УДК 81

Е.М. Коваленко
Южный федеральный университет
г. Ростов-на-Дону, Россия
emkovalenko@sfedu.ru

КОРПУСА ТЕКСТОВ В ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ ПЕРЕВОДЧИКА

[*Kovalenko E.* Corpora in the professional activity of the translator]

The paper is devoted to the application of linguistic corpora in the translator's professional activity. It deals with the collection of English corpora created by Mark Davies, linguistics professor of Brigham Young University. The opportunities to use the largest three of them as an instrument for translator's professional activity and English language researches are described. Some examples of learning the English dialect variants with the aid of the Corpus of Global Web-Based English are offered.

Key words: Corpora, Corpus Linguistics, translation.

Информационные технологии в переводческой профессиональной деятельности чаще всего рассматриваются в контексте использования систем машинного перевода. Но кроме этих систем есть еще другие лингвистические ресурсы в сети Интернет, которые могут помочь в работе переводчика и проведении исследований в области переводоведения. К таким ресурсам следует отнести, в первую очередь, корпуса текстов.

Сеть Интернет предоставляет возможность работать пользователю с большими массивами данных, однако поисковые системы, которые в некотором смысле «умеют» работать с текстами, не предназначены для решения лингвистических задач, т.к. не дают возможности получить лингвистическую информацию из этих данных. Лингвисту нужны тексты со специальной разметкой, которая содержит информацию о языковых свойствах этих текстов. Решением этой задачи занимается корпусная лингвистика, которая разрабатывает общие принципы создания и использования лингвистически размеченных корпусов текстов с применением компьютерных технологий [1, с. 7]. Корпусная лингвистика стремится дать максимально объективное описание языковой системы на основе изучения реальных текстов, вырабатывая собственный «особый способ отражения речевого материалы в корпусе текстов» [2], при этом под лингвистическим корпусом текстов понимается «большой, представленный в машиночитаемом виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [1, с. 7]. Корпуса текстов позволяют гарантировать типичность данных и обеспечить полноту представления языковых явлений при условии достаточно большого (репрезентативного) объема корпуса, а естественная контекстная форма данных разного типа дает возможность их объективного изучения.

С точки зрения использования корпусов в переводческой деятельности несомненный интерес представляют корпуса параллельных текстов (КоПарТ), в которых можно найти контексты слов, имеющих несколько переводных эквивалентов, подобрать эквиваленты терминологических и фразеологических словосочетаний. Однако и монолингвальные корпуса также представляют интерес для контрастивных исследований в области переводоведения, например, для поиска лексических вариаций, синонимов, коллокаций, слов и/или фраз, имеющих культурную маркированность; для изучения различительных признаков типов текстов, скрытых моделей использования лексики, развития концептов во времени.

Среди доступных для свободного использования корпусов особый интерес представляют корпуса английского, испанского и португальского языков, созданные профессором лингвистики Марком Дэвисом из университета Бригама Янга (США). Он разработал архитектуру и интерфейс для коллекции корпусов, в которую входит вариант Британского национального корпуса (англ. *British National Corpus*, сокр. *BYU-BNC*) [3], созданного в 1980-х – начале 1990-х годов в Oxford University Press. Этот корпус содержит 100 миллионов слов из текстов разных жанров – разговорного, художественной литературы, журнального, газетного, научного, и позволяет вместе с другими корпусами английского языка изучать различные варианты английского языка в их естественном контексте, становясь незаменимым помощником переводчика и исследователя в сфере переводоведения. В этой коллекции представлены также три крупнейших корпуса английского языка: корпус современного американского английского языка (англ. *Corpus of Contemporary American English*) (220 тыс. текстов; 520 млн. слов) [4], корпус исторического американского английского (англ. *Corpus of Historical American English*) (115 тыс. текстов; 400 млн. слов) [5] и Интернет-корпус английского языка (англ. *Corpus of Global Web-Based English*) (1,8 млн. текстов, 1,9 млрд слов) [6].

