

Как пользоваться Башкирским поэтическим корпусом

(сост. Б. В. Орехов)

Вопросы, на которые отвечает эта инструкция:

1. Что такое корпус?
2. Что такое поэтический корпус?
3. Как найти Башкирский поэтический корпус в Интернете?
4. Что искать в Башкирском поэтическом корпусе?
5. Как искать в корпусе?
 - 5.1. поиск точной формы
 - 5.2. лексический поиск
 - 5.2.1. по башкирской лемме
 - 5.2.2. по русскому эквиваленту
 - 5.3. грамматический поиск
 - 5.3.1. поиск по грамматической категории
 - 5.3.2. поиск по аффиксу
 - 5.4. поиск с учётом стиховедческих параметров
 - 5.4.1. метр
 - 5.4.2. рифма
 - 5.5. поиск с учётом авторства и датировки
6. Как читать результаты поиска?
7. Как сохранять результаты поиска?
8. Как интерпретировать результаты поиска?
9. Какие тексты входят в корпус?
10. Как ссылаться на корпус в научной работе?
11. Как при создании корпуса решена проблема авторских прав?
12. Что делать, если в корпусе нашлась ошибка?
13. Кто делал корпус и кому он принадлежит?
14. Какие технологии использованы при создании корпуса?

1. Что такое корпус?

Корпус — это электронная поисковая система. Поиск в ней осуществляется по коллекции текстов, а результатом поиска обычно являются контексты употребления слов или других языковых явлений, а также данные о частотности этих явлений в коллекции.

Обычно создатели корпусов стремятся к тому, чтобы текстовые коллекции были *представительными*, то есть достаточно большими, чтобы отражать устройство языка. В большом хорошем корпусе встретится всё, что есть в языке (слова, их значения, грамматические конструкции и т. д.).

Создатели корпусов стремятся также к тому, чтобы их корпуса были *сбалансированными*, то есть отражали бы реальное соотношение жанров в языке. Таким образом, если в хорошо сбалансированном корпусе какое-то слово (или грамматическое явление) встречается чаще, чем другое слово, то это значит, что оно встречается чаще и в языке.

Тексты, помещённые в корпус, специальным образом обрабатываются, чтобы из них можно было извлекать лингвистически значимую информацию. Такая обработка называется *разметкой*. Она позволяет искать не просто слова, но и определённые грамматические формы и другие языковые явления.

Корпуса создаются, чтобы избавить лингвиста от рутинной работы по сбору материала. Раньше при изучении какого-то явления специалист вынужден был потратить много времени на то, чтобы собрать языковые примеры с интересующим его явлением. Такие примеры выписывались на карточки. На эту работу уходили годы. Сейчас всё нужное (контексты употребления и данные о частоте вхождений) лингвист с помощью корпуса получает за секунду.

Корпус не подменяет собой лингвиста, он просто облегчает его работу, высвобождая время и силы на исследовательскую деятельность. Корпус не делает выводов, он только предоставляет языковой материал для обдумывания.

Данные, извлекаемые из корпуса, могут быть полезны не только лингвистам, но и специалистам в других гуманитарных областях: филологам, литературоведам, культурологам, историкам, социологам, фольклористам, антропологам.

Важным свойством применения корпуса как инструмента исследования становится количественный фактор. На сбалансированные корпуса можно опираться при подсчётах тех или иных языковых явлений (например, при создании частотных словарей). Благодаря корпусам филология приближается по своим методам к точным наукам.

Корпуса обычно представлены в Интернете и пользоваться ими можно в любой точке мира, где есть доступ к Сети.

Корпуса создаются для разных языков. Одним из лучших лингвистических корпусов является Национальный корпус русского языка (НКРЯ): <http://ruscorpora.ru/>. Большое значение имеют Восточноармянский

национальный корпус: <http://www.eanc.net/>, Британский национальный корпус: <http://www.natcorp.ox.ac.uk/>, Национальный корпус чешского языка: <http://ucnk.ff.cuni.cz/> Ведётся систематическая работа по созданию корпусов языков народов РФ и других народов мира: <http://web-corpora.net/>

2. Что такое поэтический корпус?

Помимо больших корпусов, отражающих реальность языка, существуют ещё и специализированные корпуса, необходимые для изучения какого-то аспекта языка и культуры. Создаются диалектные корпуса (отражающие систему диалектной речи), газетные корпуса (отражающие динамичные изменения в языке последнего времени, фиксируемые газетами), параллельные корпуса (отражающие межъязыковые соответствия).

Одной из разновидностей корпуса является *поэтический корпус*. Такого рода система позволяет не только искать нужные слова и другие языковые явления, но и делать это с учётом специфики стихотворного текста. Скажем, в поэтическом корпусе можно найти слово в строке, написанной определённым метром, или слово в позиции рифмы.

Долгое время существовал только один поэтический корпус: поэтический подкорпус в составе НКРЯ: <http://ruscorpora.ru/search-poetic.html>. [В литературе есть также упоминание о том, что в 2004 году существовал корпус китайских поэтических текстов эпохи династии Тан, но сейчас этот корпус недоступен: *San Duanmu A Corpus Study of Chinese Regulated Verse: Phrasal Stress and the Analysis of Variability // Phonology. Vol. 21, No. 1 (2004). P. 47.*

На данный момент вторым в мире действующим поэтическим корпусом (и первым в мире корпусом для башкирского языка) является Башкирский поэтический корпус (**БПК**). Он создан на базе обширной коллекции башкирских стихотворных текстов. На момент запуска в октябре 2013 года в состав корпуса входило около 17 тыс. стихотворений 101 поэта, что в совокупности составляло более 1,7 млн. словоупотреблений и около 450 тыс. стихотворных строк.

