Московский государственный университет им. М.В.Ломоносова
Филологический факультет
Кафедра русского языка
Лаборатория общей и компьютерной лексикологии и лексикографии

Читаем и исследуем

Работа с корпусами текстов с помощью информационно-исследовательской системы КИИСа

Выпуск 1:

Корпус «Поэзия и драматургия А.С.Пушкина» и

«Путеводитель по Пушкину»

Москва 2010

Без всякого труда
Оно в тебе находит
Концы моих стихов
И верность выраженья,
То звуков или слов
Нежданное стеченье,
То едкой шутки соль,
То Правды слог суровый,
То странность рифмы новой,
Неслыханной дотоль.

А.С.Пушкин, «К моей чернильнице»

Оглавление

Общие сведения о системе КИИСа

КИИСа представляет собой достаточно универсальный и доступный всем желающим инструмент филологического исследования текстов. Это предназначенная для работы с корпусами текстов информационно-исследовательская система. Она дает возможность работать с организованной и размеченной разными типами информации коллекцией текстов (корпусом) как в обычном, полнотекстовом режиме, так и в режиме просмотра конкордансов – специализированных словарей, из которых можно осуществлять вход в нужные места текстов.

Система также включает в себя в качестве отдельного компонента связанную с единицами корпуса справочную базу, в которой могут содержаться дополнительные сведения, имеющие отношение к текстам корпуса и полезные для изучения их единиц.

КИИСа позволяет специалистам (прежде всего филологам) и всем интересующимся не только пользоваться уже готовыми результатами изучения какого-либо корпуса текстов, но также проводить самостоятельный его анализ и получать новые данные – алфавитные и частотные словари произведений и их единиц, отобранных по разным признакам и др.

Система разработана в лаборатории общей и компьютерной лексикологии и лексикографии (ЛОКЛЛ) кафедры русского языка филологического факультета МГУ им. М.В.Ломоносова и предназначена в первую очередь для работы с корпусами, подготавливаемыми в ЛОКЛЛ.

Принципы информационно-исследовательской системы КИИСа были предложены А.А.Поликарповым. Создание и разметка корпуса - О.В.Кукушкина. Программирование – В.В.Федотов. Дизайн диска – А.А.Варламов.[1]

Первый продукт, представляемый с помощью данной системы, это диск, содержащий корпус текстов «Поэзия и драматургия А.С.Пушкина»[2], набор словарей по этим текстам и связанную с корпусом справочную базу данных «Путеводитель по Пушкину» (см. подробнее ниже).

В настоящее время содержание Интернет-версии этого диска может быть просмотрено по адресу http://www.philol.msu.ru/~lex/kiisa.html . Интернет-версия диска представлена в виде системы «клиент-сервер».

[1] При графическом оформлении диска и данного издания использовались рисунки Пушкина.
[2] Как и другие корпуса лаборатории, этот корпус подготовлен и размечены с помощью разработанной в ЛОКЛЛ системы автоматического анализа текстов и словарей «Dictum».

В настоящее время к изданию и распространению на основе КИИСы готовится также "Ядерный корпус газетных текстов русского языка" (1,3 млн. словоупотреблений).

Для работы с диском необходимо знать следующее:

Требования к операционной системе: Windows 2000/XP.

Для корректного отображения французского текста необходимо предварительно установить в Windows шрифт Kiisarf.ttf, который находится на диске в директории Font.

Запуск программы осуществляется с помощью файла: Kiisa.exe.

О корпусе текстов «Поэзия и драматургия А.С.Пушкина»

В данный корпус вошли все поэтические и драматические произведения А.С.Пушкина[3], исключая черновики, не опубликованные Пушкиным варианты, наброски и редакции, тексты с пометой “Dubia” (сомнительные)[4]. Тексты произведений соответствуют академическому изданию (Пушкин. Полное собрание сочинений. – М.: Издательство академии наук СССР, 1937).

Количественные характеристики корпуса:

Тип единиц	*Количество*
Тексты	880
Словоупотребления	200995 (из них 1224 на других языках).
Разные словоформ	37721 (из них 708 на других языках)
Разные лексемы	15301 [5].

С текстами корпуса связана справочная словарная база «Путеводитель по Пушкину», содержащая, в частности, пушкинские примечания и сведения об авторах коллективных произведений и др. (см. ниже).

Тексты корпуса размечены информацией 15-ти типов, представляющей интерес как для литературоведов, так и для лингвистов. Эта информация может использоваться также в обучающих целях (см. «***Типы информации, имеющиеся в корпусе «Поэзия и драматургия А.С.Пушкина***»»).

В Академическом издании и, соответственно, в корпусе используются следующие условные знаки:

- в [ ] заключаются слова, зачеркнутые Пушкиным;

- в < > заключаются дополнения, сделанные редакцией, а именно «… дополнение букв недописанного слова, случайно пропущенные слова, пополняющие контекст, без них неполный, и т.д.»[6]

[3] В том числе написанные А.С. Пушкиным тексты на французском языке. При передаче французского текста, а также немецких, французских и итальянских слов, используемых в русскоязычных произведениях А.С.Пушкина, принимается упрощенная система – буквы с надстрочными знаками заменяются на их эквиваленты без этих знаков.

[4] Исключение сделано только для стихотворных набросков неоконченных поэм («Поэма о гетеристах», «Актеон», «Бова», «Русская девушка и черкес»).

[5] Словоупотребления на других языках представлены в корпусе только словоформами. Исключение составляют только написанные латиницей существительные, грамматически связанные с русскими словами и являющиеся частью русского текста (ср. «Открыт Casino» и пр.).

[6] Пушкин. Сочинения., т. I, с. XIII

В корпусе сохраняется и орфография данного издания. Для лингвистов будет полезна информация о принятых в нем правилах передачи текстов[7], приводимая ниже.

«Все тексты Пушкина печатаются по ныне принятой орфографии. Однако, учитывая научное значение настоящего издания, которое должно давать материал не только для чтения, но и для изучения Пушкина, в отступление от общепринятых орфографических норм сохранены некоторые особенности, отражающие живой язык Пушкина с точки зрения его произношения, грамматической и лексической системы.

