КОРПУС ТЕКСТОВ – НОВЫЙ ТИП СЛОВЕСНОГО ЕДИНСТВА

В.В. Рыков
Московский Физико-Технический Институт
Rykov2000@mail.ru

Ключевые слова: корпус текстов, корпусная лингвистика, общая филология, репрезентативность, фактура речи.

Разнообразие жанров и областей применения корпусов текстов и методов корпусной лингвистики породило проблему более точного определения и описания корпуса текстов как нового типа словесного единства. Определение, ставшее уже общепризнанным, описывает корпус текстов следующим образом - расположение на машинном носителе, все тексты корпуса получены специальными процедурами отбора для того, чтобы корпус стал репрезентативным, размечены на машинном носителе однородным образом для удобства обработки его компьютером, а также что весь корпус имеет конечный размер. В результате получается четыре минимальных базовых качества, делающих собрание текстов корпусом – расположение на магнитном носителе (machine readable form), процедуры отбора (sampling) и репрезентативность (representativeness), единство разметки или представления корпуса на этом носителе, а также конечный размер.

Это определение никем не оспаривается, но иногда понимается по-разному, что приводит к неадекватному использованию и пониманию роли корпуса в лингвистических исследованиях и даже к неверной интерпретации их результатов. Обсуждению этой проблемы посвящена настоящая публикация.

Первый компьютерный корпус был создан в США (так называемый Брауновский корпус текстов) вот уже почти сорок лет назад. За это время было созданы другие корпуса текстов – похожие и не похожие на Брауновский. Созданные корпуса текстов стали использоваться в самых разнообразных исследованиях. Соответственно, появилось много публикаций, описывающих не только результаты этих исследований, но и свойства корпуса текстов как нового типа словесного единства. Появилась новая наука – корпусная лингвистика. Получили названия разнообразные жанры корпусов текстов – двуязычные, учебные и т.п.

Однако разнообразие жанров и областей применения породило проблему более точного описания термина «корпус текстов» и, соответственно, определения и описания этого нового типа словесного единства. Определение, ставшее уже общепризнанным, наделяет корпус текстов следующими качествами - расположение на машинном носителе, все тексты корпуса получены специальными процедурами отбора для того, чтобы корпус стал репрезентативным и размечены на машинном носителе однородным образом для удобства обработки его компьютером, а также что весь корпус имеет конечный размер. В результате получается четыре минимальных базовых качества, делающих собрание текстов корпусом – расположение на магнитном носителе (machine readable form), процедуры отбора (sampling) и репрезентативность (representativeness), единство разметки или представления корпуса на этом носителе и конечный размер [5].

Это определение никем не оспаривается, но иногда понимается по-разному, а иногда, похоже, и понимается не совсем верно. Можно сделать вывод, что существует проблема интерпретации того, что такое корпус. Возможно это происходит потому, что корпус текстов все еще настолько новый филологический феномен, что количество и качество публикаций с адекватным описанием этого необычного типа словесного единства оставляют тем не менее простор для достаточно неадекватных суждений.

Это можно иллюстрировать многими примерами. До сих пор можно встретиться с вопросом – «Я насобирал на своем компьютере гигабайты хороших текстов – как мне из них сделать корпус?». Очень трудно воспримнимается ответ, что для хорошего корпуса навряд ли сгодится даже один текст.Часто корпусом называют произвольное собрание текстов на дискете, связанное любой общей идеей. Но даже, казалось бы, четыре описанных выше и

хорошо известных базовых свойств корпуса могут пониматься по-разному. Прежде всего, это относится к такому свойству, как репрезентативность. Простейшие связанные с этим свойством проблемы не выглядят простыми. Действительно, может ли двуязычный учебный корпус текстов использоваться для разработки систем машинного перевода? Насколько он будет адекватен (репрезентативен) для этой задачи? Или - отражает ли составленный по всем канонам корпус текстов газетных политических метафор все речевое многообразие газетной прозы? Будет ли этот корпус репрезентативен для любых лингвистических исследований газетной прозы?

Очевидно, что правильные ответы на эти вопросы имеют не только теоретическое, но и практическое значение. Поэтому в этой работе делается попытка более глубокого анализа и описания корпуса текстов, как нового типа словесного единства, исходящее как из традиционного его определения, так и из всего многообразия практики его реализации.

Расположение текстов корпуса на машинном носителе выглядит как наиболее тривиальное требование или свойство. В отечественной филологической традиции существует простая, понятная, однако не слишком широко известная парадигма научного описания этого свойства – система понятий общей филологии [2]. Одним из изначальных понятий этой науки считается фактура речи, которая рассматривается как материал речи, соединенный с орудиями речи. Каждая фактура речи формирует свой род словесности. Из четырех фактур речи первые три уже давно известны. Это устная, письменная и печатная. У четвертой фактуры речи орудием письма является компьютер, а материалом – машинный носитель. Сейчас мы все видим, что в четвертой фактуре речи формируется довольно новый род словесных произведений. Многие жанры этой фактуры имеют прототиы или аналоги в исторически более ранних ранних фактурах. Например – электронные книги, письма. Но многие – нет.

Обратив теперь внимание на корпус текстов, можно утверждать, что это один из жанров нового рода словесности, возникший в четвертой фактуре речи и не имеющий аналогов в устной, письменной или печатной речи. Он появился впервые именно на машинном носителе, записанный и подготовленный особым образом при помощи компьютера как орудия речи. Но, если продолжить это рассуждение, то обнаружится, что не только создание (написание) этого особого текста существенно отличается от написания рассказа или научной статьи.

