Цай Ванифань,
аспирантка Высшей школы перевода (факультета)
МГУ имени М.В. Ломоносова; e-mail: lovedirk41@hotmail.com

РАЗВИТИЕ ИССЛЕДОВАНИЯ КОРПУСА В КИТАЕ — ОТ ЧАСТОТНОГО СЛОВАРЯ К ЛИНГВИСТИЧЕСКОМУ КОРПУСУ

В данной статье рассматривает развитие исследования корпуса в Китае — от частотного словаря к лингвистическому корпусу. Частотный словарь появляется в Китае для популяризации «гражданского образования». В конце XX века компьютеры значительно сократили время необходимое для сбора и организации языковых материалов, и рождается машиночитаемый корпус китайского языка. Сейчас в Китае три крупных корпуса, каждый содержит более 100 миллионов реальных материалов. Умение использовать корпус считается одним из важных навыков для исследования языка.

***Ключевые слова**:* частотный словарь, корпус, статистическая лингвистика.

Когда люди совершают акты коммуникативной деятельности через язык, его компоненты, используемые в речи, являются неопределёнными. Некоторые учёные предлагали квантитативный подход для изучения лингвистических объектов. Б.Н. Головин подчеркнул в своей работе: «Одним из реальных оснований для применения квантитативных методов в изучении языка и речи нужно признать объективную присущность языку количественных признаков и характеристики» [Головин, 1971: 11]. Аналогичное мнение предлагал В. Ярцевой: «Учёт частотности любого языкового явления — полезный приём при анализе» [Ярцева, 1970].

Статистическая лингвистика (统计语言学), также известная как квантитативная лингвистика (计量语言学), является разделом математической лингвистики. Цель данного предмета — изучить вероятность и частоту встречаемости языковых компонентов с помощью математических способов, таких как теория вероятностей и статистика, и попытаться объяснить правила языка. Результаты применения данного метода можно представить в виде частотных словарей [Фэн Чживэй, 2002]. В «Толковом переводческом словаре» указывается, что квантитативная лингвистика может в целом

рассматривается как техника лингвистического наблюдения и описания, обработки данных наблюдения; метод исследования языка и речи, не обязательно противоположный сопоставленному, сравнительно-историческому и другими методам языкознания и концепция как система количественных идей и представлений об объекте лингвистической науки [Нелюбин, 2003: 76].

Первые идеи о составлении частотного словаря в Китае появились во время революции 1911 года. В то время передовые учёные начали задумываться о важности популяризации «гражданского образования», понимая, что большинство китайцев не могли участвовать во внутренних делах страны из-за того, что они не получили образования и были неграмотны. Чтобы сделать движение за грамотность более эффективным, Чэнь Хэцинь, преподаватель Нанкинского педагогического университета, поднимал следующие вопросы: каков общий словарный запас китайского языка? Сколько иероглифов должны выучить ученики начальной школы? Сколько иероглифов должны знать взрослые? Для решения этих проблем Чэнь Хэцинь с 1920 года начал составлять "语体文应用字汇" (сборник прикладного языка).

В этом сборнике Чэнь Хэцинь и его помощники выделяли шесть категорий языковых материалов, в частности: детская литература, газеты, женские журналы, внеклассные работы школьников, древние и современные романы и другие, в общей сложности в словаре было 554 478 иероглифов. В конце словаря автор предлагал список иероглифов, который был составлен в соответствии с частотой их употребления. До публикации лингвистические достижения и находки сборника уже использовались в качестве основы для «平民千字课 (урок о тысячах иероглифов для народов)» и учебников начальной школы. «语体文应用字汇 (сборник прикладного языка)» Чэнь Хециня — это первый в своём роде частотный словарь китайского языка, а также прототип лингвистического корпуса в Китае.

В 1979 году Пекинский институт языка и культуры решил воспользоваться компьютерами для вычисления частоты употребления среди больших объёмов китайских слов. Общее количество подсчитанных слов составило 1 135 752 единицы, с 31 159 разными словами. Выбранные материалы можно разделить на следующие категории:

1. Газета и политическая статья: 440 000 слов, составляет 24.4% всего корпуса;

2. Технологии и научно-популярные статьи: 290 000 слов, составляет 11,8% всего корпуса;

3. Разговорный материал: 200 000 слов, составляет 11,1% всего корпуса;

4. Литература; 890 000 слов, составляет 48,7% всего корпуса.