... В настоящем издании сохраняется подлинная орфография Пушкина во всех тех случаях, когда замена подлинного написания современным создавала бы для нынешнего читателя неверное представление о звуковом составе или грамматической форме слова».

А именно, **сохраняются:**

- такие написания «...за которыми с несомненностью следует предполагать книжное произношение.., например, *ея* и *оне* (с заменой «ять» через «е»)»;
- «...своеобразные грамматические окончания, не совпадающие с современными, напр. *бревны, селы, ... в шинеле, в шале, ... дышет, слышет*; сохраняются особенности в написаниях суффиксов, напр. *Олинька, маминька*»;
- «вообще все написания, показывающие, что произношение Пушкина в данных словах отличалось от современного, напр. *оспоривать, дальный, диравый, незапный, щедушный, впроччем, приближиться, скрыпка, впрям* и т. п.»;
- «колебания в написании одного и того же слова, засвидетельствованные подлинниками.... напр. *цаловать* и *целовать, верьхом* и *верхом*»;
- «прописные буквы в тех случаях, где они придают особую выразительность слову, напр. в отвлеченных словах, как *Безумие, Злодейство* и т.п.».

Не сохраняются и заменяются современными:

- «такие написания Пушкина, как «щастие», «щитать», вместо которых печатается «счастие», «считать»;
- «встречающиеся у Пушкина окончания именительного падежа мужского рода прилагательных *–ой* в тех случаях, где теперь пишется *–ый,* напр. вместо «гибельной позор», «голодной прейскурант», «забытой сон» печатается: «гибельный позор», «голодный прейскурант», «забытый сон». Однако окончание *–ой* после букв *к, г, х,* свидетельствующее об их твердом произношении, сохраняется...»;
- «*-ь-* после шипящих там, где он не пишется теперь, напр. написания: *Угличь*, *царевичь, ужь* заменяются современными...» ;

[7] Пушкин. Сочинения. ,т. I, с. XIII-XIV

- строчные буквы прописными в тех случаях, когда использование большой буквы можно объяснить «… орфографической условностью или религиозными, политическими и бытовыми традициями, напр. в словах, обозначающих религиозные понятия, в обозначении лиц высокого ранга, в наименованиях народностей и т.п.».

Устройство корпуса и логика работы с ним

Доступ к текстам корпуса и их исследование осуществляется через режим «Конкорданс».

В этом режиме систему можно использовать и для простого чтения текстов произведений. Для этого нужно выбрать тип информации «Названия произведений» и перейти в закладку «Текст» (см. ***«Переход в текст и перемещение по нему»)***.

Однако при исследовании текстов основным инструментом их изучения служат конкордансы. Конкорданс – словарь особого вида, в котором каждый элемент словника связан с теми контекстами, в которых он употребляется. Компоненты конкорданса – словник и контексты – располагаются в разных окнах. Они связаны между собой, и при перемещении по словнику меняется набор контекстов. Каждая строчка окна контекстов представляет собой особый контекст, иллюстрирующий конкретный случай употребления выбранной единицы. Длина контекста по желанию может увеличиваться и уменьшаться.

Каждый из конкордансов группирует материал корпуса особым образом, в соответствии с выбранным типом информации. Выбор нужного типа информации и, соответственно, конкорданса осуществляется в окне “Тип информации”, закладка “Конкорданс**"** (см. ***«Типы информации, имеющиеся в корпусе»***).

При выборе конкретного типа информации открывается словник нужного конкорданса, для каждого из элементов которого можно получить:

- список его контекстов;
- сведения о его частотности;
- адрес элемента, т.е. текст, из которого взят его контекст.

Единицы конкорданса могут быть отсортированы трояким образом – по алфавиту, по частоте (в порядке ее убывания), а также по концам (обратная алфавитная сортировка) (см. ***«Сортировка словника», «Сортировка контекстов»***). Это дает возможность получать словари разных типов – алфавитно-частотные, частотные и обратные - по всем тем единицам, которые выделены в корпусе.

Контексты, тексты, а также словники (объемом не более 1000 единиц) могут быть помещены в буфер и сохранены в текстовый файл (см. ***«Копирование текста в***

файл», «Копирование контекстов в файл», «Копирование в файл единиц словника»).

Кроме того, для единиц ряда конкордансов («Слова», «Гиперслова», «Адресаты») автоматически выдается дополнительная информационная справка – словарная статья из “Путеводителя по Пушкину” (см. ниже).

Помимо уже готовых конкордансов, выбираемых из меню «Тип информации», система позволяет пользователю получать новые конкордансы. Для этого используется функция «Фильтр». Установка **Фильтра** на одном из элементов словника конкорданса позволяет «пересечь» информацию, содержащуюся в конкордансах разного типа, поэтому фильтры дают богатые возможности для самостоятельного исследования корпуса. С их помощью, например, можно получить список всех слов или словоформ одного конкретного произведения, список произведений, написанных в конкретный период, перечень всех произведений, написанных в определенном году, всех произведений с одним и тем же адресатом и т.п. (см. ***«Установка фильтра и уточнение запроса»***).

