Нина Борисовна Мечковская, д. филол. н.
Белорусский государственный университет,
Минск, Беларусь
э-почта: nina.mechkovskaya@gmail.com

Nina Borisovna Mechkovskaya, Dr. of Sc.
(Philology)
Belarusian State University, Minsk, Belarus
e-mail: nina.mechkovskaya@gmail.com

КОРПУСНАЯ ЛИНГВИСТИКА: ОСОБЕННЫЕ ЧЕРТЫ, ЗНАЧИМЫЕ ДОСТИЖЕНИЯ, ТЕНДЕНЦИИ РАЗВИТИЯ

Показаны предметные и организационные особенности корпусных исследований, обусловленные спецификой Big Data и алгоритмическим разноаспектным препарированием языковых данных в целях различных исследований, в том числе за пределами гуманитарного знания.

Ключевые слова: частотный словарь; синтез толкового и частотного словарей в Macmillan 2007; семантический компонентный анализ 100-тысячного словника; высокочастотные слова как семантические множители; частотная авторская лексикография; проблемы фразеологии в свете корпусных данных; словоупотребление (=c/y).

CORPUS LINGUISTICS: SPECIAL FEATURES, SIGNIFICANT ACHIEVEMENTS, DEVELOPMENT TRENDS

The article shows the subject and organizational features of corpus studies, caused by the specifics of Big Data and algorithmic multi-aspect preparation of language data for the purposes of various studies, including those outside the humanities.

Key words: frequency dictionary; synthesis of explanatory and frequency dictionaries in Macmillan 2007; semantic component analysis of a 100,000-word dictionary; high-frequency words as semantic multipliers; frequency author's lexicography; problems of phraseology in the light of corpus data; word-usage.

Электронный корпус текстов, представляя собой программно обеспеченную кроссплатформенную лингвистическую среду, предназначенную для извлечения из корпуса, с помощью корпусных менеджеров, нового знания о текстах и языке корпуса, а также разнообразного нелингвистического и негуманитарного знания, по предоставляемым пользователю познавательным возможностям, многократно превосходит докорпусные методы и техники. Потенциал корпусных исследований только начинает раскрываться, однако и на начальном этапе истории корпусной лингвистики прослеживаются некоторые ее характерные черты и тенденции развития.

В лингвистических корпусных проектах исследуются, как правило, большие массивы языковых фактов – или синхронных, или диахронических, – поэтому обычно это коллективные разработки. Наиболее крупные и общественно значимые результаты корпусных исследований принадлежат истории языков (симптоматично в этой связи обращение к истории английского, испанского, португальского языков М. Дэвиса, выдающегося корпусного лингвиста, создавшего более 10 знаменитых корпусов разных языков, включая Corpus of Historical American English (COHA, 2008–2010). COHA содержит более 475 млн с/у за 200 лет (1810–2010 гг.), что в 50–100 раз больше объема других исторических корпусов английского языка. М. Дэвис руководил также созданием исторических корпусов испанского (2001–2002) и португальского (2009–2011) языков. По всей вероятности, в близком будущем корпусные возможности приведут к крупным достижениям в диахронической лингвистике. В синхроническом изучении современных языков приоритетной областью корпусной лингвистики становятся статистические исследования языкового варьирования.

В качестве примера выдающегося новаторства корпусной лексикографии могут быть названы два достижения: 1) синтез толкового и частотного словаря в словаре Macmillan 2007, а затем и в ряде других знаменитых английских словарей, таких как Collins, Longman; 2) компонентный семантический анализ лексем 100-тысячного словника Macmillan 2007, проведенный в форме первых (в словарной статье) дефиниций, лексика которых не выходит за пределы первых по частоте 2.500 слов (которые, таким образом, выступают как семантические компоненты, используемые для описания семантики 100 тыс. лексем словаря).

