

Рыков В.В.

Москва

Понятие "корпус текстов" давно уже применяется в отечественной и зарубежной лингвистике. Обычно это понятие определяется как множество текстов, собранное и организованное в соответствии с какой-либо целью или замыслом (Рыков, 1999). Конечно, обычно предполагается, что эти тексты записаны на машинный носитель. Однако, это не мешает нам говорить о корпусе текстов Пушкина, невзирая на то, что не все эти тексты доступны машинной обработке. Более того, с появлением Интернет, не является даже обязательным, чтобы тексты корпуса были собраны в одном месте. Достаточно говорить о логике цели или замысла, объединяющего тексты в корпус, т.е. о его логической организации (как в случае с текстами Пушкина). Действительно, отдельные тексты корпуса могут быть записаны в памяти разных компьютеров, соединенных в сеть. Главное, чтобы они были доступны для обработки и исследования. Для этого нужно только знать физический адрес каждого текста, составляющего корпус.

Тогда мы можем говорить о корпусе текстов как о некотором логически организованном целом. Есть все основания говорить о первой метафоре или категории корпусной лингвистики - метафоре корпусообразующей логической дедукции. Речь идет о совокупности логических процедур, при помощи которых происходит отбор текстов для включения их в корпус.

Понятно, что деятельность по созданию и использованию корпусов текстов неизбежно требует разработки и тщательного обсуждения соответствующего набора категорий (Баранов, 1998). Тогда можно будет точнее обсуждать реальные и потенциальные результаты любого исследования корпуса. И здесь неизбежно приходится внимательнее изучить как внутренние свойства корпуса, содержащего исходный речевой материал исследования, так и внешние обстоятельства и условия его создания. Конечно, если исследователя заботит научная достоверность результатов его работы.

Остановимся на нескольких показательных примерах. Классическим примером, работающим в пользу революции в лексикографии, произведенной корпусной лингвистикой, является пример со словоупотреблением глагола "drive". Раньше никто не сомневался, что хорошим примером его употребления в речи может служить словосочетание "to drive a car". Но оказалось, что это не совсем так. В первом, ставшем уже классическим, Брауновском корпусе такое словосочетание встречается

только один раз на миллион слов. Гораздо чаще встречаются другие. Как следует оценивать этот результат? Насколько Брауновский корпус является представительным для суждений такого рода? Нужно ли менять соответствующие словарные статьи? Приходится признать, что мы имеем дело с другим понятием или метафорой - представительности или репрезентативности корпуса текстов.

Другим интересным примером может послужить нам употребление слова "berserk". Количественные результаты его употребления в современных английских газетных текстах, наиболее оперативно отражающих живую речь, озадачивают. Были проведены сравнения и оказалось, что частота употребления этого слова в различных английских газетах отличается, причем во много раз (Рыков, 1999). Причины таких значительных расхождений достаточно интересны. Они обусловлены многими факторами, в том числе и прагматического характера. Образно говоря, авторы, пишущие в ту или иную газету, ориентируются на образовательный уровень своей читательской аудитории.

Здесь же нас интересует только одна сторона проблемы – точность определения понятий однородность (гомогенность) и репрезентативность корпуса текстов. Многие лингвисты, работающие в области корпусной лингвистики, с сожалением отмечают недостаточность и неточность ее терминологии, в особенности в вопросах, касающихся определения именно этих понятий (Kilgariff, 1999).

Действительно, из приведенного примера со словом "berserk" видно, что необходимы точные, логически верные критерии вывода степени однородности и репрезентативности любого корпуса текстов. Для этого недостаточно однородности внешней процедуры отбора. Как можно увидеть из этого примера, корпус текстов (например, газетных) может иметь существенные отличия в зависимости от того, из какого источника (газеты) взяты его тексты.

Это явление может быть описано другими словами. В корпусной лингвистике пока остается недостаточно разработанным вопрос о стратегии описания и сравнения корпусов (Kilgariff, 1999). Пока что они описываются на основании внешних признаков - откуда взят текст. Можно привести один показательный пример. При статистическом исследовании Брауновского корпуса был обнаружен текст, принадлежавший жанру "газетный репортаж". По статистическим показателям, отражающим его лексический состав и грамматические особенности, он был однороден с текстами художественной прозы того же корпуса. Действительно, содержание текста (инаугурация президента Кеннеди) определило и объяснило это противоречие между внутренними речевыми свойствами этого текста и его принадлежностью, согласно внешним (издательским) критериям отбора, к "чужому" жанру (Рыков, 1984).