Полезной функцией системы является возможность анализа совместной встречаемости элементов. В данном варианте КИИСы эта возможность ограничена получением списка всех элементов, находящихся на определенном месте слева или справа от изучаемой единицы (см. ***«Анализ сочетаемости»***). Для пушкинских текстов анализ сочетаемости имеет особое значение. Именно с ней во многом связано их волшебство. Ср., например, каким бывает **«взор»** в поэтических текстах А.С.Пушкина:

```
благосклонный 1
блестящий 1
болезненный 1
больной 1
боязливый 1
бродящий 1
быстрый 1
важный 2
вдохновенный 1
вдохновенья 1
ведьмы 1
веселый 2
внимательный 3
волшебниц 1
волшебный 1
вольный 1
смелый 1
впалый 1
врага 1
враждебный 1
гибели 1
глаз 2
гордый 2
грозный 2
лукавый 1
любови 1
любопытный 2
милый 5
молящий 4
навыкате 1
наглый 1
насмешливый 2
невинный 2
невольный 2
недвижимый 1
недвижный 2
недостойных 1
недружеский 1
нежный 9
огненный 3
огромный 1
озлобленный 1
опасный 1
орлиный 1
ослепленный 1
острый 1
открытый 1
отчаянный 1
прощальный 1
равнодушный 1
ревнивый 1
робкий 2
светлый 4
скорбный 1
слабый 1
слезливый 1
смелый 2
смутный 2
смущенный 3
сожаленья 1
спокойный 2
страстный 2
страшный 3
строгий 2
стыдливый 1
суровый 1
таинственный 1
тихий 5
томительный 1
томный 10
туманный 3
тусклый 2
```

девицы 1
девы 2
дерзкий 1
деспотизма 1
дикий 3
довольный 1
единый 2
жадный 1
живой 1
зависти 1
завистливый 1
задумчивый 1
знатоков 1
исполненный тоской 2
косвенно-внимательный 1
красноречивый 1
красоты 1
кровавый 1
очей 7
первый 1
печальный 3
пламенный 1
плененный 1
полный (*неги 1, недоуменья 1, хитрой лести 1, огня надежды 1,славы 1, слез* 1)
полузакрытый 1
последний 1
потупленный 3
потухший 1
праздный 1
презрительный 1
прекрасный 2
приветный 2
привидений 1
прислуги 1
угасший 3
угрюмый 2
ужасный 3
умиленный 1
умильный 4
унылый 4
хитрый 2
хладный 3
хозяйки 2
царей 1
черни 1
чудесный 1
чудный 2
ясный 3

Типы информации, имеющиеся в корпусе «Поэзия и драматургия А.С.Пушкина»

Единицы корпуса снабжены информацией разного типа, представляющей интерес как для лингвистов, так и для литературоведов.

Доступ к этой информации осуществляется через окно «Тип информации». В этом окне можно выбрать один из готовых конкордансов. В названии конкорданса отражается состав единиц его словника.

В данном варианте корпуса имеются конкордансы со следующими названиями:

1. *Словоформы:* единицы словника - словоформы, встречающиеся в корпусе.

Имена подающих реплики персонажей драматургических произведений в этом виде конкордансов, а также в конкордансе с единицей «Слова» даются разрядкой, чтобы не искажались данные о степени употребительности.

Этот вид конкордансов удобно использовать, если нужно исследовать, например, имеющиеся в произведениях варианты форм слова.

2. *Слова:* единицы словника – лексемы.

Данный конкорданс позволяет получить список лексических единиц с разрешенной омонимией, т.е. собственно словарь. С помощью этого вида представления информации можно получать, например, словари конкретного произведения, типа текстов, периода творчества и пр.

3. *Части речи:* единицы словника - части речи.

Выбрав этот тип информации и установив Фильтр на нужной части речи, можно получить затем полный словник слов или словоформ этой части речи. Для этого нужно

просто поменять тип информации (т.е. выбрать в окне «Тип информации» пункт «Слова» или «Словоформы).

4. *Варианты слов*: единицы словника - варианты одного слова.

В одной строке словника представлены варианты одного слова, зафиксированные в текстах создателями «Словаря языка А.С.Пушкина». Поскольку в этом словаре используются данные и о прозаических текстах, т.е. материал всего творчества, то в словнике могут встречаться и варианты, не содержащиеся в текстах данного корпуса.

Находясь в конкордансе «Слова» и установив Фильтр на нужное слово, можно проверить, есть ли у этого слова варианты. Для этого нужно поменять тип информации на «Варианты слов».

5. *Семантические классы*: единицы словника – некоторые тематические группы слов, выделенные при исследовании корпуса.

В представленном на диске варианте доступны три класса: (1) Города, страны, народы; (2) Персонажи и лица; (3) Цвето- и светообозначения. При использовании данного типа информации нужно учитывать, что в класс «Цвето- и светообозначения» объединены случаи употребления слов именно в «цветовых» и «световых» значениях. Такие переносные употребления, как «черные дни», «черная печаль», «светлая грусть» и пр. , в него не включались.

Выбрав тип информации «Семантические классы» и поставив Фильтр на нужном классе, можно затем поменять тип информации на «Слова» и получить полный список слов, относящихся к этому классу. Можно также, поменяв тип информации на «Гиперслова», получить список гиперслов нужной тематики, и т.п. Тем же способом можно, например, изучить список произведений и периодов творчества Пушкина, в которых данный семантический класс представлен наиболее активно. Для этого нужно использовать такие типы информации, как «Названия произведений», «Дата», «Период творчества».

6. *Гиперслова*: единицы словника – группировки однокоренных, а также связанных семантической производностью слов. В одно гиперслово сводились также разные способы именования одного и того же объекта. Объединение наименований лиц и географических объектов производилось на основе указателей имен собственных, имеющихся в Собрании сочинений.

Гиперслова последовательно выделялись только для трех указанных выше семантических классов. Гиперслово - особая информационно-поисковая единица. Ее использование позволяет найти по одному запросу все случаи указания на тот или иной

цвет (например, по гиперслову «**белый**» - *«белый», «белизна», «беловатый», «белоснежный», «мраморный*» и пр.), упоминания о той или иной стране и ее обитателях (например, по гиперслову «**Россия**» - *«рос», «русский», «по-русски», «полунощный флаг», «держава», «отечество», «отчизна»* и др.), о персонажах и лицах (например, по гиперслову **«Ларина Татьяна**» - «*Таня», «Татьяна», «Ларина*»; по гиперслову «**Хвостов**» - *«Свистов», «Отец зубастых голубей», «Хвостов», «Хлыстов», «Графов»* и пр.). Гиперсловные группировки позволяют получать частотные словари особого, когнитивного типа и исследовать реальную частоту обозначения того или иного понятия, а не просто частоту употребления конкретного слова.

