавание старой орфографии настроили так же, как это сделано сейчас в «Яндексе»,— распознанные тексты представляются пользователю в новой орфографии, а на картинке, под которую «подложен» текст в новом правописании, виден текст в старой орфографии. Для этих книг, набрав в запросе «сын», вы слово «сынъ» найдете. Но в этом массиве вы не сможете отделить книги в старой орфографии от книг в новой, потому что с точки зрения распознанного текста они одинаковы.

Результат: десятки тысяч книг распознаны тремя разными способами. Чтобы в одной книге найти слово «сѣнь», вам надо набрать его с ятем, а «сынъ» — с ером, а в другой, аналогичной,— без ятя и без ера. Чтобы что-то всерьез искать, нужно представлять себе, как были организованы разные этапы распознавания разных текстов в *Google Books*, какие могли возникать типичные ошибки распознавания, и только тогда оттуда можно будет что-то вытащить. Но для этого нужна работа диггера, а не традиционного профессионала-филолога. С другой стороны, тщательная вычитка распознанных текстов замедляет процесс оцифровки в сотни, если не в тысячи раз. Что делать? Все эти вещи требовали рефлексии.

А не возникало ли проблем, связанных не с разными языками, а с разными шрифтами?

Конечно же, возникал и этот вопрос: мы воспроизводим книгу, а нужно ли воспроизводить ее шрифтовые особенности? Возьмем любую книжку. Казалось бы, какая разница: с засечками шрифт или без засечек? Если он на бумаге с засечками, а у нас будет без засечек, в общем случае ничего не изменится. Но есть огромное количество книг, где шрифт семантизирован и где разными шрифтами набраны разные тексты или фрагменты текста — как, например, в романе Умберто Эко «Маятник Фуко». Зна-

чит, в таких случаях нужно воспроизводить различия. А с какой точностью это воспроизводить? Положим, шрифт с засечками можно противопоставлять шрифту без засечек. А внутри категории «шрифтов с засечками»? Нужно ли отличать при электронном воспроизведении *Garamond* от *Times*? Зависит от того, имеется ли это различие в издании и как оно нагружено — семантически? эстетически? И нужно решить, что мы воспроизводим — текст? внешний вид издания? Это разные вещи. Главное, что такой упрощенной номенклатурой шрифтов все отнюдь не исчерпывается. Вот есть задача, которая до сих пор не имеет стандартного решения. Немецкий, готический шрифт, фрактура,— исторически сложившийся графический вариант латиницы; это не другой алфавит, а другая гарнитура. Привычный нам вариант латиницы, именуемый «антиква», и фрактура отличаются не составом букв, а начертанием.

Кроме некоторых знаков типа умлаута…

Да, умлаут может обозначаться как точками, так и отдельной буквой «e» (*ä* = *ae*, *ü* = *ue* и т. д., причем иногда «e» располагается не после, а над «a» или «u»), имеются особые лигатуры (слитные написания двух букв), но это дела не меняет. Дело в том, что в Юникоде нет специального диапазона, где находились бы готические буквы: у них те же самые кодировки, только гарнитура другая. Появляется шрифтовая зависимость, то есть в *одном* шрифте мы не можем воспроизвести разницу между готикой и антиквой, а она иногда нужна. Известно, что немецкие книги могли вплоть до конца Второй мировой войны печататься на выбор — готикой или антиквой. И скорописи было тоже две, иногда эта разница семантизировалась. Но допустим, что в таких случаях разница имеет чисто оформительский характер (хотя это не всегда так).

Но вот другой пример. У Пушкина в «Борисе Годунове», в сцене битвы на равнине, идет разговор на трех языках — по-русски, по-французски и по-немецки. И в оригинальном издании, и в современных академических и научно-массовых изданиях используются три шрифта: для русской речи — кириллица, для французской — антиква, для немецкой — готика. В электронных изданиях без специальной шрифтовой подгрузки воспроизвести эту специфику невозможно, и все тексты «Бориса Годунова» в интернете эту особенность теряют. Будем во всем винить технику? Нет, просто нужно понимать, что́ нужно воспроизво-

дить, а чем можно пожертвовать. Кстати, в переводах «Годунова» на другие языки это соответствие между шрифтами и языками часто теряется. И есть даже научные работы, толкующие, как Пушкин изображает столкновение трех языков, используя только два шрифта.

Как организована взаимосвязь между филологами и программистами? Вы ставите перед ними задачи?

