А. А. Гойло
Минск, МГЛУ

ПРИМЕНЕНИЕ КОРПУСОВ ТЕКСТОВ ДЛЯ СОПОСТАВЛЕНИЯ НАЦИОНАЛЬНЫХ РАЗНОВИДНОСТЕЙ АНГЛИЙСКОГО ЯЗЫКА

В статье сравниваются два корпуса текстов, отражающие национальные разновидности английского языка, – Global Web-Based Corpus of English (GloWbE) и International Corpus of English (ICE). Рассматриваются преимущества и недостатки изучаемых корпусов. Так, GloWbE имеет гораздо больший объем материала (1,9 млрд слов), а ICE, в свою очередь, имеет более широкий охват источников языкового материала. Кроме того, в работе предлагается классификация корпусов текстов и отмечаются сферы практического применения данного источника материала за пределами сферы научных исследований (например, в обучении иностранным языкам).

Поиск материала – важный этап любого научного исследования. И если раньше лингвистам приходилось вручную отбирать материал из множества разрозненных источников (книг, журналов, статей или видео), на что могло уходить значительное количество времени, то теперь материал можно найти гораздо быстрее при помощи корпусов текстов.

Корпуса текстов представляют собой совокупность специально отобранных языковых данных, хранящихся в электронном виде. Эти данные, как правило, особым образом размечены (аннотированы). Базовым способом аннотации является разметка слов по частям речи (POS-tagging). Благодаря разметке появляется возможность гибкого поиска определенных фактов языка в корпусах.

Корпуса текстов можно классифицировать по нескольким признакам (в качестве примеров в основном представлены корпуса английского языка):

- по количеству языков: одноязычные (British National Corpus) и многоязычные (Europarl);
- по направленности: общие (Brown Corpus) и специальные (например, они могут давать информацию о просодических явлениях, отражать только устную речь или определенный речевой жанр);
- по временной соотнесенности: синхронические (Corpus of Contemporary American English) и диахронические (например, Corpus of Historical American English охватывает период с 1810 по 2009 год).

В связи с процессом глобализации и распространением английского языка по всему миру появляется интерес к изучению национальных разновидностей английского языка, которые развились в странах, где английский или является одним из официальных языков, или получил широкое распространение среди населения. Национальные разновидности имеют ряд отличительных черт, которые можно исследовать с помощью специальных корпусов текстов. В данной работе мы рассмотрим два наиболее известных и объемных корпуса вариантов английского языка: International Corpus of English (ICE) и Global Web-Based English (GloWbE).

ICE на данный момент включает в себя 13 подкорпусов разновидностей английского языка. Каждый подкорпус содержит 500 текстов устной и письменной английской речи размером примерно в 2000 слов, т.е. каждый сегмент, отражающий национальную разновидность английского языка, включает в себя примерно 1 миллион слов. При этом бóльшая часть текстов представляет собой записи устной монологической и диалогической речи. Например, из 300 текстов устной речи в каждом подкорпусе 180 диалогов и 120 монологов, которые могут включать в себя записи телефонных звонков, парламентских выступлений, школьных уроков, подготовленной и неподготовленной речи [1].

Каждый подкорпус создавался отдельной командой исследователей из соответствующей страны. Несмотря на это, все сегменты корпуса полностью совместимы друг с другом и имеют общую структуру, а также схему аннотации [2].

ICE не имеет веб-интерфейса, и все материалы корпуса доступны для скачивания непосредственно на жесткий диск. Для многих подкорпусов доступны для скачивания также аудиофайлы с записями устной речи, расшифровки которых используются в качестве текстов в соответствующем подкорпусе.

GloWbE включает в себя 1,9 миллиарда слов и 20 разновидностей английского языка. При этом разновидности представлены неравномерно: больше всего слов насчитывают сегменты, в которых отражена американская и британская разновидности (около 380 миллионов), а, например, в сингапурском, малайзийском и нигерийском подкорпусах около 40 миллионов. Текстовые данные для корпуса были отобраны из 1,8 миллиона уникальных веб-страниц из 20 англоговорящих стран. При этом создатели корпуса убрали повторяющиеся элементы веб-страниц (юридическая информация и др.). Сайты были отобраны с помощью поисковых запросов в Google с применением поиска по региону. Принадлежность сайта к определенной стране определялась в первую очередь по доменному имени. В случае, если доменное имя не принадлежит определенной стране (например, *.net*), принадлежность сайта устанавливалась при помощи IP-адреса, информации о геолокации на странице и ссылок на страницу. Таким образом, можно утверждать, что выдача результатов по поисковым запросам в данном корпусе с очень высокой долей вероятности действительно относится к определенной стране [3].