Пример 1: Начало частотного списка гиперслов с семантикой «Города, страны, народы» по корпусу «Поэзия и драматургия А.С.Пушкина» (частота 50 словоупотреблений и выше)

1. Россия 257
2. Москва 105
3.Франция (Галлия), Париж 63
4. Иудея, Израиль, Евреи, Иерусалим 62
5. Казаки (Козаки) 53
6 Парнас 52
7. Польша 51.

Пример 2: Начало частотного списка гиперслов с семантикой «Цвето- и светообозначения»: по корпусу «Поэзия и драматургия А.С.Пушкина» (частота 50 и выше)

1. Темный 462
2. Блестящий 198
3. Светлый 172
4. Бледный 139
5. Белый 134
6. Черный 131
7. Седой 101
8. Золотой 73
9. Синий 59
10. Сияющий 57
11 Ясный 54
10. Красный 53.

Пример 3: Начало частотного списка гиперслов с семантикой «Персонажи, лица» по корпусу «Поэзия и драматургия А.С.Пушкина» (частота 100 и выше)
1. Ларина Татьяна 147
2. Онегин Евгений 144
3. Дон-Гуан 140.

7. *Названия произведений*: единицы - названия произведений, составляющих корпус.[8]

Этот тип информации удобно использовать для просмотра текстов произведений. Установив Фильтр на нужном названии, можно также получить словарь произведения, узнать, имеется ли у него адресат, когда оно было написано и пр.

[8] Названия, не написанные большими буквами, представляют собой первые строки произведения.

8. *Адресаты* – лица, которым точно или предположительно адресовано то или иное произведение или о которых в нем идет речь .

9. *Дата* – время написания произведения. Дата дается по академическому изданию.[9]

10. *Период творчества:* единицы - временные периоды творчества А.С.Пушкина. [10]

11. *Тип текста*: единицы – типы произведений (роман, поэма, стихотворение, отрывок или набросок поэмы или стихотворения, драматургическое произведение, коллективное стихотворение и пр.).

Этот тип информации позволяет собрать, например, словарь всех стихотворений, словарь драматургических произведений А.С.Пушкина и т.п.

12. *Части произведений:* единицы – композиционные части крупных произведений – главы, сцены, части, вступления, эпилоги, авторские примечания, предисловия, посвящения. С помощью этого типа информации можно уточнить данные об адресе контекста, например, о номере главы «Евгения Онегина», из которой взят контекст.

13. *Композиционные элементы:* единицы – вставные композиционные элементы произведений - эпиграфы, письма героев, песни.

14. *Иноязычные тексты:* единицы – целые произведения и элементы произведений, написанные на иностранных языках. Этот тип информации нужно использовать для получения перевода – именно для единиц конкорданса этого типа осуществляется автоматическая выдача статьи «Путеводителя», содержащей русский переводной эквивалент.

15. *Концы стихотворных строк*: единицы – словоформы, стоящие в конце строк стихотворных (рифмованных и нерифмованных) произведений. При выборе этого типа информации появляется список таких словоформ. Для удобства исследования рифм его нужно отсортировать с помощью «обратной сортировки». Если же установить фильтр на нужном произведении, выбрав затем данный тип информации и «обратную сортировку», то можно изучать систему рифм конкретного произведения.

---

[9] Многоточие между датами обозначает интервал, в который могло быть написано произведение. Тире между датами обозначает интервал, в который произведение писалось.

[10] При возможном попадании произведения с неизвестной точно датой в разные периоды для него указывается интервал, в которое оно могло быть написано (через многоточие), и рядом ставятся знаки вопроса. В отдельный период объединены произведения 1824-1825 года.

Пересекая разные типы информации с помощью Фильтров и меню «Тип информации», можно провести самостоятельное исследование и получить большое количество новых данных.

Конкретные сведения о работе с системой

I. Компоненты системы

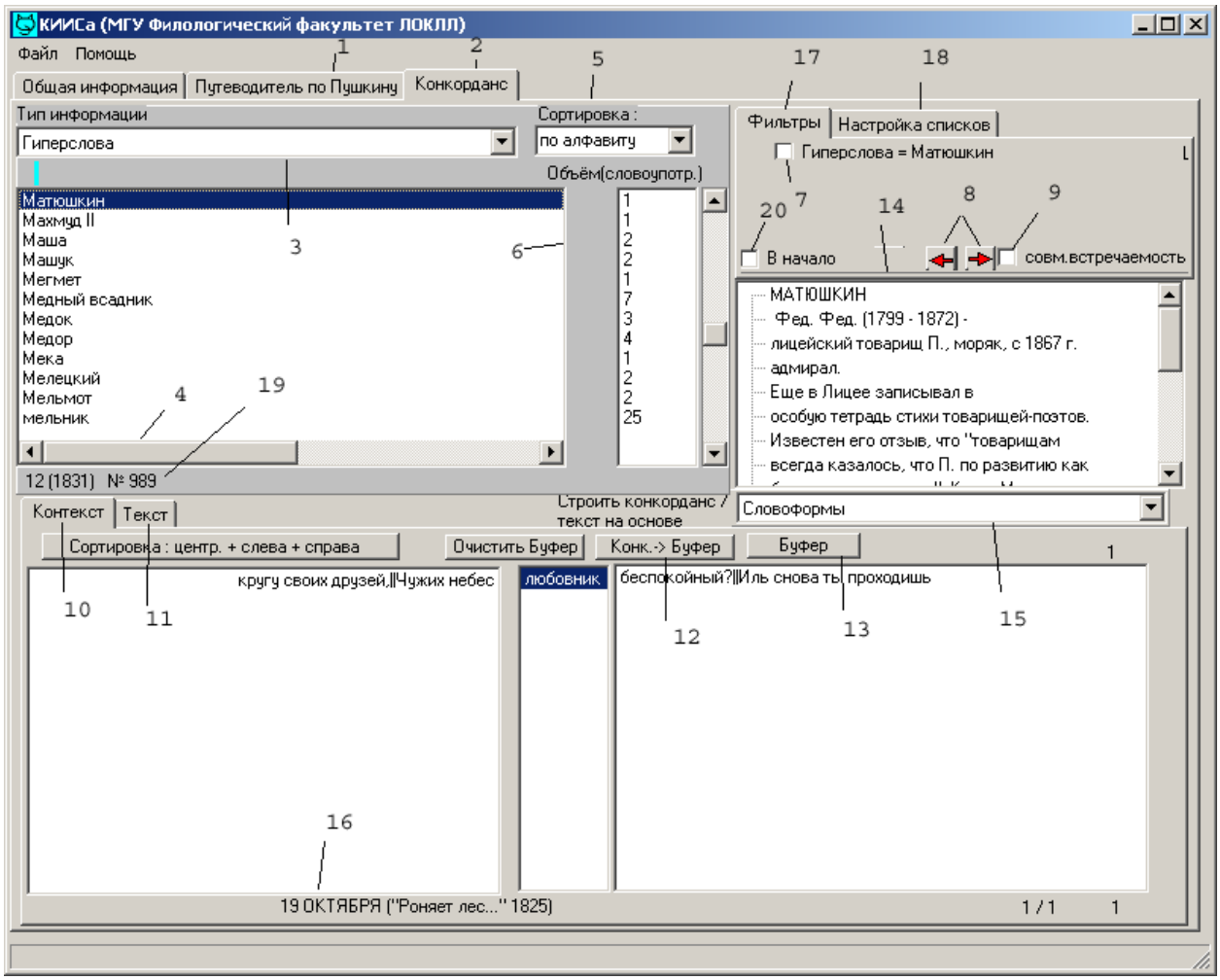
В системе имеются два основные компонента - “Конкорданс” и “Путеводитель по Пушкину”.