Да, задачи всегда должны ставить те, для кого создается информационная система. К примеру, поиск в старой орфографии доступен сейчас всем пользователям, а до нашего заказа (то есть заказа очень узкой группы людей для очень конкретной цели) такой задачи не стояло. Мы с самого начала наметили цель — аутентичное воспроизведение печатного издания: структуры, пагинации, номенклатуры шрифтов, орфографии. Но при этом какие-то элементы мы все же можем опускать, если примем соответствующее решение (ну вот мы, например, не воспроизводим колонтитулы). Но нас постоянно преследовали и будут преследовать вопросы: что сохранять, а чем допустимо пожертвовать? Например, разбивка на строки в поэзии обязательна, а в прозе она обычно не считается обязательным элементом. Но в некоторых академических изданиях (например, в изданиях Пушкина и Чехова) комментарии даются к такой-то строке на такой-то странице. Если строки не воспроизвести, комментарий повиснет в воздухе. Это один из тех многочисленных вопросов, которые встают перед специалистом по электронной текстологии.

Иные, но сходные вопросы встают относительно воспроизведения изображений. Конечно, понимание самого принципа аутентичности меняется. В конце 1990-х — начале 2000-х она осуществлялась через «глубокую» разметку (*TEI*, «глубокий» *HTML*): это логическая и семантическая разметка текстов, способная одновременно воссоздавать визуальный облик печатного издания — создавать представление о нем, близкое к исходному. Сейчас тренд другой — совмещение текстового и имиджевого представления: поиск идет по тексту, а результаты отображаются на картинку, факсимильно воспроизводящую исходную страницу. Так устроена коллекция *Google Books*, так устроена цифровая коллекция Национальной библиотеки Франции — *Gallica*. При таком подходе многие проблемы снимаются, но одновременно появляются новые, то есть это вовсе не панацея. Возьмем уже обсуждавшуюся проблему поиска русского текста в старой ор-

фографии. Как искать? Нужно, чтобы я набирал слово в старой орфографии и чтобы поиск осуществлялся только в старой орфографии? Но пользователь может не знать старой орфографии, и тогда он не сможет корректно сформулировать запрос. Иди знай, как пишется «осѣненнаго»; есть ли там ять и где именно? Да и дореформенная орфография существовала в нескольких вариантах — все возможные написания сможет перечислить разве что специалист по истории орфографии. Или набирать в одной орфографии, чтобы машина при этом искала в разных? Для этого был придуман алгоритм редукции старой орфографии к новой (он использован в ФЭБ и на большом «Яндексе»). Старая орфография более информативна, чем новая, она отражает этимологию (в заимствованных словах «ф» пишется в соответствии с *f* и *ph*, а «ѳ» — в соответствии с *th*), она различает омонимы типа «всѣ — все» (сейчас графическое противопоставление «все — всё» возможно, но не обязательно), «миръ» как отсутствие войны и «міръ» как вселенная (это искусственное графическое различение, давно введенное, притом что эти понятия связаны друг с другом: ср. латинское *pax* в значении *universum*).

А у Толстого как? Это какой-то вечный спор…

Толстой колебался. Думал, как назвать роман: «Война и универсум», «Война и перемирие»? Напечатал под названием «Война и миръ». А в 1913 году Бирюков издал роман под названием «Война и міръ», а ведь «міръ» — это ведь не только универсум, но и общество. На эту тему есть исследование у Эвелины Ефимовны Зайденшнур, крупного текстолога-толстоведа. Максим Шапир много писал о недопустимости модернизации орфографии в научных изданиях. У нас с Николаем Перцовым есть статья в «Вопросах языкознания» за 2011 год о том, почему научные издания классики XIX века должны вернуться к воспроизведению аутентичной орфографии (авторской либо современной автору). Там разбирается в том числе и вопрос с «войной и миром». Почему это важно? Потому хотя бы, что, например, у Маяковского поэма называется «Война и міръ» (война и общество) — то есть не так, как в абсолютном большинстве изданий Толстого. А узнать об этом мы можем только из набранного петитом однострочного комментария в академическом собрании сочинений.

Игорь, вот вы показали, как важный, необходимый инструментарий предлагается сначала немногочисленным специалистам, потом всем гумани-

тариям, потом взыскательным читателям, а потом и «обычным» читателям-пользователям. Осыпают ли вас коллеги только благодарностями? Какие вообще отношения складываются с цехом? Ощутимо ли противодействие с его стороны? Например, под девизом: «Книжки надо читать, а вы потакаете тенденции читать с экрана».

