***Корпусная лингвистика*** – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий. Под названием ***лингвистический, или языковой, корпус текстов*** понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. В понятие «корпус текстов» входит также система управления текстовыми и лингвистическими данными, которую в последнее время чаще всего называют *корпусным менеджером* (или корпус-менеджером) (англ. corpus manager). Это специализированная поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме.

Целесообразность создания и смысл использования корпусов определяется следующими предпосылками:

1. 1) достаточно большой (репрезентативный) объем корпуса гарантирует типичность данных и обеспечивает полноту представления всего спектра языковых явлений;
2. 2) данные разного типа находятся в корпусе в своей естественной контекстной форме, что создает возможность их всестороннего и объективного изучения;
3. 3) однажды созданный и подготовленный массив данных может использоваться многократно, многими исследователями и в различных целях.

Можно сказать, что все современные лингвистические исследования и работы по составлению словарей и грамматик так или иначе ориентированы на использование представительных корпусов текстов. Развитие современных интеллектуальных программных систем, предназначенных для обработки текстов на естественном языке, также требует большой экспериментальной лингвистической базы. Спрос на корпусные данные совпал с появлением соответствующих технических возможностей.

Первые лингвистические корпусы текстов появились в 60-е гг. прошлого столетия. В 1963 г. в Брауновском университете (США) впервые был создан большой корпус текстов на машинном носителе (Brown Corpus). Авторы корпуса У. Френсис (W. Francis) и Г. Кучера (H. Kucera) спроектировали его как набор из пятисот двухтысячесловных прозаических печатных текстов американского варианта английского языка. Тексты принадлежали пятнадцати наиболее массовым жанрам англоязычной печатной прозы США и были напечатаны в 1961 г. Корпус сопровождался большим количеством материалов его первичной статистической обработки — частотный и алфавитно-частотный словарь, разнообразные статистические распределения. Появление Брауновского корпуса вызвало всеобщий интерес и оживленные дискуссии. Прежде всего они коснулись принципов отбора текстов и состава потенциально решаемых на таком корпусе задач. Затем последовали Ланкастерский корпус английского языка (Lancaster-Oslo-Bergen Corpus, LOB), Уппсальский корпус русского языка. Среди современных корпусов английского языка наиболее известны Британский национальный корпус (British National Corpus), Международный корпус английского языка (International Corpus of English), лингвистический Банк английского языка (Bank of English) и др. В настоящее время корпусы созданы для многих языков мира (см. Приложение 1). Ведется работа и над созданием Национального корпуса русского языка.

В первой половине 90-х гг. корпусная лингвистика окончательно сформировалась как отдельный раздел науки о языке. При этом она тесно взаимодействует с компьютерной лингвистикой, используя ее достижения и в свою очередь обогащая ее.

Поиск в корпусе данных позволяет по любому слову построить конкорданс – список всех употреблений данного слова в контексте со ссылками на источник. Корпусы могут использоваться для получения разнообразных справок и статистических данных о языковых и речевых единицах. В частности, на основе корпусов можно получить данные о частоте словоформ, лексем, грамматических категорий, проследить изменение частот и контекстов в различные периоды времени, получить данные о совместной встречаемости лексических единиц и т.д. Представительный массив языковых данных за определенный период позволяет изучать динамику процессов изменения лексического состава языка, проводить анализ лексико-грамматических характеристик в разных жанрах и у разных авторов, и т.д. Корпусы призваны служить также источником и инструментом многоаспектных лексикографических работ по подготовке разнообразных исторических и современных словарей. Данные корпусов могут быть использованы для построения и уточнения грамматик и в целях обучения языку.

Можно сказать, что корпусная лингвистика имеет своим предметом теоретические основы и практические механизмы создания и использования представительных массивов языковых данных, предназначенных для лингвистических исследований в интересах широкого круга пользователей.

**Репрезентативность**

Задача создателей корпуса – собрать как можно большее количество текстов, относящихся к тому подмножеству языка, для изучения которого корпус создается. Но главное не только и не столько в количестве языкового материала, сколько в его пропорциональности. Можно сказать, что корпус – это уменьшенная модель языка или подъязыка. Важнейшее понятие корпусной лингвистики – репрезентативность. Под *репрезентативностью* понимается необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т.п. Имеются разные подходы к определению репрезентативности, можно сказать, что применительно к общеязыковому (национальному) корпусу это понятие невозможно рассчитать и описать строго математически, однако к этому можно и нужно стремиться, как на этапе проектирования корпуса, так и на этапе его эксплуатации.

**Размер корпуса**

Термин «корпус» обычно обозначает собрание текстов конечного фиксированного размера. С течением времени объем и состав корпуса может меняться, однако эти изменения должны или не менять его репрезентативность, или менять обоснованно. Объем первых корпусов составлял 1 млн словоупотреблений (Брауновский корпус, Уппсальский корпус русского языка). В настоящее время считается, что объем общеязыкового корпуса должен быть не меньше 100 млн словоупотреблений.

