МИНОБРНАУКИ РОССИИ

**ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССОНАЛЬНОГО ОБРАЗОВАНИЯ**

“ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИСТЕТ”

*Факультет романо-германской филологии*

*Кафедра немецкой филологии*

*Лингвистические корпусы. Корпусная лингвистика*

Курсовая работа

45.03.02 Теория и методика преподавания иностранных языков и культур

Немецкое отделение

Студент Усачева В.С. 01.10.18

Руководитель Донина О.В, к.ф.н., 01.10.18

Воронеж, 2018

**Оглавление**

[**ВВЕДЕНИЕ** 2](#_Toc526167715)

[**1.** **Корпусная лингвистика как раздел прикладной лингвистики** 4](#_Toc526167716)

[**1.1** **Определение корпуса и корпусной лингвистики** 4](#_Toc526167717)

[**1.2 Определение разметки корпуса** 6](#_Toc526167718)

[**2.** **История зарождения Национальных лингвистических корпусов** 8](#_Toc526167719)

[**2.1 Международные лингвистические корпуса** 8](#_Toc526167720)

[**2.2 Национальный Корпус русского языка** 9](#_Toc526167721)

[**3.** **Корпусная лингвистика как способ обучения языку** 12](#_Toc526167722)

[**3.1** **Основные проблемы в развитии корпусной лингвистики** 12](#_Toc526167723)

[**3.2** **Основные перспективы развития корпусной лингвистики** 13](#_Toc526167724)

[**ЗАКЛЮЧЕНИЕ** 16](#_Toc526167725)

[**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**: 18](#_Toc526167726)

**ВВЕДЕНИЕ**

Корпус языка и наука, которая с этим связана, корпусная лингвистика, - это область, которая стала активно развиваться и ворвалась в жизнь лингвистов примерно в самом конце XX — начале XXI века. Лингвистика корпусов по определению является суперсовременной; это именно та область лингвистики, которая изучена не до конца, тем не менее признание роли корпусов в настоящее время стремительно растет.

Текст, как устный, так и письменный, является одним из основных источников языкового материала, необходимого для проведения лингвистических и филологических исследований. На основе анализа множества текстов можно сделать вывод об интересующем исследователя языковом явлении, например, о поведении грамматической конструкции, использования выразительных средств в языке и т.п. Совокупность текстов, собранных в соответствии с определёнными принципами, размеченными по определённому стандарту и обеспеченные специализированной поисковой системой, являются лингвистическими корпусами [7, С. 21].

Одной из важнейших проблем, отмечаемых в лингвистике сегодня, является нехватка адекватных педагогических текстовых материалов и актуальных вокабуляров при изучении языка или решении лингвистических задач, проведении исследований. В данной области значительную помощь может оказать привлечение методов корпусной лингвистики: автоматизированное извлечение информации, обучение на основе данных, текстовые поиски в крупномасштабных корпусах с использованием методов обработки естественного языка.

В последние годы создание корпусов и корпусно-ориентированные исследования стали неотъемлемой частью деятельности лингвистов. Направление по изучение корпусов, как национальных, так и иностранных, является актуальным в наши дни, что особенно связано с развитием техники. С появлением сети Интернет создалась принципиально новая ситуация в системе образования и сфере науки. Информационный потенциал Интернета просто неисчерпаем.

Таким образом, цель курсовой работы: рассмотреть основные функции корпусной лингвистики и выявить возможности использования корпусной лингвистики в процессе обучения русскому языку как в средних, так и высших учебных заведениях.

Задачи курсовой работы:

1. Изучить определения основных понятий «корпуса языка» и корпусной лингвистики; выявить их типологию.
2. Рассмотреть общее историческое развитие корпусов в мире и выявить основные этапы развития корпусной лингвистики в России.
3. Дать краткую характеристику крупнейших электронных библиотек.
4. Описать, каким образом можно использовать материалы Национального корпуса русского языка при изучении лексики, морфологии, синтаксиса.