Через компонент «Конкорданс» осуществляются основные виды операций с корпусом – работа со словниками, контекстами и переход в полный текст произведений. Этот компонент активируется по умолчанию после загрузки программы (см. рис 1).

II. Работа с конкордансами и полным текстом

Все операции с корпусом осуществляются из закладки “Конкорданс”, открывающейся после загрузки. Общий вид экрана см на рис.1.

Рис.1. Вид системы после открытия (режим «Конкорданс»).



1 - закладка словарной справочной системы «Путеводитель по Пушкину»; 2 - открытая закладка «Конкорданс»; 3 - меню «Тип информации», через которое осуществляется выбор типа конкорданса и формы представления результатов при фильтрации; 4 - окно словника (на рис. фрагмент словника с единицами типа «Гиперслова», отсортированного по алфавиту) ; 5 - меню типа сортировки; 6 – данные о количестве случаев употребления каждой единицы (в данном случае гиперслова) в корпусе; 7 – окно установки фильтра (справа показывается элемент словника, на котором стоит курсор и на который будет поставлен фильтр в случае установки флага в окне «Фильтр»); 8 – кнопки возврата к прежнему состоянию; 9 – окно, в котором выставляется флаг перед началом анализа совместной встречаемости; 10 – открытая закладка «Контекст» и экран контекстов для того элемента словника, на котором стоит курсор (в центральной части экрана – элемент словника); 11 – закладка «Текст», по которой можно перейти в полный текст произведения, из которого взят выделенный цветом контекст (название этого произведения, см. 16); 12 - очистка буфера, в который были скинуты контексты; 12 – кнопка записи контекстов в буфер; 13 – просмотр записанного и запись в файл; 14 – всплывающая справка из «Путеводителя»; 15 – меню типов единиц, которые будут отображаться в экране контекстов (в нормальном виде это словоформы); 16 – название произведения, из которого взят выделенный цветом контекст; 17 – открытая закладка «Фильтры»; 18 – закрытая закладка настройки списков; 19 -

количественная характеристика словника; 20 – окно для настройки попадания в начало словников и текстов.

В этом режиме возможны следующие виды действий: операции со словниками, операции с текстами, операции с контекстами (см. ниже)

II.1.Операции со словниками

Выбор нужного типа информации (словника)

В левом углу экрана находится окно «Тип информации». При переходе в него появляется меню, содержащее перечень основных типов информации (конкордансов), которые доступны пользователю (см. выше). При выборе одного из пунктов («Слова», «Словоформы», «Части речи», «Названия произведений» и пр.) в окне словника появляется алфавитно-частотный список, элементы которого отображают нужный тип информации. Так, при выборе пункта «Части речи» появится словарь частей речи, элементами которого будут все представленные в корпусе части речи; при выборе пункта «Слова» будет получен алфавитно-частотный словарь, отображающий весь лексический состав корпуса и т.д.

Выбранный словник является основным рабочим инструментом. На каждый из его элементов можно получать контексты, данные о сочетаемости с другими элементами, а также характеристику по любому типу информации. Так, установив фильтр на нужном названии стихотворения, можно получить список слов этого стихотворения с частотой для каждого слова или же информацию (если она есть) о предполагаемом или установленном его адресате. Для этого необходимо просто выбрать тип информации «Слова» (в первом случае) или же «Адресаты» (во втором).

Сортировка словника

Единицы любого словника могут быть отсортированы тремя способами: по алфавиту, по частоте (по убыванию) и по “концам” (т.н. обратная алфавитная сортировка). Чтобы отсортировать словник, нужно выбрать нужный тип сортировки в окне “Сортировка”.

Обратная сортировка имеет смысл прежде всего для таких словников, как «Словоформы» и «Слова». С помощью их обратной сортировки можно исследовать аффиксальную структуру словоформ и слов, систему рифм и пр.

Поиск нужной единицы в словнике

Поиск нужного элемента в словнике можно осуществить простым набором букв на клавиатуре. Искомое отображается голубым цветом под окном «Тип информации». Специально устанавливать курсор в эту позицию не нужно.

Удаление и отмена напечатанного - Del, Esc.

