КОРПУСНАЯ ЛИНГВИСТИКА: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

Ю.А. ВОЛОСНОВА, *доц. каф. перевода МГУЛ, канд. филол. наук*

В настоящее время в российском и зарубежном языкознании выделилось направление, известное под названием ***корпусная лингвистика (corpus linguistics)***. Корпусная лингвистика ставит целью изучение речевых закономерностей на материале больших текстовых объемов (корпусов), которые были предварительно обработаны, размечены и систематизированы в электронной форме. Отличие таких корпусов текстов от обычных электронных библиотек в том, что в них при помощи специальных программ можно искать необходимые фрагменты текстов по заданным параметрам, а затем обобщать и анализировать полученные данные. При этом само исследование занимает время, измеряемое в секундах.

Отличие корпусной лингвистики от традиционной заключается не только в применении компьютерных технологий, но и в общем подходе к изучаемым явлениям. Корпусная лингвистика направлена на изучение речи, воплощенной в текстах конкретных языков. Она идет от речевых данных к выводам, т.е. использует индуктивный эмпирический метод, при этом опирается на квантификативные исследования [1].

Интерес исследователей к корпусной лингвистике обусловлен широким развитием компьютерных технологий и программного обеспечения, которое предоставило исследователям исключительные возможности для получения и обработки информации. Тот объем сведений и данных, на который исследователь раньше тратил месяцы и годы, сидя в библиотеках и выписывая от руки нужную информацию, сейчас оказался буквально «на кончиках пальцев» и может быть получен и обработан за несколько минут.

Однако информационный взрыв порождает не только блага, но и проблемы. Парадоксально, но самой главной из них является обилие информации. С этой проблемой общество сталкивалось уже во второй половине двадцатого века, когда информационные технологии еще не получили такого развития. В 60-х гг. прошлого века в индустриально развитых странах существовало правило: если требовалось внедрить какой-либо проект стоимостью в пределах 50 000 долларов, который уже существовал, но на поиск информации о нем требовалось время, то дешевле было разработать аналогичный проект заново, чем заниматься поиском информации о нем.

С развитием Интернета процесс получения информации стал неизмеримо легче, быстрее и дешевле, но возникла новая проблема: как в огромном потоке сведений отобрать нужные именно вам. Так, например, на запрос «***корпусная лингвистика***» русскоязычная поисковая система Яндекс выдает 2033 страницы и 373 сайта по теме, а англоязычная система Yahoo по запросу «***corpus linguistics***» выдает 938 000 упоминаний данного словосочетания в Интернете, причем время поиска в Yahoo занимает 0,14 секунды. Хотя сравнение страниц и сайтов, с одной стороны, и количества упоминаний, с другой, не совсем корректно, тем не менее данная статистика дает представление, во-первых, об объеме информации по данному вопросу и, во-вторых, о сравнительном развитии данного направления в России и на Западе.

Но если рассматривать весь объем полученных ссылок, то время, затраченное на просмотр всей информации, будет несоизмеримо больше времени, проведенного в библиотеке в 20-м веке. При этом большая часть ссылок является перекрестной, т.е. авторы ссылаются друг на друга, либо упоминание слов запроса относится к рекламе публикаций или заказных рефератов и диссертаций, а также является анонсами семинаров и конференций. Иными словами, большая часть ссылок не несет научной, лингвистической информации и является «информационным шлаком», который надо отсеивать, на что тоже требуется время.

В такой ситуации вступает в действие так называемый Первый Закон Муэрса: «*потребитель уклоняется от использования поисковой системы, когда искать информацию более хлопотно, чем обойтись без нее*» [3]. Подобно тому, как промышленность выпускает все новые и новые бытовые приборы, чтобы облегчить домашний труд хозяйкам, так и программисты работают над созданием все новых и новых программных продуктов, призванных облегчить рутинный поиск информации и побудить пользователя к дальнейшим научным исследованиям. Но «*для точного, научно обоснованного поиска нужной информации необходимо строить специализированную семиотическую (знаковую) систему*» [2]. В.В. Рыков, пионер российской корпусной лингвистики, предлагает для создания такой системы создать информационно-поисковый язык (ИПЯ), под который составляется так называемый поисковый образ документа (ПОД), соответствующий концептуальному содержанию документа [2]. Это концептуальное содержание зачастую не выражено в пользовательском словесном запросе информации, что порождает огромный объем предоставляемой информации, который блокирует дальнейшую работу с ней.