Все тексты имеют морфологическую и стиховедческую разметку, то есть позволяют задавать поиск по разнообразным лингвистическим и стиховедческим параметрам: части речи, именные и глагольные словоизменительные категории, стихотворные метры, рифменная позиция, а также по русскому семантическому эквиваленту слова (подробнее см. § 5 «Как искать в корпусе?»).

Пользоваться корпусом можно свободно и бесплатно.

3. Как найти Башкирский поэтический корпус в Интернете?

Чтобы воспользоваться БПК, нужно иметь компьютер, подключённый к

Интернету. На таком компьютере необходимо запустить *браузер* — программу для навигации по Сети Интернет. Распространёнными браузерами являются Google Chrome, Windows Internet Explorer, Mozilla Firefox, Opera.

В браузерах имеется *адресная строка*, то есть такое поле, куда вводится интернет-адрес (URL) некоторой страницы. Обычно это поле располагается вверху окна:

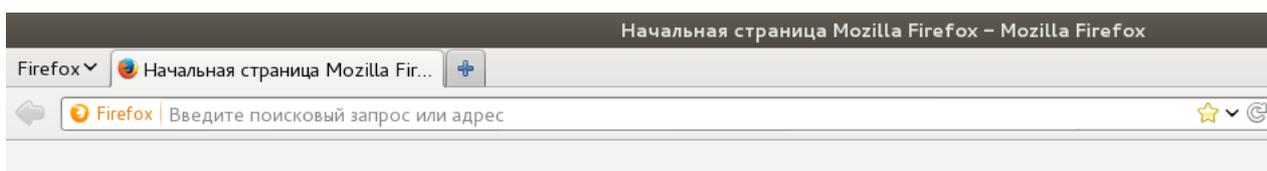


Рис. 1.

В этом поле нужно ввести адрес: <http://web-corpora.net/bashcorpus/search/> и нажать клавишу Enter.

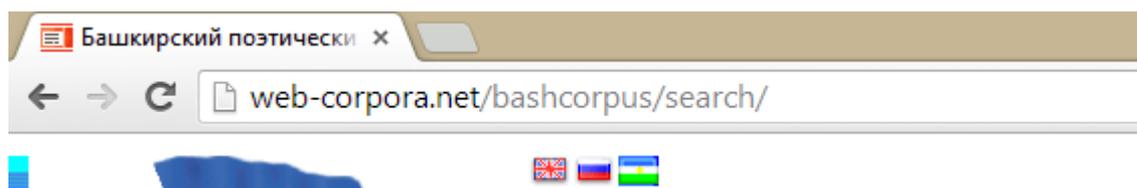


Рис. 2.

Воспользоваться корпусом можно и при помощи имеющего выход в Интернет мобильного устройства (смартфона или планшетного компьютера). Принцип работы тот же: нужно ввести адрес <http://web-corpora.net/bashcorpus/search/> в адресную строку браузера.



Рис. 3

Если вы ввели правильный адрес и соединение с Интернетом активно, вы попадёте на страницу, на которой можно ввести поисковый запрос. Здесь вы найдёте различные поля и формы, с помощью которых можно обратиться к корпусу.

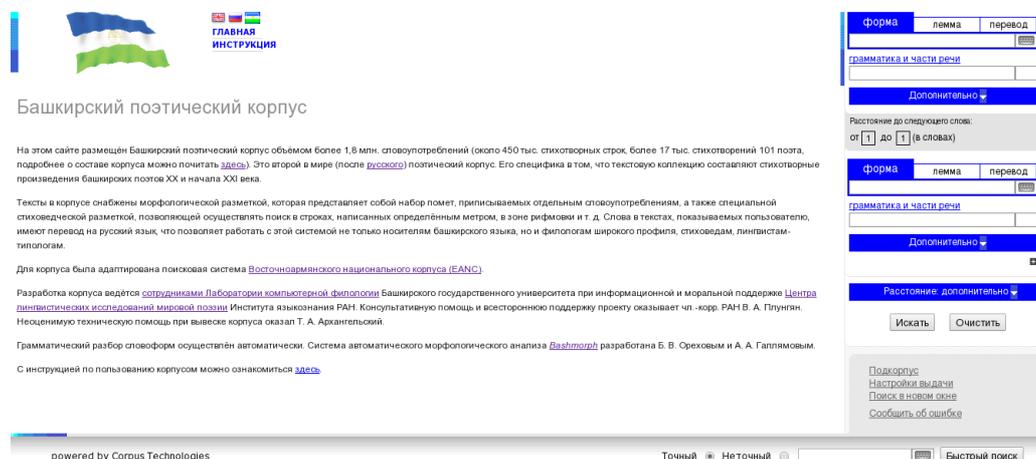


Рис. 4.

4. Что искать в Башкирском поэтическом корпусе?

БКК — это мощный научный инструмент, из него можно извлекать различные данные, пригодные для исследовательских целей.

Возможные на данный момент поисковые запросы к корпусу делятся на два типа:

- а. лексический поиск;
- б. грамматический поиск.

Лексический поиск — это поиск слов. Если мы хотим выяснить, как функционирует какое-нибудь слово (например, *йондоз*) в башкирской поэзии, мы прибегаем именно к этому типу поиска. При этом можно искать как точную форму этого слова (скажем, *йондоззар*), так и *лемму*, то есть слово во всех его словоформах (тогда мы найдём и *йондоз*, и *йондоззар*, и *йондозо*, и *йондозом*).

