

Александр Баркович (Минск)

КОРПУС ТЕКСТОВ КАК ВТОРИЧНАЯ СЕМИОТИЧЕСКАЯ СИСТЕМА

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 [Баранов 2001, 115].

В 60-е годы 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 [Рыков 1999].

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 собственной 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 польского, в том числе и интересен языковыми экспериментами. Таким образом, это не что иное, как антология: сюда попадут все польские 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. Отсюда задача: стандартный, эталонный корпус национального языка; языка лишённого по возможности сознательных стилевых и лексических экспериментов, тем не менее "гладкого" и "профессионального". Возможно 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 [Сичинава 2001].

Работа над созданием корпусов текстов по польскому языку начата сравнительно недавно. Первые разработки в этой области относятся к началу 1970-х годов. Первый польский корпус – *Słownik frekwencyjny polszczyzny współczesnej* (SFPW) включал 500 тыс. словоформ из пяти стилевых регистров. Реально используемых корпусов польского языка сравнительно немного. В первую очередь здесь следует упомянуть создаваемые локально, в университетских лабораториях, корпусы текстов в Варшаве, Кракове, Вроцлаве и Познани. Созданные при помощи разных программных средств, они, к сожалению, несовместимы и не могут взаимно обмениваться данными.

Сегодня наиболее известен корпус текстов, создаваемый в Академии наук Республики Польша, – *Korpus Języka Polskiego PWN*. В корпусе 100 млн. словоупотреблений; до недавнего времени он распространялся только на CD-носителях, а сейчас доступен через Интернет (www.pwn.pl).

Существуют также специализированные корпуса текстов – как правило, они или создаются как подсистемы Korpus'a Języka Polskiego PWN или вливаются в него после создания. К таким корпусам можно отнести Корпус для лексикографа (Korpus dla leksykografa с Korpus Redakcji Słowników PWN) – двухмиллионный подкорпус, созданный отдельно и влитый в общий корпус PWN, или Корпус "Rzeczpospolitej", созданный на основе одноименной газеты и включающий более 3,5 млн. слов.

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 корпуса 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 [McEnery, Wilson 1999].

Действительно, корпус текстов все чаще рассматривается как **вторичная семиотическая система**, свойства которой не ограничиваются свойствами составляющих ее текстов, – ему присуща эмергентность. И выявление характеристик этой вторичной семиотической системы – задача отнюдь не тривиальная, как нетривиальна любая обработка результатов и их интерпретация. Более того, через характеристику этой вторичной семиотической системы возможна характеристика субъекта – ее 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.

Идея о том, что корпус текстов – система, которой присуща эмергентность, т. е. свойства, которыми не обладает совокупность составляющих систему элементов, тесно связана с мыслью о том, что такой же системой является словарь. Словарь – особый макротекст, в котором содержится больше информации, чем совокупная информация словарных статей. В Интернете доступны разные словари польского языка – толковые, переводные, энциклопедические (www.ling.pl). Доступность фактического материала позволяет плодотворно исследовать свойства словарей как корпусов текстов.

Таким образом, в Польше уделяется немалое внимание как прикладным проблемам создания корпусов, так и теоретическим проблемам корпусной лингвистики. Например, этому посвящена была очередная сессия Меж-

дународной школы гуманитарных наук Восточной и Центральной Европы (Варшава, 19-31 января 2004 г.).

ЛИТЕРАТУРА

Баранов А.Н. Введение в прикладную лингвистику. М, 2001.

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>).

Сичинава Д.В. К задаче создания корпусов русского языка в Интернете // <http://corpora.narod.ru/article.html>.

McEnery T., Wilson A. Corpus Linguistics. Edinburgh, 1999.