Тексты в корпусе разделены на две категории: общие и блоги. Корпус включает в себя различные типы текстов, представленных в Интернете: текстовое наполнение веб-страниц разнообразной тематики, статьи, пользовательские блоги, комментарии пользователей на различных ресурсах (например, форумах). Размер корпуса позволяет успешно искать примеры употребления редких конструкций или примеры, отражающие сочетаемость малоупотребительных слов, что может быть затруднительным, если использовать корпуса меньшего размера.

Корпус имеет веб-интерфейс, позволяющий выполнять различные поисковые запросы, например, поиск по частям речи, поиск сочетаемости, поиск слов с заданным контекстом. Веб-форма также позволяет задать определенные параметры вывода информации, например, есть возможность группировать выдачу, показывать частотность и т.п. Данные особенности корпуса делают возможным выявление различных лексических, морфологических и синтаксических особенностей 20 разновидностей английского языка. Поиск по ближайшему контексту позволяет сравнивать семантику слов в различных разновидностях английского.

Таблица

Результаты сравнения двух корпусов

Корпус	ICE	GloWbE
Размер	13 миллионов слов	1,9 миллиарда слов
Разновидности английского языка	13	20
Источник материала	Различные (например, научные статьи, записи телефонных разговоров, телерепортажи)	Веб-страницы
Типы текстов	Устная и письменная речь различных видов	Разнообразные тексты, представленные в Интернете
Веб-интерфейс	Нет	Есть
Скачивание материалов корпуса	Доступно	Доступно

Как можно заметить, оба корпуса охватывают значительное количество национальных разновидностей английского языка, однако GloWbE намного больше, чем ICE, по размеру. Больший размер корпуса является очевидным преимуществом, так как дает возможность поиска наименее употребительных языковых единиц. Однако ICE представляет бóльшее разнообразие в источниках языковых данных и намного более широкий охват различных речевых жанров, в то время как GloWbE в качестве источника использует только данные с интернет-страниц.

Сфера применения корпусов текстов выходит за рамки академических исследований. В частности, корпуса текстов можно активно использовать в сфере образования – в преподавании иностранных языков или при составлении учебных пособий (например, с помощью параллельных кор-

пусов текстов отбирать «рабочие» варианты перевода определенных речевых оборотов, встречающихся в текстах общественно-политической или других направленностей).

Таким образом, очевидны перспективы использования корпусов текстов в области контрастивных исследований языков и их разновидностей.

ЛИТЕРАТУРА

1. International Corpus of English (ICE) [Electronic resource]. – Mode of access : http://ice-corpora.net/ice. – Date of access : 01.10.2017.
2. UCL Survey of English Usage [Electronic resource]. – Mode of access : http://www.ucl.ac.uk/english-usage/projects/ice.htm. – Date of access : 01.10.2017.
3. Corpus of Web-Based Global English [Electronic resource]. – Mode of access : https://corpus.byu.edu/glowbe. – Date of access : 01.10.2017.

The author provides grounds for using text corpora in comparative study of the varieties of English. Two of the existing corpora of the English language varieties are then compared. Commentary on the strengths and weaknesses of both corpora and examples of application of corpora outside the academic field are also provided.

Р. В. Детскина
Минск, МГЛУ

СПОСОБЫ ПЕРЕДАЧИ НА РУССКИЙ ЯЗЫК СЕМАНТИКО-СИНТАКСИЧЕСКИХ ОТНОШЕНИЙ МЕЖДУ КОМПОНЕНТАМИ НЕМЕЦКИХ СЛОЖНЫХ ПРИЛАГАТЕЛЬНЫХ

В статье рассматриваются некоторые способы передачи семантико-синтаксических отношений между отдельными компонентами немецких сложных прилагательных с точки зрения первого компонента.

В ходе исследования мы выполнили перевод на русский язык 1000 немецких сложных прилагательных. В качестве материала для исследования использовались сложные прилагательные, взятые из произведений E. M. Remarque «Drei Kameraden», Torey L. Hayden «Meine Zeit mit Sheila», Werner Kopacka «Der Schneepalast».

Особые трудности вызывает перевод немецких сложных определительных прилагательных. Эта группа прилагательных обладает наибольшим разнообразием семантико-синтаксических отношений между ее компонентами и менее изучена. Новое прилагательное образуется не как результат прибавления одной основы к другой, а как результат некоторого структурного преобразования одной формы наименования в другую, более компактную. В этом и проявляется синтаксичность рассматриваемого словообразовательного процесса. Синтаксичность словообразовательных процессов вообще является универсальным свойством, определяющим характер и типологию