МИНОБРНАУКИ РОССИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет теории и методики преподавания

Кафедра романской филологии

Курсовая работа

«Лингвистические корпусы. Корпусная лингвистика»

Студент \_\_\_\_\_\_\_\_\_\_\_ Пименова С. 25.09.18

Руководитель \_\_\_\_\_\_\_\_\_ Донина. О.В к.ф.н. 22.10.18

Воронеж 2018

Оглавление

[Введение 3](#_Toc6920541)

[Глава 1.1Введение в понятие корпусной лингвистики. 4](#_Toc6920542)

[1.2 История корпусной лингвистики 5](#_Toc6920543)

[1.3 Основные понятия корпусной лингвистики 7](#_Toc6920544)

[1.4 Типология корпусов 10](#_Toc6920545)

[Глава 2. Технология создания корпусов. 14](#_Toc6920546)

[Глава 3. НКРЯ. 16](#_Toc6920547)

[Доступ 17](#_Toc6920548)

[Заключение 17](#_Toc6920549)

[Список литературы: 19](#_Toc6920550)

# Введение

Развитие корпусной лингвистики, а также построение корпусов является одной из актуальных проблем современного языкознания. На данный момент использование корпусов играет ведущую роль при проведении большинства лингвистических исследований. Хотя еще в первой половине ХХ века было возможно построение корпусов только вручную. Это отнимало массу времени, сил и затрат. Поэтому обращение к корпусам текстов было минимальным и лишь в том случае, когда большое количество людей было в этом заинтересовано. Но благодаря развитию корпусной лингвистики использование корпусов стало возможным повсеместно. Структурировать и синхронизировать материал стало гораздо легче, а затраты резко снизились.

Важность развития корпусной лингвистики не вызывает никаких сомнений. В ней объединены многие положительные свойства гуманитарных и технических наук. Но возникнув относительно недавно, она не успела полностью сформироваться, и ее статус как независимой науки еще не установлен. Так, многие ученые рассматривают корпусную лингвистику в качестве подобласти традиционной. Они утверждают, что она имеет лишь прикладное применение, а  теоретические обоснования отсутствуют. По их мнению, корпусная лингвистика является улучшенной методикой сбора и обработки материала, новым информационным ресурсом. С другой стороны, если есть прикладная корпусная лингвистика, почему не может быть и теоретической, т.е. дисциплины, обладающей собственным предметом, методом и теорией. Отсутствие теоретической базы на данный момент не влияет на возможность ее разработки в будущем.

Задачи данной работы следующие:

а) рассмотреть основные понятия корпусной лингвистики и принципы создания лингвистических корпусов;

б) описать технологию создания лингвистических корпусов.

в) описать, каким образом можно использовать материалы Национального корпуса русского языка при изучении лексики, морфологии, синтаксиса;

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.

# Глава 1.1Введение в понятие корпусной лингвистики.

**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) 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.

Одним из основных источников языкового материала, необходимого для проведения лингвистических и филологических исследований, является текст, письменный или устный. На основе анализа множества текстов можно сделать вывод об интересующем исследователя языковом явлении, например, о поведении грамматической конструкции, использовании выразительных средств в языке и т.п. Развитие вычислительной техники способствовало тому, что большое количество текстов стало доступно в электронном виде. Для того чтобы можно было работать с такими объемами текстов, извлекая из них нужную информацию, во всем мире стали создаваться лингвистические корпусы, т.е. коллекции текстов, специально отобранных, размеченных по различным лингвистическим параметрам и обеспеченных системой поиска. Увеличение объема материала исследования потребовало применения новых методов анализа лингвистических данных, включая их статистическую обработку.

# 1.2 История корпусной лингвистики

Можно сказать, что первые корпуса и корпусные методы появились задолго до возникновения корпусной лингвистики как научного направления. По сути, любое лингвистическое исследование, основанное на сопоставлении и анализе контекстов, является корпусным. Еще в конце XIX - начале XX в. в целях усовершенствования средств связи, а также в целях быстрого обучения языку создавались частотные словари. Частотность слов оценивалась по специальным выборкам текстов.

Первые корпуса в строгом смысле этого слова появились в 60-х гг. ХХ в. Прообразом для них послужили словарные картотеки - собрания фрагментов текстов, обычно в виде карточек, содержащих то или иное слово, и систематизированные относительно описываемого слова (в основном, по алфавиту).

В 1963 г. в Брауновском университете (США) для создания частотного словаря американского варианта английского языка был создан большой корпус на цифровом носителе (Brown Corpus), включающий 1 млн. слов. При оценке частоты некоторого слова в языке возникает проблема "сбалансированности" выборки.