Проведение данной работы необходимо, так как корпусы языка - информационно-справочная системы, основанные на собрании русских текстов в электронной форме. Корпусы помогают в решении лингвистических и даже шире, филологических, задач.

1. **Корпусная лингвистика как раздел прикладной лингвистики**
   1. **Определение корпуса и корпусной лингвистики**

Одним из основных источников языкового материала, необходимого для проведения лингвистических и филологических исследований, является текст. Развитие вычислительной техники способствовало тому, что большое количество текстов стало доступно в электронном виде [1, С. 6]. 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.

Возможность пользоваться обширными электронными ресурсами значительно облегчила процесс сбора материала в лингвистических исследованиях. Однако такая доступность языковых данных радикально изменила требования к доказательной базе лингвистических исследований: в наше время мало привести несколько экзотических примеров в подтверждение выдвигаемой концепции, необходимо представить достаточно полную выборку данных того или иного языка. Однако работать с мегабайтами «сырых» текстов не менее трудно, чем искать примеры вручную.

Таким образом, корпусная лингвистика включает два аспекта: во-первых, создание и разметка[[1]](#footnote-1) (аннотирование) корпусов текстов и разработка средств поиска по ним и, во-вторых, собственно лингвистический аспект - экспериментальные исследования на базе корпусов. Необходимость и возможность обработки большого массива текстов для извлечения из них лингвистических, литературоведческих и других данных обусловили бурный рост электронных ресурсов, включая лингвистически аннотированные корпуса текстов. Использование корпусов текстов дает возможность наблюдать поведение интересующих исследователя языковых единиц (слов, словосочетаний, грамматических категорий, синтаксических конструкций и т.д.) в естественной языковой среде, т.е. в реально существующих, а не искусственно сконструированных контекстах [6, С. 6].

Другими словам, корпус языка — это электронное собрание текстов на данном языке. «Если представить себе, что тексты — это сложная игрушка, собранная из разных деталей, то ребенок может с игрушкой играть, развлекаться, а лингвист поступает как плохой ребенок, который разбирает текст на части, ничего другого ему не надо. Потому что нормальному человеку от текста нужен смысл, понимание того, что ему сказали, а лингвист сразу начинает разбирать его на части…» ‑ именно такое определение корпусам текста и их предназначению дает доктор филологических наук В.А. Плунгян [7, С. 14].

Наличие большого количества текстов в электронной форме существенно облегчило задачу создания больших представительных корпусов размером в десятки и сотни миллионов слов, но не ликвидировало проблем: сбор тысяч текстов, снятие проблем с авторскими правами, приведение всех текстов в единую форму, балансировка корпуса по темам и жанрам отнимают много времени.

Поэтому корпус некоторого языка - это, в первом приближении, собрание текстов на данном языке, представленное в электронной форме и снабженное научным аппаратом. Аппарат, «встроенный» в корпус, обычно называется «разметкой», или «аннотацией», корпуса; корпус тем лучше, чем полнее и совершеннее его аннотация. Собственно, наука о корпусах (а таковая давно уже существует и называется «корпусная лингвистика») - это прежде всего наука о том, как сделать хорошую разметку корпуса [7, С. 23].

**1.2 Определение разметки корпуса**

Под разметкой понимается приписывание текстам и их компонентам специальных меток: внешних, экстралингвистических, структурных и собственно лингвистических, описывающих лексические, грамматические и прочие характеристики элементов текста [Захаров, 2005. С. 6]. Метаразметка включает в себя сведения об авторе и о самом тексте. Морфологический вид разметки является наиболее распространенным в существующих корпусах, при этом учитывается не только признак части речи, но и признаки грамматических категорий. Морфологическая разметка осуществляется с помощью специальных программ автоматического морфологического анализа. Например, в небольшой части Национального корпуса русского языка[[2]](#footnote-2) (объемом 6 млн словоупотреблений) произведено ручное снятие морфологической омонимии и дополнительная коррекция результатов работы программы автоматического морфологического анализа [4, С. 11].

