

РОЛЬ КОРПУСНОЙ ЛИНГВИСТИКИ В ИЗУЧЕНИИ РАЗГОВОРНОГО АНГЛИЙСКОГО

Одной из характерных особенностей развития современной лингвистики является системный подход к изучению языка, который рассматривается как целостный объект, состоящий из взаимосвязанных элементов. Совершенно очевидно, что для проведения объективного исследования языка необходим сплошной анализ по возможности большего материала, позволяющего получить полную системную картину состояния изучаемого объекта.

Постоянно развивающийся национальный язык является обширно распространенным во времени и пространстве и, следовательно, любое лингвистическое исследование должно опираться на анализ достаточно полного и широкого языкового материала, чтобы быть достоверным. Именно этим определяется тот факт, что сбору материала придается такое большое значение. Если в прошлом сбор языковых данных и их дальнейшая обработка осуществлялись вручную, то в настоящее время развитие современных технологий, в частности современной компьютерной лингвистики, значительно облегчило сбор данных и позволяет собрать обширный материал, представляемый на машинных носителях, который далее подвергается разнообразным лингвистическим исследованиям.

Средством организации языкового материала для его системного анализа и объективного представления является создание корпусов данных (в частности корпусов текстов), объединяющих колоссальный языковой материал практически по любой тематике. Языковой корпус (от латинского *corpus* — *body*, основная часть) — это собрание лингвистических данных, в виде письменных текстов или транскрипций записанной речи, это большое количество «естественного», «реального» языка, собранное из различных источников и систематизированное на компьютерной основе.

Однако совокупность большого количества материала, представленного на машинных носителях, отнюдь не является корпусом, правильно отражающим реальное функционирование языковой системы. Корпус — это не просто отобранные по определенной методике и представленные в электронном виде тексты конкретных сфер употребления языка, но и их такая организация, которая позволяет правильно отразить исследуемые явления функционирования языковой системы. Подобная информация должна быть организована в виде некоторой базы данных, что облегчает ее практическое использование и научный анализ. Саммерс дает следующее определение корпуса: «Большое количество (собрание) естественного реального языка, взятое из различных источников, объединенного в единую компьютеризованную систему таким образом, чтобы предоставить максимум удобства при изучении языка» (Summers, 1992;14).

В последние десятилетия получила развитие новая научная дисциплина — корпусная лингвистика, которая дает возможность не только изучить большое количество языковых данных за значительно короткий промежуток времени, но и исследовать, систематизировать и, как справедливо отмечает Грейнджер, "раскрыть совершенно новые факты о языке." (Granger, 1998)

Проблемам и задачам корпусной лингвистики посвящены фундаментальные работы Синклера (1991), Свартвика (1990), Гринбаума (1996), Кверка (1992), Гвишиани (1997) и др. Как подчеркивается в их исследованиях, главной целью корпусной лингвистики является лингвистическое описание языковой системы, а также подтверждение, проверка гипотез о языке, как, например, определение различий использования того или иного звука, слова или синтаксической конструкции.

Как известно, лингвистические исследования в самом широком смысле проводятся для изучения либо языкового аспекта, либо речевого аспекта, как реализации реального употребления языка. Соответственно основной задачей корпусной лингвистики является полное и системное отражение содержательного общения на языке, а также анализ представленных конкретных фактов использования языка и на их основе представление и обоснование функционирования языка как системы. Также в

центре внимания корпусной лингвистики оказывается языковая личность, т.е. ее речевая деятельность, массовая коммуникация, проблема ее описания.

Главной особенностью направления исследования является подход к рассмотрению прикладных проблем лингвистики строго в коммуникативных процессах. При этом в центре внимания оказывается не только (естественный) язык как система, но и процесс содержательного общения на языке, и по возможности точное его описание. Корпусная лингвистика уточняет результаты и выводы проведенных ранее исследований речи; позволяет произвести новые более широкие и системные по охвату речевого материала лингвистические исследования. Составляемый корпус основывается на фиксации реальных языковых реализаций, на реальном функционировании языка в речи и представляет собой совокупность естественно — языковых выражений, реализуемых в речи.

Профессиональная или академическая работа с корпусом требует понимания теорий и предположений, которые лежат в основе строения и анализа корпуса. Подготовка корпуса не может быть изолирована от целей, намерений его использования, т.к. последние определяют выбор текстов, количество материала, систему кодировки и т.д. Источники данных, так же, как и ограничение видов использованных текстов предоставляют различный материал для дальнейшего рассмотрения и анализа широкого круга аспектов языка.

Так, базовыми единицами корпуса могут быть отдельные слова, короткие фразы, предложения, словосочетания (синтагмы), фрагменты текстов и прочее, в зависимости от целей проводимого исследования, т.е. предполагается ли фонетический, лексический, грамматический, синтаксический, жанровый и т.д. анализ собранного материала.