В области фразеологии своды коллокаций и конкордансов (в частности, в Национальном корпусе русского языка) укрепляют концепцию фразеологии, разработанную И. А. Мельчуком (1960; 1995; 2007 и др.). По И. А. Мельчуку, признаки устойчивости и идиоматичности сочетаний лексем взаимно независимы и градуальны (ср. характерные термины *полуфразама*, *полуидиома*, *полная идиома* и т. п.). Коллокации, представленные в корпусе в десятках и сотнях повторений, делают признак «воспроизводимости» (который до недавнего времени рассматривался как единственно обязательный признак фраземы) ненадежным и побуждают видеть во фразеологическом фонде континуум сочетаний лексем с разной степенью устойчивости и идиоматичности, а во фразеологии – науку не о идиомах и клише, но о лексической сочетаемости. Ее предмет составляет необозримое множество воспроизводимых и по-разному несвободных сочетаний лексем. Их число в речи огромно. «Люди говорят не словами, а фразами <...>. Количественно фраземы превосходят слова в словарях примерно в соотношении 10 к 1» [1, с. 218].

Поистине бурное развитие в русистике авторской корпусной лексикографии привело к созданию ряда частотных словарей языка русских классиков; ожидаются исследования, призванные на основе BigData открывать новые горизонты в лингвистической поэтике и литературоведении. Пока это в основном дело будущего, однако есть и вполне конкретные впечатляющие результаты. Таковы, в частности, исследования А. Я. Шайкевича: сопоставление лексики корпусов поэзии А. С. Пушкина и А. Б. Мицкевича; его же анализ статистических данных по корпусу текстов Достоевского, напечатанный во Введении к изданию: Шайкевич А. Я., Андрищенко В. М., Ребецкая Н. А. «Статистический словарь языка Достоевского» (М., 2003).

Для корпусной индустрии на всем протяжении ее истории характерен стремительный рост объемов корпусов. Один из старейших корпусов – Deutsches Referenzkorpus (Мангеймский корпус, открыт в 1964 г.), в настоящее время его объем превышает 57,6 млрд с/у. Новейшие миллиардные корпусы по объемам обрабатываемой и предоставляемой информации конкурируют с поисковиками ранга Google, AOL, Yahoo!, Яндекс. Два примера: 1) мониторинг мегакорпус «мирового английского» “News on the Web” (акроним NOW), разработан в 2010 г. М. Дэвисом, объем 19,6 млрд с/у, аккумулирует тексты сетевых газет и журналов на английском в 21 стране; 2) корпус англоязычных сегментов всемирной паутины (iWeb), разработанный в 2013 г. М. Дэвисом, объем 14 млрд с/у.

Однако с распространением корпусов в разных странах появилась необходимость в адаптации корпусных данных к запросам и возможностям рядовых гуманитариев – преподавателей языков, лингвистов-исследователей, лексикографов, издателей, журналистов. Одна из задач адаптации состояла в сокращении объемов языкового материала, включаемого в адаптированные версии корпусных разработок. Необходимость сокращать корпусные BigData ставит вопрос о назначении миллиардных корпусов. Фактически корпусы текстов предстали как продукция тройного назначения.

Корпусы первого поколения (объемом 100 млн с/у), называемые или осознаваемые как «национальные» или «государственные», сохраняют относительную сбалансированность подкорпусов и широкую социально-гуманитарную адресацию. Более поздние корпуса специализируются по двум направлениям: 1) мониторные мегакорпусы газетно-журнальных текстов (включая сетевые) адресованы социологам, политикам, этнологам, демографам, журналистам – всем, кто в силу профессии и социальной ответственности стремится держать руку на пульсе сегодняшней жизни, чтобы понимать или предвидеть, что будет со страной и человечеством завтра; 2) сематически безграничные (неизбирательные) корпуса, аккумулирующие наличные оцифрованные тексты, используемые в информатике как сырье для «обработки языка» (language processing): машинного предобучения нейронных сетей и создания статистических алгоритмов самосвязываемости слов в адекватные словесные реакции искусственного интеллекта.

ЛИТЕРАТУРА

1. Иорданская Л. Н., Мельчук И. А. Смысл и сочетаемость в словаре. М. : ЯСК, 2007. 666 с.