Лексический поиск можно задавать не только по башкирскому слову, но и по его русскому семантическому эквиваленту (например, *звезда*).

Область поиска можно ограничить. То есть мы можем искать слово со значением *звезда* не во всей башкирской поэзии, а только в стихах А. Игебаева или только в стихах, написанных восьмисложным метром, или только в стихах, написанных в 1960-х годах, или только в позиции рифмы (в *клаузуле*).

Грамматический поиск — это поиск форм по определённым грамматическим признакам. Мы прибегаем к нему, когда нам безразлично, какие именно слова (*йондоз* или *Урал*) нам попадутся, а важно, в каких грамматических формах они стоят. Скажем, мы можем искать слова только в местно-временном падеже (и найдутся формы *өйзә*, *мөйөштә*, *кылдарза*). Или

слова только в форме единственного числа глагольного лица (в результате получим формы *белам, шаярам, юндым*).

Так же, как и с лексическим поиском, мы можем искать формы с определёнными грамматическими признаками только в стихах определённых поэтов, написанных определёнными метрами и (или) в определённый временной период.

Лексический и грамматический типы поиска могут совмещаться. Это происходит в том случае, если мы ищем определённое слово, стоящее в определённой грамматической форме. Например, если нас интересует слово *йондоз* в форме множественного числа, мы зададим лексико-грамматический поиск по корпусу и получим контексты и данные о частотности употребления слова *йондоз* в форме множественного числа.

В рамках этих типов поиска можно решать любые исследовательские задачи. Например, с помощью корпуса мы можем быстро ответить на вопросы, подобные этим:

- какое слово (X или Y) чаще встречается в башкирской поэзии?
- какое слово (X или Y) чаще встречается в поэтических текстах автора Z?
- какая грамматическая форма (X или Y) преобладает в творчестве поэта Z?
- встречается ли в башкирских стихах, написанных в период с t_1 по t_2 , слово (или грамматическая форма) X?
- слово (или грамматическая форма) X чаще встречается в стихах, написанных метром Y или метром Z?
- встречаются ли (и насколько часто) слова X и Y в одной строке в башкирских поэтических текстах?

Исследовательская интерпретация этих вопросов уже лежит в сфере компетенции самого исследователя.

С помощью корпуса можно ставить и решать и более сложные исследовательские задачи. На базе поэтического корпуса в составе Национального корпуса русского языка проведены исследования, отражённые в сборнике «Корпусный анализ русского стиха» (М., 2013).

5. Как искать в корпусе?

На странице по адресу <http://web-corpora.net/bashcorpus/search/> находятся поисковые формы, позволяющие задать запрос к корпусу. Сформировав запрос, нужно нажать кнопку «Искать» на сайте или клавишу Enter на вашей клавиатуре. Для удобства работы с поисковыми формами имеется также кнопка «Очистить», сбрасывающая все внесённые изменения.

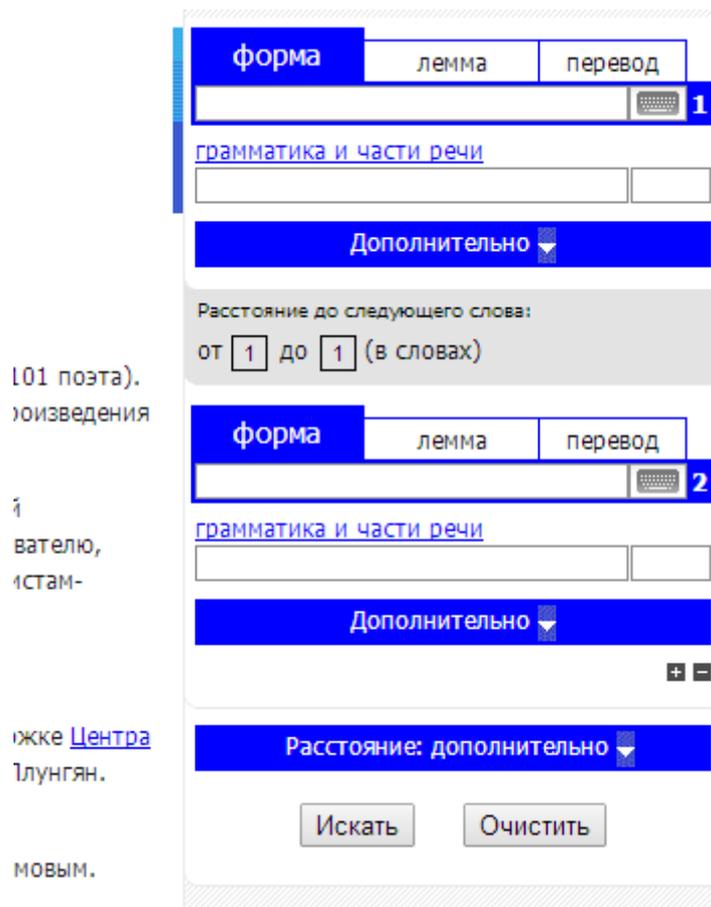


Рис. 5

Результаты поиска появятся в левой части экрана. Это контексты, в которых встречаются искомые слова или конструкции. Если контекстов много, они могут быть разделены на несколько страниц выдачи, для навигации по ним в нижней части экрана имеются ссылки на 1, 2, 3 и т. д. страницы.

Лексический поиск может выполняться как для точной формы слова, так и при поиске по лемме.