Хорошая разметка, в частности, позволяет быстро и эффективно найти в корпусе те слова, формы и конструкции, которые нужны исследователю. Ведь в обычном тексте нет никаких сведений, например, о грамматической характеристике входящих в него слов. Если требуется найти слово (например, слово «диван» во всех его формах), то с этой задачей неплохо справится и обычный текстовый редактор: достаточно написать в окне поиска цепочку букв «диван», и результат окажется вполне приемлемым. Несколько хуже, правда, дело будет обстоять в том случае, если мы захотим искать, например, все формы русского слова «рожь». У этого слова есть беглая гласная, поэтому в некоторых формах его основа выглядит как «-рож-», а в некоторых – как «-рж-». Но если мы напишем в окне поиска только эти две буквы (РЖ), результат окажется неприемлемым: слишком много других русских слов тоже начинаются на РЖ- («ржавый», «ржаветь», «ржать», «Ржев» и т. п.). Стало быть, программа поиска должна понимать как минимум то, какие формы в тексте относятся к одному и тому же слову (например, ржи и рожью, но не ржал и Ржев), т. е. хотя бы частично «понимать» грамматическую структуру данного языка.

Тем более это понимание необходимо, если требуется искать не слова, а формы. Например, найти в достаточно длинном тексте все формы дательного падежа единственного числа. Никакой текстовый редактор с этой задачей, конечно, не справится. Для того чтобы грамматические формы можно было автоматически найти в тексте, этот текст как раз и необходимо предварительно разметить. Иначе поиск возможен только вручную, а это процедура долгая и трудоемкая.

Это только один, самый простой пример, который показывает, для чего может быть нужна разметка. На самом деле хорошо размеченный текст для специалиста оказывается совершенно бесценным. Ведь в своей исследовательской работе лингвисты зависят прежде всего от количества и качества собранного материала. На то, что у исследователей предшествующих поколений уходили недели или месяцы напряженного труда, с помощью корпуса можно сделать за минуты [7, С. 16].

Разумеется, для этого надо сначала необходим корпус. Итак, корпус - это электронное собрание текстов, размеченное таким образом, чтобы в нем можно было быстро найти слова и конструкции с заданными грамматическими и другими интересными лингвисту свойствами.

1. **История зарождения Национальных лингвистических корпусов**

**2.1 Международные лингвистические корпуса**

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.

Впервые определение «Национальный лингвистический корпус» появилось в названии Британского национального корпуса (British National Corpus, BNC), созданного в 1990-е годы в Великобритании специалистами лексикографами; это не самый первый электронный корпус, созданный в мире, но один из лучших, крупнейших и наиболее известных. Для британцев слово «национальный» означало в первую очередь «характеризующий британский национальный вариант английского языка» (в отличие от американского, австралийского и т. п.), но поскольку этот корпус очень быстро стал практически эталоном корпуса вообще, то значение слова «национальный» незаметно изменилось. Национальным корпусом стали называть просто самый большой и представительный корпус, характеризующий язык данной страны в целом. Таким, например, оказался Чешский национальный корпус (Český narodní korpus), содержащий практически исчерпывающую коллекцию текстов на чешском литературном языке разных периодов.

**2.2 Национальный Корпус русского языка**

Огромное значение для развития прикладной лингвистики в России играет открытие Национального Корпуса русского языка. Национальный корпус русского языка — крупнейшее электронное собрание текстов, включающее более 500 млн словоупотреблений. Это один из основных источников, к которому обращаются лингвисты, исследующие русский язык. Проект был запущен 9 лет назад, в апреле 2004 года. Корпус пополняется и поддерживается силами многих организаций, среди которых Институт русского языка имени В. В. Виноградова РАН, Институт проблем передачи информации РАН, МГУ и многие другие. С самого начала Яндекс разрабатывал адаптированную для НКРЯ версию поисковой машины для поиска текстов с разного рода разметкой [7, С. 15].