Издавна лингвисты стремились установить связь между содержанием (концептами) и его лингвистическим выражением (словами), выявить закономерности и зависимости между содержанием и формой. Однако такие исследования зашли в тупик, поскольку концепции относятся к области философии и психологии, а слова – к области лингвистики. Одна и та же концепция может быть выражена разными словами, а в одно и то же слово может быть вложено разное содержание. Как говорил Шалтай-Болтай в «Алисе в Зазеркалье», «*когда я беру слово, оно означает то, что я хочу, не больше и не меньше*». Но несмотря на расхождение концепций, стоящих за словами, исследователи не оставляют попыток установить закономерности выражения содержания формой. Например, Муэрс считает, что каждое понятие может быть описано особой лексической единицей – дескриптором [4]. Он рассматривает людей, «привязанных к словам» (*word-bound*), как «невосприимчивых к идеям» (*idea-blind*), и полагает, что такому типу людей недоступен информационный поиск, забывая при этом, что даже самые абстрактные идеи непременно должны выражаться в тех или иных словах, а в компьютерных запросах слов тем более не избежать.

Метод дескрипторов, являющихся переходной формой от лексического запроса к понятийному и пригодных для компьютерного поиска, был описан в работе Р.П.Футрелла и С.Гауч [5]. Само название работы, включающее слово *bootstrapping* («самообеспечение»), предполагает, что переход от словесного описания запроса к понятийному происходит на основе слов все того же языка, только ограниченных в своем составе и организованных по-иному. Процесс перехода они называют расширением запроса (*expanding query*). Запрос расширяется и становится «концептуальным» (*conceptual*), используя обращения к оперативному корпусу необходимой лексики (*online database*) и так называемым специализированным «матрицам подобия» (*similarity matrix*). Таким способом устраняется бессмысленное лексическое сравнение (*word matching*) при поиске и осуществляется переход к концептуальному поиску (*conceptual retrieval*). Их исследование проводится на материале корпуса из 200 000 слов, который они при помощи предварительной разметки делят на классы и группы подобия (*simsets*), образующие иерархически расположенные гнезда (*clusters*). Данные компьютерных исследований были подтверждены на материале тезауруса Roget's. Таким образом, результаты компьютерного анализа почти не отличались от результатов эмпирического анализа Roget's, но само компьютерное исследование заняло несравнимо меньше времени, чем традиционное. Увеличение эффективности информационного поиска идет через сравнение семантических структур в поисковом запросе и текстах корпуса.

В западной лингвистике проблема языковых корпусов была успешно решена уже давно. Еще в 1967 г. Г. Кучера и Н. Френсис опубликовали работу «Вычислительный анализ современного американского английского языка», которая была проведена на Брауновском корпусе, состоящем из текстов современного американского варианта объемом 1 млн слов. Словарь «Американское наследие», опубликованный в 1960 г., стал первым словарем, составленным на основе текстового корпуса. Вслед за ним появился британский словарь Collins' COBUILD, основанный на корпусе «Банк английского языка». Далее последовали корпусы LOB (британский английский 1960-х гг.), Kolhapur (индийский английский), Wellington (новозеландский английский), ACE (австралийский английский), Frown Corpus (американский английский начала 90-х гг. XX в.), FLOB Corpus (британский английский начала 90-х годов XX в.). В 90-х гг. при участии консорциума издателей, Оксфордского и Ланкастерского университетов, Британской Библиотеки был составлен Британский Национальный корпус объемом 100 млн слов. Были также разработаны корпус латинских текстов «Персей», чешский корпус Карлова университета и др. В 1992 г. была создана организация Европейская корпусная инициатива (ECI), предусматривающая создание около 40–50 корпусов текстов на европейских языках, каждый объемом от 12 тысяч до пяти млн слов.

Что касается русского языка, то отечественная лингвистика отстает в этом от-

ношении от зарубежных исследований. Уппсальский корпус русских текстов создан еще в 1960-е гг. и остается единственным завершенным и активно используемым проектом такого рода. Он не удовлетворяет современным требованиям из-за устарелых материалов и ограниченности объема (1 млн словоупотреблений). Он также не является лингвистически аннотированным или размеченным (в нем не указаны морфологические, синтаксические, семантические свойства тех или иных сегментов текста, что затрудняет поиск по нему), в то время как современная лингвистика оперирует в основном аннотированными корпусами (*treebanks*).