5.1. поиск точной формы

У вас есть возможность искать в корпусе точную форму слова (например, *йондоз* или *өйзэ*). В этом случае найдутся только контексты употребления именно этих форм (то есть формы *йондоззар* или *өй* в результатах значиться не будут).

Для того, чтобы выполнить такой поиск по корпусу, необходимо ввести нужную форму в поле поиска (верхняя строка в правой части экрана) при активной вкладке «форма».

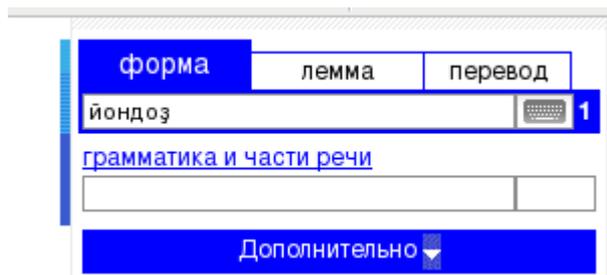


Рис. 6

Если заменить часть вводимого слова знаком «звёздочка» (*), то в поиске будут отображены все слова, в которых на месте этого знака будут любые символы. Например, при поиске формы «йондоз*» будут показаны результаты с формами йондозоң, йондозға, йондоззар, йондоззо и др.

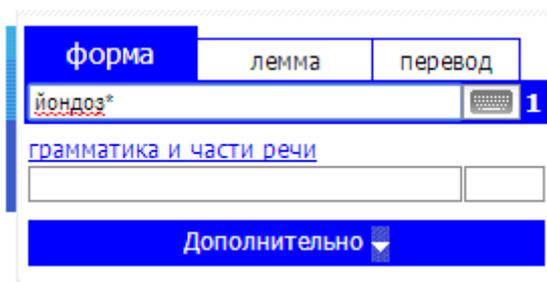


Рис. 7

5.2. лексический поиск

5.2.1. по башкирской лемме

Лексический поиск в полную силу проявляет себя только при поиске по лемме. При таком типе поиска в результатах будут показаны не только начальные формы слов, но и все их словоформы (не только *йондоз*, но и *йондоззар*, и *йондозом*). Для того, чтобы выполнить такой поиск по корпусу, необходимо ввести нужную лемму в поле поиска (верхняя строка в правой части экрана) при активной вкладке «лемма».

Важно! В поле поиска по лемме необходимо вводить слово в его словарной (исходной) форме. То есть *йондоз*, но не *йондоззар*. Для глаголов исходной формой считается чистая основа (то есть форма повелительного наклонения во втором лице ед. ч.) *тукта*, а не *туктау* и не *туктарга*.

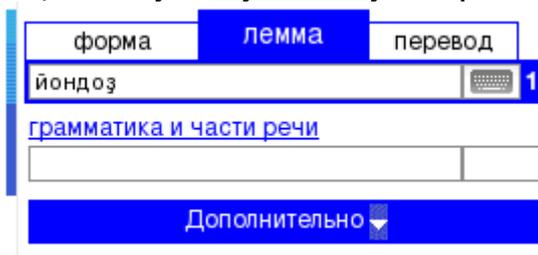
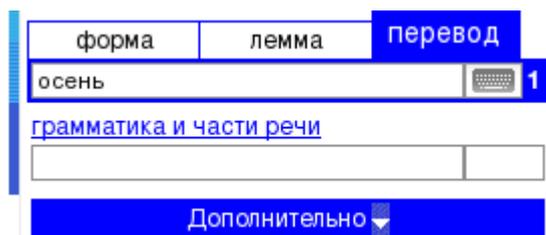


Рис. 8

5.2.2. по русскому эквиваленту

Вы можете искать башкирские слова, введя их русский семантический эквивалент (перевод) в соответствующее поле в форме поиска. Для того, чтобы выполнить такой поиск по корпусу, необходимо ввести нужное русское слово в поле поиска (верхняя строка в правой части экрана) при активной вкладке «перевод». Например, «осень».

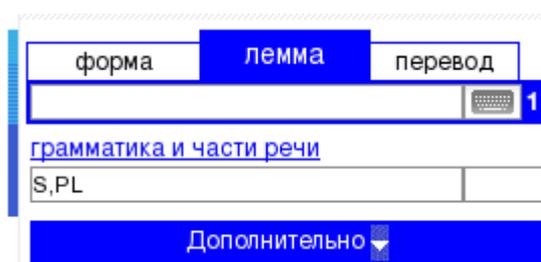


The screenshot shows a search interface with three tabs: «форма», «лемма», and «перевод». The «перевод» tab is active. Below the tabs is a search input field containing the word «осень». To the right of the input field is a small icon of a keyboard and the number «1». Below the search field is a link «грамматика и части речи». At the bottom is a blue button labeled «Дополнительно» with a downward arrow.

Рис. 9

5.3. грамматический поиск

Грамматический поиск позволяет искать формы слов, соответствующие некоторым морфологическим признакам. Для того, чтобы выполнить этот тип поиска, поле «лемма» нужно оставить пустым (если ввести в этом поле лемму, то поиск будет производиться только для форм этой леммы), а в разделе морфологических признаков указать требуемые характеристики. Например, существительное во множественном числе.



The screenshot shows the same search interface as in Figure 9, but with the «лемма» tab selected. The search input field is empty. Below the search field is a link «грамматика и части речи». Below the link is a text input field containing «S,PL». At the bottom is a blue button labeled «Дополнительно» with a downward arrow.

Рис. 10

Чтобы не запоминать все обозначения грамматических категорий, можно воспользоваться вспомогательным окном, всплывающим по щелчку на ссылке «грамматика и части речи». В открывшемся окне можно поставить галочки на всех нужных пунктах грамматических категорий и нажать «Ok», в строку поиска они будут перенесены автоматически.