Морфологическая разметка текстов корпуса осуществляется автоматически с помощью разработанной в Яндексе программы MyStem[[3]](#footnote-3). Однако у ряда текстов вручную «снята омонимия», то есть для каждого слова указаны его правильная словарная форма и грамматические характеристики. Для многих компьютерных лингвистов эта часть корпуса представляет особый интерес, так как её можно использовать при разработке морфологических анализаторов и для разных вычислительных экспериментов.

Понятно, что Национальный корпус должен быть прежде всего большим: его объем измеряется сотнями миллионов словоупотреблений Иначе говоря, он должен содержать все типы текстов, представленные в данном языке в данный исторический период, и при этом содержать их в правильной пропорции.

Именно поэтому Национальный корпус русского языка не ограничивается, например, только произведениями художественной литературы, сколь бы важны они ни были для изучения русского языка. Он содержит и газетные и журнальные статьи разной тематики (от общественно-политических до, например, спортивных), и специальные тексты (научные, научно-популярные и учебные по разным отраслям знания), и рекламу, и частную переписку и дневники. Словом, в Корпус попадают образцы практически любого существующего в русском языке письменного дискурса. Более того, составители Корпуса хорошо понимали, что для полного и адекватного представления о том, что происходит в современном русском языке (или, если угодно, с современным русским языком) необходимо еще в большей степени расширить рамки Корпуса и включить в него, наряду с письменными текстами, также и записи устной речи.

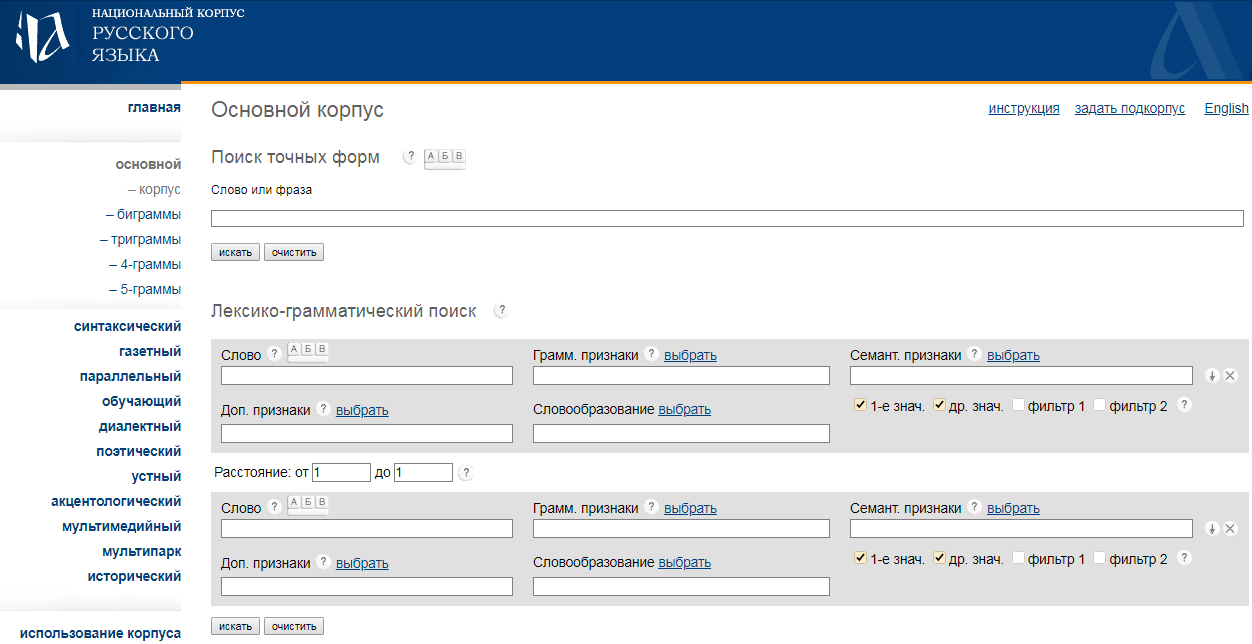


Рис. 1 Осуществление поиска необходимой лексемы на главной странице Национального Корпуса русского языка