В языке частотность многих слов обусловлена тематикой текстов. Так, например, слово «переменная» будет чрезвычайно частотно в математических текстах. Вероятность же встретить данное слово в художественной литературе очень мала.

Для обеспечения корректности данных относительно частоты употребления слов создатели корпуса (У. Френсис и Г. Кучера) разработали строгую процедуру отбора текстов: в корпус вошли 500 фрагментов прозаических текстов, относящихся к 15 наиболее массовым жанрам и напечатанных в 1961г. Возникновение корпусных методов связано с бурным развитием компьютерных технологий во второй половине ХХ в. Возможность сканирования и распознавания текста (перевод в текстовый формат), появление баз данных и систем управления базами данных сделали возможным сбор, хранение и обработку огромных массивов текстовых данных.

Не последнюю роль в развитии корпусной лингвистики сыграла популяризация мировой сети Интернет, т.к. корпуса стали доступны широкому кругу пользователей, значительно расширились возможности их наполнения. С тех пор накоплен значительный опыт разработки и применения корпусов. Ежегодно публикуется множество работ по данной тематике.

В России разработкой и исследованием корпусов занимаются специалисты Центра лингвистической документации при Независимом Московском университете, отдела экспериментальной лексикографии Института русского языка им. В. В. Виноградова РАН, Института языкознания РАН, Института проблем передачи информации РАН, Института лингвистических исследований РАН в Санкт-Петербурге и др.

1.3 Основные понятия корпусной лингвистики  
  
Центральное понятие корпусной лингвистики - **лингвистический корпус.** Тексты в корпус выбираются не случайным образом, а в соответствии с проблемной областью, т.е. областью реализаций интересующих исследователя языковых явлений.

Проблемная область имеет два аспекта: языковой и речевой.

*Языковой* аспект - это само изучаемое явление, а *речевой* - это множество контекстов, в которых это явление представлено. Проблемная область может быть как очень широкой (все произведения Достоевского Ф.М.), так и достаточно узкой (случаи согласования сказуемого с количественной группой по числу).

Одним из принципиальных вопросов является вопрос о том, какие тексты и в каком объеме необходимо отобрать в корпус. С одной стороны, хотелось бы, чтобы исследуемое явление, как бы оно ни было редко в языке, нашло отражение в корпусе.

Одним из требований, предъявляемым к составу и структуре корпуса является требование **полноты**. Данное требование входит в противоречие с другим важным принципом создания корпуса - требованием **репрезентативности**.

Задача создателей корпуса - собрать как можно большее количество текстов, относящихся к тому подмножеству языка, для изучения которого корпус создается. Каким бы специфичным ни был феномен, ни один корпус не может содержать все его реализации. Поэтому корпус - это всегда определенная выборка из проблемной области, которая осуществляется на основе некоторых критериев, устанавливаемых исследователем в зависимости от задачи. Такая выборка должна отражать те или иные параметры исследуемого языкового явления в той же пропорции, что и в языке вообще или в некотором исследуемом подмножестве языка. С точки зрения отбора текстов в корпус различают сбалансированные корпусы и мониторные корпуса.

В сбалансированных корпусах в корпус включаются тексты, представляющие разные модусы **дискурса** (включая как письменные, так и устные тексты), разнообразные по жанрам, стилям и тематике. При разработке корпуса устанавливаются пропорции, в которых должны быть представлены тексты разных жанров, стилей и т.п. Такие корпусы имеют фиксированный объем. Пополнение таких корпусов происходит только после тщательной процедуры отбора новых текстов. Мониторные корпусы постоянно пополняются новыми текстами на данном языке, при этом баланс текстов разных модусов, стилей и жанров не соблюдается. Создатели мониторных корпусов считают, что "статистическая" обоснованность данных, полученных из корпуса, будет достигнута за счёт объема корпуса, исчисляемого в миллиардах слов.

Важным параметром корпуса является его объем. Если первые корпусы достигали миллиона слов (точнее, словоупотреблений или текстоформ), то объем современных корпусов исчисляется сотнями миллионов (например, объем Национального корпуса русского языка на данный момент составляет около 360 млн. словоупотреблений) или миллиардами (например, объем корпуса английского языка Bank of English превышает 2,5 миллиарда слов). Как уже отмечалось, для решения различных лингвистических задач необходимо, чтобы тексты и отдельные языковые единицы внутри текстов содержали дополнительную лингвистическую и металингвистическую информацию - **разметку** (аннотацию).