Когда все начиналось, приходилось многое выслушивать:

> Зачем нам вся эта инженерия? Нужно ходить в архив, чувствовать запах старой бумаги. Надо обязательно листать, не доверять никаким копиям, пощупать первое издание...

Ну, я как человек, который достаточно много работает в архивах, знаю, как важно и листы потрогать, и на чернила посмотреть, и первопечатные издания в руках подержать, и не доверять никаким копиям, а посмотреть в оригинал и, если возможно, в несколько экземпляров печатного оригинала, ибо всякое бывает. Но это одна сторона вопроса. С другой же стороны, потратить два-три дня на поиски забытой цитаты в 16-томнике Пушкина — это роскошь недопустимая, и для этого совершенно не нужно нюхать запах корешков, а лучше 16-томник грамотно оцифровать и эту цитату быстро найти. О том, что в данном случае значит «грамотно», мы уже говорили. Ведь можно оцифровать и так, что ничего нельзя будет найти, а найдя — идентифицировать, и тогда эта оцифровка окажется для профессионала бесполезной. Я бы сказал так: читать или просто перелистывать удобнее бумажное издание, а работать удобнее с электронным.

Не сталкивались ли вы с содержательными, интересными возражениями от коллег?

Вначале возражения всегда шли «от консерватизма»: вот есть некоторые сложившиеся, опробованные и апробированные практики, а эти новые инструменты их уничтожают. На самом же деле новые практики не уничтожают старые, а расширяют и дополняют их. Потому что никто не только не запрещает коллегам идти в архив, но и все, что делается в сфере *Digital Humanities*, никак не отменяет архивной и традиционной библиотечной работы. Это первое.

Второе. Я до этого только об электронных библиотеках говорил, а ведь электронные библиотеки — не единственный способ работы с текстами, в том числе с литературными. Существуют

лингвистические ресурсы. Например, создан Национальный корпус русского языка, существуют другие, ранее созданные корпусы — Британский, Чешский. Они позволяют исследовать литературные тексты. В них есть специальные подкорпусы — так, существуют русский и чешский поэтические подкорпусы. С их помощью можно работать с формальными лингвопоэтическими параметрами поэтических текстов, там совершенно другие задачи, другой инструментарий. Речь не идет о жестком противопоставлении или-или: корпусы *или* библиотеки. Специалисту требуется и то и другое (а также третье — электронные архивы, и четвертое — электронные каталоги, и многое другое). Исторически сложилось так, что я больше занимаюсь электронными библиотеками, хотя корпусный подход мне тоже близок и нужен. С другой стороны (если взять историю создания русских филологических и лингвистических электронных ресурсов) возможны и нередки случаи обмена информацией и компетенцией. Например, поэтические тексты XVIII века были в свое время переданы в Национальный корпус русского языка из РВБ.

Наконец, есть области, для которых этот инструментарий недостаточен. В последние годы я много говорю и пишу о том, что для авторской лексикографии и истории стиха ни электронные библиотеки, ни языковые корпусы не пригодны. Разумеется, не я один об этом говорю. Действительно, если вы хотите создать словарь языка писателя, то важно охватить весь корпус его текстов и дифференцировать его хронологически. А обычное представление текстов в научном или научно-массовом издании (возьмем, например, Жуковского в издании «Библиотеки поэта») таково, что текст дается в последней авторской редакции, а датируется годом создания первой редакции. Большое количество слов, которых не было в ранней редакции, появятся в поздней, а при таком представлении они окажутся датированы более ранним периодом. В случае Жуковского разрыв этот может составлять 30–35 лет — срок огромный. А 1820–1830-е годы — это годы ломки языка, очень быстрого его изменения. Зафиксировано словосочетание впервые в 1815 или же в 1845 году — разница принципиальна. В корпусе могут быть приведены некоторые разночтения, некоторые редакции. Формального сопоставления между разными редакциями нет нигде. В электронной библиотеке разные редакции есть, но они «есть» так же, как они есть на книжной полке: у вас имеется научное издание, в нем печатаются некоторые редакции, другие редакции печатаются в виде разночтений, в некоторых изданиях собран более полный кор-

пус, в других — менее полный, следующее издание содержит поправки к предыдущему и т.д. В электронной библиотеке все это есть в электронном виде, но это все не формализовано. Вы не можете сделать качественный словарь языка писателя, просто взяв пять книг с полки. Точно так же вы не сможете сделать его, просто взяв тексты из электронной библиотеки. Но и на основе Национального корпуса вы тоже не сможете создать такой словарь. Требуется новый, специально препарированный корпус особого типа.

