УДК 81-139

О.А.Мартынюк

КОРПУСНАЯ ЛИНГВИСТИКА И НОВЫЕ ВОЗМОЖНОСТИ ЛИНГВИСТИЧЕСКОГО ИССЛЕДОВАНИЯ

Аннотация. Современные информационные технологии и технические средства открывают новые возможности для лингвистического исследования на базе языковых корпусов. В статье представлено описание крупнейших корпусов, история их создания, вскрыта их роль и целесообразность использования в лингвистическом исследовании.

Ключевые слова: лингвистическое исследование, корпус, корпусная лингвистика

Несмотря на то, что корпусной лингвистикой к настоящему времени накоплен серьезный опыт разработки корпусов для различных языков, широкое применение последних в лингвистических исследованиях до сих пор отсутствует. В связи с этим весьма актуальным является обращение к вопросу о языковых корпусах и возможностях их использования профессиональными лингвистами.

Получивший широкое распространение в последнее время термин «корпусная лингвистика» связан с разделом языкознания, занимающимся разработкой, созданием и использованием текстовых корпусов. В Википедии лингвистический корпус определяется как «совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой». В определении корпусной лингвистики Г.Р.Беннет акцент сделан на еще одной характеристике корпуса – содержащиеся в корпусах тексты представляют собой живую речь: “A corpus is a large, principled collection of naturally occurring examples of language stored electronically” [2].

Само по себе обращение исследователей к анализу реальных языковых фактов на основе текстовой выборки не является новым. Реальные примеры использования отдельных слов в литературных источниках для иллюстрации их значения приводил еще Самуэль Джонсон в своем словаре “A Dictionary of the English Language”, изданном в 1755 году. Помимо того, что такого рода иллюстрация значения слова была отличительной инновационной чертой этого словаря, любопытно отметить, что общее количество цитат (114,000) , используемых для иллюстрации значений слов более чем в два с половиной раза превышало количество включенных в словарь слов (42,773), а в качестве

источников использовались художественные произведения Шекспира, Свифта, Мильтона, Драйдена и др.

Отобранные вручную выборки из печатных литературных источников широко использовалось в исследовательских целях как зарубежными, так и советскими лингвистами еще с середины XX века. Использование же образцов устной речи для анализа языка в целом для того периода – явление достаточно редкое. В качестве яркого примера можно привести Чарльза К.Фриза, который был убежден в том, что лингвистическое исследование должно основываться на образцах живой речи, реально используемой носителями языка в речевой коммуникации. Его однофамилец, Питер Х. Фриз, в своей работе, посвященной Чарльзу Фризу, отмечает, что во всех своих исследованиях Фриз "made every attempt to discover the characteristics of the language spoken in the communities while the speakers were focused on what they were communicating, not how they were communicating it" [3]. Вполне закономерно, что при написании своего известного труда "The Structure of English" Фриз использовал корпус текстов объемом 250 000 слов, составленный на основе записи около 50 часов телефонных разговоров, участники которых не знали, что их разговоры записывают.

Стремительное развитие компьютерных технологий во второй половине века дало толчок для создания в 60-х годах двадцатого столетия первого компьютерного корпуса, получившего название Брауновского корпуса и послужившего в качестве образца для последующих создателей корпусов текстов на различных языках. Этот корпус был создан в частном американском университете Брауна (штат Род-Айленд) и включал 500 фрагментов по 20 тысяч слов каждый. В качестве источников авторы корпуса У.Френсис и Г. Кучера использовали случайно отобранные на основе особой вероятностной процедуры прозаические тексты американского варианта английского языка, принадлежащие пятнадцати наиболее массовым жанрам печатной прозы США.

Бурное развитие компьютерных технологий позволило многократно увеличить объем текста, а появившееся новое лингвистическое направление корпусной лингвистики внесло свой вклад в развитие качественных характеристик корпусов. В.П. Захаров [1] выделяет три основные предпосылки целесообразности создания и использования корпусов:

1) достаточно большой (репрезентативный) объем корпуса гарантирует типичность данных и обеспечивает полноту представления всего спектра языковых явлений;

2) данные разного типа находятся в корпусе в своей естественной контекстной форме, что создает возможность их всестороннего и объективного изучения;

3) однажды созданный и подготовленный массив данных может использоваться многократно, многими исследователями и в различных целях.

Корпуса могут иметь разную степень организации и структурированности. Наиболее представительными и полными по уровню научной обработки текстов являются национальные корпуса языков. Общепризнанным образцом национального корпуса является, Британский национальный корпус (BNC), на который ориентированы многие другие современные корпуса.

Британский национальный корпус включает 100 миллионов словоупотреблений, входящих в тексты, отобранные на основании следующих трех основных критериев: (1) область/сфера; (2) время; (3) источник, к которым также применялись определенные требования.