5.3.1. поиск по грамматической категории

Важно не допускать противоречащих друг другу грамматических характеристик, потому что в этом случае результат поиска будет нулевым. Так,

имя прилагательное не может иметь глагольных морфологических признаков, а у глагола не может быть степени сравнения.

В некоторых случаях грамматическая система, заложенная в корпусе, расходится с общепринятой системой, положенной в описание башкирского языка. Так, в БПК не проводится различий между наречиями и прилагательными (оба эти класса слов объединены в единый класс «адъективов»). Введены отсутствующие в школьной и академической грамматике абессивный падеж и выделительные числительные.

В настоящее время в корпусе доступен поиск по следующим грамматическим признакам (в этом списке представлены только те формы, которые образуются синтетическим способом и имеют формальные показатели): части речи: имя существительное, глагол, адъектив, имя числительное, местоимение, частица, послелог, союз, междометие; число, лицо, наличие клитик неуверенности, предположительности, просьбы и вопроса, сказуемость, принадлежность, неличная принадлежность, основной падеж, родительный падеж, дательный падеж, винительный падеж, исходный падеж, местно-временной падеж, абессив, уподобление, разряды числительных: выделительные, собирательные, порядковые, приблизительные, разделительные, сравнительная степень адъективов, неперIODическое действие, отрицание, деепричастие; наклонения: повелительное, намерения, условное, желательное; залoги: возвратный, страдательный, взаимный, понудительный; времена: настоящее, прошедшее определённое, прошедшее неопределённое, будущее неопределённое, будущее определённое; инфинитив, имя действия.

Кроме того, есть возможность искать слова, образованные с помощью словообразовательных аффиксов со значением абстрактного существительного и деятеля.

Все эти признаки можно выбрать (отметить «галочками») во вспомогательном окне или ввести вручную в строке поиска по грамматическим категориям (вторая строка в верхней левой части экрана).

	разряды числительных	нефинитные формы глагола	Время
	<input type="checkbox"/> выделительное <input type="checkbox"/> собирательное <input type="checkbox"/> порядковое <input type="checkbox"/> приближительное <input type="checkbox"/> разделительное	<input type="checkbox"/> деепричастие <input type="checkbox"/> инфинитив <input type="checkbox"/> имя действия	<input type="checkbox"/> настоящее время <input type="checkbox"/> прошедшее определённое время <input type="checkbox"/> прошедшее неопределённое время <input type="checkbox"/> будущее определённое время <input type="checkbox"/> будущее неопределённое время
	прочие именные категории <input type="checkbox"/> со значением уподобления <input type="checkbox"/> сравнительная степень	залог <input type="checkbox"/> возвратный залог <input type="checkbox"/> страдательный залог <input type="checkbox"/> возвратный или страдательный залог <input type="checkbox"/> взаимный залог <input type="checkbox"/> понудительный залог	другие глагольные категории <input type="checkbox"/> отрицательная форма <input type="checkbox"/> неперIODическое действие
	Лицо <input type="checkbox"/> первое лицо единственного числа <input type="checkbox"/> второе лицо единственного числа <input type="checkbox"/> первое лицо множественного числа <input type="checkbox"/> второе лицо множественного числа <input type="checkbox"/> третье лицо множественного числа	Наклонение <input type="checkbox"/> наклонение намерения <input type="checkbox"/> повелительное наклонение <input type="checkbox"/> условное наклонение <input type="checkbox"/> желательное наклонение	клитки <input type="checkbox"/> неуверенность <input type="checkbox"/> предположительность <input type="checkbox"/> вопросительность <input type="checkbox"/> просьба

Рис. 10

5.3.2. поиск по аффиксу

В корпусе вы можете искать формы, задавая поиск по форме аффикса. Например, можно найти все формы, образованные при помощи аффикса мн. ч. *-тар*:

Найдено: **22 887** вхождений, **8 542** документов

1. **Сыңрау торна ((Драма))** Мәхмүт һибәт 1992—1993
Улар, етекләшәп, биш кайын үскән ергә йүгәрә. Был мәлдә уларзың **шатлыктары** оҫҡотарғ.
2. **Юбилей шиғыры (Башкорт дәүләт академия драма театрының утыз йыл юбилеи)** Нигмәти 1950
Бәхет һәм **уңыштар** теләнем.
3. **Куркаклықтың үз-үзенә мәзһиәһе (Мин йәшәйем һезҙең йәнегезҙә —)**
«Һез **актармы**, әллә кызылдармы»

Рис. 11

Для этого нужно ввести нужный аффикс в строку поиска по грамматическим категориям (вторая строка в левой верхней части экрана). При таком запросе мы найдём формы *шатлыктары*, *актармы*, *уңыштар* и др.

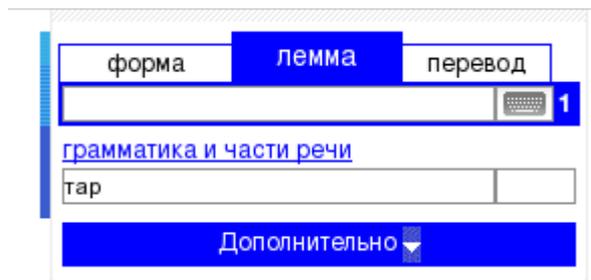


Рис. 12

5.4. поиск с учётом стиховедческих параметров

Так как БПК имеет дело с особым образом организованными текстами (стихотворными), в нём имеется возможность искать слова и другие языковые явления в их связи с такими стиховедческими параметрами как метр и рифма.