Корпус современного американского английского языка (англ. *Corpus of Contemporary American English*, сокр. *COCA*) [4] содержит 220 тыс. текстов и 520 млн. слов – это самый большой и единственный общедоступный корпус американского английского. Корпус регулярно обновляется (каждый год добавляется порядка 20 млн. слов), поэтому позволяет исследовать текущие изменения в языке. В равных долях представлены 5 жанров:

1. *Устный* – 109 млн. слов (состав: транскрипция спонтанной речи из почти 150 телевизионных программ и радиопередач).
2. *Художественная литература* – 105 млн. слов (состав: короткие рассказы и пьесы из литературных, детских и популярных журналов, первые главы первого издания книг с 1990 по настоящее время, сценарии кинофильмов).
3. *Популярные журналы* – 110 млн. слов (содержит статьи из почти 100 журналов различной тематики – новости, здоровье, дом садоводство, женские, финансовые, религиозные и спортивные журналы).
4. *Газеты* – 106 млн. слов (содержит статьи из 10 газет США, тексты взяты из различных разделов (местные, спортивные, финансовые новости)).
5. *Научные журналы* – 103 млн. слов (содержит статьи из почти 100 различных рецензируемых научных журналов, которые представляют все научные направления согласно системе каталога библиотеки Конгресса США, как в целом, так и по количеству слов в год).

Корпус исторического американского английского языка (англ. *Corpus of Historical American English*, сокр. *COHA*) [5] содержит 115 тыс. текстов и более 400 млн. слов – это самый большой размеченный корпус исторического английского языка. Корпус сбалансирован по жанрам и поджанрам с разделением на десятилетние периоды, например: на тексты художественной литературы приходится 48–55% от общего объема текстов корпуса для каждого десятилетия (с 1810-х по 2000-е гг.), с учетом выделения поджанров художественной литературы – проза, поэзия, драма, и т.д. С точки зрения создателей корпуса такой подход позволяет изучать реальные исторические изменения в английском языке, а не изменения в структуре жанров в тот или иной исторический период.

В корпусе представлены 4 типа текстов:

1. *Художественная литература* (источники: тексты из Project Gutenberg (1810–1930), исторические книги проекта Making of America Корнельского университета (1810–1900), книги (1930–1990), театральные и киносценарии, тексты из корпуса *COCA* (1990–2010)).
2. *Журналы* (источники: журналы проекта Making of America Корнельского университета (1810–1900), журналы (1900–1990), тексты из корпуса *COCA* (1990–2010)).
3. *Газеты* (источники: не менее пяти газет для каждого десятилетия (1850–1980), тексты из корпуса *COCA* (1990–2010)).
4. *Публицистика* (источники: тексты из Project Gutenberg (1810–1900), www.archive.org (1810–1900), книги (1900–1990), тексты из корпуса *COCA* (1990–2010)). Тексты сбалансированы по составу в соответствии с каталогом библиотеки Конгресса США.

Интернет-корпус английского языка (англ. *Corpus of Global Web-Based English*, сокр. *GloWbE*) [6] содержит около 2 млрд. слов, собранных автоматически с 1,8 млн. веб-страниц на 340 тыс. сайтов из 20 различных англоязычных стран мира – это один из крупнейших интернет-корпусов английского языка. Этот корпус можно использовать для изучения различий между диалектами английского языка. Например, для изучения особенностей употребления лексики в диалектах английского языка можно сравнить частоту употребления любого слова или фразы для 20 диалектов. В частности, согласно корпусу такие формы, как *fortnight*, *trousers*, *rained off*, *on holiday*, *at university*, [*be*] *different to*, *rather more* + [*adj. (прилаг.)*] чаще встречаются в британском английском, чем в американском диалекте. В ирландском диалекте распространены такие формы как *jackeen*, *banjax*, *culchie*, *childer*, *soft day*, в австралийском английском – *bikkies*, *thongs*, *rockmelon*; в сингапурском диалекте – *rakyat*, *makan*, *hand phone*, на Ямайке используют *ackee*, *bammy*, *guinep*, *callaloo*. Можно проводить также сравнительный анализ между группами стран, например, в Южной Азии распространены такие формы как *out of station*, *eve teas*, *be elder to*, *keep in view* [6].

