МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ

РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГБОУ ВО «Тверской государственный университет»

Факультет иностранных языков и международной коммуникации

Направление «Лингвистика»

Кафедра английского языка

РЕФЕРАТ

на тему: «Корпус данных»

по дисциплине: «Информационные технологии в лингвистике»

|  |  |
| --- | --- |
|  | Выполнила:  студентка 11 группы, дневного отделения  Коньшина Олеся Александровна |
|  |  |
|  | Проверила:  к.ф.н, доцент,  доцент кафедры английского языка  Масленникова Е.М. |

ТВЕРЬ

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) 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.

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.

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.

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.

**Содержание корпусов.**

Нынешние исследования в области корпусов имеют, как можно судить по материалам конференции, состоявшейся в 2001 году в Лувене (Бельгия) (“LINC 2001”), явно выраженный “аннотационный крен”. В фокусе внимания находится формат аннотации (разметки) корпусов, методы борьбы с ошибками, даже психологическая реальность аннотированного корпуса (доклад А. Абейе и др.), а ведь проблема того, что именно за тексты должны быть представлены в корпусе и какие типы корпусов могут (должны?) различаться по этому параметру, не менее важна.

Основная задача заключается в репрезентативности корпуса. Общепризнанно, что письменный и устный язык функционирует во многих регистрах (жанрах и пр. – термины в разных традициях различны), от литературы религиозной до технической, от публичной политической речи до неформальной беседы. Коль скоро эти регистры различаются на многих языковых уровнях, то и в корпусе должны быть представлены все или большинство из них. Как гласит нынешнее присловье, в Интернете можно найти только то, что туда положили; то же относится и к корпусу. Решение этой задачи, несомненно, связано с охватом текстов. Принцип “чем больше, тем лучше” здесь работает; частотность и релевантность любого лингвистического явления проще проследить на объёме в сто миллионов словоупотреблений, чем в сто тысяч. Такая задача стояла и перед создателями первых корпусов, в т. ч. Уппсальского, в который включены, наряду с художественными произведениями, публикации журналов и газет.

Задача фиксации устной речи в корпусе вполне решаема, и зарубежный опыт (прежде всего немецкий) это показывает, но это задача для отдельного корпуса; необходимость транскрипции и желательность аудио- или видеодублирования неизбежно сообщают устному корпусу иной формат, а известные языковые особенности разговорной речи ставят её отдельно от письменной. Если мы вынесем устную речь за скобки, то мы должны вычленить из всего “моря” письменных русских текстов некоторые подмножества в зависимости от наших задач

**Источники текстов**

Источники текстов для корпусов в настоящее время весьма обширны. Если в 1960-х годах всякий текст приходилось ad hoc представлять в электронном виде, то в настоящее время практически все жанры русского письменного языка обширно представлены в Интернете. Особенностью Рунета являются библиотеки текстов, содержащие огромные коллекции как художественных, так и технических, правовых, публицистических и проч. произведений. На европейских и американских Интернет-сайтах не так легко найти коллекцию художественных текстов; зачастую это запрещено авторскими правами. У нас же имеются еженедельно обновляющаяся библиотека Максима Мошкова (www.lib.ru), коллекции “Общий текст” (www.textshare.da.ru), “Русский текст” (www.russiantext.com), а также такие сайты, как www.klassika.ru, www.divanchik.net и множество других. Как правило, тексты сканируют “на общественных началах” пользователи Интернета и присылают их администраторам библиотек. Таким образом, в распоряжении создателя корпусов находятся целые массивы художественных текстов в неразмеченном электронном виде – от Пушкина и Достоевского до Пелевина и Акунина – из которого только надлежит выбрать нужное. Немало интересного для составителя корпуса содержат и сайты литературных журналов, в том числе таких, как “Новый мир”, “Знамя” (“Журнальный зал” на www.infoart.ru). Нехудожественные тексты, прежде всего публицистика, новости, в меньшей степени – научные статьи также обильно представлены в русском секторе Всемирной сети. Это и официальные сайты газет и политических журналов, и “новостные” серверы, и различные образовательные ресурсы.

