А. А. Соловьёва (*Минск, БГУ*)

ЯЗЫКОВЫЕ КОРПУСЫ В КОНТРАСТИВНЫХ ИССЛЕДОВАНИЯХ

Данная статья посвящена проблеме использования корпусов в лингвистических исследованиях. Рассматриваются основные понятия корпусной лингвистики, виды корпусов и принципы их формирования. Помимо этого обсуждается вопрос о потенциале корпусной лингвистики в контрастивных исследованиях.

В заключении рассматриваются параллельные корпуса как подходящие источники данных для исследования различий и сходств между языками, а также понятие эквивалентности перевода как методология для контрастивного анализа.

В данной статье мы не случайно останавливаемся на проблемах использования корпусов при проведении контрастивного лексико-семасиологического анализа и составлении словарных статей. Корпусная лингвистика в России активно развивается, а Национальный корпус русского языка по праву может считаться одним из лучших и наиболее репрезентативных корпусов в мире, находящихся в свободном доступе. Языковым корпусом принято считать информационно-справочную электронную систему, состоящую из подобранной и обработанной по определенным критериям и правилам совокупности текстов, предназначенных для исследования языка. В английском языке термину корпус соответствует слово *corpus*. Форма множественного числа – *corpora*. В русскоязычной научной литературе исследований, в которых бы проводился обзор возможностей и репрезентативности английских корпусов, ранее не осуществлялось, поэтому данная работа может представить интерес для исследователей, занимающихся проблемами английской философии и языкознания.

В последние десятилетия бурно развивается корпусная лингвистика, имеющая в качестве своего предмета корпус как большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач [1]. Именно корпус позволяет получать данные, недоступные при традиционных методах лингвистического анализа (интроспекция, анкета, опрос информантов), а выводимые обобщения имеют статус не интроспективной догадки, как при традиционном подходе, а эмпирически наблюдаемого факта [Там же]. По критериям репрезентативности и отбора текстов различаются два основных типа корпусов:

– корпуса, относящиеся ко всему языку;

– сознательно смещенные корпуса (У. Э. Френсис), относящиеся к какому-либо подъязыку (жанр, стиль, язык определенной социальной группы и т.п.) [2].

Корпуса первого типа строятся на основе принципа дедукции – движения от общего к отражающему это общее частному корпусу текстов. Они универсальны и имеют целью отражение всего многообразия речевой деятельности, существующей независимо от исследователя. Такие корпуса доступны полностью или частично всем заинтересованным лицам через Интернет.

Среди современных корпусов наиболее известны традиционные: Британский национальный корпус – British National Corpus (http://thetis.bl.uk), насчитывающий около 100 млн словоупотреблений, и Мангеймский корпус немецкого языка (около 1 млрд словоупотреблений) (http://corpora.ids-mannheim.de/~cosmas). В последнее время все более популярным среди германистов становится и создаваемый Лейпцигским университетом корпус современного немецкого языка (http://wortschatz.uni-leipzig.de). В России в последние годы создание корпусов также осознается как одна из актуальных задач компьютерной лингвистики. Уже функционирует Национальный корпус русского языка, представленный по адресу http://ruscorpora.ru. Ведется работа по созданию представительного национального корпуса русского языка (Большой корпус русского языка), объемом не менее чем в 100 млн словоформ (bokrcorpora.narod.ru).

Корпуса второго класса строятся специально для отражения некоторого лингвистического или культурного феномена. Здесь критерий отбора текстов для корпуса задается его создателем в зависимости от целей практической или научной деятельности. Методологии построения данного типа корпусов индуктивны и занимаются проблемой корректности отражения описываемого феномена в корпусе текстов, призванным отразить в себе этот феномен [2]. К русскоязычным корпусам данного типа можно отнести, например, Компьютерный корпус текстов русских газет конца XX века (http://www.philol.msu.ru/~lex/corpus) и Корпус политических метафор.

В зависимости от материала, положенного в основу корпуса, и способа его организации, а также конкретных целей его использования существуют несколько классификаций корпусов. Например, в зависимости от классифицирующего признака выделяются: по типу данных письменные, речевые и смешанные корпусы, по признаку параллельности – одноязычные, двуязычные и многоязычные массивы и т.д. [3]. С точки зрения использования лингвистами, наиболее значимыми считаются исследовательские, иллюстративные, статические, динамические виды корпусов, а также корпусы параллельных текстов [Там же].

