А.А. Баркович *(Минск)*

КОРПУС ТЕКСТОВ КАК ИСТОЧНИК И БАЗА НОВЫХ НОМИНАЦИЙ

Для развития современной лингвистики и науки в целом характерным является системный подход. Только сплошной, без субъективного изъятия анализ материала позволяет получить полную, системную картину состояния изучаемого объекта.

Таковым объектом, обширно распределенным во времени и пространстве и является тот или иной национальный язык, постоянно развивающийся и представленный колоссальным разнообразием случаев его употребления в виде текстов, построенных в разных условиях и с разными целями.

Корпус текстов – это вид корпуса данных, единицами которого являются тексты или их достаточно значительные фрагменты, включающие, например, какие-то полные фрагменты макроструктуры текстов данной проблемной области [1, с. 115].

В 60-е годы XX в. в Брауновском университете (США) впервые был создан большой корпус текстов на машинном носителе. Его авторы У. Френсис и Г. Кучера спроектировали его как набор из пятисот двухтысячесловных прозаических печатных текстов американского варианта английского языка. Тексты принадлежали пятнадцати наиболее массовым жанрам англоязычной печатной прозы США. Брауновский корпус а) стал своеобразным стандартом для создания других подобных корпусов; б) послужил импульсом для создания новой науки – корпусной лингвистики; в) область применения корпуса текстов и методов корпусной лингвистики оказалась намного шире и разнообразнее, чем ожидали его создатели.

Как объект исследования корпус текстов представляет собой сверхсложную многоуровневую динамическую систему. Отношения в такой системе, возникающие на уровне «корпус текстов – отдельный его текст», а также другие проблемы, так или иначе связанные с корпусом текстов как языковым объектом, стали изучаться новой наукой – *корпусной лингвистикой* (corpus linguistics). Спустя полвека стало ясно, что корпусная лингвистика – новое, быстро растущее направление, имеющее свои традиции, признанных лидеров, свои научные центры, методы и проблематику [2].

Сколько бы ни декларировалась независимость лингвистики от эстетико-культурного подхода, письменный язык – это прежде всего язык культуры. Поэтому собрание культурно значимых текстов на данном языке также представляет собою обладающий соб-

ственной ценностью источник для аннотированного корпуса. В практическом преломлении «культурная значимость», во-первых, означает, что текст является потенциальным источником расхожих цитат (что важно при оценке частотности того или иного языкового явления). Во-вторых, текст признаётся принадлежащим данному слою, если он вносит какой-то вклад в историю языка, например русского, в том числе и интересен языковыми экспериментами. Таким образом, это не что иное, как антология: сюда попадут все русские прозаики первого-второго ряда (в перспективе возможно также – и культурно значимые переводы, в том числе и Библия). Такой *корпус* можно условно назвать *репрезентативным.*

Но культурная значимость не гарантирует «стандартности» авторского языка. От корпуса требуется соответствие узусу и языковой компетенции его образованного носителя. Между тем тексты таких авторов, как Зощенко, Платонов или даже Гоголь, изобилуют фразами, которые не будут грамматически правильными с точки зрения современной авторам нормы. Отсюда задача: *стандартный, эталонный корпус* русского языка; языка лишённого по возможности сознательных стилевых и лексических экспериментов, тем не менее «гладкого» и «профессионального». Сюда не попадут, конечно, ни Зощенко, ни Платонов, ни Солженицын, однако «пойдут» такие писатели второго ряда, как Трифонов или Рыбаков, язык которых может почитаться достаточно «нейтральным» и «правильным», и даже, возможно, некоторые представители массовой литературы (такие тексты, как детективы, любовные романы и проч.). Поэтому возможно и расширение стандартного корпуса за рамки художественной литературы, с включением, например, публицистики.

Для некоторых задач, например этимологических, необходим **полный корпус** языка, в него входят все тексты (хотя бы печатные) на данном языке; единственным параметром его может быть время [3].

Работа над созданием корпусов текстов по русскому языку начата сравнительно недавно. Первые разработки в этой области относятся к началу 1970-х годов. Реально используемых корпусов сравнительно немного. В первую очередь здесь следует упомянуть «Уппсальский машинный фонд русского языка», создававшийся с 1987 г. в Уппсальском университете. Общий объем корпуса – около 1 млн словоупотреблении. В корпус отбирались художественные тексты с начала 1960 г., специальные журнальные тексты – с начала 1985 г. и газетные статьи – с начала 1987 г.