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;

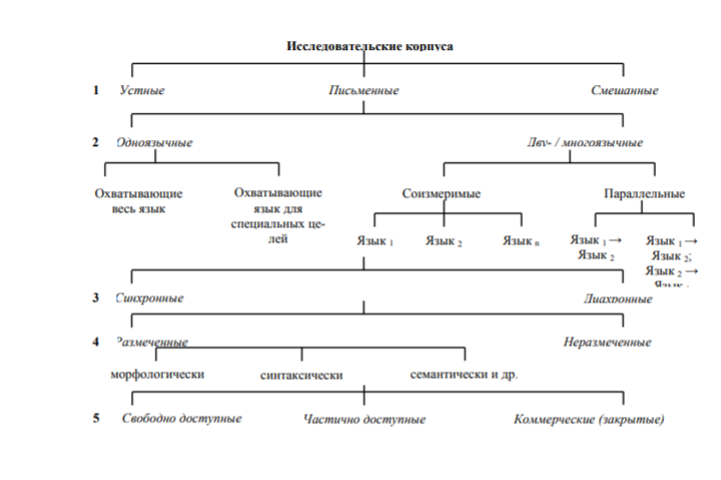
**· анафорическая разметка**. Фиксирует референтные связи, например, местоименные;

**· 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.

# 1.4 Типология корпусов

Среди существующего многообразия исследовательских корпусов подчас очень сложно ориентироваться, ведь цели и задачи, стоящие перед лингвистом, очень часто совпадают в общем, но разнятся в отдельных отраслях и областях. Правильный выбор соответствующего корпуса – это первый шаг, который должен осуществить исследователь при анализе «изучаемых объектов».

Все разнообразие существующих корпусов «определяется многообразием исследовательских и прикладных задач, для решения которых они создаются» и может быть представлено следующей схемой (см. рисунок). 1.



* Устные – письменные – смешанные. Под **устным корпусом** подразумевается структурированная совокупность речевых фрагментов, которая обеспечена программными средствами доступа к ним.

Первые устные корпуса появились в начале 80-х гг. прошлого века на материале американского варианта английского языка. Позже возникли специальные координационные центры по сбору, хранению, распространению и созданию устных корпусов.

1. Например, LDC (Linguistic Data Consorcium, <http://www.ldc.upenn.edu>),
2. CSLU (Center for Spoken Language Understanding , http://www. cslu.ogi.edu),
3. ELRA (European Language Resources Association, http://www.elra.info). Большинство из существующих корпусов относятся к письменным (например, находящаяся в свободном доступе часть Мангеймского корпуса немецкого языка, http:// [www.ids-mannheim.de](http://www.ids-mannheim.de)) либо смешанным, однако доля лингвистически размеченных устных текстов даже в смешанных корпусах (чаще всего это национальные корпуса какого-либо языка, например: русского языка –
4. http://www.ruscorpora.ru, американского варианта английского языка – http://corpus. byu.edu/coca,
5. английского языка – http:// corpus.byu.edu/bnc) ничтожно мала по отношению ко всему массиву корпуса.

* . Одноязычные – двуязычные / многоязычные.

Среди одноязычных корпусов можно выделить две группы: с одной стороны, корпуса, охватывающие весь язык, с другой – охватывающие только язык для специальных целей. Например, Corpus of Early English Medical Writing (CEEM)– корпус медицинских текстов на английском языке 1375–1750 гг. общим объемом около 1,5 млн слов, в котором содержатся теоретические работы, справочники, стихотворные тексты на медицинские темы.

В двуязычных и многоязычных корпусах тексты могут быть представлены соизмеримо или параллельно.

Так, в 1992 г. была создана Европейская корпусная инициатива (European Corpus Initiative (ECI)) – международная организация, занимающаяся созданием огромного многоязычного корпуса для научных целей ( см.: http://www.elsnet. org/resources/eciCorpus.html). В данном соизмеримом корпусе содержатся в основном тексты европейских языков, а также тексты на таких языках, как турецкий, китайский, японский, русский и др., общим объемом более 98 млн слов, данный корпус носит коммерческий характер. Корпус лельных текстов предназначены, в первую очередь, для сопоставительного анализа текстов «оригинал – перевод» в целях обучения методам и приемам перевода. Удачный пример такого вида корпусов – European Parliament Proceedings Parallel Corpus 1996–2011 (http://www.statmt.org/europarl), где представлены параллельные тексты заседания Европейского парламента на всех европейских языках с переводом на английский.

* Синхронный – диахронный. Синхронные корпуса предполагают представление текстового материала для рассмотрения состояния языка как системы в определенный момент времени.