Тем не менее, в российских институтах ведется обширная работа по созданию корпусов текстов. Начиная с 1980–1990 гг., работа над созданием компьютерных баз данных по русскому языку ведется в рамках Машинного фонда русского языка при Институте русского языка РАН под руководством В. М. Андрющенко. В 1995 г. был возобновлен Междисциплинарный семинар ДИАЛОГ. Это самое представительное российское мероприятие, целиком посвященное компьютерной лингвистике и ее приложениям, собирает каждый год большое число ведущих специалистов в области интеллектуальных языковых технологий из компьютерных фирм, вузов и научных институтов со всей России и из-за рубежа. Постоянными участниками ДИАЛОГа являются Московский Государственный Университет, РГГУ, Институт Языкознания РАН, компании ABBYY, Яndex, РосНИИ Искусственного Интеллекта, Институт Проблем Информации РАН. Научная программа семинаров ДИАЛОГ охватывает основные направления фундаментальных исследований и коммерческих разработок, находящихся на пересечении лингвистики, методов представления и обработки знаний и самых современных информационных технологий.

Кафедра математической лингвистики филологического факультета Санкт-Петербургского государственного университета (СПбГУ) совместно с Институтом лингвистических исследований (ИЛИ РАН) и кафедрой прикладной лингвистики Российского государственного педагогического университета им. А.И. Герцена (РГПУ) 10–14 октября 2006 г. провела международную научную конференцию «Корпусная лингвистика-2006». В ходе конференции обсуждались проблемы корпусной лингвистики, имеющие как теоретическое, так и прикладное значение. Тематика конференции позволяет охватить проблемы репрезентативности корпусов и отбора источников, лингвистической и экстралингвистической разметки, многие другие аспекты создания корпусов – лингвистические, программные, технологические. Отдельный аспект корпусной лингвистики, предлагаемый к обсуждению, – проблемы использования корпусов для проведения лингвистических исследований и разработки специализированного инструментария, а именно лингвистических поисковых систем (корпус-менеджеров), обеспечивающих удобный интерфейс для пользователей. Особое внимание планируется уделить координации различных разработок в области корпусной лингвистики в России и выработке стандартов.

Что касается разработки самих корпусов, то на филологическом факультете МГУ на базе лаборатории общей компьютерной лексикологии и лексикографии был создан Компьютерный корпус текстов русских газет конца XX века. Для этого был осуществлен подбор обширного газетного материала для корпуса (тексты общим объемом более 11 млн словоупотреблений) на основе принципов включения в него полных номеров 13 российских газет на русском языке за отдельные даты 1994–1997 гг. Его полная версия (более 1 млн словоупотреблений) готовится к представлению в Интернете. В настоящий момент на сайте лаборатории можно ознакомиться с тестовым фрагментом корпуса общим объемом более 200 тыс. словоупотреблений. Поиск по корпусу может проводиться по словам, корням слов и по различным типам информации, характеризующим русские лексемы, словоформы и тексты в целом. Особый интерес в данном корпусе представляет система маркировки газетных текстов маркерами конкретных жанров и жанровых типов. Изучение литературы по теме и проведенный анализ текстов позволили выявить круг основных жанрообразующих факторов [6].

В Институте лингвистики РГГУ ведется несколько проектов по созданию специализированных корпусов. Исследовательским коллективом под руководством С.И. Гиндина, объединяющим преподавателей, сотрудников и студентов кафедры теоретической лингвистики и Кафедры математики, логики и интеллектуальных систем, создана и продолжает развиваться гипертекстовая филологическая информационная система по творчеству В.Я. Брюсова. Система включает полный структурированный электронный корпус текстов Брюсова и его филологическое сопровождение. В Центре типологии ведется работа над созданием мультимедийного корпуса русских разговорных текстов.

Таким образом, корпус текстов, с одной стороны, это исходный речевой материал для корпусной лингвистики и для других лингвистических дисциплин; с другой стороны, результат деятельности корпусной лингвистики.

В настоящее время перед российской прикладной лингвистикой стоит более широкая задача создания Национального корпуса русского языка. Этот проект поддержан Российской академией наук в рамках программы «Филология и информатика». В нем участвуют лингвисты многих научных учреждений и вузов.

Национальный корпус – это собрание текстов в электронной форме, представляющих данный язык на определенном этапе его существования, отображающий данный язык во всем многообразии жанров, стилей, социальных и территориальных диалектов и т.п. Корпус должен быть представительным, т.е. содержать по возможности все типы письменных и устных текстов, представленных в языке, и все эти тексты должны входить в корпус по возможности пропорционально их доле в языке соответствующего периода. Это возможно только при значительном объеме корпуса (сотни миллионов словоупотреблений). Планируемый составителями объем Национального корпуса русского языка – 200 млн слов.

