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: «Параметры поиска в онлайновых текстовых корпусах»

по дисциплине «Информационные технологии в лингвистике»

|  |  |
| --- | --- |
|  | Выполнила: студентка 1 курса группы 12 дневной формы обученияКотлярова Александра Евгеньевна |
|  |  |
|  | Проверила:к.ф.н., доцент, кафедры английского языкаМасленникова Евгения Михайловна |

ТВЕРЬ

2021

Содержание

|  |  |
| --- | --- |
| 1.Поиск в корпусе | 3 |
| 2. Поиск точных форм. Что такое лексема и словоформа? | 3 |
| 3.Лексико-грамматический поиск | 4 |
| 4.Поле «Слово» | 4 |
| 5.Поле «Грамматические признаки» | 4 |
| 6.Поле «Семантические признаки» | 5 |
| 7.Как расширить форму запроса | 6 |
| 8.Расстояние между словами | 6 |
| 9.Снятая и неснятая грамматическая омонимия | 7 |
| 10.Основные параметры текста | 7 |
| 11.Жанр и тип текста | 8 |

# Поиск в корпусе

Возможен поиск форм и расширенный (лексико-грамматический) поиск примеров по корпусу. На основании созданного пользователем поискового запроса выдается список контекстов, то есть фрагментов текстов, содержащихся в Корпусе и содержащих искомые слова и грамматические формы.

Часть текстов Корпуса представляет собой тексты с неснятой грамматической омонимией, то есть в них не разрешена неоднозначность форм с одинаковым написанием (например, *печь* — инфинитив глагола и существительное, *печи* — форма родительного, дательного и предложного падежа ед. ч., именительного и винительного мн. ч.). Все такие формы получают весь возможный набор разборов, таким образом, поисковая выдача по этим текстам содержит некоторое количество «шума». Возможно задать ограничение на поиск по корпусу только со снятой или только с неснятой грамматической омонимией. Кроме того, пользователь может быть заинтересован в ограничении поисковой базы по тематике, жанру, году создания текстов и т. п. Возможность задания пользовательского подкорпуса доступна на странице «[Мой корпус](https://ruscorpora.ru/mycorpora-main.html)».

### Поиск точных форм. Что такое лексема и словоформа?

Поиск точных форм доступен через соответствующее окно на странице [«Поиск в корпусе»](https://ruscorpora.ru/new/search-main.html). Запрос, состоящий из одного слова (например, *отечества*), находит все контексты, содержащие точную форму данного слова (но не *отечеством, отечеств* и т. д. ). Каждая из этих форм (образующихся при помощи грамматического словоизменения) называется словоформой, а всё их множество, объединенное общим значением, называется лексемой; лексему принято обозначать по «словарной» форме (например, говорят, что *отечества* — это словоформа родительного падежа единственного числа лексемы *отечество*; *сладок* — это словоформа краткой формы мужского рода единственного числа лексемы *сладкий*).

Запрос в поиске точных форм, содержащий несколько слов (*И дым отечества нам сладок и приятен*), находит все контексты, содержащие все словоформы запроса в пределах одного предложения — следующие не обязательно в том же порядке, что и в запросе (например, найдется контекст *Отечества и дым нам сладок и приятен*).

Поиск контекстов, содержащих определенные знаки препинания или цифры, пока не поддерживается.

### Лексико-грамматический поиск

В лексико-грамматическом поиске можно задавать последовательность лексем и/или словоформ, обладающих определенными грамматическими и/или семантическими характеристиками. Каждая строка формы расширенного поиска соответствует лексеме или словоформе. В одной строке можно указать только конкретную лексему (словоформу), или только грамматические признаки, или только семантические признаки, или же любые попарные сочетания этих значений, или, наконец, и то, и другое, и третье.

### Поле «Слово»

В поле «Слово» указывается искомая лексема (обязательно в словарной форме: **отечество**, а не **отечества**) или словоформа (в кавычках: **"отечества"**). В поле «Слово» действуют символ **\***, обозначающий любую последовательность символов в начале или конце лексемы или словоформы, а также логические операторы «ИЛИ» и «НЕ»: для поиска слов любовь или ненависть нужно ввести **любовь|ненависть**, для поиска любого слова, кроме без, нужно ввести **-без**. Эти символы можно комбинировать при помощи логического «И», задаваемого знаком **&** или одиночным пробелом. Так, запрос **\*вед -швед** найдет все лексемы, заканчивающиеся на вед (языковед, искусствовед...), но не швед; запрос **мой&"его"** найдет все вхождения лексемы мой и словоформы его. Чтобы искать все лексемы, кроме нескольких определенных, следует поставить в начале запроса звёздочку (например, **\* -папа -мама**).