Довольно показательным выглядит именно это свойство корпуса. Действительно, это приготовленное достаточно сложным образом словесное произведение, строго говоря, никто не читает в обычном смысле этого слова. Конечно, есть достаточно много жанров печатной речи, которые крайне редко читают подряд – например словари или энциклопедии. Но для корпуса это свойство оказывается существенно усиленным. В данном случае компьютер выступает даже не просто как средство визуализации текста на машинном носителе. Между его читателем (пользователем) и его текстами стоит достаточно сложный программный интерфейс, позволяющий сделать выборку словесного материала из корпуса по разнообразно сформулированным запросам. Так выглядят и происходят выглядевшие ранее тривиально в других фактурах речи процессы написания (создания) и чтения в приложении к такому филологическому киберчуду конца XX века как корпус текстов.

Как можно видеть, приложение парадигмы общей филологии позволяет более четко и вполне адекватно осмыслить даже такие казалось бы простые на вид понятия, как написание и чтение текста. Гораздо сложнее, как уже говорилось ранее, обстоит дело с другими свойствами корпуса – процедурами отбора (sampling) при его создании и репрезентативностью как конечным результатом этого процесса. Здесь они выступают как одно свойство в диалектическом единстве.

Действительно, здесь характерно именно то, что при создании отбор текстов в корпус производится по ясно описанным и четко выполненным критериям. Эти критерии конструирования корпуса (так называемые design criteria) и логически вытекающие из них процедуры отбора текстов для корпуса должны адекватно (репрезентативно) отразить в составе его текстов то, ради чего этот корпус создавался. В нашем примеревшийся уже классическим Брауновский корпус текстов (далее – БК) создавался для того, чтобы отразить специфические особенности печатной прозы США 60-х годов XX века. Этот специальный набор признаков и процедур, использующихся для создания корпуса текстов с целью отражения определенной лингвистической реалии, описывается парой взаимосвязанных признаков – отбором и репрезентативностью. Корпус для того, чтобы считаться корпусом, а не архивом или библиотекой, должен быть особым образом построен (отобран) и отвечать критерию качества - репрезентативности.

Мы видим, что, как уже было сказано выше, репрезентативность – это и есть то свойство, которое делает корпус корпусом, отличает его от более аморфных родственных образований, расположенных также на машинном носителе - например электронного архива или библиотеки. Репрезентативность (representativeness) – это название того набора принципов или требований, на основе которых был организован или составлен корпус. Для БК таким принципом репрезентативности была задача адекватно отразить лингвистические особенности печатной прозы США 1961 года. Исходя из этой главной цели выстраивался процесс его составления.

Это свойство тоже не является тривиальным. Действительно, мы не можем по своему усмотрению отбирать тексты для корпуса текстов Пушкина или газеты «Известия». Но мы можем отбирать и заменять тексты для того корпуса

текстов, как особого типа словесного единства, который мы для себя определяем и строим. В нашем корпусе вполне возможно, что один текст, скажем из жанра газетный репортаж, может быть легко заменен другим таким же текстом из того же жанра, при условии, что он соответствует критерию отбора.

Итак, в начале создания корпуса, как и любого другого текста, лежит замысел [2]. Для реализации этого замысла составляются критерии и процедуры отбора текстов в создаваемый корпус. При ясном понимании этого факта корпус текстов можно уподобить сложному зеркалу, отражающему внешнюю по отношению к нему речевую деятельность или некоторый представительный ее фрагмент (печатную прозу, устную разговорную речь). Такой корпус текстов можно назвать универсальным.

Однако практика составления и использования корпусов текстов (далее КТ) дает основания утверждать, что существет много жанров КТ, построенных по несколько другому принципу. Эти принципы основаны на том, что из доступного составителям множества текстов составляется КТ, отвечающий какой-либо специфической потребности его составителя (отладка системы машинного перевода, обучение иностранному языку и т.п.). Такие КТ можно назвать специальными. Очевидно, что использоваться они должны, как правило, в тех целях, для которых они спроектированы [1,3,4]. Вообще говоря, нельзя быть уверенным в надежности лингвистического исследования, основанного на изучении многообразия лексического состава какого-либо языка, если материалом для него послужил учебный корпус или другой специальный корпус. Специальный корпус не всегда может быть объективным отражением внешней по отношению к нему речевой деятельности. Вообще говоря, он предназначен для использования его только для тех целей, которые соответствуют замыслу его составителя., т.к. отбор текстов для КТ происходит в соответствии с этим замыслом. Надежные результаты в любых исследованиях может дать только универсальный КТ, т.к. он создается для отражения внешней по отношению к нему речевой деятельности.

Литература

1. Баранов А.Н. Проблема репрезентативности корпуса текстов // Труды Международного семинара Диалог-2001 по компьютерной лингвистике и ее приложениям. – Аксаково, 2001.
2. Рождественский Ю.В. Общая филология. – М., 1996.
3. Рыков В.В. Корпус текстов как отражение состояния русского языка // Труды Международного конгресса «Русский язык: исторические судьбы и современность». – Москва: МГУ, 2001.
4. Рыков В.В. Корпус текстов как реализация объектно-ориентированной парадигмы // Труды Международного семинара Диалог-2002. – М.: Наука, 2002.
5. McEnery T., Wilson A. Corpus Linguistics. – Edinburgh, 1997.