Разметка – главная характеристика корпуса. Она отличает корпус от простых коллекций (или «библиотек») текстов, в изобилии представленных в современном Интернете. Чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса. В Национальном корпусе русского языка в настоящее время используется четыре типа разметки: метатекстовая, морфологическая, акцентная и семантическая; в ближайшее время планируется внедрение синтаксической разметки. Система разметки постоянно совершенствуется.

Создание Национального корпуса дает огромные возможности для всех направлений лингвистических исследований. Возможность массовой статистической обработки текстов позволяет математически подтверждать или опровергать гипотезы, составлять грамматики и словари. Например, ни один английский толковый словарь не составляется сейчас без опоры на Британский Национальный корпус. Основными пользователями национальных корпусов в первую очередь являются лингвисты. Однако статистические данные об определенных языковых периодах могут заинтересовать литературоведов и историков, а также преподавателей и студентов высших учебных заведений. Составление учебников и словарей также невозможно без опоры на корпус.

Разрабатываемый Национальный корпус русского языка будет охватывать прежде всего период от начала XIX до начала XXI в. Этот период представляет как язык предшествующих эпох, так и современный, в разных социолингвистических вариантах – литературном, разговорном, просторечном, отчасти диалектном. В корпус включаются оригинальные (непереводные) произведения художественной литературы (проза и драматургия, в дальнейшем также поэзия), имеющие культурную значимость, а также представляющие интерес с точки зрения языка. Помимо художественных текстов, в корпус в большом количестве включаются и другие образцы письменного (а для современного этапа – и устного) языка: мемуары, эссеистика, публицистика, научно-популярная и научная литература, публичные выступления, частная переписка, дневники, документы и т. п.

Исследование с помощью национальных корпусов предоставляет исследователям

многочисленные возможности и направления. Кроме определения частотности словоформ, корпус дает представление о конкордансе. ***Конкорданс*** – список словоформ, встречающихся в тексте, расположенных в алфавитном порядке. В противоположность словарю слово здесь дается с его словесным окружением.

Сходным, но не тождественным, является исследование ***коллокации***, последовательности слов или терминов, встречаемость которых превышает процент случайных совпадений. Понятие коллокации относится к правилам употребления сочетаний слов, таких как сочетания определенных предлогов и глаголов или глаголов и существительных. Таким образом, корпусные исследования позволяют изучать речевые единицы в контексте, не выделяя их из естественного окружения.

Работа с корпусами текстов позволяет поднять на новый уровень дистрибутивную методику, которая направлена на определение совокупности:

1) всех линейных окружений данной языковой единицы;

2) всех сочетаний исследуемой языковой единицы.

Дистрибутивный метод может включать также контекстуальное исследование семантики, но прежде всего он заключается в возможности прогнозирования в речи (тексте) одних элементов на основании знания других. Например, фразеологически связанная единица или грамматическая форма многих слов часто дает возможность стопроцентного предсказания другого слова или грамматической формы. В остальных случаях распределения слов прогнозирование подчиняется вероятностным закономерностям, т.е. можно предсказать слово или грамматическую форму лишь с известной долей вероятности. Корпусная лингвистика позволяет выразить эту вероятность в точных цифрах.

Сопоставление двух или нескольких корпусов текстов позволяет создать компьютерные программы, так называемые накопители переводов, интерактивные инструменты, позволяющие переводчику накапливать в специальной базе данных эквивалентные текстовые фрагменты на двух языках, чтобы в дальнейшем быстро находить образцы для перевода новых текстов. В качестве фрагментов могут выступать слова, словосочетания и целые фразы. При работе над текстами, близкими по жанру и тематике, такие инструменты по мере пополнения базы данных все больше упрощают и ускоряют перевод. Программа сегментирует переводимый текст (выделяет фразы или обособленные обороты) и сличает полученные сочетания с элементами базы данных, в случае совпадения предлагая переводчику подставить готовый перевод фрагмента в конечный текст (или же подставляя его автоматически). В наиболее совершенных программах имеются встроенные морфологические модули и средства проверки орфографии для нескольких языков, а также относительно интеллектуальные средства сопоставления параллельных текстов с целью автоматического формирования парных фрагментов на двух языках. Процедуры сопоставления в разных программах различны, но, как правило, они включают элементы диалога и иногда требуют модификации одного из параллельных текстов.