5.4.1. метр

Чтобы задать поиск по метру, нужно выбрать из списка во вспомогательном окне (ссылка «метры и аффиксы» в правой верхней части окна) количество слогов, соответствующее силлабическому метру:

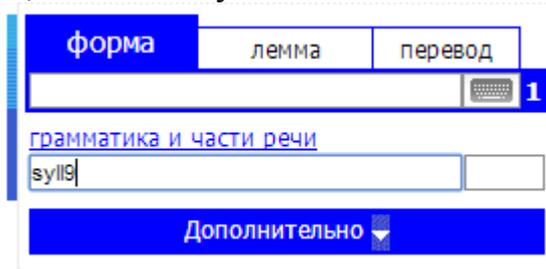


Рис. 13

или вручную написать syllX в поле поиска по грамматическим категориям (вторая строка в левой верхней части экрана), где вместо X будет нужное количество слогов в стихотворной строке.

В результатах будут показаны те слова или другие языковые явления, которые встречаются только в строках заданной длины.

5.4.2. рифма

Рифма обычно появляется на концах строк. Позицию на конце строки правильнее называть *клаузулой*. Вы можете искать слова на концах строк, выставив соответствующую настройку в поисковой форме (выпадающий список «позиция в строке» в меню «Дополнительно»). В результатах будут показаны соответствующие заданным поисковым критериям слова, находящиеся в позиции конца строки.

5.5. поиск с учётом авторства и датировки

Корпус позволяет задавать поисковые запросы таким образом, чтобы были

найжены слова или другие языковые явления, приходящиеся на тексты одного автора или временного периода. Они задаются в специальном окне, всплывающем по ссылке «Подкорпус» в нижней левой части экрана.

Все указанные типы поиска (кроме поиска точной формы, не сочетающегося с собственно лексическим поиском и с грамматическим поиском) могут комбинироваться между собой. Таким образом, вы можете искать некоторые слова (по башкирской лемме или по русскому значению), одновременно задавая их грамматические признаки и ограничивая область поиска датировкой, автором или метром.

Примечание: поисковые запросы, использующие максимум дополнительных параметров могут работать медленнее «простых» запросов.

6. Как читать результаты поиска?

Результаты поиска будут показаны в отдельном окне браузера.

Результаты содержат количество и контексты для найденных языковых явлений (слов или грамматических форм):



Найдено: **1 518** вхождений, **1 045** документов

Рис. 14

Само найденное слово в примере отмечено подсветкой:

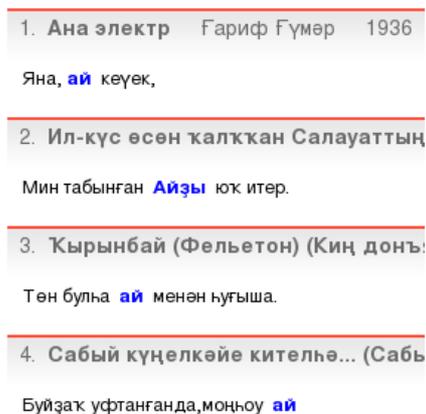


Рис. 15

Каждое слово в результатах содержит *морфологическую разметку*. То есть можно установить, какой грамматический разбор существует для этого слова в корпусе. Для этого нужно навести курсор мыши на слово и его разбор появится в виде всплывающей подсказки:

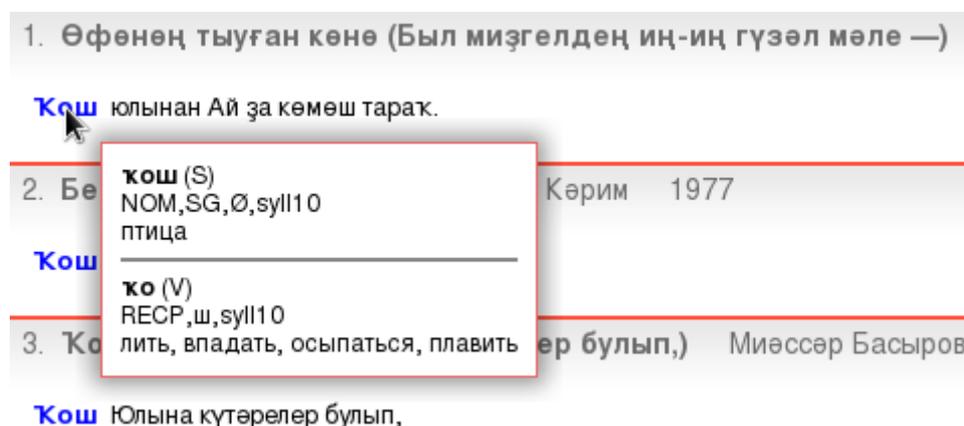


Рис. 16

Грамматические пометы здесь даны в виде сокращений, соответствующих общепризнанным Лейпцигским правилам глоссирования.