Тексты из литературных источников представлены в корпусе лишь в 25%, в то время как остальные 75% корпуса составляют письменные тексты из информативных источников в приблизительно равных долях представленности таких областей, как прикладные и естественные науки, искусство, финансы и торговля, досуг, политика и т.д. По принадлежности к источнику текста, все тексты распределились следующим образом: 60% составили тексты, отобранные из книг; 25% – из периодических изданий, остальные тексты – из других видов печатной продукции (брошюры, реклама и т.д.) (от 5 до 10%), а также неопубликованные материалы (письма, дневники, эссе, меморандумы (от 5 до 10%). Источниками менее 5% текстов послужили материалы, предусматривающие озвучивание (речи политиков, пьесы, радиоматериалы и т.д.)

Временной критерий в Британском Национальном корпусе ориентирован на современность, поэтому в корпус вошли тексты, опубликованные после 1975 года, за исключением некоторых художественных произведений, которые были напечатаны после 1964 года, но продолжают оставаться популярными, оказывая определенное влияние на развитие языка.

Британский национальный корпус включает также аудио корпус, который по объему составляет около 10% всего корпуса. Аудио корпус BNC состоит из двух частей – демографической и контекстно-обусловленной. Демографическая часть корпуса основана на текстах, записанных 124 волонтерами – мужчинами и женщинами, проживающими в разных уголках Великобритании и представляющими различные социальные группы населения. Волонтеры записывали на магнитофон все свои разговоры в течение 2-3 дней и письменно фиксировали информацию о каждом разговоре в отдельную тетрадь. Также, по возможности, записывалась информация об участниках разговора (пол, возраст, род занятий и т.д.). Во вторую, контекстно-

обусловленную часть вошли транскрипции записей, сделанных во время различных публичных мероприятий, таких как лекции, новостные передачи, собрания, консультации, интервью, проповеди, выступления политиков, членов парламента и т.д.

С момента выпуска третьего издания BNC в 2007, к основному корпусу были добавлены два подкорпуса – BNC Sampler, в котором представлен один миллион наиболее частотных слов письменной речи и один миллион наиболее частотных слов устной речи и BNC Baby, содержащий репрезентативные образцы четырех жанров – художественного, газетного, академического и разговорного – по одному миллиону словоупотреблений каждый.

Созданный после BNC, Национальный корпус русского языка превышает BNC по объему, представляя более развитую структуру подкорпусов, что открывает новые возможности для анализа. Перечень подкорпусов Национального корпуса русского языка представлен в следующей таблице с указанием их объемов:

Подкорпус	**Число текстов**	**Число предложений**	**Число словоупотреблений**	**% словоупотреблений**
Основной корпус	76 882	17 574 752	209 198 275	57.3%
- в том числе со снятой омонимией	2 147	516 852	5 944 188	1.6%
Газетный корпус	181 175	8 553 495	113 292 003	31.0%
Диалектный корпус	197	20 273	194 283	0.1%
Обучающий корпус	229	65 666	664 751	0.2%
Параллельный корпус	370	1 609 609	24 022 437	6.6%
Поэтический корпус	41 448	638 861	6 738 474	1.8%
Устный корпус	3 034	1 604 626	10 122 579	2.8%
Мультимедийный корпус	31 741	148 619	648 576	0.2%
Всего:	**335 076**	**30 215 901**	**364 881 378**	**100%**

В основной корпус входят прозаические письменные тексты XVIII — начала XXI века. Часть его представляет собой глубоко аннотированный корпус, в котором для каждого предложения построена полная морфологическая и синтаксическая структура или дерево зависимостей. В

газетном корпусе представлены статьи из средств массовой информации 1990-2000-х годов. Параллельные корпуса дают возможность найти все переводы для определенного слова или словосочетания на русский язык или с русского языка. Параллельные корпуса имеются для таких языков, как английский, немецкий, французский, испанский, итальянский, украинский, белорусский. Корпус диалектных текстов включает запись диалектной речи различных регионов России с сохранением их грамматической специфики и предусматривает специальный поиск с учётом диалектной морфологии. Поиск в поэтическом корпусе возможен не только на основе лексики или грамматики, но и по специфическим для стиха признакам. Обучающий корпус включает исключительно тексты со снятой грамматической омонимией, и разметка его ориентирована на школьную программу. В корпус устной речи входят расшифровки магнитофонных записей публичной и частной устной речи, а также транскрипты кинофильмов. Мультимедийный корпус представлен фрагментами кинофильмов 1930—2000-х годов, снабженными видео- и аудиорядом. Возможен поиск не только по произносимому тексту, но и по жестам (кивание головой, похлопывание по плечу и т. п.) и типу речевого действия (согласие, ирония и т. п.)