В частности, в некоммерческой версии Британского национального корпуса (http://corpus.byu. edu/bnc) представлены лишь тексты конца ХХ в. – с 1980 по 1993 г.

Для рассмотрения исторического развития какого-либо языкового явления либо всей языковой системы в целом существуют диахронные корпуса. Например, Thesaurus Indogermanischer Text- und Sprachmaterialien (http://titus.uni-frankfurt.de), в котором представлены индогерманские тексты различных эпох.

* Неразмеченные – размеченные.

**Неразмеченный корпус** – это массив текстов, которые содержат определенное количество упоминаний искомого элемента. При этом результаты поиска, предоставляемые в неразмеченных корпусах, могут быть использованы в лингвистических исследованиях, но только с чисто статистической точки зрения. Размеченные (морфологически, синтаксически, просодически и др., см. выше) корпуса предоставляют намного больше возможностей для проведения лингвистического анализа.

# Глава 2. 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.

# Глава 3. НКРЯ.

Корпус - это информационно-справочная система, основанная на собрании текстов на некотором языке в электронной форме. Национальный корпус представляет данный язык на определенном этапе (или этапах) его существования и во всём многообразии жанров, стилей, территориальных и социальных вариантов. Национальный корпус создается специалистами по корпусной лингвистике для научных исследований и обучения языку. Большинство крупных языков мира уже имеет свои национальные корпуса (различающиеся по полноте и уровню научной обработки текстов). Общепризнанным образцом является, в частности, Британский национальный корпус: на него ориентированы многие другие современные корпуса. Среди корпусов славянских языков выделяется также Чешский национальный корпус, созданный в Карловом университете Праги.  
  
**Состав корпуса**

В кор­пус вхо­дят как пись­мен­ные тек­сты (ху­до­же­ствен­ные, ме­му­а­ры, пуб­ли­ци­сти­ка, на­уч­ная, ре­ли­ги­оз­ная ли­те­ра­ту­ра, по­все­днев­ная пе­чат­ная про­дук­ция), так и за­пи­си уст­ных тек­стов (пуб­лич­ной речи и част­ных бесед).

В кор­пус также вхо­дят под­кор­пу­сы [по­э­ти­че­ских](https://wiki2.org/ru/%D0%9F%D0%BE%D1%8D%D0%B7%D0%B8%D1%8F) и [диа­лект­ных](https://wiki2.org/ru/%D0%94%D0%B8%D0%B0%D0%BB%D0%B5%D0%BA%D1%82) тек­стов, [кор­пу­сы па­рал­лель­ных тек­стов](https://wiki2.org/ru/%D0%91%D0%B8%D1%82%D0%B5%D0%BA%D1%81%D1%82)(для сле­ду­ю­щих язы­ков до­ступ­ны па­рал­лель­ные кор­пу­сы с рус­ским: ан­глий­ский, ар­мян­ский, бе­ло­рус­ский, бол­гар­ский, бу­рят­ский, ис­пан­ский, ита­льян­ский, ки­тай­ский, ла­тыш­ский, немец­кий, поль­ский, укра­ин­ский, фран­цуз­ский, швед­ский, эс­тон­ский и муль­ти­я­зыч­ный), от­дель­ный га­зет­ный кор­пус (ма­те­ри­а­лы СМИ на­ча­ла XXI века), цер­ков­но­сла­вян­ский кор­пус (бо­го­слу­жеб­ные тек­сты, со­вре­мен­ные (XIX—XX век) и более ран­них пе­ри­о­дов), ис­то­ри­че­ский (в том числе древ­не­рус­ский, ста­ро­рус­ский, бе­ре­стя­ных гра­мот), син­так­си­че­ский, ак­цен­то­ло­ги­че­ский, муль­ти­ме­дий­ный и обу­ча­ю­щий под­кор­пу­сы.

**Объём корпуса**

Объём ос­нов­но­го кор­пу­са на 7 ап­ре­ля 2018 года со­став­лял 283 млн сло­во­упо­треб­ле­ний, а общий объём кор­пу­сов пре­вы­ша­ет 600 млн словоупотреблений..

Тек­сты снаб­же­ны ме­та­раз­мет­кой (по дате со­зда­ния, ав­то­ру, жанру и т. п.); сло­во­фор­мы в текстах снаб­же­ны ав­то­ма­ти­че­ской мор­фо­ло­ги­че­ской и се­ман­ти­че­ской раз­мет­кой; па­рал­лель­ные тек­сты вы­ров­не­ны; тек­сты по­э­ти­че­ско­го кор­пу­са снаб­же­ны также осо­бой мет­ри­че­ской раз­мет­кой.