А в перспективе есть возможность делать новые типы справочных изданий? Например, словарь языка поэта по состоянию на такой-то год?

Я хотел бы их делать: хронологически дифференцировать корпусы с представлением всех дошедших до нас вариантов произведения, чтобы все редакции и варианты текста были представлены в одной общей системе и они были бы идентифицированы и датированы. Тогда у нас появится возможность использовать такие дифференцированные корпусы и для работ по авторской лексикографии, и для работ по истории стиха. Мы будем подсчитывать те или иные ритмические формы или языковые конструкции по конкретной редакции, точному году, а сейчас это требует дополнительной филологической работы. Мы вынуждены предварять профессиональный лингвистический или стиховедческий анализ традиционной работой с источниками, поскольку электронные ресурсы для этого пока не приспособлены. Но появляются новые способы подготовки и представления информации, новые виды разметки, новые инструменты анализа, хранения и поиска информации. Все это дает профессионалу новые возможности.

Во-первых, компьютерные средства позволяют ускорить рабочий процесс, поскольку автоматизируются рутинные операции. Заполнение карточек с примерами или разметка ритмических форм (я нарочно беру примеры из разных сфер) занимают очень много времени, а дальше ведь довольно часто приходится все делать вручную заново, поскольку нет общего хранилища, нет общей системы сопоставлений между разными фрагментами данных. Как только у нас появляется такого рода инструментарий, вместо 80% рабочего времени, уходящего на обработку материала, и 20% времени уходящего на его осмысление, мы можем получить обратное соотношение. У нас будет гораздо больше времени на размышления и выводы.

Во-вторых, это масштабирование задач. Оно становится возможно благодаря ускорению. Мне очень близок сциентистский пафос, скажем, Бориса Исааковича Ярхо, который боролся с импрессионизмом в науке, с «безответственным словоупотреблением». Давайте, говорил он, посчитаем то, что можно посчитать, а потом будем разбираться с остальным. А когда говорят, что вот такое описание вот такого-то цветочка характерно для такого-то автора — это не наука. Наука — это когда мы посмотрим по текстам максимально доступного числа авторов, как именно они описывают цветочек, а дальше мы увидим, отличаются эти описания цветочка у разных авторов или нет. Это уже будет основа для научного исследования, потому что фактов будет много, мы начнем их систематизировать и они будут верифицируемы, а наши выводы — фальсифицируемы. Не любой подсчет является наукой, но и заявление, что такое-то слово, скажем, у Баратынского звучит свежо и необычно, тоже не является наукой. Иногда оказывается, что оно звучит необычно только для невежественного исследователя, а для современников поэта это было вполне стандартное словоупотребление.

То есть в каком-то смысле появилась техническая возможность относительно легко выполнить заветы Ярхо? Или даже: исчезла возможность отговорки, чтобы их не выполнять?

Ярхо своим «точным литературоведением» вообще во многом предвосхитил *Digital Humanities*. Что касается его методики, то некоторые виды работ (например, анализ прозы) были весьма трудоемкими. Сам он называл этот процесс муравьиной работой. А возражали ему так: зачем советские люди будут производить такое количество разнообразных подсчетов, если еще неизвестно, какой из них пригодится в народном хозяйстве? Сейчас ясно, что эти возражения снимаются хотя бы потому, что многое можно сделать с помощью компьютера и работа займет гораздо меньше времени — ее можно будет сделать в десятки, а то и в сотни раз быстрее. И формализованные результаты можно будет использовать в том числе на практике — они не будут оставаться какой-то игрой ума. Ведь и много позже одним из возражений против работ Андрея Анатольевича Зализняка по формальному описанию русского языка было такое: зачем нам формализованное описание русской грамматики? У нас же есть школьная грамматика. Зачем нам знать, что у русского существительного не три парадигмы склонения, а гораздо боль-

ше? Но сегодня нам ясно, что без формализованной грамматики никакого поиска с учетом морфологии в интернете просто бы не было. Мы не смогли бы набирать слово в одном падеже и получать в результатах поиска то же слово в других падежах.