Корпус может предоставить данные по грамматическим различиям между Британским Английским и Американским Английским, например: параллельные предлоги

AmE

It's twenty of four

It's five after eight

BrE

It's twenty to four

It's five past eight

It's in back of the building
I haven't seen her in ages
или частота их использования

It's behind the building
I haven't seen her for ages

Prep	AmE	BrE
of	36,432	35,287
in	20,870	20,250
to	11,165	10,876

Существуют также национальные, исторические, географические, диалектические и прочие различия. Корпус может содержать современные или устаревшие тексты, тексты могут быть на одном или нескольких языках. Корпусы различаются и по своим размерам, периоду составления, жанру, доступности и т.д.

Первым современным электронно - читаемым корпусом был Брауновский Корпус Стандартного Американского Английского (The Brown Corpus of Standard American English), созданный в Брауновском университете в 1962-1963гг. под руководством У. Френсиса. Корпус состоит из 1млн. слов американского английского. Он содержит 500 текстов, каждый из которых включает 2 000 словоупотреблений. В корпусе содержатся тексты, относящиеся к литературному варианту английского языка. Тексты были образованы в различных пропорциях из 15 разных категорий текстов. Сегодня этот корпус считается маленьким и несколько устаревшим, хотя на его примере был построен LOB Корпус Британского Английского (Lancaster - Oslo - Bergen), включающий тексты приблизительно того же периода и тех же жанров.

По аналогии Брауновского корпуса был создан и Kolhapur Корпус Индийского Английского.

Появление Брауновского корпуса стало основой для разработки LIMAS корпуса в Германии, создававшегося в рамках проекта системы немецко-английского машинного перевода, но LIMAS корпус отличается от Брауновского корпуса жанровой классификацией текстов. Корпус состоит из 500 подкорпусов по 2000 словоупотреблений каждый, представляющих тексты различных типов, которые были опубликованы в 1968г. Корпус содержит 1млн. словоупотреблений.

Внушительным по своим размерам является Банк Английского Языка(BoE), основанный в 1991г. Главной целью корпуса является

обеспечение текстовой информации для составления словарей и для изучения языка. В отличие от других корпусов Банк Английского Языка представляет собой постоянно пополняемый новым материалом корпус. ВоЕ является контролируемым корпусом. На сегодняшний день он включает более 320 млн. слов.

В 1995г. был основан еще один большой корпус BNC Британский Национальный Корпус (British National Corpus). В нем содержатся как письменные, так и разговорные материалы. Сейчас этот корпус содержит 100 млн. слов современного британского английского.

В двадцати центрах по всему миру собирается материал для международного корпуса ICE (The International Corpus of English). Каждый ICE корпус будет состоять из 1млн. слов национальных вариаций английского (как письменного, так и разговорного).

Само слово "английский" является свидетельством новых параметров, в которых язык существует в двадцатом веке (Питерс, 1995). Уже принятым фактом является то, что английский язык охватывает весь современный мир, чтобы достичь всех уголков мира, людей разных национальностей, различных этнических групп. Его роль усиливается с каждым днем, т.к. современный мир движется в направлении "глобализации". Слово "английский" наиболее точно подходит прилагательному "глобальный"; что может быть более глобальным, чем всемирно распространенный английский язык (Гвишиани, Баранов, 2002).

Для нашего исследования особый интерес представляют разговорные корпусы. В этой связи следует особо подчеркнуть опыт создания разговорных корпусов в Германии (Баранов, Добровольский, 1997), где создание подобных корпусов началось довольно давно и уже накоплен значительный опыт методики сбора материала, организации и описания материала в корпусе, выработаны определенные принципы отображения изучаемой проблемной области, разработаны различные параметры обработки материала и их кодификация для соответствующего представления в последнем.

Одним из первых корпусов английской разговорной речи является Лондонско-Лундский Корпус Английского Разговорного Языка, основной целью которого была фиксация и отражение

употребления английского языка. Лондонско-Лундский Корпус Разговорного Английского (The London-Lund Corpus of Spoken English) — это корпус британского английского, но он отличается эксклюзивным содержанием транскриптов разговорного материала. Корпус был основан в 1959г. Кверком, Гринбаумом и Свартиком. Целью проекта является обеспечение источников точного описания грамматических особенностей английского языка в речи взрослого образованного носителя. Корпус содержит 287 текстов письменной и устной форм английского языка.

Еще одним проектом по созданию разговорного корпуса явился так называемый CRISTINE Корпус. В отличие от LLC, CRISTINE Корпус создан для изучения грамматических и других структурных признаков спонтанной, неформальной речи англичан разных классов и регионов Британии. Главной целью корпуса было развитие детальных, исчерпывающих, подробных стандартов английского языка именно в том виде, в котором он используется в реальной жизни. Это не первый структурно-анализированный корпус разговорного английского, но он отличается несколькими достоинствами. Он имеет общественную представительность, использует ту же аналитическую схему, что и созданный в рамках того же проекта SUSANNE Корпус, хорошо оформленный формат, по возможности доступный и для нетехнических пользователей. Это придает особую ценность при исследованиях.

В корпусе представлены 147 участников. 40 текстов корпуса приблизительно равны по своей длине, в среднем по 2800 строк каждый. Тексты корпуса содержат около 112 000 строк, соответствующих 80 500 «полным словам», не считая элементов сомнения и пауз.