На данном этапе (7 апреля 2018 г.) общий объем национального корпуса русского языка составляет более 600 млн слов. Количество словоупотреблений в основном корпусе превышает 283 млн [4]. В устном подкорпусе содержится 11,3 млн лексических единиц. Для нашего исследования особый интерес представляет газетный подкорпус русского языка, так как выбранная нами для исследования лексико-семантическая группа «финансы» очень широко используется в текстах публицистического характера, причем текстов современных. Объем данного подкорпуса 173 521 766 слов. Это корпус материалов СМИ 2000-х годов. Он велик по объему и тем самым очень удобен для статистических наблюдений над языком XXI века, но не может быть присоединен к Основному корпусу без нарушения сбалансированности. В корпусе представлены следующие издания: «Известия», «Комсомольская правда», «Новый регион 2», «РБК Daily», «РИА Новости», «Советский

спорт», «Труд-7». Составителями выбраны именно эти издания, так как они довольно достоверно отражают разнообразие основных газетных жанров. Также в целях сбалансированности тексты равномерно распределены в периоде с 2000 по 2010 годы.

Особенного внимания заслуживает параллельный корпус, а именно пары английский – русский, русский – английский. Если раньше английский подкорпус находился в стадии разработки, а его объем составлял 20,2 млн словоупотреблений, то в декабре 2015 г. было проведено пополнение, и теперь объем корпуса составляет 24,6 млн словоупотреблений.

Однако соотношение английских аутентичных текстов и переводных неравномерно. Так, подкорпус дает возможность исследователю обратиться к текстам английских авторов XX–XXI вв. с параллельным переводом на русский. Объем этой части подкорпуса составляет всего лишь 725,6 тыс. словоупотреблений. Переводы же произведений русской литературы на английский язык составляют 3,3 млн слов. Таким образом, использование национального корпуса русского языка может стать хорошим подспорьем для контрастивных исследований в области английского языка, однако разница в наполнении между двумя частями английского подкорпуса значительна, и для изучения определенных лексико-семантических групп такой объем может быть недостаточным. Поэтому для нашего исследования мы были вынуждены прибегнуть к помощи английских корпусов. Большинство крупных языков мира уже имеет свои национальные корпуса (различающиеся по полноте и уровню научной обработки текстов). Общепризнанным образцом является, в частности, Британский национальный корпус: на него ориентированы многие другие современные корпуса.

Британский национальный корпус представляет собой сборник из 100 миллионов слов из письменного и устного языка из широкого круга источников, предназначенных для представления широкого поперечного сечения британского английского языка конца XX века как устного, так и письменного. Последнее издание – BNC XML Edition – выпущено в 2007 году.

Письменная часть Британского национального корпуса (90 %) включает, например, выдержки из региональных и национальных газет, специализированных периодических изданий и журналов для всех возрастов и интересов, академических книг и популярной литературы, опубликованных и неопубликованных писем и меморандумов, школьных и университетских эссе, среди многих других видов текста. Устная часть (10 %) состоит из орфографических транскрипций неписанных неформальных разговоров (записанных добровольцами, выбранными из разных возрастных, региональных и социальных классов демографически сбалансированным способом) и лексики разговорного языка, собранной в разных контекстах, начиная от официальных деловых или правительственных встреч и заканчивая радиопередачами и телефонными звонками.

Корпус кодируется в соответствии с Руководством Инициативы кодирования текста, чтобы представлять как результат CLAWS (автоматический определитель части речи), так и множество других структурных свойств

текстов (например, заголовки, абзацы, списки и т.д.). Полная классификация, контекстная и библиографическая информация также включаются в каждый текст в виде заголовка, соответствующего кодирования текста.

Работа по строительству корпуса началась в 1991 году и была завершена в 1994 году. После завершения проекта никаких новых текстов не было добавлено, но корпус был слегка пересмотрен до выпуска второго издания BNC World (2001) и третьего издания BNC XML Edition (2007). С момента завершения проекта два субкорда с материалами из Британского национального корпуса были выпущены отдельно: BNC Sampler (общая коллекция из миллиона письменных слов, один миллион разговорных) и BNC Baby (четыре миллиона миллионных образцов из четыре разных жанра).