Источником устных текстов для нас являются, например, записи радиопередач, интервью, круглых столов и т. п., а также тексты, полученные в более неформальной обстановке - например, беседы социологов или маркетологов с целевыми группами респондентов на определенную тему (качества какого-либо товара, социальные проблемы, личные обстоятельства участников беседы, и т. п.): в таких ситуациях люди, как правило, говорят достаточно свободно и естественно. Но, не ограничиваясь этим, мы ведем также записи бытовых разговоров (диалогов и монологов) на улице, в учреждениях, в домашней обстановке. В таких записях участвуют группы добровольцев из разных регионов России, так как в Корпусе, конечно, должны быть образцы речи не только жителей крупных столичных городов. Сразу следует добавить, что отдельной проблемой является включение в Корпус образцов не общерусского языка (пусть и в его разговорном варианте), а настоящей диалектной речи [3, С. 25].

|  |  |  |
| --- | --- | --- |
| Название | Состав | Доступ |
| Национальный корпус русского языка, http://www.ruscorpora.ru | Более 600 млн. слов | Свободно доступный |
| Хельсиновский аннотированный корпус русских текстов, www.slav.helsinki.fi/hanco | Содержит тексты журнала «Итоги» | Свободно доступный |
| Regensburg Russian Diachronic Corpus, http://rhssl1.uni-regensburg.de/SlavKo/korpus/rrudi-new | Содержит тексты на церковнославянском (древнерусском) | Свободный доступ для выполнения исследовательских задач предоставляется после подписания лицензионного соглашения |

Таб. 1 Наиболее полные национальные корпусы русского языка

1. **Корпусная лингвистика как способ обучения языку**
   1. **Основные проблемы в развитии корпусной лингвистики**

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[[4]](#footnote-4)) со статистической оценкой их значимости.

Также важно, что корпус состоит из конечного числа текстов, но он призван адекватно отражать лексикограмматические феномены, типичные для всего объёма 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[[5]](#footnote-5)), 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.

**3.2 Основные перспективы развития корпусной лингвистики**

Достаточно большой опыт работы с корпусами текста показывает, что с их помощью можно по-новому решить целый ряд лингвистических задач:

1. В лексикографии и лексикология (составление различных словарей, определение значений многоязычных слов, выявление ассоциативных связей слов в тексте, выявление терминов и терминологических логических словосочетаний и т.д.).

2. В грамматике (определение частоты употребления грамматических морфем в текстах различного типа, выявление наиболее употребляемых типов словосочетаний и предложений, определение частоты употребления классов слов и т.д.).

3. В лингвистике текста (дифференциация типов текста, создание конкордансов, выявление связи между предложениями в абзацах и между абзацами и т.д.).

4. При автоматическом переводе текстов (поиск контекстов слов, имеющих несколько переводных эквивалентов, поиск переводных эквивалентов терминологических и фразеологических словосочетаний параллельных текстах и т.д.).

5. В учебных целях (выбор цитат, отдельных произведений, примеров, используемых в процессе создания учебников и учебных пособий и т.д.).

Несмотря на столь большой интерес к проблемам корпусной лингвистики, до сих пор в решении ее задач остается много проблем. Так, не выработаны единые подходы к классификации корпуса текстов. Такие классификации приводится по различным основаниям зависимости от типа тексов, их языка, способом их организации на магнитных носителях, степени предварительной подготовки текстов [3, С. 25].