Игорь, вы говорили о масштабировании. Почему это важно?

Потому что постановка некоторых вопросов обретает смысл, только если ответ получен на большом материале, чтобы результат был статистически значим. Определить частотность какого-то параметра в маленьком рассказике под силу девятикласснику. А сравнить по такому же параметру, скажем, раннего Толстого с поздним — задача и по-настоящему интересная, и разрешимая только тогда, когда под рукой есть соответствующий инструментарий. И потом, как только появляются ускорение и масштабирование, возникает новая, большая система, а большие системы, как мы знаем, обладают свойством эмерджентности: в них появляются новые, не планировавшиеся при их создании связи, и благодаря этим свойствам появляются возможности постановки новых задач, которые раньше не ставились и которые никому не приходили в голову. Можно это описать как имманентный способ развития систем. Это уже вопрос, относящийся не только к «цифровой филологии», но и к философии.

К философии медиа, во всяком случае.

Да. Вот у нас есть электронная среда, а раньше была печатная. Она нас меняет или мы ее меняем? Конечно, и то и другое. Возьмите спор Фридриха Киттлера и Маршалла Маклюэна о том, что такое электронная среда: расширение возможностей человека или некая автономная саморазвивающаяся сущность? Это напоминает старый вопрос, занимавший русских формалистов: что такое литература? Одна из социальных функций человека или же некоторый имманентный процесс? Разумеется, и то и другое. Это и имманентный процесс: попробуйте написать русское стихотворение, чтобы оно было стихотворением и чтобы оно вообще ни на что не опиралось, чтобы в нем не действовали законы русского языка, правила русской версификации. Это невозможно, какие-нибудь до вас существовавшие законы в нем обязательно начнут работать. С другой стороны, если бы мы могли это все описать как полностью имманентный процесс, то мы действительно смогли бы запрограммировать следующе-

го «Онегина», составить такую генеративную грамматику, которая могла бы писать гениальные стихи.

Но ведь уже делаются какие-то эксперименты в этом направлении.

Да, у нас есть, условно говоря, генеративная грамматика Тютчева. Но попробуйте написать с ее помощью новое стихотворение Тютчева или хотя бы пародию на нее! Используйте какой угодно алгоритм, хоть в голове его прокручивайте, хоть в машине... Ясно, что можно создать генератор текстов (и множество таких генераторов давно уже создано), но это не значит, что вы создадите генератор культурно значимых текстов. Это совсем другое. Пока что роботы умеют писать только центоны — даже не полноценные пародии (хотя некоторые центоны у «автоматического поэта» «Яндекса», который составляет стихи из поисковых запросов, совершенно замечательные).

Что-то подобное происходит и с медиа. Ясно, что когда культура перешла от рукописи к книге, многое изменилось. Помимо элитарной литературы появилась массовая литература, которая при этом оказалась отделенной от фольклора. Появилось огромное количество профессионалов печатного дела — литераторы, журналисты, газетчики. Вещь зажила по своим законам да еще начала надиктовывать нам наше поведение. Кто-то хочет совершить такой поступок, чтобы об этом написали в газете. Кто-то хочет прочесть автобиографию знаменитости. Дети могут захотеть стать космонавтом, журналистом, автором детективов... Ребенку XIV века и в голову бы не пришло стать ни автором детективов, ни космонавтом.

В чем отличие новых медиа от медиа эпохи Гутенберга? Принципиально изменилась ситуация с правом на републикацию. Если для эпохи Гутенберга авторское право — это «копирайт», право на изготовление копии, то сейчас право на изготовление копии поставлено под вопрос. Не только потому, что копирование уже не занимает столько времени и сил, как раньше, но и потому что изменился его статус. Что такое копирование? Когда я смотрю на экране в удаленном режиме какой-то текст, я его уже фактически копирую. Идея копирайта, как она была сформулирована в 1830-е годы и окончательно кодифицирована в 1970-е, в течение 1990-х рухнула, стала абсурдом, тормозом, бревном на пути к прогрессу — хотя бы потому, что нарушает право человека на информацию. Из механизма защиты писательского меньшинства она превратилась в авторитарный

механизм именно из-за изменения среды, а у этой среды есть своя внутренняя логика, независимая от ее субъективного осознания. Ведь любое знание не существует как нечто полностью доступное субъективному сознанию. У физики как науки есть своя логика развития, но при этом нет ни одного человека, который знал бы «всю физику», или института, все члены которого в совокупности знали бы «всю физику». Есть то, что Поппер называл *objective knowledge* — знание, которое превосходит все, что знают отдельные физики. Попперовский подход к знанию — один из способов обозначить то, что́ мы сами делаем и что́ одновременно оказывает обратное воздействие на нас. Точно так же любая культурная деятельность оказывает обратное воздействие на деятеля.