Полная техническая документация, охватывающая все аспекты Британского национального корпуса, включая ее дизайн, разметку и содержание, приводится в Справочном руководстве для British National Corpus (XML Edition).

На сегодняшний день существует несколько подкорпусов в Британском национальном корпусе.

Одноязычный: речь идет о современном британском английском, а не о других языках, используемых в Великобритании. Однако небританские слова английского и иностранного языка встречаются в корпусе.

Синхронный: Он охватывает британский английский конца двадцатого века, а не историческое развитие данного языка до сегодняшнего дня.

Общий: он включает в себя множество разных стилей и разновидностей и не ограничивается какими-либо конкретными предметными областями, жанром или регистром. В частности, он содержит примеры как устного, так и письменного языка.

Образцовый: Для письменных источников образцы из 45 000 слов берутся из разных частей текстов одного автора. Более короткие тексты объемом до 45 000 слов или тексты с несколькими авторами, такие как журналы и газеты, полностью включены.

В заключении несколько слов о потенциале корпусной лингвистики в контрастивных исследованиях. Наиболее перспективным в этом направлении представляется разработка параллельных корпусов текстов ПКТ), состоящих из множества исходных тестов (оригиналов) и их переводов на один или несколько языков. Использование ПКТ, помимо преимуществ одноязычного корпуса при изучении отдельного языка, создает практически оптимальные условия для проведения исследования проблем передачи различных языковых значений в сопоставляемых языках, поиска использующихся в переводческой практике эквивалентов.

На необходимость подобного рода исследований указывал уже В. Г. Гак, когда утверждал, что, «сравнивая переводы с подлинником, мы сплошь и рядом обнаруживаем такие лексические замены, которые не предусматриваются никакими словарями и никак не могут быть объяснены с их помощью» [2], а «речевые параллели можно выявить лишь с помощью лингвистического эксперимента…, либо сравнивая переводы» [4].

Осознавая тот факт, что работа с электронными корпусами открывает новые возможности и, безусловно, повышает уровень объективности лингвистического исследования, мы должны всегда помнить о том, что когда «целью формирования корпуса является лексический анализ, приходится отказаться от всех надежд на полное отображение лексики. Лексикон языка настолько велик, настолько огромно, почти бесконечно, число возможных сочетаний, что мы не в состоянии представить себе корпус, который вместил бы все это…. Лексикон, напротив, фактически отрытая система. Как бы мы долго не расширяли выборку, мы по-прежнему будем встречать еще не представленные в ней слова» [1].

Таким образом, исходя их описания всех видов языковых корпусов, можно сделать вывод, что имеется довольно хорошая база для проведения исследования. Стоит отметить, что языковой корпус необходимо использовать для выявления и сравнения частотности употребления той или иной лексической единицы в английской и русской публицистической речи. Помимо этого обращается внимание на контекстность употребления слова, а также при помощи примеров из корпуса происходит наблюдение за значениями, в которых употребляются единицы из выбранной лексико-семантической группы.

ЛИТЕРАТУРА

1. *Гак, В. Г.* Сопоставительная лексикология / В. Г. Гак. – М. : Международные отношения, 1977. – 264 с.
2. *Поликарпов, А. А.* Об одной рецензии / А. А. Поликарпов. – Режим доступа : //http:// www.linguide.com.ua/content. – Дата доступа 09.01.2017.
3. *Падучева, Е. В.* Высказывание и его соотнесенность с действительностью (референциальные аспекты семантики местоимений) / Е. В. Падучева. – М. : Едиториал УРСС, 2004. – 228 с.
4. *Кубрякова, Е. С.* Язык и знание: На пути получения знаний о языке: Части речи с когнитивной точки зрения. Роль языка в познании мира / Е. С. Кубрякова – М. : Языки славянской культуры, 2004. – 560 с.

This article is devoted to the problem of corpora use in linguistic research. The basic concepts of language corpora, their types and principles of their formation are considered. In addition, the question of the potential of corpus linguistics in contrastive studies is discussed. In conclusion, parallel corpus is considered as a suitable source for studying the differences and similarities between languages, and also the notion of equivalence of translation as a methodology for contrastive analysis.