Ниже приводятся принятые в корпусе грамматические обозначения и их соответствия в русской и международной языковедческих традициях:

помета	международное соответствие	русское соответствие
S	substantive	существительное
V	verb	глагол
ADJ	adjective	адъектив
NUM	numeral	числительное
SPRO	pronoun	местоимение
PART	particle	частица
POST	postposition	послелог
CONJ	conjunction	союз
INTJ	interjunction	междометие
PL	plural	множественное число
SG	singular	единственное число
1,2,3SG	first (second, third) person, singular	первое (второе, третье) лицо единственного числа
1,2,3PL	first (second, third) person, plural	первое (второе, третье) лицо множественного числа
UNCERT	uncertainty	клитика со значением

		неуверенности
UNDEF	undefiniteness	клитика со значением предположительности
INTERROG	interrogative	клитика со значением вопросительности
REQUEST	request	клитика со значением просьбы
PRED	predicative	сказуемость
POSS	possessive	принадлежность
IPOSS	impersonal possessive	неличная принадлежность
NOM	nominative case	основной падеж
GEN	genitive case	родительный падеж
DAT	dative case	дательный падеж
ACC	accusative case	винительный падеж
ABL	ablative case	исходный падеж
LOC	locative case	местно-временной падеж
ABE	abessive case	абессив
DERIV.ABSTR	abstract noun	абстрактное существительное
DERIV.AGENS	agens noun	со значением деятеля
ASSIM	assimilation	уподобление
DIST	distinctive	выделительное (числительное)
COL	collective	собирательное
ORD	ordinary	порядковое
APRX	approximative	приблизительное
DIVIS	divisive	разделительное
COMP	comparative	сравнительная степень
ACYCL	acyclic	непериодическое действие
GER	gerund	деепричастие
IMP	imperative mood	повелительное наклонение
REFL	reflexive genus	возвратный залог
PASS	passive genus	страдательный залог
RECP	reciprocal genus	взаимный залог
CAUS	causative genus	понудительный залог
NEG	negation	отрицание
PRES	present tense	настоящее время
PST.DEF	past definite tense	прошедшее определённое

PST.INDF	past indefinite tense	прошедшее неопределённое
FUT.INDF	future indefinite tense	будущее неопределённое
FUT.DEF	future definite tense	будущее определённое
DESI	desiderative mood	наклонение намерения
COND	conditional mood	условное наклонение
OPT	optative mood	желательное наклонение
INF	infinitive	инфинитив
SUP	supin	имя действия

Каждое уверенно разобранное программой автоматического анализа морфологии слово содержит перевод на русский язык.

Если слово имеет два возможных грамматических разбора, эти разборы разделяются горизонтальной чертой.

Кроме стихотворной строки, содержащей найденное слово или грамматическую форму, пользователь может также просмотреть ближайший контекст из нескольких стихотворных строк, предшествующих искомой, и следующих за ней. Для этого нужно щёлкнуть по ссылке «Расширить контекст» справа от найденного текста.

Над поэтическим контекстом приводится название стихотворения, имя автора и год (если он указан):

1. Дустар һейләй (Дустар һейләй, дустар шулай һейләй:) Дауыт Юлтый 1923
Кандар түктөм, йәндәр үртөнөм;

2. Куркаклыктың үз-үзенә мәзһиәһе (Мин йәшәйем һеззәң йәнегеззә —) Факиһа Туғызбаева —
Үзүзәрә өсөн курка улар,

3. Начальник (колхозсы-бригадир хаты) (Без, иҫкене күргән кешеләр,) Ғ. Сәләм 1934
Кейгән **кабыктары** тире тун.

Рис. 18

7. Как сохранять результаты поиска?

В настоящий момент найденные в БПК контексты можно сохранить себе на компьютер, скопировав их со страницы в текстовый документ. Для этого нужно выделить интересующие примеры на странице результатов поиска. Нажать Ctrl+C или правую кнопку мыши и в появившемся контекстном меню выбрать пункт «Копировать». Затем в открытом текстовом документе на компьютере вставить скопированное с помощью комбинации клавиш Ctrl+V или пункта

«Вставить» в контекстном меню, вызываемом нажатием правой кнопки мыши.

Кроме того найденный контекст можно сохранить в txt-файл из окна расширенного контекста (вызывается по ссылке справа от найденного текста). Для этого нужно щёлкнуть по ссылке «Сохранить в файл»:

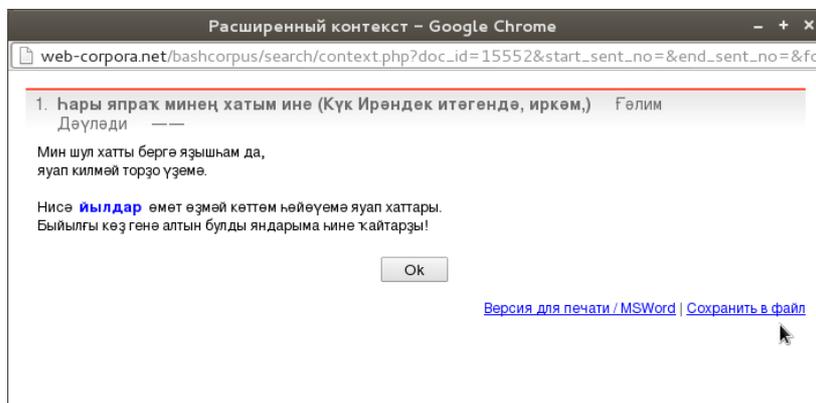


Рис. 19

8. Как интерпретировать результаты поиска?

Если по заданному вами запросу ничего не нашлось, это может означать либо логическую ошибку, допущенную в вашем запросе (скажем, одновременно выбранные не сочетающиеся в языке грамматические признаки, как, например, «существительное» и «деепричастие»), либо то, что явление, которое вы искали, встречается редко и именно поэтому не зафиксировано в корпусе. Возможно, если вы выберете менее строгие условия поиска (выберете для поиска не одного автора, расширите временной интервал, снимите требование искать в строках определённого метра), то получите ненулевой результат.