Давайте вернемся к вопросу об эмерджентности.

Да, это очень важно: появление новых свойств, которых раньше не было, и возможность постановки новых задач. Например, некоторые модели речевой деятельности раньше не имели практического применения. Если в нашей модели происходит перебор каких-то вариантов и теоретически таких вариантов — сотни и тысячи, то без практической возможности быстро перебрать эти тысячи вариантов такая модель остается отвлеченной игрой ума. Но такая модель может стать реальным инструментом познания с появлением автомата, который быстро исчисляет эти варианты. Это мы уже видели на примере грамматического словаря Зализняка. На следующем этапе мы начинаем моделировать такие процессы, о которых раньше никто не думал, потому что раньше их нельзя было смоделировать. Ну, один человек за пятьдесят лет мог бы просчитать тысячу вариантов. Но какой в этом смысл, если для получения результата надо сравнить десять или сто тысяч? Здесь простой переход количества в качество. Некоторый результат появляется только тогда, когда вы сравнили сто тысяч вариантов, а вручную вы можете сделать только тысячу-две.

Значит, сегодня конкретным специалистам — лермонтоведам, тургеневедам — дан в руки невиданный дотоле в истории науки арсенал средств. Дало ли это неслыханные или просто ощутимые результаты?

Даже если говорить только о русской литературе и только о десятке ресурсов — *Google Books*, РВБ, ФЭБ, Библиотека Мош-

кова, электронная библиотека Пушкинского Дома, библиотека *ImWerden*, Национальный корпус русского языка,— они уже кардинально изменили филологическую практику. Все ли этим богатством пользуются? Нет, не все. Человек — существо консервативное. До него доходит не сразу, а когда доходит, он не сразу бросается менять свои привычки. Интересно, что нередко до провинции новшества доходят быстрее, потому что провинциальному исследователю могут быть недоступны традиционные формы библиотечного обслуживания — в его городе может просто не быть достойной «традиционной» библиотеки. То же касается зарубежных коллег — тех, у которых под рукой нет большой славянской библиотеки. А дальше все зависит от установки. Если есть установка на то, чтобы фундировать каждое утверждение конкретными примерами из текстов, тогда вероятно, что появится желание и обработать-посчитать. Но есть и такие «исследователи», которым достаточно почесать в затылке и сообщить миру, что Пушкин — русский гений.

Так заметен ли всплеск с появлением новых средств в этих областях?

И да и нет. С одной стороны, появились целые новые области знания, новые дисциплины — появилась компьютерная лингвистика, корпусная лингвистика, практикующая совершенно новые подходы к работе с языковым материалом. С другой стороны, многое закостенело в своей консервативной нетронутости. То есть существенные результаты, несомненно, есть, но их меньше, чем могло бы быть, потому что представления специалистов в какой-либо области о возможностях автоматизации труда в этой самой области часто сводятся к формуле: «пусть придут и мне дадут». Но жизнь устроена так, что никто не может прийти и дать. Как программист может прийти и дать вам инструмент, если он вообще не знает, чем вы занимаетесь? А если он попытается себе это представить, то результат будет плачевным. Как в примере с библиографами и инженерами из *Google Books*: инженеры все продумали, но не учли специфику библиографической работы.

То есть гуманитарий должен уметь сформулировать задачу на понятном программисту языке?

Да, мне кажется, я после многих лет работы в той сфере, которую сейчас начали называть *Digital Humanities*, стал своего рода специалистом по «переговорам» между филологами и програм-

мистами. Я помогаю коллегам с двух сторон этого дигитального «фронта» найти взаимопонимание.

А вот практически? Можно отличить статью, написанную коллегой инертным и консервативным, работающим по старинке, от статьи, написанной по результатам машинной обработки большого корпуса?

Это сложный вопрос. Когда речь идет о верифицируемых данных и фальсифицируемых утверждениях (например, «такое-то выражение употреблялось в 1830-х годах крайне редко или, скажем, реже, чем в 1860-е»), то теперь это более-менее легко проверить: человек взял это из головы, из небольшой картотеки или посчитал по огромному массиву. Иногда исследователи склонны к неправомерным обобщениям результатов, полученных на ограниченном материале.