Существует также корпус текстов словаря языка Достоевского. В качестве основного текстового источника использовалось академическое полное собрание сочинений Ф.М. Достоевского.

«Компьютерный корпус газетных текстов русского языка конца 20-го века» (http://www.philol.msu.ru/~lex/corpus.html), подготовленный в течение 2000–2002 гг. в Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ, позволяет получить объективную картину состояния современного русского газетного языка (а отчасти и картины состояния всего языка в целом, если иметь в виду то, что в наборе газетных жанров присутствуют многие жанры других родов словесности). Для этого был осуществлен подбор обширного газетного материала для корпуса (тексты общим объемом более 11 млн словоупотреблений) на основе включения в него полных номеров 13 российских газет на русском языке за отдельные даты 1994–1997 гг., ежедневных и неежедневных, «левых» и «правых», центральных и местных, общих и профессионально ориентированных. Эти принципы позволяют получить относительно объективную и надежную картину соотношения в газетном материале текстов различного типа, например различных жанров и жанровых типов, их единиц и отношений между ними. Это обеспечивает также возможность анализа в дальнейшем не только общих для всего газетного языка характеристик, но и жанровой специфики лексико-фразеологической, грамматической и иной информации, поскольку тексты и единицы корпуса автоматически и полуавтоматически маркируются различного рода маркерами.

Важнейшее свойство корпуса текстов – его *репрезентативность* по отношению к проблемной области. Под репрезентативностью понимается способность корпуса текстов отражать все свойства проблемной области, релевантные для данного типа лингвистического исследования, в определенной пропорции, определяемой частотой явления в проблемной области. Другими словами, частота явления в корпусе должна быть близка частоте в проблемной области. Например, текстовые корпусы должны содержать в соответствующей пропорции тексты с базовыми, наиболее типичными макроструктурами, имеющимися в данной проблемной области.

Репрезентативность, конечно, не исчерпывается перечисленными параметрами. Так, в каждом конкретном случае может оказаться необходимым учесть стилистическую, временную, авторскую и другие составляющие текстового массива проблемной области.

Требование репрезентативности в самом простом варианте отражается в пропорциональном сужении проблемной области. В этом случае можно говорить о «пропорциональной стратегии» организации корпуса текстов.

Репрезентативность превращает набор текстов на машинном носителе в уникальное словесное единство – корпус текстов. Это свойство корпуса настолько важно, что иногда говорят о репрезентативности как о результате процесса самоорганизации корпуса, рассматриваемого, безусловно, как метафора. Тогда по идее, лежащей в основании корпусной лингвистики, *корпус текстов отразит объективную картину речевой деятельности независимо от воли его создателя* [4].

Действительно, *корпус текстов* все чаще рассматривается как *вторичная семиотическая система,* свойства которой не ограничиваются свойствами составляющих ее текстов, – ему присуща эмержентность. И выявление характеристик этой вторичной семиотической системы – задача отнюдь не тривиальная, как нетривиальна любая обработка результатов и их интерпретация. Более того, через характеризацию этой вторичной семиотической системы возможна характеризация субъекта – её пользователя. Сегодня в центре внимания корпусной лингвистики все чаще оказывается языковая личность, т. е. её речевая деятельность, массовая коммуникация, проблема её описания. Например, контент-анализ даст надежные результаты только на материале обширного корпуса текстов.

Равным образом корпус текстов – надежное обоснование для введения новой лексики. В современном белорусском языке, где сосуществуют две нормы, а словари сильно отстают и в фиксации новых слов, и в регулировании процесса пополнения лексикона, доказательным фундаментом для устойчивого существования нового слова является именно представительный корпус текстов. В частности, нам представляется обоснованным предположение о том, что слово (словоформа), которое неоднократно встречается в трёх белорусскоязычных газетах разной направленности в течение года-двух, вполне может быть введено в лексикон и стать составляющей словника новых слов белорусского языка.

Список литературы

1. Баранов А.Н. Введение в прикладную лингвистику. М.: Эдиториал УРСС. 2001. – 358 с.

2. Рыков В.В. Прагматически ориентированный корпус текстов // Тверской лингвистический меридиан: Теоретический сборник / Ред. И.П. Сусов. – Тверь. 1999. – Вып. 3. – С. 89–96 (http://www.tol.tversu.ru/Meridian3.htm).

3. Сичинава Д. В. К задаче создания корпусов русского языка в Интернете // http://corpora.narod.ru/article.html. (Сентябрь 2001).

4. McEnery T., Wilson A. Corpus Linguistics. – Edinburgh: Edinburgh University Press. 1999.