Однако создание текстовых корпусов не решает всех проблем, стоящих перед исследователями. Важной составной частью работы является создание программ, позволяющих эффективно работать с корпусами. В первую очередь это так называемые корпус-менеджеры, поисковые системы на базе определенного корпуса. В качестве одного из примеров можно привести программу «Экспертная лингвистическая система», разработанную в СПГУ. Программа ЭЛС может использоваться для изучения лексического состава текстов корпуса, для поиска контекстов лексем или словоформ в корпусе (в частности при создании wordnet-тезауруса), для создания всевозможных частотных словарей (например словарей языка писателей).

Основной задачей программы ЭЛС является анализ частотности лексем в текстах корпуса. В выходном списке можно просматривать контекст либо для всех лексем, либо только для графических омонимов, по желанию пользователя. Также имеется возможность выбора варианта сортировки полученного списка. Кроме того, для выбранных пользователем файлов программа ЭЛС про-

изводит подсчет некоторых дополнительных статистических данных: количества предложений, количества слов в предложении и т.д. В возможности программы входит автоматическое деление слов на слоги.

Составители корпусов в работе сталкиваются с рядом проблем. Прежде всего следует отметить проблему ***репрезентативности*** корпусов, т.е. способности отражать все свойства проблемной области. Репрезентативность определяется фонетическими, морфологическими, синтаксическими, стилевыми параметрами. Создатель корпуса сначала ставит вопрос, какой корпус и для кого он создает. Невозможно представить компьютеру все тексты или все разговоры данного языка, поэтому создатели корпуса ориентируются на исследователей, которым этот корпус предназначен.

Лексикографам хватает большого корпуса с примерами малоупотребительных слов и/или их форм, но для специализированных исследований грамматистов, стилистов и др. надо охватить разные стили и жанры национального языка. И здесь составители сталкиваются с проблемой ***отбора текстов*** для корпуса. Например, У.Френсис и Г.Кучера ставили целью представить корпус текстов, отвечающих ясным и четким критериям отбора:

1. Происхождение и состав текста (автор должен был быть урожденным носителем американского варианта английского языка, диалог должен был занимать менее половины объема текста).

2. Синхронизация (включены были тексты, впервые изданные в 1961 г.).

3. Продуманное соотношение численной представленности различных жанров и отбор отдельных текстов при помощи особой вероятностной процедуры.

4. Доступность для компьютерной обработки (специальные пометы для передачи графических особенностей текста и т.п.).

Объем отдельного текста должен статистически достоверно отражать его стилевые особенности, а численный состав и соотношение жанров должны адекватно представлять стилевые особенности жанров и их относительный вес. При этом представляется противоречивым тот факт, что объективный программный инструмент, каковым является корпус, в основе своей построен на субъективном отборе и определяется человеческим фактором.

Подводя итоги данного обзора, можно выделить перспективы корпусной лингвистики и вытекающие из них проблемы. Во-первых, огромное количество информации затрудняет ее поиск, в результате чего требуется создание специальных корпусов и дополнительных поисковых инструментов, предназначенных именно для целей данного корпуса. Во-вторых, концептуальный поиск, призванный сделать более эффективным пользовательский запрос, в действительности сводится к поиску синонимов и систематизации их по семантически гнездам, т.е. относится скорее к области семантики, чем логики, и с трудом поддается формализации. В-третьих, корпуса строятся с целью дать объективную информацию, поэтому перед создателями корпуса стоит задача уменьшить фактор субъективности при отборе текстов для корпуса и разработать строгие и четкие критерии отбора. Все это не снижает значимости и перспективности исследований в области корпусной лингвистики и указывает на все новые направления прикладных исследований.

Библиографический список

1. Рыков, В.В. Персональный сайт курса лекций по корпусной лингвистике Рыкова В.В. rykov-cl.narod.ru/c.html
2. Клименко, С.В., Рыков, В.В. Корпусная лингвистика и информационный поиск. Доклад на конференции «Диалог-2000. http://www.dialog-21.ru/Archive/2000/Dialogue %202000-2/171.htm
3. Mooers C.N. «Mooers» law, or why some retrieval systems are used and other are not // American Documentation. – 1960. – Vol.11, N.3.
4. Mooers C.N. Descriptors // Encyclopedia of library and information science / A.Kent and H.Lancour, eds. – Vol.7. – New York, 1972. – Vol.7. – P. 31-45.
5. Futrelle R.P., Gauch S. Experiments in syntactic and semantic classification and disambiguation using bootstrapping // Acquisition of Lexical Knowledge from Text.– Columbus, OH. Assoc. Computational Linguistics, 1993. – P. 117-127.
6. Сайт лаборатории общей и компьютерной лексикологии и лексикографии. Филологический факультет МГУ им. М.В.Ломоносова. http://www.philol.msu.ru/~lex/corpus/corp_descr.html