1,5 % тек­стов снаб­же­ны [мор­фо­ло­ги­че­ской](https://wiki2.org/ru/%D0%9C%D0%BE%D1%80%D1%84%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F_(%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0))  и [се­ман­ти­че­ской](https://wiki2.org/ru/%D0%9B%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D1%81%D0%B5%D0%BC%D0%B0%D0%BD%D1%82%D0%B8%D0%BA%D0%B0) раз­мет­кой со сня­той вруч­ную омо­ни­ми­ей («дез­ам­би­гу­и­ро­ван­ный под­кор­пус»).

## Доступ

В на­сто­я­щее время *сво­бод­ным* и *бес­плат­ным* яв­ля­ет­ся *толь­ко поиск* по кор­пу­су. Сайт кор­пу­са и поиск по нему под­дер­жи­ва­ют­ся ком­па­ни­ей «[Ян­декс](https://wiki2.org/ru/%D0%AF%D0%BD%D0%B4%D0%B5%D0%BA%D1%81)», со­труд­ни­ки ко­то­рой при­ни­ма­ли уча­стие также в раз­ра­бот­ке про­грамм­но­го обес­пе­че­ния кор­пу­са. До­ступ ко всему кор­пу­су невоз­мо­жен в связи с за­ко­ном об ав­тор­ских пра­вах. Для по­лу­че­ния до­сту­па к 1/6 раз­ме­чен­ной части под­кор­пу­са необ­хо­ди­мо за­ре­ги­стри­ро­вать­ся и при­нять ли­цен­зи­он­ное соглашение.

# Заключение

Таким образом, корпус – это представленный в электронном виде, как правило, размеченный для анализа в лингвистических целях, обеспеченный сравнительно простой в использовании поисковой системой репрезентативный массив неотредактированных текстов, представляющих как можно большее количество «вариантов» языка. В период зарождения корпусной лингвистики вопросов компьютеризации данного направления не ставилось, и «исследователи указывали на возможность пренебречь языковой вариативностью, т. е. территориальной, социальной, профессиональной, возрастной, гендерной, индивидуальной и тому подобной дифференциацией языка]. Сегодня же, пренебрегая ею, мы сознательно ограничиваем себя различными рамками при изучении текстов определенного языка, что ставит под вопрос объективность подобного рода исследования. С появлением электронных корпусов многообразие форм существования языка стало более наглядным, возможности исследования языковых данных расширились. Современный лингвистический корпус содержит сотни миллионов словоупотреблений, а то, что с помощью электронного корпуса результаты примеров словоупотреблений можно получить за считанные доли секунд, существенно упрощает задачу лингвистам. Представленная типология корпусов, не претендуя на всеохватность, показывает нам существующее многообразие корпусов текстов и позволяет сориентироваться в нем для последующего проведения научного исследования

# Список литературы:

1. Захаров В. П. Корпусная лингвистика. СПб., 2005.
2. Кибрик А. Е., Брыкина М. М., Леонтьев А. П., Хитров А. Н. Русские посессивные конструкции в свете корпусно-статистического исследования // Вопросы языкознания. 2006. Вып. 1. С. 16–45.
3. Кривнова О. Ф. Области применения речевых корпусов и опыт их разработки // Тр. XVIII Сессии Российского акустического общества РАО. Таганрог, 2006.
4. Мельников Г. П. Системная типология языков: Принципы, методы, модели / Отв. ред. Л. Г. Зубкова. М.: Наука, 2003.
5. Плунгян В. А. «Интегрум» и Национальный корпус русского языка в лингвистических исследованиях // Integrum: точные методы и гуманитарные науки. М., 2006. С. 76–84.
6. Рыков В. В. Прагматически ориентированный корпус текстов // Тверской лингвистический меридиан. Тверь, 1999. Вып. 3. С. 89–96
7. Кольцова, Л.М., Грачева, Ж.В. Ресурсы Национального корпуса русского языка как основа подготовки научных и учебно-методических работ по современному русскому языку / Л.М. Кольцова, Ж.В. Грачева // Национальный корпус русского языка и проблемы гуманитарного образования. - М.: Тэис, 2007. - С. 6 - 10.
8. Полат, Е.С. Интернет в гуманитарном образовании: учеб. пособие для студ. высш. учеб. заведений / Е.С. Полат. - М.: ВЛАДОС, 2001. - 272 с.
9. Абрамова, С.В. Использование Национального корпуса русского языка в учебно-исследовательской работе школьников / С.В. Абрамова // Национальный корпус русского языка и проблемы гуманитарного образования. - М.: Тэис, 2007. - С. 5 - 7.