Но привносит ли электронная обработка данных в статью нечто, что ее сразу отличает от той, которая написана без такой обработки?

Это зависит от ученого. Возьмем пример из историко-литературной сферы. Проблема формульности, цитации, интертекстуальности, взаимосвязей между текстами. Установление цитаты в художественном тексте может ведь полностью перевернуть наши представления о нем: либо кардинально изменить его интерпретацию, либо бросить новый свет на историю текста. У любителей охотиться за цитатами бывает трудно понять, пользовались ли они машинным поиском. Я вот больше половины своих находок такого рода сделал в доинтернетную эпоху. Другое дело, что тогда я мог тратить годы на поиски одной цитаты, а теперь многое могу находить легко и быстро. Но это не значит, что в сети легко все можно найти, об этом мы уже говорили. В общем, я думаю, что такого четкого водораздела — между теми, кто пользуется в своих исследованиях современными информационно-поисковыми технологиями, и теми, кто работает «по старинке»,— нет, по крайней мере пока.

Трансформируется ли исследовательское поле под влиянием нового инструментария?

Конечно. Нерефлективное «знаточество» теряет смысл: ученый, который просто знает много разрозненных фактов, уже не так интересен. С другой стороны, поскольку добывать разрозненные

факты стало легко, а определенную ценность они все-таки имеют, то нередко происходит подмена науки псевдонаукой. Ведь раньше огромная коллекция фактов, собранных традиционными методами, имела больший смысл, потому что благодаря «древесной» структуре знаний факты отбирались человеком исходя из некоторой общей картины мира: факт — это листик на побеге, который растет на ветке, которая, в свою очередь, растет на большом дереве единого дисциплинарного знания. Пусть эти факты разрозненные, но это листики одного дерева. При новых подходах имеется возможность сразу же набрать эти листики, не зная, на каких ветках и каком дереве они растут. Но это общая проблематика современного знания: данные стало добывать легче, но умение анализировать данные, умение ставить вопросы сходит на нет.

Человек, который вручную отбирает эти листики, отбрасывает случайные совпадения, тогда как при массовой обработке неизбежна погрешность.

Это верно, есть плюсы и минусы, но они общие для разных видов знаний, не только для филологического. Доступность разного рода энциклопедической информации по разным аспектам знания из плюса легко становится минусом, потому что вместе с фактом тебе не предоставлен путь к этому факту. Раньше это знание добывалось так: ты вырабатывал методику его получения, и по дороге к факту — от дерева к ветке, от ветки к листику — встречал много интересного. Ну а когда «ствола больше нет», познающему непонятно, как один листик связан с другим. Но даже если это знание какого-то нового типа — скажем, «ризомное», а не «древесное», — то у него должны быть свои способы организации и аккумуляции. Иначе знание опять «распадется» на множество мелких фактов и перестает быть знанием.

Хаос, который творился в оцифровке на заре информационной эры, отчасти продолжается. Необходимо ли вмешательство государства или академических институтов для регламентации оцифровки, для введения каких-то норм, для запрета хулиганства, пиратства, любительства?

Это большой, сложный, многоаспектный вопрос. Нужно ли поддерживать государственные и профессиональные начинания разного рода? Думаю, что нужно. Но когда государство начинает что-то поддерживать, оно регламентирует. И не всегда эта регламентация приносит благо. В России был период, когда, например, создание электронных библиотек поддерживали и государство,

и научные фонды, государственные и частные. Было много разных инициатив, которые искали и находили себе поддержку. Видимо, рубеж столетий был кратковременным золотым веком для этого направления *Digital Humanities*. Многое из того, что появилось тогда, продолжает существовать поныне. Но затем поддержка крайне сузилась. Так, разработка и создание электронных библиотек до определенного момента считались научной проблемой, а теперь переведены в ранг прикладных, инженерных задач. В результате профессиональные электронные библиотеки оказались практически брошены без системной поддержки. В России, в отличие от западной «цифровой гуманитаристики», я не вижу поддержки крупномасштабных проектов — исключения единичны. Что касается пиратства, то его запрещать, на мой взгляд, бессмысленно и даже вредно — нужно, как мы уже говорили, менять нормы в сфере авторского права. А любительство запретить и вовсе невозможно — можно только противопоставить аматерству качественный профессионализм.