Нужно, однако, помнить, что этот способ поиска доступен только при **алфавитном порядке** сортировки словника.

Получение количественной информации

Каждая единица словника сопровождается информацией о количестве словоупотреблений, на нее приходящихся. Так, например, цифра около названия произведения обозначает количество словоупотреблений в произведении с этим названием.

Цифры под окном словника обозначают объем показываемого на экране фрагмента словника (он зависит от размера шрифта), общее количество его единиц (в скобках) и номер строки словника, на которой стоит курсор. Например, 12(553) N 1 означает, что в словнике 553 единицы, они выводятся на экран порциями по 12 единиц, курсор стоит на 1-ой единице словника.

При использовании фильтра справа от единиц словника появляется не одна, а две колонки с цифрами. Фильтр позволяет определить, какие именно значения того или иного типа может иметь «фильтруемая» единица и в каком количестве случаев она имеет то или иное значение. Однако при этом бывает важно знать, сколько вообще единиц с таким значением имеется в корпусе. Поэтому при фильтрации дается две характеристики: количество случаев употреблений, в которых «фильтруемая» единица имеет данное «значение» (см. левую колонку), и общий объем употреблений с этим значением в корпусе (см. правую колонку).

Приведем несколько примеров.

Пример 1. При установке фильтра на гиперслово «Аид (Айдес)» и выборе типа информации «Семантический класс» в словнике конкорданса появятся данные
Персонажи, лица /6 / 6486
Страны, города, народы /4 / 2145

Это означает, что выбранное гиперслово относится к двум семантическим классам – в 6-ти случаях оно обозначает персонаж (бог Плутон, Аид) в 4-х – подземное царство. Всего же словоупотреблений семантического типа «Персонажи, лица» 6486, а «Страны, города, народы» - 2145.

Пример 2. При установке фильтра на названии текста «Кинжал» и выборе типа информации «Слова» появится словник этого стихотворения, при каждом из слов которого в левой колонке будет указано, сколько раз слово употреблено именно в стихотворении «Кинжал», а во второй – сколько раз оно употреблено во всем корпусе. Ср. ***безглавый 1*** | ***3*** (в «Кинжал» - 1 , во всех произведениях - 3).

Пример 3. При установке фильтра на семантический класс «Страны, города, народы» и выборе типа информации «Слова» у слова «Альбион» появятся цифры **5|5,** что означает, что у слова «Альбион» пять употреблений данного семантического класса и пять употреблений в корпусе (т.е. во всех случаях употребления это слово относится к классу «Страны, города, народы»).

Пример 4. При составлении запроса, показывающего, как упоминания о Наполеоне распределены по годам (Фильтр на гиперслово: «Наполеон», «Тип информации» = «Дата») появится список дат, при каждой из которых будут стоять две цифры: сначала число упоминаний имени Наполеона, имеющихся в произведениях конкретного года, затем – общий объем всех словоупотреблений в произведениях этого года.

При фильтрации к характеристикам словника, показываемым под ним, добавляется еще один показатель – номер единицы словника в общем, «нефильтрованном» списке единиц данного типа информации. Этот показатель стоит последним и в скобках.

Анализируя количество употреблений единиц того или иного типа, можно получить интересные данные об особенностях текста и его автора. Так, например, отношение общего количества словоупотреблений корпуса к числу разных лексем позволяет оценить степень лексического разнообразия поэтического языка Пушкина. Оно непосредственно соотносится с разнообразием тем, лиц, идей, понятий и пр., отраженных в текстах, и хорошо объясняет, почему пушкинские произведения мы воспринимаем как «энциклопедию русской жизни». В качестве меры лексического разнообразия может быть взято предложенное Г. Херданом отношение логарифма объема словаря к логарифму объема текста. На материале поэзии и драматургии Пушкина этот показатель равен 0,79 [11]. Если устранить вариативность лексем, сведя варианты в одну лексему, то этот показатель будет немного ниже, однако и при этом он будет очень высок для поэтического языка, что характеризует пушкинские тексты как лексически очень насыщенные, богатые.