Еще большим объемом характеризуется Корпус современного американского английского (Corpus of Contemporary American English, сокращенно COCA), включающий 450 млн. словоупотреблений. С ним сопряжены Исторический корпус американского английского (Corpus of Historical American English, COHA) объемом 400 млн словоупотреблений и **Glo**bal **W**eb-**B**ased **E**nglish (GloWbE), основанный почти на 1,9 миллиарда слов, представленных на 340,000 веб-сайтах в двадцати англо-говорящих странах. Совместное использование этих корпусов предоставляет уникальную возможность исследования вариативности английского языка в области диалекта и жанра на разных этапах их исторического развития. Возможности извлечения информации на основе корпусов практически не ограничены и зависят от цели исследования. Это обусловлено в первую очередь тем, что как сами корпуса, так и инструменты работы с ними постоянно развиваются, а вместе с ними развивается и корпусная лингвистика. Основными сферами практического применения корпусной лингвистики являются такие области, как лексикография, грамматика, социолингвистика, перевод, изучение языка, обучение языку, стилистика, диалектология, историческая лингвистика.

О важной роли корпусной лингвистики для исследований на материале корпусов свидетельствует образование в Великобритании Центра по исследованию корпусов, под эгидой которого раз в два года, начиная с 2001 года проводятся конференции на базе трех университетов – Бирмингемского,

Ланкастерского и Ливерпульского. В США создана Американская Ассоциация Корпусной Лингвистики (AACL), которая проводит ежегодные конференции с 1998 года. Проводятся конференции и в других странах, где также имеются ассоциации корпусной лингвистики. В 2014 году международные конференции, посвященные корпусной лингвистике, состоятся в Великобритании, США, Испании, Гонконге. Имеется у этого нового направления и свой международный журнал – Corpus Linguistics and Linguistic Theory, издаваемый при Университете Калифорнии.

Роль языкового корпуса для лингвистического анализа трудно переоценить. Использование корпуса позволяет не только многократно ускорить исследование, но и существенно повысить его эффективность и достоверность за счет возросшего охвата исследуемого материала. Снижение трудоемкости исследования на основании статистических методов привело к расширению возможностей корпусного анализа по сравнению с докорпусным периодом. Создание диахронических корпусов, в которых собраны тексты, относящиеся к временному промежутку в несколько столетий, сделало возможным выявить изменения и механизмы развития языка. Национальные корпуса языков представляют практическую ценность для исследований в области языка не только на всех его уровнях, но и в аспекте изучения различных жанров. Наличие во многих национальных корпусах подкорпусов параллельных текстов расширяет возможности проведения сопоставительных исследований языков и предоставляет ценнейший материал для исследований в области перевода.

В целом можно говорить об образовании новой дисциплины – корпусной лингвистики, имеющей свой особым образом отобранный и размеченный материал и соответствующий специфический инструментарий, который используется для его анализа.

Ссылки:

(1) Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие. – СПб., 2005 – 48 с.

(2) Bennet, Gena R. Using Corpora in the Language Learning Classroom. – http://press.umich.edu/pdf/978472033850-part1.pdf

(3) Fries, Peter H. Charles C.Fries. Linguistics and Corpus Linguistics", p.99 - http://icame.uib.no/ij34/Fries.pdf

Анотація. Мартинюк О.А. Корпусна лінгвістика і нові можливості лінгвістичного дослідження. – Стаття.

Сучасні інформаційні технології і технічні засоби відкривають нові можливості для лінгвістичного дослідження на базі мовних корпусів. У статті

представлений опис найбільших корпусів, історія їх створення, розкрита їх роль і доцільність використання в лінгвістичному дослідженні.
Ключові слова: лінгвістичне дослідження, корпус, корпусна лінгвістика

Summary. Martynyuk O. Corpus Linguistics and New Possibilities for Linguistic Research. – Article.
Modern information and computer technologies open new possibilities for linguistic research on the basis of language corpora. Presented in the article is a description of the largest corpora and their development, as well as their role and potential for linguistic research.
Key words: linguistic research, corpora, corpus linguistics.

УДК 811.111’06’367.63

Мойсеєнко Н. Г.

(Одеса)

ГРАМАТИЧНЕ ЗНАЧЕННЯ ЗАЙМЕННИКА У СУЧАСНІЙ АНГЛІЙСЬКІЙ МОВІ У СВІТЛІ ТЕОРІЇ ПРОТОТИПІВ

Анотація. Стаття показує наявність розбіжностей у поглядах лінгвістів на визначення категоріальної семантики займенників та обґрунтовує необхідність розробки концептуальних категорій займенникової співвіднесеності, що дасть змогу встановити частиномовне значення мовних одиниць, що традиційно об’єднуються у групу займенників.
Ключові слова: частина мови, концепт, прототип, займенник, категорія.

У сучасному мовознавстві взагалі і, в англістиці, зокрема, не існує єдиної точки зору щодо частиномовної семантики займенника. Категоріальний статус одиниць, що традиційно відносяться до займенника, по-різному визначається дослідниками англійської мови [1-14; 16-40]. Це обумовлює актуальність нашого дослідження.

Мета даної статті зводиться до аналізу найбільш поширених точок зору щодо категоріальної семантики англійського займенника, виділенню проблем, пов’язаних із її встановленням, а також розробці шляхів визначення концептуальних категорій займенникової співвіднесеності, під якими ми розуміємо ієрархічно структуровані об’єднання концептів із концептом -