Какая разница между профессионалом и диггером? Раньше профессионал и был диггером…

Да он и сейчас иногда вынужден им быть. Есть же профессионалы разного типа. Есть революционеры, а есть собиратели. Есть те, кто совершает прорыв, кому важно наметить новое направление в целом и продвинуться, посылая к черту подробности. А есть те, которые хотят как раз этих подробностей. Видимо, эти два типа профессионалов всегда были и останутся в будущем, как всегда останется значимым куновское противопоставление «научной революции» «нормальной науке». Другое дело, что само по себе нахождение подробностей стало иным, чем раньше. Но вместе с тем многие новые методы — это всего лишь новая форма старых. Человек, который не владеет методикой поиска информации в традиционной библиотеке, и в электронной библиотеке мало что найдет.

Мы говорили об автоматизации рутинных процедур, об экономии времени. Не происходит ли простого перераспределения: сэкономленное время теперь должно уходить у профессионалов на освоение новых информационно-компьютерных методов?

Иногда происходит. Вот пример из личного опыта. То количество времени, которое я вложил в создание коллекций электрон-

ных текстов по Пушкину и пушкинской эпохе, конечно, вернулось сторицей, потому что с их помощью я добыл интересные результаты, получение которых без этого инструментария заняло бы гораздо больше времени. А ведь кроме этого был создан и дополнительный продукт, которым пользуются не только его создатели, но и другие люди. Традиционный «гутенберговский человек» постоянно что-то забывал и искал заново, а сейчас ситуация однажды найденного и удобно хранящегося решается все-таки гораздо лучше, чем пятьдесят лет назад.

Но сегодняшний человек гораздо хуже все запоминает…

А это уже обратная сторона легкости! Найти легко, но найденное уже не помнится.

К тому же запоминаются не только факты и истины, но и ошибки.

Не только запоминаются и сохраняются, но и множатся. Любой ошибочный текст размножается с той же скоростью, что и правильный. Это огромная проблема.

Кто-то работает с ошибками в базах?

В конкретных библиотеках такие специалисты есть. Тексты проверяют, ошибки находят и исправляют. О каких-то ошибках сообщают пользователи. Но в целом корректировка данных — это отдельная (и непростая) задача.

В заключение пара персональных вопросов от члена редколлегии «Логоса» Игоря Чубарова. Какие изменения в силу этих процессов претерпел художественный перевод?

Как практикующий переводчик, могу сказать, что новых полезных инструментов уже создано очень много. Я, например, пользуюсь корпусами параллельных текстов — они могут ускорить знакомство с традицией перевода того или иного фрагмента памятника. Благодаря электронным библиотекам всегда можно посмотреть употребительность тех или иных слов и выражений. Например, когда я переводил с итальянского на русский трактат о Петрарке поэта-символиста Вячеслава Иванова, я должен был учитывать русские аналоги его итальянских оборотов. Поэтому его следовало переводить не на русский язык «вооб-

ще», а на русский язык Вячеслава Ивановича Иванова. Поскольку к тому времени на РВБ уже появилась электронная версия брюссельского собрания сочинений Иванова, я имел возможность найти в его русских сочинениях сходные контексты и выбрать те синонимы, которые с высокой вероятностью выбрал бы и он. Да и от программ автоматического перевода есть прок, хотя еще десятилетие назад казалось, что большой практической пользы от них не будет, и некоторые специалисты предрекали закат машинного перевода.

Не могли бы вы прокомментировать рефлексию на современные способы хранения, переработки и трансляции данных, например, у Киттлера?

Мне предельно близок пафос Киттлера: изгнать *Geist* из *Geisteswissenschaften*. И не потому что в них нет *Geist*'а, а потому что *Geist* должен определяться апофатически. При исследовании продуктов творческой деятельности всегда образуется какой-то «остаток», с которым мы не можем справиться посредством доступных научных методов. Но, по-моему, нужно сначала применить доступные методы, затем попытаться разработать новые и только потом то, что останется, объявить, если захочется, не подлежащим естественно-научному изучению *Geist*'ом. А когда у нас все *Geist*, это может означать только одно: людям лень заниматься наукой и они свои вольные фантазии выдают за результаты исследований. Ярхо, которого мы не раз сегодня вспоминали, называл таких гуманитариев скорострельными интуитивистами и добавлял с сожалением: «Не любят люди работать».

Спасибо вам большое!