### Поле «Грамматические признаки»

В поле «Грамматические признаки» указываются искомые морфологические характеристики лексемы и/или словоформы. Эти признаки можно указать при помощи служебного окна, открывающегося по ссылке «Выбрать», или же набрать на клавиатуре.

В поле «Грамматические признаки» действует логическое «ИЛИ», задаваемое при помощи знака **|**, а также логическое «И», задаваемое знаком **&** или одиночным пробелом (разрешается использование любого числа незначащих пробелов вокруг знаков **|** и **&**). При записи сложных выражений используются круглые скобки. Например, **S & (nom|acc)** означает существительное в именительном или винительном падеже. Инвентарь обозначений грамматических признаков и необходимые комментарии см. в разделе «[Морфология](https://ruscorpora.ru/corpora-morph.html)». Но проще выбрать грамматические признаки (или задать несколько альтернативных лексем) в служебном окне.

Служебное окно, открывающееся по ссылке «Выбрать», содержит перечень морфологических характеристик, разбитых по грамматическим категориям: часть речи, падеж, род, залог, число и т. п. Отметьте «галочками» те характеристики словоформ, которые вы хотите найти. Внутри каждой категории признаки в нормальном случае являются взаимоисключающими (например, глагол не может стоять в первом и третьем лице одновременно), поэтому «галочки», отмечающие грамматические признаки одной категории, трактуются как логическое «ИЛИ» (невзаимоисключающие пометы одной категории при одном слове также найдутся), а отмечающие признаки разных категорий — как логическое «И». Чтобы инвертировать отмеченные признаки одной категории (изменить выбранные на невыбранные, и наоборот), щелкните левой кнопкой мыши на названии этой категории. Это способ выразить логическое «НЕ»: например, чтобы искать любой падеж, кроме творительного, выберите позицию «творительный падеж», а затем инвертируйте список значений категории падежа.

### Поле «Семантические признаки»

В поле «Семантические признаки» указываются искомые семантические и словообразовательные характеристики лексемы. Эти признаки можно задать при помощи служебного окна, открывающегося по ссылке «Выбрать», или же набрать на клавиатуре.

Служебное окно, открывающееся по ссылке «Выбрать», содержит перечень семантических характеристик, разбитых по нескольким категориям: разряд, таксономия, мереология, оценка, словообразование и т. п. При наведении курсора на любой пункт этого перечня во всплывающей подсказке появятся примеры слов, имеющих данные характеристики. Отметьте «галочками» (checkbox) те характеристики словоформ, которые вы хотите найти. «Галочки», отмечающие семантические признаки внутри одной категории, трактуются как логическое «ИЛИ» (например, в категории «Таксономия» можно задать поиск глаголов «движения» и «физического воздействия»). Галочки, отмечающие признаки разных категорий, трактуются как логическое «И» (например, чтобы найти приставочные глаголы движения, нужно отметить «галочками» признак «движение» в категории «Таксономия» и признак «приставочные глаголы» в категории «Словообразование».

Как правило, семантические характеристики имеют иерархическую структуру, например:

бытийная сфера

существование
начало существования
прекращение существования

Отметив «галочкой» основной класс, вы найдете слова всех вложенных подклассов. Чтобы исключить из поиска какой-либо из вложенных подклассов, уберите «галочку» при названии этого подкласса. Обратите внимание, что словник основного класса чаще всего не исчерпывается словами вложенных подклассов (так, класс глаголов «физического воздействия» включает в себя не только глаголы создания и уничтожения, но и, например, глаголы обработки типа *вымыть*). Чтобы найти слова, входящие только в основной класс, уберите «галочки» при всех вложенных подклассах.

Вы можете задать поиск по определенной части речи, а для имен существительных - по одному из [семантических разрядов](https://ruscorpora.ru/new/corpora-sem.html#razriad). В верхней части формы располагаются закладки "Имена предметные", "Прилагательные", "Глаголы" и т. д. Щелкнув по закладке, вы перейдете к форме с семантическими признаками, относящимися к нужной Вам части речи.

