ОБЗОР ОСНОВНЫХ ЛИНГВИСТИЧЕСКИХ КОРПУСОВ АНГЛИЙСКОГО ЯЗЫКА

Лесная М. И.
Харьковский национальный университет им. В. Н. Каразина
г.Харьков, пл. Свободы, 4, тел. 7075136,
e-mail: marianna1983@yandex.ru

Для современных филологов лингвистические корпусы являются эффективным инструментом анализа устной и письменной текстовой информации. Лингвистический корпус понимаем как большой, представленный в машиночитаемом виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач [1, с. 7]. Наибольшее количество корпусов создано на основе английского языка, что объясняется его распространенностью, а также высоким уровнем развития корпусной лингвистики в США и Великобритании.

Первый корпус был создан в 1960-е гг. в Брауновском университете (США) У. Френсисом и Г. Кучерой. Корпус содержал около 500 текстов объемом 2000 печатных знаков каждый, написанных на американском варианте английского языка, и включал морфологическую и синтаксическую разметку.

Британский национальный корпус (BNC – British National Corpus) на сегодняшний день считается эталонным, поскольку по его образцу создавалось большинство современных корпусов. Данный корпус был разработан в Оксфордском университете при участии Ланкастерского университета и Британской библиотеки с 1991 по 1994 год. Объем корпуса – свыше 100 млн. словоупотреблений, 90% из которых соответствуют письменным текстам, 10% – устным. BNC включает как метатекстовую, так и морфологическую разметку, и является синхронным корпусом общего назначения. Поиск по данному корпусу доступен на сайтах http://corpus.byu.edu/bnc и http://www.natcorp.ox.ac.uk/ и осуществляется с помощью корпусного менеджера XAIRA [2].

Оксфордский корпус английского языка (Oxford English Corpus) является самым большим из когда-либо созданных. Он содержит свыше 2 млрд. словоупотреблений и отражает состояние современного английского языка на всей территории его распространения. В корпусе представлены тексты, созданные с 2000 года, основную часть составляют материалы, размещенные во Всемирной Паутине. Также в Oxford English Corpus вошел ряд текстов на бумажных носителях, в частности, технические инструкции, статьи из газет и журналов, произведения художественной литературы и т.п. Данный корпус используется сотрудниками Oxford University Press, в частности для составления словарей [6].

Корпус современного американского английского (The Corpus of Contemporary American English, COCA) является самым большим корпусом

английского языка, находящимся в свободном доступе (сайт http://corpus.byu.edu/coca/). Он был создан М. Дэвисом (Brigham Young University, США) в 2008 году. COCA является корпусом смешанного типа, поскольку в нем представлены и письменные тексты (художественная проза, популярные журналы, газеты, научная литература и пр.), и устная речь. Корпус современного американского английского содержит 445 млн. словоупотреблений и включает тексты с 1990 года по настоящее время. Поисковый интерфейс предоставляет широкие возможности: поиск слов, словосочетаний, лемм, грамматических форм, синонимических рядов и т.п. Корпус обновляется два раза в год и удобен для отслеживания динамики лингвальных изменений [8].

Национальный корпус американского английского (The American National Corpus, ANC) создается по образцу Британского Национального корпуса (BNC) и призван отразить американский вариант современного английского языка. В корпусе представлены тексты, созданные начиная с 1990 г. На сегодняшний день объем корпуса составляет 22 млн. словоупотреблений, фрагмент объемом 15 млн. словоупотреблений доступен для свободного скачивания с сайта http://www.americannationalcorpus.org/OANC/index.html [7].

«Банк английского языка» (The Bank of English) является составной частью одной из крупнейших языковых баз Collins Corpus, которая используется для создания современных словарей. Данный корпус содержит свыше 650 млн. словоупотреблений, 65-70% из которых соответствуют британскому варианту английского языка, 25-30% – американскому. В состав корпуса входят различные типы письменных текстов и устной речи. Корпус включает метатекстовую разметку, а также частеречную разметку. В общедоступной версии корпуса, размещенной на сайте http://www.collinslanguage.com/content-solutions/wordbanks существует возможность выбора подкорпуса: британские книги, газеты, журналы, радиопередачи и др [5].

Кембриджский международный корпус (Cambridge International Corpus) создавался как база для составления учебных материалов и словарей английского языка. В корпус вошли британские и американские тексты разных типов, записи устной речи носителей британского и американского вариантов английского языка общим объемом свыше 700 млн. словоупотреблений. Отдельный подкорпус образуют тексты экзаменационных работ студентов из разных стран, изучающих английский язык в качестве иностранного [3].

Международный корпус английского языка (International Corpus of English, ICE) является совокупностью национальных подкорпусов, отражающих словоупотребление в различных вариантах английского языка (Австралия, Великобритания, Гонконг, Индия, Ирландия, Канада, Кения, Малайзия, Новая Зеландия, Сингапур, США, Танзания, Филиппины, Шри-Ланка, Южная Африка, Ямайка). Каждый подкорпус включает письменные и устные тексты и имеет объем 1 млн. словоупотреблений. В настоящее время International Corpus of English находится на этапе разработки. Полностью подготовлен Британский компонент корпуса (ICE-GB), его тексты снабжены морфологической и синтаксической разметкой. На сайте

http://www.ucl.ac.uk/english-usage/projects/ice.htm предоставляется свободный доступ к фрагменту корпуса объемом 20 тыс. словоупотреблений [4].

Благодаря репрезентативности, большому объему, разнообразию жанров, наличию как устных, так и письменных текстов созданные корпусы английского языка предоставляют филологам богатое поле для исследования языка. Систематический анализ корпусных данных позволяет эффективно отслеживать изменения в языке, создавать точные лексикографические описания, верифицировать лингвистические гипотезы.

Список литературы

1. *Захаров В.П., Богданова С.Ю.* Корпусная лингвистика: учебник для студентов гуманитарных вузов / *В.П. Захаров, С.Ю. Богданова*. – Иркутск: ИГЛУ, 2011. – 161 с.

2. *British National Corpus* [Electronic Resource]. – Way of access : http://www.natcorp.ox.ac.uk/corpus/index.xml. – Title from the screen.

3. *Cambridge Language* [Electronic Resource]. – Way of access : http://www.cambridgelanguage.com/content-solutions/wordbanks. – Title from the screen.

4. *International Corpus of English* [Electronic Resource]. – Way of access : http://ice-corpora.net/ice/. – Title from the screen.

5. *MyCobuild.com* [Electronic Resource]. – Way of access : http://www.mycobuild.com/about-collins-corpus.aspx. – Title from the screen.

6. *Oxford Dictionaries* [Electronic Resource]. – Way of access : http://oxforddictionaries.com/words/the-oxford-english-corpus. – Title from the screen.

7. *The American National Corpus* [Electronic Resource]. – Way of access : http:// http://americannationalcorpus.org/. – Title from the screen.

8. *The Corpus of Contemporary American English* [Electronic Resource]. – Way of access : http://corpus.byu.edu/coca/. – Title from the screen.