Если вы получили на свой запрос много примеров (психологический порог для «много» устанавливается интуитивно), то, скорее всего, это значит, что вас интересует какое-то частотное языковое явление, проявляющее себя и в поэтических текстах.

Если какое-то слово чаще встречается в текстах одного поэта, чем в текстах другого поэта, то это должно свидетельствовать об особенностях их индивидуальных стилей, образных рядов, поэтических стратегий.

Найденные с помощью корпуса контексты дают понять, в каком окружении появляются в поэтических текстах те или иные слова, какое значение приобретают.

С более серьёзными научными задачами, решаемыми с помощью корпуса, можно ознакомиться в работах, созданных на материале НКРЯ: сборники статей «Корпусный анализ русского стиха» (М., 2013), «Корпусные исследования по русской грамматике» (М., 2009).

9. Какие тексты входят в корпус?

Состав корпуса на момент запуска насчитывал около 17 тыс. стихотворений и с тех пор постоянно пополняется. Подробно с содержимым корпуса можно ознакомиться на странице <http://lcph.bashedu.ru/index.php?go=corpus>

10. Как ссылаться на корпус в научной работе?

Если вы использовали БПК для своей научной работы, следует сослаться на БПК, указав его в качестве источника, а также приведя имя автора и название цитируемого произведения. Например:

Боролоп күз тегәнем

Тарих битгәрен битләнем

Таныш эзәр эзләнем

Сәләх Кулибай «Үтелгән юл (Поэманан)», БПК.

Если требования к оформлению текста предполагают иные стандарты ссылок, то оговорить опору на БПК можно в свободной форме в начале статьи («Материалом исследования послужил Башкирский поэтический корпус (БПК)»), а далее приводить уже только сами контексты (всегда следует указывать автора и заглавие):

Юк, Һин түгел май сәскәе,

Түгел көзгәң алмаһы ла.

(Самат Ғабидуллин. *Ергә һин тыуган өсөн*)

Ссылку на корпус можно поместить в сноске (Абдулхак Игебаев. Мин — тәбиғәт балаһы (Кобайыр). БПК). Однако сокращение «БПК» обязательно нужно раскрыть либо в начале статьи, либо в специальном списке сокращений.

Если вы хотите привести ссылку на сам БПК, оформленную по правилам ГОСТа, то в списке литературы этот электронный ресурс должен быть оформлен так:

БПК — Башкирский поэтический корпус [Электронный ресурс] [2013—2014]. Дата обновления: 27.11.2013. URL: <http://web-corpora.net/bashcorpus/search/> (дата обращения: 15.04.2014).

Даты в этом библиографическом описании обновляются.

Рис. 23

11. Как при создании корпуса решена проблема авторских прав?

Корпус не стоит путать с электронной библиотекой. В его задачи не входит электронное представление текста в Интернете. Корпус предназначен только для поиска контекстов и не обязан показывать текст целиком. Корпус рассчитан на использование учёными. Поэтому он подпадает под действие статьи 1274 главы 70 Гражданского кодекса РФ (принят ГД ФС РФ 24.11.2006): «Допускается без согласия автора или иного правообладателя и без выплаты вознаграждения, но с обязательным указанием имени автора, произведение которого используется, и источника заимствования <...> цитирование в оригинале и в переводе в *научных*, полемических, критических или информационных целях правомерно обнародованных произведений в объеме, оправданном целью цитирования...» (выделено мной — Б. О.)

Иными словами, на БПК не распространяются ограничения, связанные с авторскими правами.

12. Что делать, если в корпусе нашлась ошибка?

Ошибки при оцифровке текстов неизбежны. Невозможно избежать их и при подготовке такой большой системы, как БПК, на момент запуска имевший текстовую базу более чем в 1 млн. словоупотреблений.

Если вы обнаружили ошибку в текстах, то сообщите о ней создателям корпуса, мы её исправим. Лучше всего сделать это по электронной почте, написав письмо по адресу lcph@bashedu.ru. В теле письма нужно указать текст, в котором содержится ошибка (автор, название, дата, если она указана), ошибочную форму и форму, на которую, как вы считаете, нужно её исправить. Желательно также скопировать в письмо *строку запроса*, в результатах которого вам попала ошибка. Она приведена в окне, всплывающем по ссылке «Сообщить об ошибке» в левой нижней части экрана.

13. Кто делал корпус и кому он принадлежит?

Корпус создавался силами сотрудников Лаборатории компьютерной филологии факультета башкирской филологии и журналистики Башкирского государственного университета. Тексты оцифровывались, вычитывались и размечались к. ф. н. А. А. Галлямовым и частично — З. Ш. Цыгановой. Программы автоматической обработки написаны к. ф. н. Б. В. Ореховым. Для поиска по базе адаптирована поисковая система Восточноармянского национального корпуса, созданная компанией Corpus Technologies.

Корпус располагается на сервере Института русского языка им. В. В. Виноградова РАН. Все авторские права на поэтические тексты остаются за башкирскими поэтами.

14. Какие технологии использованы при создании корпуса?

Оцифровка башкирских поэтических текстов производилась с помощью программы распознавания изображений ABBYY FineReader. Далее тексты проходили ручную стиховедческую разметку и автоматическую морфологическую разметку с помощью программы машинного анализа башкирской морфологии, разработанной в Лаборатории компьютерной филологии Башкирского государственного университета Б. В. Ореховым и А. А. Галлямовым. Онлайн-версией программы можно воспользоваться на специальной странице: <http://lcpb.bashedu.ru/cgi-bin/parser.pl>

Поисковые механизмы БПК созданы в компании Corpus Technologies.