Результаты статистики были собраны в «现代汉语频率词典 (частотный словарь китайского языка)». И хотя словарь меньше других словарей (стоит отметить, что и выбор материала не являлся равномерным), он всё же достиг больших успехов и предложил четыре списка составленные по частоте употребления: список слов в алфавитном порядке; список слов в порядке убывания частоты; список слов в порядке убывания использования и высокочастотный список слов различных материалов.

В конце XX века компьютеры значительно сократили время, необходимое для сбора и организации языковых материалов. Это привело к возникновению совершенного нового ответвления языкознания — корпусной лингвистике. В отличие от квантитативной лингвистики, корпусная лингвистика в основном изучает машиночитаемый сбор текста на естественном языке, его хранение, поиск, грамматические аннотации, синтаксический и семантический анализ, пытаясь устроить корпус с этими функциями для использования в областях количественного анализа языка, лексикографии, анализа стиля, понимания естественного языка и машинного перевода.

Изначально машиночитаемый корпус китайского языка включал в себя следующие: корпус китайских современных литературных произведений, созданный Уханьским университетом (около 5 270 000 иероглифов), корпус китайских учебников средней школы, созданный Пекинским педагогическим университетом (около 1 068 000 иероглифов), современный китайский корпус, созданный Пекинским университетом аэронавтики и космонавтики (около 20 000 000 иероглифов) и т.д.

Ранние корпусы обнажали две проблемы:

1. Несмотря на то, что компьютеры используются, в большинство из них информацию надо было вводить вручную. Корпус имеет небольшие размеры и слабый норматив.

2. В области автоматической сегментации слов, из-за отсутствия единообразных норм, разные корпусы используют разные методы, что приводит к разным результатам.

Чтобы решить эти проблемы, в октябре 1990 года Китай разработал национальный стандарт GB-13715 «Современная сегментация китайского слова для обработки информации» и предложил принцип определения сегментации китайского слова, который является основой для автоматической сегментации китайских слов.

В 1991 году Национальный комитет языка начал составлять масштабный сбалансированный корпус китайского языка —

语料库在线 (корпус онлайн / cncorpus.com). В основе данного корпуса содержатся языковые материалы с 1919 года по настоящее время. Целью создания корпуса является содействие изучению лексического, синтаксического, семантического и прагматического китайского языка, а также предоставление языковых ресурсов для обработки информации на китайском языке. Составление корпуса был закончено в 1998 году, и Министерство образования создало исследовательскую группу под названием «Глубокая переработка китайского корпуса», чтобы обработать материалы для корпуса.

На сегодняшний день в корпусе содержится около 100 миллионов знаков. Среди них около 70 миллионов символов-материалов, написанных до 1997 года. Все они были опубликованы в виде бумажных книг и введены в корпус вручную. Материал после 1997 года составляет около 30 миллионов знаков, половина вводится вручную, а половина берётся из электронных версий. Собранные материалы можно разделить на следующие группы:

1. Гуманитарные и социальные науки (составляет 50% всего материала): политические, исторические, экономические и литературные и художественные материалы;

2. Естественные науки (составляет 30% всего корпуса): сельское хозяйство, промышленность, медицина и технология;

3. Комплексная категория (составляет 20% всего корпуса): деловая речь и материалы, которые трудно завершить.

Кроме того, корпус онлайн разработал два подраздела — аннотированный корпус и корпус древнего китайского языка. Аннотированный корпус включает в себя приблизительно 50 миллионов знаков. Разметка относится к сегментации слова и характеру слова, она была вычитана вручную 3 раза, а её точность превышает 98%. Корпус древнего китайского языка содержит текстовые материалы в период с династии Чжоу (770 до н.э.) по династию Цин (1912 г.), в общей сложности — около 70 миллионов знаков.

Примерами других крупных корпусов в Китае являются корпус BBC, созданный Пекинским университетом языка и культуры, и корпус CCL, созданный Пекинским университетом.