Кроме лексических исследований с помощью данного корпуса можно изучать идиоматические выражения, частота употребления которых «чувствительна» к объему корпуса. Например, изучение в корпусе выражений, связанных с «*head*», показало, что такие идиомы, как *in over ~ head*, *head start*, *heads or tails*, *talking* [*head*], (*like*) *a deer in the headlights*, *cooler heads* более распространены в американском и канадском варианте английского языка, а частота использования таких выражений, как *price on ~ head, head over heels (in love), head and shoulders above, two heads are better (than one), from head to toe,* равномерно распределена по всем диалектам английского языка (согласно данным корпуса) [6].

Корпус GloWbE позволяет также проводить морфологические и синтаксические исследования. Например, согласно корпусу, такие формы, как [*be*] *spoilt* (vs *spoiled*) и [*have*] *learnt* (vs. *learned*) меньше распространены в американском и канадском варианте английского, в которых чаще встречаются такие формы как *dove* (vs *dived*), чем в других диалектах. Размер корпуса позволяет сравнивать редко встречаемые синтаксические конструкции в разных диалектах, например: в американском и канадском английском конструкции [*stop*] *someone* V-*ing* и [*prevent*] *someone* V-*ing* (*they stopped* / *prevented him going*) встречаются довольно редко по сравнению с британским, ирландским, австралийским, новозеландским диалектом, а такой дискурсивный маркер, как "*that said*" чаще встречается в американском варианте английского языка [6].

С помощью корпуса можно изучать коллокации (частотно устойчивые словосочетания) в разных диалектах английского языка, например: *scheme* в американском английском используется в более негативных сочетаниях, чем в британском (например, *evil, fraudulent, nefarious*); в британском английском использование *cupboards* не ограничивается только кухней, как, например, в американском диалекте; в британском английском *boost* (глагол) используется прежде всего как "increasing" something – *увеличение чего-то* (финансов, количества), в то время как в американском английском значение расширяется до "improvement" – *улучшение, повышение* (настроения, безопасности) [6].

Одним из наиболее интересных применений корпуса является возможность сравнивать частоты употребления слов и словосочетаний в разных странах. Например, согласно данным корпуса, слова *Quran* или *Allah* чаще всего встречаются в Пакистане и других мусульманских странах, слово *Buddh* – в Шри-Ланке, а *feminism* – в светских англоязычных странах; сочетание [*adj. (прилаг.).*] + *book* в азиатских странах чаще употребляется по отношению к религиозным текстам (*divine, revealed, Buddhist*), чем в светских англоязычных странах; слово *belief* в Южной Азии чаще сочетается (слева) с прилагательными *Hindu, corrupt, wrong, Islamic, heretical*, а в светских англоязычных странах – с прилагательными *silly, contradictory, liberal* и *Catholic*; слово *wife* в светских странах употребляется гораздо реже, чем в странах Востока, с такими прилагательными как *chaste, temporary, obedient, Muslim, virtuous* [6].

Кроме описанных корпусов коллекция Марка Дэвиса содержит еще корпус новостных текстов, автоматически собираемых в сети Интернет, корпуса текстов британского варианта Википедии, Британского парламента, сценариев американских сериалов, небольшой корпус канадского варианта английского языка и корпус на основе Google Books (американский, британский английский и испанский языки). Таким образом, эта коллекция дает уникальные возможности для изучения различных вариантов английского языка, что представляет несомненный интерес не только для лингвиста-исследователя, но и переводчика, преподавателя языка, а также для всех, кто изучает английский язык.

ЛИТЕРАТУРА

1. *Захаров В.П., Богданова С.Ю*. Корпусная лингвистика: Учебник для студентов гуманитарных вузов. Иркутск: ИГЛУ, 2011.
2. *Рыков В.В.* Курс лекций по корпусной лингвистике. URL: http://rykov-cl.narod.ru/c.html (дата обращения 2.11.2016).
3. British National Corpus (BYU-BNC) URL: http://corpus.byu.edu/bnc/ (дата обращения 2.11.2016).
4. Corpus of Contemporary American English (COCA). URL: http://corpus.byu.edu/coca/ (дата обращения 2.11.2016).
5. Corpus of Historical American English (COHA). URL: http://corpus.byu.edu/coha/ (дата обращения 2.11.2016).
6. GloWbE: Corpus of Global Web-Based English. URL: http://corpus.byu.edu/glowbe/, http://corpus.byu.edu/glowbe/help/dialects.asp (дата обращения 2.11.2016).

14 ноября 2016 г.