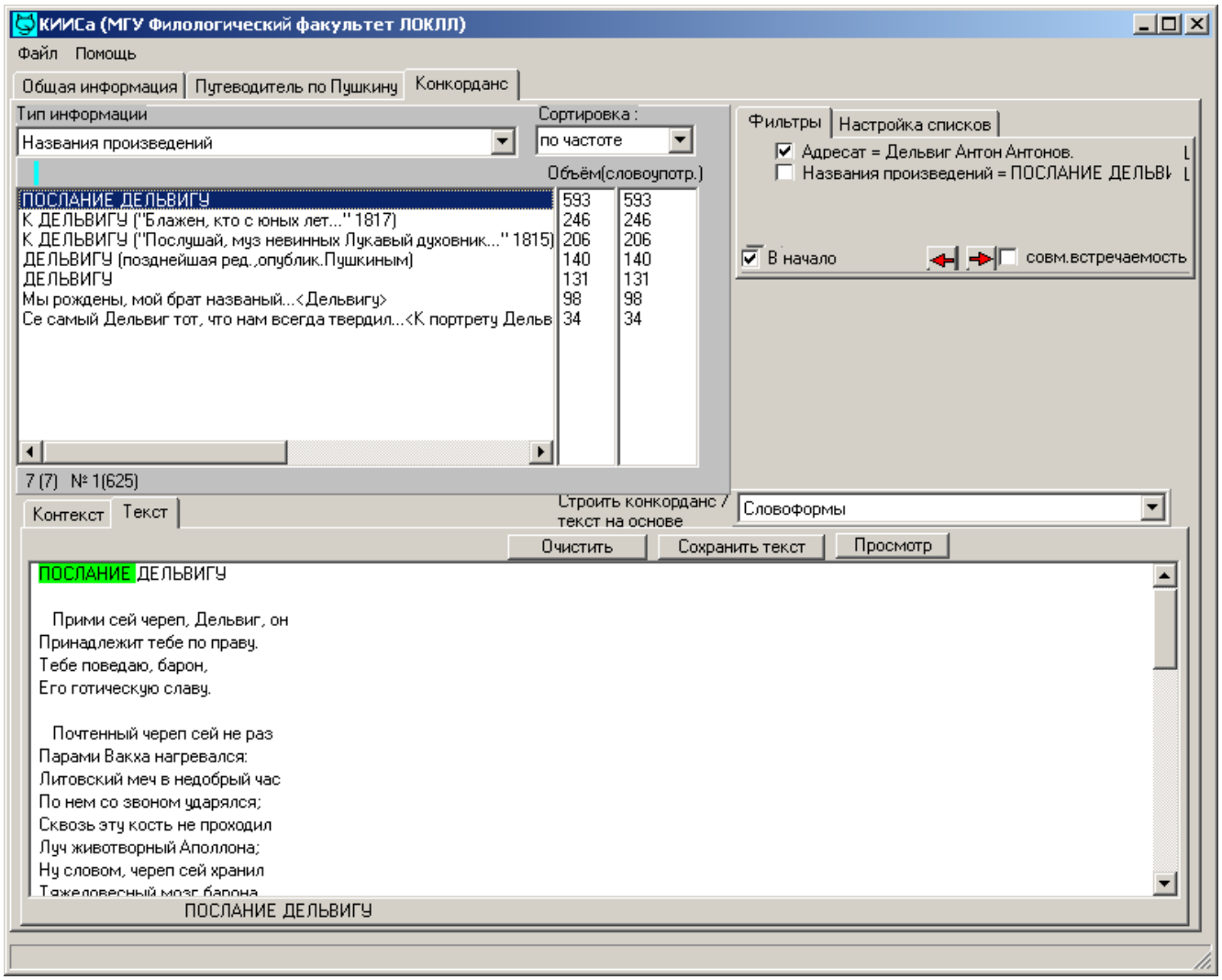
Установка фильтров и уточнение запроса

Выбрав нужную единицу в словнике, нужно установить флаг в окошке «Фильтр». После его установки происходит автоматическая смена типа информации в окне словника. Если появившийся тип информации не соответствует тому, что требуется, необходимо поменять его на нужный в окне «Тип информации».

[11] При наличии в корпусе двух вариантов одного текста (см. позднейшие напечатанные А.С.Пушкиным варианты лицейских стихотворений) в объем текстов корпуса не включались словоупотребления более позднего варианта. Из общего объема текста также исключались иноязычные словоупотребления.

При необходимости можно сделать дальнейшее уточнение, установив Фильтр на какую-либо единицу нужного типа информации. Все действия будут отображаться в правом верхнем углу в виде записей выбранных вами элементов. См. рис. 2.

Рис. 2. Результаты выполнения запроса «Стихотворения, адресованные Дельвигу» :



Конкорданс получен в результате выполнения следующих действий:

1. Выбран тип информации «Адресат».
2. Из списка адресатов выбран курсором «Дельвиг Антон Антонов.» и в окошке «Фильтр» поставлен флаг.
3. Выбран тип информации «Названия произведений». В результате в окне словника список адресатов заменился на список произведений, адресованных Дельвигу.
4. Произведена частотная сортировка этого списка. В результате стихотворения оказались расположены в порядке убывания объема слов.

Копирование в файл единиц словника

Для словников объемом менее тысячи единиц возможно копирование в файл. Оно осуществляется в окне словника с помощью нажатия правой клавиши мыши. Курсор при этом должен стоять на каком-либо слове. Если копирование возможно, то

появится пункт меню «Копировать список». При нажатии список будет занесен в буфер. Его можно просмотреть, нажав кнопку «Буфер», и затем записать в файл.

Получение дополнительной информации из окна словника

При перемещении по словникам некоторых видов конкордансов («Слова», «Гиперслова», «Адресаты», «Иноязычные тексты») может появляться окно со статьей из «Путеводителя по Пушкину». Оно возникает, если соответствующая единица описана в «Путеводителе». Справочная статья появляется **только** в том случае, если открыта закладка «**Контекст»,** а не «Текст».

Если в статье содержится отсылка «см.», то, щелкнув на ней левой клавишей мыши, можно перейти к нужному слову.

Если нужно получить информацию о единице, отсутствующей в актуализованном конкордансе, можно войти в закладку «Путеводитель» и попробовать поискать в этой базе данных.