**Разметка**

Для решения различных лингвистических задач мало лишь наличия массива текстов. Требуется также, чтобы тексты содержали в себе явным образом разного рода дополнительную лингвистическую и экстралингвистическую информацию. Так в корпусной лингвистике возникла идея размечен­ного корпуса. *Разметка* (tagging, annotation) заключается в приписывании текстам и их компонентам специальных меток (tag, tags): внешних, экстралингвистических (сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика; сведения об авторе могут включать не только его имя, но также возраст, пол, годы жизни и многое другое. Это кодирование информации имеет название *метаразметка*), структурных (глава, абзац, предложение, словоформа) и собственно лингвистических, описывающих лексические, грамматические и прочие характеристики элементов текста. Набор этих метаданных во многом определяет возможности, предоставляемые корпусами исследователям. При выборе этих данных необходимо руководствоваться целями исследования и потребностями лингвистов, а также возможностями по внесению в текст тех или иных дополнительных признаков. Среди лингвистических типов разметки выделяются:

* ***морфологическая*** разметка. В иностранной терминологии употребляется термин part-of-speech tagging (POS-tagging), дословно – частеречная разметка. В действительности морфологические метки включают не только признак части речи, но и признаки грамматических категорий, свойственных данной части речи. Это основной тип разметки: во-первых, большинство крупных корпусов являются как раз морфологически размеченными корпусами, во-вторых, морфологический анализ рассматривается как основа для дальнейших форм анализа – синтаксического и семантического, и, в-третьих, успехи в компьютерной морфологии позволяют автоматически размечать корпусы больших размеров;
* ***синтаксическая***разметка, являющаяся результатом синтаксического анализа, или *парсинга* (англ. parsing), выполняемого на основе данных морфологического анализа. Этот вид разметки описывает синтаксические связи между лексическими единицами и различные синтаксические конструкции (например, придаточное предложение, глагольное словосочетание и т.п.);
* ***семантическая*** разметка. Хотя для семантики нет единой семантической теории, чаще всего семантические тэги обозначают семантические категории, к которым относится данное слово или словосочетание, и более узкие подкатегории, специфицирующие его значение;
* ***анафорическая***разметка. Фиксирует референтные связи, например, местоименные;
* ***просодическая*** разметка. В просодических корпусах применяются метки, описывающие ударение и интонацию. В корпусах устной разговорной речи просодическая разметка часто сопровождается так называемой *дискурсной* разметкой, которая служит для обозначения пауз, повторов, оговорок, и т.д.

Существуют и другие типы разметки.

**Технология создания корпусов**

Технологический процесс создания корпуса можно представить в виде следующих шагов или этапов.

1. Определение перечня источников.

2. Оцифровка текстов (преобразование в компьютерную форму). Следует сказать, что насколько раньше задача ввода текстов в компьютер была тяжела и трудоемка, настолько сегодня эта проблема решается довольно легко, по крайней мере, что касается современных текстов и в современной орфографии. Эта легкость базируется на успехах в оптическом вводе (сканирование) и распознавании текстовой информации и на глобальной компьютеризации современной жизни, в том числе и в областях, связанных с обработкой текстовой информации. Тексты в электронном виде для создания корпусов могут быть получены самыми разными способами — ручной ввод, сканирование, авторские копии, дары и обмен, Интернет, оригинал-макеты, предоставляемые составителям корпусов издательствами и проч.

3. Предобработка текста. На этом этапе все тексты, полученные из разных источников, проходят филологическую выверку и корректировку. Также осуществляется подготовка библиографического и экстралингвистического описания текста.

4. Конвертирование и графематический анализ. Некоторые тексты проходят также через один или несколько этапов предварительной машинной обработки, в ходе которых осуществляются различного рода перекодировка (если требуется), удаление или преобразование нетексто­вых элементов (рисунки, таблицы), удаление из текста переносов, «жёстких концов строк», обеспечение единообразного написания тире и проч. Как правило, эти операции выполняются в автоматическом режиме. Обычно на этом же этапе осуществляется сегментирование текста на его структурные составляющие.

5. Разметка текста. Разметка текста заключается в приписывании текстам и их компонентам дополнительной информации (метаданных). Метаописание текстов корпуса включает как содержательные элементы данных (библиографические данные, признаки, характеризующие жанровые и стилевые особенности текста, сведения об авторе), так и формальные (имя файла, параметры кодирования, версия языка разметки, исполнители этапов работ). Эти данные обычно вводятся вручную. Структурная разметка документа (выделение абзацев, предложений, слов) и собственно лингвистическая разметка обычно осуществляются автоматически.

6. На следующем этапе осуществляется корректировка результатов автоматической разметки: исправление ошибок и снятие неоднозначности (вручную или полуавтоматически).

7. Заключительный этап – конвертирование размеченных текстов в структуру специализированной лингвистической информационно-поисковой системы (corpus manager), обеспечивающей быстрый многоаспектный поиск и статистическую обработку.

8. И, наконец, обеспечение доступа к корпусу. Корпус может быть доступен в пределах дисплейного класса, может распространяться на CD-ROM и может быть доступен в режиме глобальной сети. Различным категориям пользователей могут предоставляться разные права и разные возможности.

Конечно, в каждом конкретном случае состав и количество проце­дур могут отличаться от выше перечисленных, и реальная технология может оказаться гораздо сложнее.