Основными сложностями, с которыми приходится сталкиваться при превращении “текстов для читателя” в “тексты для исследователя”, является отсутствие единого стандарта подачи текста (в том числе даже в пределах одного ресурса – так, на www.lib.ru тире передаётся то дефисом в пробелах, то двумя дефисами, абзацный отступ то тремя неразрывными пробелами, то пятью), большое подчас количество опечаток (для отсканированных текстов – ошибок распознавания), сохранение в тексте переносов, номеров страниц, иногда даже оформление строчек как абзацев (так называемые “жёсткие концы”) и проч. Имеются методы полуавтоматического устранения таких неудобств, но опечатки – наиболее неприятные и трудноустранимые погрешности, серьёзно, как выяснилось, затрудняющие процесс “корпоризации” библиотечных текстов. В текстах Интернет-новостей, а также в многих других местах “неофициального” Интернета встречаются, наряду с опечатками, и просто орфографические и пунктуационные ошибки

**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 корпуса**

Разметка заключается в приписывании текстам и их компонентам специальных **тегов**: лингвистических и внешних (экстралингвистических). Выделяют следующие лингвистические типы разметки: морфологическая, семантическая, синтаксическая, анафорическая, просодическая, дискурсная и т. д. К некоторым корпусам применяются дальнейшие структурные уровни анализа. В частности, некоторые небольшие корпусы могут быть полностью синтаксически размечены. Такие корпусы обычно называют *глубоко аннотированными* или *синтаксическими*, а сама синтаксическая структура при этом является [деревом зависимостей](https://ru.wikipedia.org/wiki/%D0%94%D0%B5%D1%80%D0%B5%D0%B2%D0%BE_%D0%B7%D0%B0%D0%B2%D0%B8%D1%81%D0%B8%D0%BC%D0%BE%D1%81%D1%82%D0%B5%D0%B9).

Ручная разметка (аннотирование) текстов — дорогостоящая и трудоемкая задача. На данный момент в открытом доступе представлены различные программные средства для разметки корпусов[[2]](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%80%D0%BF%D1%83%D1%81_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%BE%D0%B2#cite_note-_888ba3dfab8467bf-2). Условно их можно разделить на **обособленные (stand-alone)** и **веб-ориентированные (web-based)**. При этом акцент разработчиков в последние годы сместился в сторону веб-приложений.

**Данные системы обладают рядом преимуществ:**

* возможность одновременной разметки одного документа несколькими людьми
* не требуют установки дополнительных программных средств, кроме браузера
* гибкое разграничение прав доступа
* отображение текущего прогресса процесса разметки
* возможность модификации размечаемого корпуса

### Проблема представительности

Корпус состоит из конечного числа текстов, но он призван адекватно отражать лексикограмматические феномены, типичные для всего объема текстов в соответствующем языке (или подъязыке). Для представительности важен как размер, так и структура корпуса. Представительный размер зависит от задачи, поскольку он определяется тем, как много примеров может быть найдено для исследуемых феноменов. В связи с тем, что со статистической точки зрения язык содержит большое число относительно редких слов (Закон Ципфа), для исследования первых пяти тысяч наиболее частотных слов (например, *убыток, извиняться*) требуется корпус размером около 10-20 миллионов словоупотреблений, в то время как для описания первых двадцати тысяч слов (*незатейливый, сердцебиение, роиться*) уже требуется корпус свыше ста миллионов словоупотреблений.

### Проблема представления результатов

В больших корпусах возникает проблема, которая ранее была неактуальной: поиск по запросу может выдавать сотни и даже тысячи результатов (контекстов употребления), которые просто физически невозможно просмотреть в ограниченное время. Для решения этой проблемы разрабатываются системы, позволяющие группировать результаты поиска и автоматически разбивать их на подмножества (кластеризация результатов поиска), либо выдающие наиболее устойчивые словосочетания (коллокации) со статистической оценкой их значимости.