Преподаватели при использовании параллельных корпусов могут создавать упражнения, заключающие в себе примеры из аутентичного материала. В задачах обучения иностранному языку такие корпусы позволяют подобрать возможные эквиваленты изучаемой лексики, проследить ее значения и функции в тех или иных контекстах. Н.Р. Добрушина отмечает тот факт, что «в типичном учебнике русского языка упражнения на 60% состоят из примеров XIX века; 30 % приходится на литературу XX века, а еще 10% составляют примеры, сконструированные автором.

Таким образом, преподаватель получает множество примеров как грамматической, так и лексической формы слова. Студент в свою очередь получает естественные примеры демонстрации тех или иных грамматических или лексических явлений, может самостоятельно проводить лингвистические исследования.

В обучении переводу параллельные тексты являются незаменимыми помощниками, и фактически ни одно практическое пособие по практике перевода не обходится без них. Параллельные корпусы текстов-образцов особенно полезны в том случае, когда переводчик работает со строго нормированными (конвенциональными) текстами. Жанровостилистическое и стилистическое оформление таких текстов практически не допускает варьирования, отступления от определенных социокультурных норм. Это тексты деловой переписки, тексты-рецепты, тексты-прогнозы погоды, тексты-контракты и т.д.

Из изложенного выше следует, что параллельные корпусы текстов могут служить наглядным примером способов переводов, образцами применения приемов перевода в дальнейшей практике. Они являются своего рода наглядными пособиями для перевода. Можно сделать вывод, что в задачах обучения переводу параллельные корпусы текстов могут рассматриваться как реферативная информация и предоставлять образцы профессионального перевода при изучении приемов и способов перевода. Практически все современные лингвистические исследования и работы по составлению словарей и грамматик так или иначе ориентированы на использование представительных корпусов текстов.



Рис. 2 Использование параллельных текстов для сопоставления русского и английского значений слова «школа» («a school»)

**ЗАКЛЮЧЕНИЕ**

Корпусная лингвистика – достаточно молодое и быстро развивающееся направление в лингвистике. Внедрение корпусных методов радикально изменило общий научный ландшафт в лингвистике. Теперь ограничений на объем анализируемого материала и скорость поиска информации в нем по существу нет, а это означает, что в распоряжении исследователя оказываются колоссальные массивы текстов самого разного типа.

Назначение корпуса текстов – показать функционирование лингвистических единиц на большом материале и в их естественном окружении – контекстной среде.

В курсовой работе мы рассмотрены основные перспективы и проблемы развития корпусной лингвистики, на основе чего были выявлены основные способы применения корпусов текста для улучшения подхода к обучению языку и проведению лингвистических исследований.

При изучении синтаксиса с помощью Корпуса можно наглядно продемонстрировать грамматическую форму и принципы построения схем словосочетаний, отработать навыки разграничения различных типов предложений. Также корпуса языка предоставляют широкие возможности для самостоятельной исследовательской деятельности лингвистов. На этапе сбора материала он позволяет набрать достаточное количество примеров для изучения того или иного явления языка, не затрачивая для этого много времени и усилий. «Возможность работать с Корпусом, включающим в себя тексты разных стилей, жанров, сфер употребления, письменных и устных форм (в записи) позволяет расширить представление о языковой норме через понимание механизма ее становления в речевой практике. Это может стать важным фактором для формирования языковой личности».

Таким образом, на основе анализа множества текстов можно сделать вывод об интересующем исследователя языковом явлении, на чем и базируется корпусная лингвистика.