Следует особо подчеркнуть, что исходной формой хранения корпусов разговорной речи является запись на магнитной ленте. Затем акустическая форма переводится в фонетическую транскрипцию. Как показывает опыт создания разговорных корпусов, полный вариант фонетической транскрипции из-за сложности транскрибирования и восприятия пользователем оказался неудачным. В результате нескольких опытов графической фиксации разговорной речи выработался неполный вариант

фонетической транскрипции, сохраняющий некоторые особенности произношения (интонация, диалектное произношение форм) и отражающий последовательность или одновременность произнесения речевых актов. Одной из форм фонетической транскрипции является так называемая "партитурная" нотация, считающаяся наиболее удачной. Она строится как музыкальная партитура, но вместо инструментов выступают участники коммуникации. Это позволяет более точно отразить процесс общения в абсолютном измерении (на временной оси) и в относительном измерении, характеризующем речь участников коммуникации в сравнении друг с другом (Mepge, 1993). Чисто технически партитурная нотация требует большей точности и более сложна в написании. Однако современные компьютерные средства позволяют упростить создание партитурных транскриптов речи.

При сборе разговорных данных для корпуса наиболее распространенными являются два подхода. Первый подход называется демографическим, т.е. намечается конкретное число говорящих, которым дается определенный промежуток времени для беседы.

Второй подход может быть назван «жанровым» (Mc Carthy) или контекстуально-регулируемым. Он заключается в том, что делается попытка выбрать не только говорящих, но и соответствующую среду, окружение и контекст. Жанровый подход пытается найти баланс между говорящим, окружением и контекстом.

Разговорный массив подвергается дальнейшему делению на монологи, диалоги, прямую и косвенную речь.

Сегодня создатели корпусов стараются включить столько разговорного материала, сколько позволяют технические и финансовые возможности. Создание разговорных транскриптов — это длительный и дорогостоящий процесс.

Из вышеизложенного становится очевидным, что для изучения разговорного английского материалы соответствующих корпусов и соответственно обработанные являются ценным источником для объективного анализа и получения достоверных результатов относительно тех проблем, которые непосредственно связаны с научным исследованием разговорного английского.

Приложение:

The Brown Corpus of Standard American English - создан в Брауновском университете в 1962-1963 гг. под руководством У. Френсиса. Корпус состоит из 1 млн. слов американского английского, содержит 500 текстов, каждый из которых включает 2 000 словоупотреблений.

Лондонско-Лундский Корпус Разговорного Английского (The London-Lund Corpus of Spoken English) - был основан в 1959 г. Кверком, Гринбаумом и Свартвиком, содержит 287 текстов письменной и устной форм английского языка.

CRISTINE Корпус - создан в 1990-ых, в корпусе представлены 147 участников, 40 текстов корпуса в среднем содержат по 2800 строк каждый, около 80 500 слов.

LOB Корпус Британского Английского (Lancaster - Oslo - Bergen) - состоит из 1 млн. слов письменного английского, включает 500 текстов в 2 000 словоупотреблений.

Kolhapur Корпус Индийского Английского - содержит 1 млн. слов.

LIMAS Корпус - был опубликован в 1968 г., состоит из 500 подкорпусов по 2000 словоупотреблений каждый, 1 млн. слов.

Банк Английского Языка (BoE) - основан в 1991 г., включает более 320 млн. слов.

BNC Британский Национальный Корпус (British National Corpus) - содержит как письменные, так и разговорные материалы, состоит из 100 млн. слов современного британского английского.

ICE (The International Corpus of English) - каждый ICE корпус будет состоять из 1 млн. слов национальных вариаций английского (как письменного, так и разговорного).

Литература:

Гринбаум О. Н. Компьютерные аспекты стилеметрии // Прикладное языкознание С Пб, 1996

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

Баранов А.Н., Добровольский Д.О. Постулаты когнитивной семантики. — Известия РАН. Сер. литературы и языка, т. 56, 1997, №1

Crystal D. The Cambridge Encyclopedia of the English language. CUP, 2000

Granger S. "The computer learner corpus: a versatile new source of data for SLA research" in learner English on computer, S. Granger, London & New York, Longman, 1998

Gvishiani N., Baranova L. MSLU. Virtual reality in ELT & Linguistics. Corpus Studies 6th International Conference

Menge H.H. Korpora gesprochener Sprache// Computereinsatz in der Angewandten Linguistik. Konstruktion und weiter Verarbeitung sprachlicher Korpora. Frankfurt am Main, 1993

McCarthy M. Spoken Language and Applied Linguistics. CUP, 1998

Peters P. "Australian English & World English", Cambridge Language Reference, News, 1995

Quirk R., Greenbaum S., A Student's Grammar of the English Language. Longman, London 1992

Sinclair J. Corpus, concordance, collocation, Oxford University Press, 1991

Summers D. "English in the Raw", Modern English Teacher, Vol.1, CUP, 1992

Svartvik J. The London-Lund Corpus of Spoken English: Description and Research. Lund University Press, 1990