В настоящее время в корпусе не разрешена неоднозначность семантических характеристик полисемичных слов (так, на запрос «человек+отрицательная характеристика» найдутся все вхождения словоформы *собака*, в том числе те, где речь идет о животном; на запрос «часть мебели» найдутся все словоформы *ножка*). В ряде случаев для повышения качества поиска можно дополнительно задать грамматические характеристики (чтобы на запрос «человек+отрицательная характеристика» не нашлись слова *мешок*, *дуб* и т. п. в прямом значении, задайте характеристику «одушевленное»)

### Как расширить форму запроса

По умолчанию в форме имеются две строки, и она рассчитана на поиск одной словоформы или сочетания двух словоформ. При необходимости работы с тремя или более словоформами следует нажать кнопку "+" внизу поискового запроса; последовательные нажатия добавляют третью, четвертую и т. д. (до 10) поисковую строку.

**Расстояние между словами**

Между поисковыми строками имеется окно «Расстояние: от до». Оно служит для того, чтобы точнее определить структуру сочетания слов, которое вы ищете. Между соседними словами расстояние считается равным одному слову; расстояние, равное нулю, означает совпадение словоформ. Так, расстояние от 1 до 3 между словом (лексемой) критика и грамматической характеристикой «существительное в родительном падеже» соответствует контекстам критика взглядов, критика чистого разума и критика буржуазных философских систем. Если задать минимальное расстояние, равное нулю, то найдутся также и все контексты, просто содержащие словоформы критики или критик, в которых признаки обеих строк сочетаются.

Во всех найденных контекстах порядок словоформ будет в точности соответствовать последовательности двух соответствующих поисковых строк. Контексты с обратным порядком элементов (для нашего примера это, например, объект замечаний и придирчивой критики) найдены не будут.Пример расширенного поискового запроса

Предположим, мы хотим найти все контексты с глаголами, начинающимися с *под-* и имеющими косвенное дополнение с предлогом *под*, который управляет винительным падежом: например, *подложил тетрадь под листок*.

Запрос формируется так:

Слово 1: лексема под\*, грамматическое значение глагол (V).
Расстояние между словами 1 и 2: от 1 до 3.

Слово 2: лексема под, грамматическое значение предлог (PR) — ведь есть и существительное под (низ печи).
Расстояние между словами 2 и 3: от 1 до 2.

# Мой корпус

В разделе «Мой корпус» пользователь может задать подмножество корпуса, по которому в дальнейшем будет вестись поиск. Подробнее о параметрах текста см. в разделе «Параметры текста».

### Снятая и неснятая грамматическая омонимия

Часть текстов Корпуса представляют собой тексты с неснятой грамматической омонимией, то есть в них не разрешена неоднозначность форм с одинаковым написанием (например, *печь* — инфинитив глагола и существительное, *печи* — форма родительного, дательного и предложного падежа ед. ч., именительного и винительного мн. ч.). Все такие формы получают весь возможный набор разборов, таким образом, поисковая выдача по этим текстам содержит некоторое количество «шума». Возможно задать ограничение на поиск по корпусу только со снятой или только с неснятой грамматической омонимией. В поисковой выдаче рядом с названием текста всегда указывается, снята в нем грамматическая омонимия или нет.

### Основные параметры текста

В этом блоке можно задать слово или слова, входящие в имя автора или название произведения, не обязательно в словарной форме или в форме, совпадающей с входящей в название: так, произведение с названием *Критика чистого разума* найдется по каждому из запросов *критика*; *критик*; *разума*; *разум*; *чистый*; *чистый разум* и т. п. По ссылке «выбрать» можно указать пол и год рождения автора (в том числе диапазон); пользователь может задать также дату создания произведения и/или объем искомого текста в словах (в том числе диапазон).

### Жанр и тип текста

В этом блоке можно задать поиск по художественным или нехудожественным текстам; в пределах каждого класса можно задать или (лучше) выбрать из списка жанр и тип текста, место и время описываемых событий (для художественных текстов); сферу функционирования, тип и тематику текста (для нехудожественных текстов). В форме запроса следует отметить искомые значения «галочкой» (checkbox), для сохранения значений следует нажать «ОК». Щелчок мыши на заголовке формы инвертирует выбранные значения (это удобно, если вы, к примеру, ищете все типы текстов, кроме романов; выберите «роман» и инвертируйте выбранные значения).