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**:

1. Баранов, А.Н. Введение в прикладную лингвистику / А.Н. Баранов. - М., 2001. - 137 с.
2. Волоснова, Ю.А. Корпусная лингвистика: Проблемы и перспективы / Ю.А. Волоснова // Лесной вестник. – 2006 – №5 – С. 42 – 49.
3. Зубов, А.В. Возможности и перспективы / А.В. Зубов // Международная научная конференция. Сборник работ. Минск, 2006. – Вып. 1 – С. 21-27.
4. Козлова, Н.В. Лингвистические корпуса: определений основных понятий и типологий / Н.В. Козлова // Вестник НГУ. Серия: лингвистика и межкультурная коммуникация. НГУ. – Новосибирск, 2013. – Вып. 11. – С. 10-20.
5. Ляшевская, О.Н., Плунгян, В.А. О морфологическом стандарте Национального корпуса русского языка / О.Н. Ляшевская, В.А. Плунгян // Национальный корпус русского языка: 2003 - 2005. Результаты и перспективы. - М., 2005. - С. 187 - 206.
6. Нагель, О.В. Корпусная лингвистика и ее использование в компьютеризированном языковом обучении / О.В. Нагель // Язык и культура. – 2008 – №8(62) – 8 с.
7. Плунгян, В.А. Зачем мы делаем Национальный корпус русского языка? / В.А. Плунгян // Отечественные записки. – 2005. – № 2. – С. 12 - 32.
8. Ушакова, Н.В. Применение корпусов текстов в обучении иностранному языку / Н.В. Ушакова, Е.П. Курчева // Вестник Белорусского Государственного Университета. – 2012. – №18(21) – 45 с.
9. Корпусная лингвистика / СПбГУ, ин.-т лингвистических иссследований РАН // Режим доступа: <http://corpora.iling.spb.ru/theory.htm>, свободный.
10. Национальный корпус русского языка – (<http://www.ruscorpora.ru>).

**ГЛОССАРИЙ**

1. **Аннотирование –** процесс аналитико-синтетической переработки информации, целью которого является получение обобщенной характеристики документа, раскрывающей его логическую структуру и наиболее существенные признаки.
2. **Коллокация -** словосочетание, имеющее признаки синтаксически и семантически целостной единицы, в котором выбор одного из компонентов осуществляется по смыслу, а выбор второго зависит от выбора первого (например, ставить условия — выбор глагола ставить определяется традицией и зависит от существительного условия, при слове предложение будет другой глагол — вносить).
3. **Корпусная лингвистика —** раздел языкознания, занимающийся разработкой, созданием и использованием текстовых корпусов. Термин введён в употребление в 1960-е годы в связи с развитием практики создания корпусов, которому начиная с 1980-х способствовало развитие вычислительной техники.
4. **Корпусы языка -** совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой.
5. **Национальный корпус русского языка (НКРЯ) —** доступный для поиска электронный онлайн-корпус русских текстов. Открыт 29 апреля 2004 года. Также доступен для поиска исторический корпус церковнославянских, древнерусских (XI—XIV века) и среднерусских (XV — начало XVIII века) текстов.

1. Разметка - процесс обработки информации, который предназначается для резюмирования статей или книг. [↑](#footnote-ref-1)
2. доступный для поиска электронный онлайн-корпус русских текстов. Открыт 29 апреля 2004 года. Также доступен для поиска исторический корпус церковнославянских, древнерусских (XI—XIV века) и среднерусских [↑](#footnote-ref-2)
3. Программа MyStem производит морфологический анализ текста на русском языке. Она умеет строить гипотетические разборы для слов, не входящих в словарь. Первую версию программы написали Илья Сегалович и Виталий Титов. [↑](#footnote-ref-3)
4. словосочетание, имеющее признаки синтаксически и семантически целостной единицы, в котором выбор одного из компонентов осуществляется по смыслу, а выбор второго зависит от выбора первого [↑](#footnote-ref-4)
5. эмпирическая закономерность распределения частоты слов естественного языка: если все слова языка (или просто достаточно длинного текста) упорядочить по убыванию частоты их использования, то частота n-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n [↑](#footnote-ref-5)