Корпус Пекинского университета (BLCU Corpus Center / BCC корпус) — это корпус китайского языка, в который также включены три подкорпуса: английский корпус, французский корпус и китайско-английский двуязычный корпус. Общее количество материала корпусов BCC составляет около 15 миллиардов слов, включая газеты (2 миллиарда), литературу (3 миллиарда), Weibo (3 миллиарда), науку и технику (3 миллиарда), комплексный (1 миллиард) и древний

китайский (2 миллиарда). Это масштабный корпус, который может в полной мере отразить языковую жизнь современного общества в Китае. Корпус представляет собой обрабатывающий корпус, что означает, что языковые материалы на современном китайском прошли процесс сегментации и разметки по характеру и синтаксису. Корпус поддерживает поиск по словам по характеру.

Корпус CCL был разработан Исследовательским центром китайского языка при Пекинском университете при поддержке Института компьютерных языков Пекинского университета и Института компьютерных технологий Академии наук Китая. Общее количество символов в корпусе составляет 783 463 175, из которых общее количество символов в современном китайском корпусе составляет 581 794 456, а в древнем китайском корпусе — 201 668 719 слов. Он содержит тексты с XI века до нашей эры по настоящее время. Корпус CCL включает в себя: запись устной речи (Пекинский диалект), материалы фильмов и телевизионных программ, материалы из интернета и письменные источники (большинство из них из газеты).

Корпус CCL поддерживает сложные выражения поиска (такие как несмежные запросы ключевых слов, указание удалённых запросов и т.д.), поддерживает запросы на пунктуацию (например, запрос «?» для извлечения всех вопросительных знаков в корпусе); Пользователь может загрузить результат запроса (текстовый файл) с веб-страницы; CCL предоставляет богатую, ориентированную на строки функцию поиска. Хотя материалы в корпусе не обрабатываются, они могут удовлетворить различные потребности в исследованиях.

Список литературы

Головин Б.Н. Язык и статистика [Текст] / Б.Н. Головин. М.: Просвещение, 1971. 189 с.

Нелюбин Л.Л. Толковый переводческий словарь. 3-е изд., перераб. М.: Флинта: Наука, 2003. 320 с.

Корпус BCC. URL: http://bcc.blcu.edu.cn/

Корпус CCL. URL: http://ccl.pku.edu.cn:8080/ccl_corpus/

Корпус онлайн. URL: http://www.cncorpus.org/

冯志伟，胡凤国 《数理语言学》，商务印书馆，2012, 491 页

冯志伟 《中国语料库研究与现状》，语言文字应用，2002, 43–62 页.

Cai Wangyifan,
Postgraduate Student at the Higer School of Translation and Interpretation, Lomonosov Moscow State University, Russia;
e-mail: lovedirk41@hotmail.com

DEVELOPING CORPUS RESEARCH IN CHINA: FROM A FREQUENCY DICTIONARY TO A LINGUISTIC CORPUS

Frequency dictionaries are published in China to popularize "civic education." At the end of the 20th century, computers reduced the time needed to collect and organize various types of language material. At the same time, a machine-readable corpus of Chinese came into being. Today, there are three large corpora in China, each containing more than 100 million real materials. The ability to use the corpus is considered one of the most important skills to research the language.

Key words: corpus, frenquecy dictionary, computatuional linguistics.

Reference

Golovin B.N. Yazyk i statistika [Language and statistics]. Moscow: Prosveshchenie, 1971. 189 p. (In Russian).

Nelyubin L.L. Tolkovyj perevodcheskij slovar' [Explanatory dictionary of translation]. 3-e izd., pererab. Moscow: Flinta: Nauka, 2003. 320 p. (In Russian).

Case BCC. URL: mode of access: http://bcc.blcu.edu.cn/ (In Russian).

Case CCL. URL: mode of access: http://ccl.pku.edu.cn:8080/ccl_corpus/ (In Russian).

Case online. URL: mode of access: http://www.cncorpus.org/ (In Russian).

冯志伟，胡凤国 《数理语言学》，商务印书馆，2012, 491 页.

冯志伟 《中国语料库研究与现状》，语言文字应用，2002, 43–62 页