Таким образом, речь идет о сопоставлении внутренних, имманентных данному тексту (корпусу текстов) свойств с его внешними атрибутами (принадлежностью к определенному жанру, газете, времени). Эта проблема не является так уж абстрактной, как это может показаться на первый взгляд. Корпусная лингвистика, как филологическая дисциплина, с самого начала ориентировалась на общественную речевую практику. И ответа на этот вопрос требует именно речевая практика. Действительно, одним из путей повышения эффективности поиска информации в Интернете, является построение речевого образа пользователя. Этот образ может быть построен на основе корпуса его запросов, идущих с конкретного компьютера. Этот корпус может быть неоднороден (за компьютером работают совершенно разные по возрасту или специализации пользователи и т.п.). Равно как и важно

определить его принадлежность уже известному, типовому образу пользователя на основании того же корпуса индивидуальных запросов.

Тогда возможны будут вопросы типа "Насколько подобны (близки) эти два корпуса текстов" ? В простейшем, редуцированном варианте этот вопрос звучит так: "Насколько подобны эти два текста" ? На уровне лексики этот вопрос может выглядеть еще проще: "Существуют ли слова, на основании которых можно различать (уподоблять) группы текстов"? Частоты (или другие характеристики) таких слов-маркеров могли бы служить критерием при решении поставленных выше проблем. Решение этой проблемы существует и она описана в специальных исследованиях (Kilgariff, 1999). Оказывается, можно найти критерии (почти всегда на уровне легко узнаваемых и интерпретируемых лексических единиц) подобия/однородности и различия текстов практически для любой четко сформулированной прагматической цели. Существуют признаки, обнаруженные на представительных универсальных корпусах текстов, отличающие мужскую речь от женской, американский английский от британского и т.п.

Тогда мы сможем перейти к формулированию последней метафоры - самоорганизации корпуса текстов. Действительно, найдя маркеры, процедуры, критерии однородности корпуса (в зависимости от поставленной прагматической задачи) мы можем задать процедуру самоорганизации (селекции, отбора) нужного нам корпуса текстов согласно требуемым параметрам. Это можно сделать вручную, можно при помощи компьютера среди текстов, записанных на машинный носитель (в так называемой генеральной совокупности текстов). Еще более наглядно и эффективно (а также и эффектно) это выглядит, если поиск нужных текстов для корпуса вести в самом большом хранилище текстов - сети Интернет (Holmes-Higgin, 1996).

Тогда выбранные и опробованные критерии принадлежности текста к корпусу и однородности его состава послужат правилами отбора и составления (компиляции) нужного корпуса текстов. Если такой отбор происходит в автоматическом режиме, то это вполне напоминает процесс самоорганизации. Такой корпус может быть построен на базе определений, заданных пользователем (user-defined corpus) или набора требований, выдвинутых в соответствующей проблемной области (problem-defined corpus). В любом случае этот алгоритм может рассматриваться как самоорганизация корпуса в соответствии с прагматической целью или самоорганизация прагматически ориентированного корпуса.

Литература

Баранов А.Н. Автоматизация лингвистических исследований: корпус текстов как лингвистическая проблема // Русистика сегодня. - Москва, 1998. - N.1-2. - С.179-191.

Михайлов М.Н. Компьютерное обеспечение корпуса текстов (Взгляд пользователя) // Русистика сегодня. - Москва, 1998. - N.1-2. -

C.192-201.

Рыков В.В. Лингвистическое исследование Брауновского корпуса // Риторика и стиль. - Москва: МГУ, 1984. - С. 102-121.

Рыков В.В. Корпусная лингвистика (научно-аналитический обзор) // РЖ: Социальные и гуманитарные науки: Зарубежная литература. - М.:ИНИОН, 1996. - N.4 - С.43-51.

Рыков В.В. Прагматически ориентированный корпус текстов // Диалог-99. - Таруса, 1999. - С. 211-220.

Holmes-Higgin P., Ahmad H. Assembling and viewing a corpus of texts: self-organisation, logical deduction and spreading activation as metaphors // EURALEX'96 Proceedings. - Stockholm, 1996. - P.109-120.

Kilgariff A. Comparing Corpora // International Journal of Corpus Linguistics. - Philadelphia: John Benjamins, 1999. - Vol. 4(2). (www.itri.bton.ac.uk/~Adam.Kilgariff/ijcl.ps.pz)