ОСНОВНЫЕ ЭТАПЫ СТАНОВЛЕНИЯ КОРПУСНОЙ ЛИНГВИСТИКИ

Гатиятуллина Г.М., Березников Д.

В данной работе проведен обзор наиболее значимых этапов развития корпусной лингвистики как отдельного направления.

Корпусная лингвистика – бурно развивающееся направление современного языкознания. Сочетая в себе современные методы статистических расчетов и обработки данных с одной стороны, и большого объема эмпирических данных с другой, корпусная лингвистика представляет большой интерес для лингвистов и специалистов в области компьютерных технологий. Цель данной работы провести обзор наиболее значимых этапов развития корпусной лингвистики как отдельного направления.

Начало создания корпуса текстов относят к XIII в. В целях упрощения поиска библейских цитат для будущих богословов монахи вручную индексировали слова в каждой строчке, изучая Библию страницу за страницей. Изучение корпуса вышло на совершенно иной уровень с началом применения в лингвистике новых технологий. Так, с 50-х по 70-е гг. XX в монах иезуитского ордена отец Роберто Буса впервые создал электронный корпус тестов полного собрания сочинений Фомы Аквинского Index Tomisticus. Данный подход на основе карточек широко использовался также при создании словарей с целью изучения контекстного употребления того или иного слова, а также выявления устойчивых словосочетаний.

Впервые электронный корпус письменного английского языка (Brown corpus) был создан в 60-е годы XX века Нельсоном Френсисом и Генри Кучерой [1,2]. Корпус содержит более миллиона словоупотреблений американского английского из документов, опубликованных в 1961г. Данный корпус послужил основой для создания целого ряда корпусов, известных под названием «Семейство корпусов Браун». Далее подобные корпусы начали создаваться и в других языках.

Следующим поворотным этапом развития корпусной лингвистики стала совместная работа шведских и английских ученых, которые корпус London-Lund Corpus of Spoken English (100 устных текстов объемом 5 000 слов каждый) с подробно фонологически затранскрибированным и аннотированным материалом [3,4]. Корпус создавался с 1959г по 1990гг и стал первым корпусом текстов устной речи. У истоков корпуса лежали два совместных проекта по изучению употребления языка в устной речи: the Survey of English Usage (SEU), начатом в 1959г. Рэндольфом Кирком (Randolph Quirk) в университете

Лондона (большое влияние на создание корпуса оказал также Дэвид Кристалл (David Crystall), и the Survey of Spoken English (SSE), начатом Яном Свартвиком (Jan Svartvik) в 1975г. в Лундском университете в Швеции. Тридцать четыре устных текста позже были опубликованы в виде отдельной книги Svartvik and Quirk (1980),также корпус послужил основой для книги Comprehensive Grammar (Quirk et al. 1985) Данный корпус доступен в базе данных корпусов текстов ICAME и OTA [3].

Три этапа развития корпуса SEC (Spoken English Corpus) сыграли важную роль в разработке методологии корпусной лингвистики, а также ее эмпирической базы. Корпус SEC (Spoken English Corpus) устного литературного британского английского языка создавался группой лингвистов под началом Taylor и Knowles в 1988г. с целью изучения организации речи в зависимости как от ее стилевой или жанровой принадлежности, так и гендерной. Так, узкожанровые тексты как поэтический, религиозный, пропагандистский, новостной дискурсы сравнивали с повседневной разговорной речью. Кроме того, все тексты сравнивались по гендерной принадлежности. Позже корпус был доработан, фонологическая метка была доработана паузами, была размечена длина слова во временном отрезке, были оцифрованы звуковое содержание, а также тоновое ударение, в результате чего в корпусе появилась опция машинного чтения с тем, чтобы услышать литературное произношение того или иного слова. Данный проект получил название MARSEC (machine readable SEC). Далее, следующей ступенью стала разработка проекта Aix-MARSEC, в которой помимо всего уже наработанного появились разметки на 9 уровнях (на фонемном, слоговом, лексическом, грамматическом, синтаксическом, стопы, ритмическом уровнях, а также уровне малого и большого коммуникативного шага). Одним из результатов создания этого корпуса стала возможность автоматического прогнозирования окончания коммуникативного шага и границ фразы с учетом просодических и синтаксических законов и правил (Claire Brierley and Eric Atwell) [2, 3, 4]. Подобная разметка позволяет производить автоматические подсчеты частоты употребления того или иного слова, а также поиск контекстного употребления или конкордансов того или иного слова.

С увеличением выбора инструментария по обработке текстов, расширились и цели изучения языка в действии. Помимо модуса – письменного и устного – корпусы также стали делить на сбалансированный и открытый. Сбалансированный или репрезентативный корпус представляет собой фиксированный объем репрезентативных текстов, на примере которых можно изучить функционирование языка в той или иной области, а также в отдельных жанрах и стилях речи. Открытые же корпусы постоянно

пополняются все новыми и новыми текстами, на основе которых можно более подробно изучить количественные и статистические данные, а также случаи исключения, редкого или ошибочного употребления того или иного феномена. Внутри устных корпусов также тексты делятся на так называемые демографические (для изучения диалектов, гендерного, возрастного различия в функционировании языка) и контекстуальные (для изучения текстов в контекстом, ситуативном аспекте) [1]. Так, в зависимости от поставленной цели стали создаваться корпусы для изучения диалектов (The Spoken Corpus of the Survey of English Dialects, The Intonational Variation in English Corpus), разговорной речи различных возрастных групп (CHILDES, The Bergen Corpus of London Teenage Language (COLT), BNC Baby), а также появился большой интерес к созданию корпусов отдельных стилей и жанров речи [2].

Таким образом, создание корпусов текстов явилось той самой эмпирической базой для изучения живого языка, а разработка методов и программ компьютерной обработки текстов с одной стороны облегчило трудоемкую работу классификации и сортирования карточек с каждым словом в тексте, дало возможность обрабатывать в более короткие сроки и более объемные корпусы текстов, но самое главное, снизила погрешность в расчетах. Более того, компьютерные программы обработки текстов позволили изучить живую речь в более широком спектре и на совершенно новом уровне.

Список литературы:

1. Meyer, Charles F. English corpus linguistics. An introduction.– Cambridge University Press. – 2002. – 186p.
1. http://www.lancaster.ac.uk/~xiaoz/papers/corpus%20survey.htm#_Toc92298863
2. Svartvik, Jan and Quirk, Randolph (1980) (eds.). *A Corpus of English Conversation* Lund: CWK Gleerup.
3. Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey and Svartvik, Jan (1985). *A Comprehensive Grammar of the English Language* London: Longman.
4. http://ota.ox.ac.uk/desc/0168
5. http://clu.uni.no/icame/
6. http://www.helsinki.fi/varieng/CoRD/corpora/index.html
7. http://www.reading.ac.uk/AcaDepts/ll/speechlab/marsec/

8. Taylor, L.J. & Knowles, G. 1988. Manual of information to accompany the SEC Corpus. Mimeo, UCREL, Lancaster University.

9. Brierley Claire, Atwell Eric Prosodic Phrase Break Prediction: Problems in the Evaluation of Models against a Gold Standard // TAL. Vol.48(1). – 2007- P. 187-206

10. http://www.reading.ac.uk/AcaDepts/ll/app_ling/internal/corpus.html