**Корпусы текстов русского языка:**

* [Национальный корпус русского языка](https://ru.wikipedia.org/wiki/%D0%9D%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9_%D0%BA%D0%BE%D1%80%D0%BF%D1%83%D1%81_%D1%80%D1%83%D1%81%D1%81%D0%BA%D0%BE%D0%B3%D0%BE_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0" \o "Ключевое слово)
* [Генеральный интернет-корпус русского языка](https://ru.wikipedia.org/wiki/%D0%93%D0%B5%D0%BD%D0%B5%D1%80%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9_%D0%B8%D0%BD%D1%82%D0%B5%D1%80%D0%BD%D0%B5%D1%82-%D0%BA%D0%BE%D1%80%D0%BF%D1%83%D1%81_%D1%80%D1%83%D1%81%D1%81%D0%BA%D0%BE%D0%B3%D0%BE_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0)
* [Русскоязычный корпус проекта Aranea](http://unesco.uniba.sk/guest/)
* [Корпус биографических текстов](https://sites.google.com/site/utcorpus)
* [RuTweetCorp](https://study.mokoron.com/)

Корпусы текстов обеспечивают возможность для исследователя опустить такой трудоёмкий этап, как сбор текстов. Корпусы помогают выявлять закономерности, феномены, опираясь на речевой материал. Развитие компьютерных технологий упростило и ускорило процессы лингвистической обработки больших текстов.

Появление различных технических средств, компьютерных технологий и необходимость совершенствования способов сбора и хранения информации обусловило развитие корпусной лингвистики. Создание корпусов текстов значительно облегчило сбор и хранение информации. Это очень ценится при создании словарей, глоссариев, лексикографических работ. Также способ хранения корпусов текстов позволяет более надежно и дольше хранить любой языковой материал, что является важной находкой и инструментом в лексикографии.

Также создание корпусов текстов позволило быстрее и качественнее производить различные лингвистические исследования и решать важные исследовательские задачи.

Но так как корпусная лингвистика относительно молодая и развивающаяся наука, в ней существуют различные проблемные вопросы.

К настоящему времени накоплен уже довольно значительный опыт построения корпусов текстов, однако информация о наличии корпусов в разных странах ограничена. Это существенно затрудняет проведение научных исследований и нередко приводит к дублированию работы.

Сложность получения информации о существующих корпусах текстов связана с тем, что практически все они создаются по локальным проектам отдельными организациями. Международная координация усилий в этой области пока еще не налажена в должной мере. Вторая проблема в области корпусной лингвистики связана с первой: локальность проектов приводит к возникновению различных форматов представления корпусов текстов. Отсутствие единого формата представления дополняется и несовместимостью программного обеспечения.

В имеющихся описаниях корпусов нет ясной информации о содержательных принципах отбора материала. Поскольку корпус является сужением проблемной области, то представительность корпуса прямо связана с принципами отбора материала. Если же нет уверенности в представительности корпуса, его заведомо нельзя использовать для многих видов лингвистической деятельности, например, для оценки частоты употребления лексем в тех или иных значениях или для составления словников некоторой проблемной области.

Таким образом, перед учеными встает задача устранения тех самых проблемных вопросов, возникших при процессе создания корпусов текстов.

**Список использованных источников:**

**URL:** [**https://elar.urfu.ru/bitstream/10995/36257/1/ittsm-2016-34.pdf**](https://elar.urfu.ru/bitstream/10995/36257/1/ittsm-2016-34.pdf) **(дата обращения 28.05.2021)**

**URL:** [**https://ru.wikipedia.org/wiki**](https://ru.wikipedia.org/wiki) **(дата обращения 28.05.2021)**

**URL:** <https://wreferat.baza> **(дата обращения 28.05.2021)**

**Захаров В.П.** Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005. – 48 с.