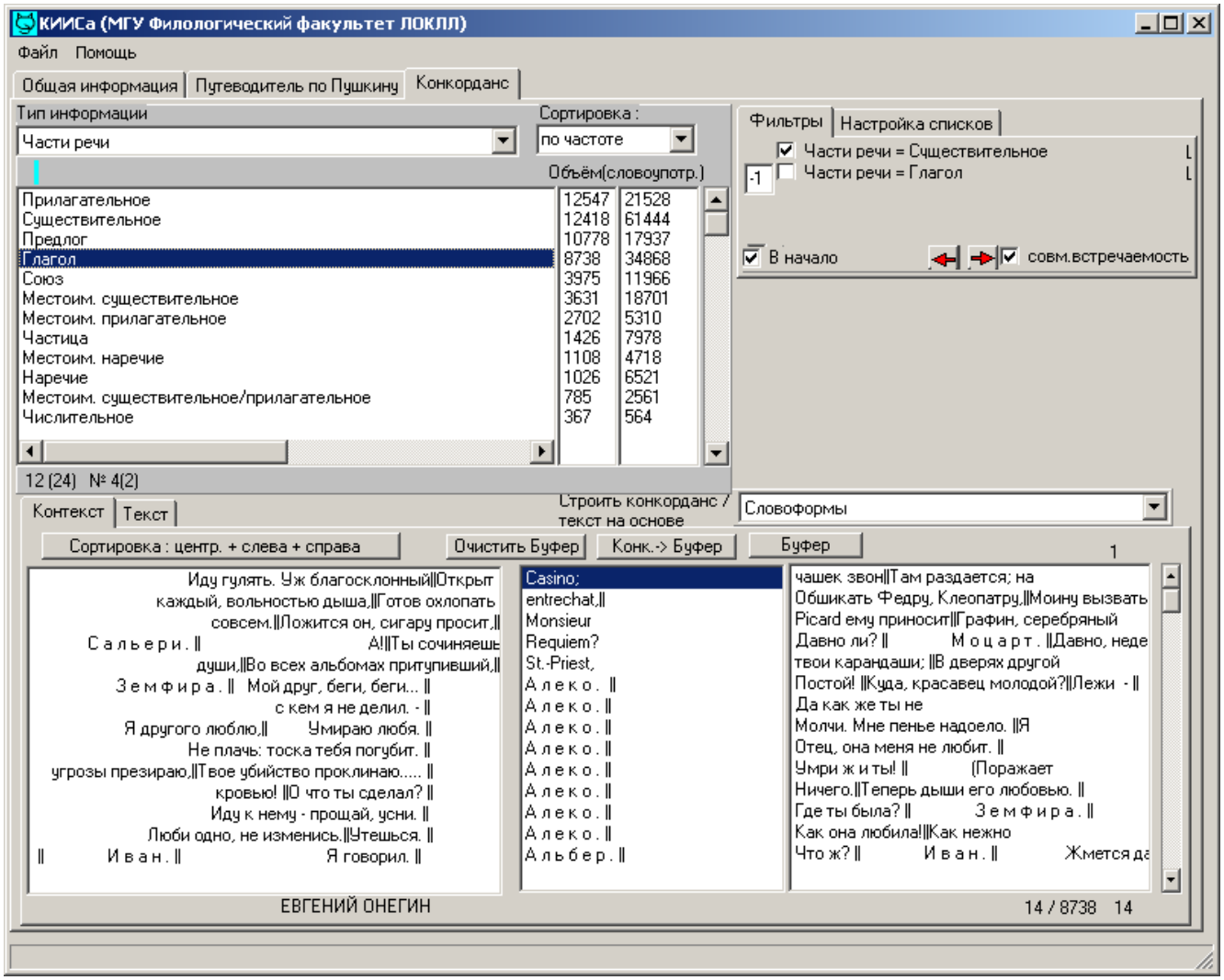
Анализ сочетаемости

В правом верхнем углу находится окно «Совместная встречаемость», позволяющее проводить анализ сочетаемости какого-либо элемента словника. При установлении в этом окне флага около фильтра возникает дополнительное окошко, в котором нужно задать позицию, сочетаемость с которой будет исследоваться. Например, введя цифру «2», можно получить все элементы, находящиеся через одно слово справа от интересующего нас элемента, а при вводе той же цифры со знаком «минус» будет получен список элементов, находящихся во второй позиции слева.

После определения позиции «соседа» необходимо выбрать исследуемый его признак (начальную форму, часть речи, семантический класс и пр.), установив фильтр на каком-либо значении нужного типа информации. В результате в окне словника появится список всех единиц нужного типа, встречающихся в заказанной позиции относительно этого элемента. Формат отображения этих единиц можно изменить, в зависимости от того, что именно нужно исследовать – сочетаемость выбранного элемента с определенными словами, частями речи и пр.

Анализ совместной встречаемости оказывается наиболее полезным при работе с такими единицами, как слова, словоформы и части речи. Ср. результат использования этой операции на рис. 3.

Рис. 3. Анализ частеречной принадлежности слов, стоящих слева от существительного.



Для получения этого вида экрана были произведены следующие действия:

1) был установлен флаг в окне «Совместная встречаемость» и значение ‘-1’ в окошке позиционирования (т.е. задан анализ позиции «первое слово слева»);
2) в «Типе информации»= «Части речи» был установлен фильтр на элементе «Существительное» (см. флаг напротив записи «Части речи=«Существительное» в закладке «Фильтры»);
3) в «Типе информации» был выбран вид «Части речи»;
4) была произведена сортировка по частоте (см. запись в окне сортировки).

В результате в окне словника отображается частотный список частей речи, стоящих слева от существительного.

Возврат к прежнему запросу

В правой части над окном контекстов находятся маленькие окошки со стрелками «<-» и «->», означающими, соответственно, «назад» и «вперед». С их помощью можно вернуться к нескольким предшествующим состояниям конкорданса.

II.2. Операции с контекстами

Переход в окно контекстов

Переход из окна словника в окно контекстов и обратно осуществляется с помощью клавиш <- и -> (на цифровой клавиатуре) или мышью. В окне контекстов слева имеются две закладки – «Контекст» и «Текст» (см. ниже).

В открывающейся по умолчанию закладке «Контекст» содержится весь набор контекстов на ту единицу словника, с которой был осуществлен выход в окно контекстов. Поскольку контексты выдаются на экран порциями и поскольку из текста можно выйти не в начало списка контекстов, а в его конец или середину, то иногда может показаться, что найдено меньше контекстов, чем нужно. В этом случае нужно использовать клавиши PgUp и PgDn, с помощью которых осуществляется обновление экрана и перемещение по контекстам.

Если в контекст попадает граница между строками, то она отображается знаком '||'.

Изменение длины контекста

Для увеличения или изменения длины контекста нужно использовать клавиши + и - (на цифровой клавиатуре). Если этого недостаточно, то можно войти в полнотекстовый режим, перейдя в закладку «Текст». Флаг «в начало» при этом не должен быть установлен, иначе на экране отобразится начало текста, а не то место, в котором находится изучаемый контекст.

Сортировка контекстов

Сортировка контекстов – удобное средство для изучения сочетаемости. Она производится с помощью кнопки «Сортировка: центр+слева+справа». После ее нажатия выпадает меню, при выборе пункта которого меняется порядок расположения контекстов. По умолчанию контексты отсортированы по алфавиту в порядке, указанном в названии кнопки сортировки контекстов, т.е. сначала центральное слово, затем слово слева, затем слово справа.

Настройка единиц конкорданса

По умолчанию в контекстах представлены **словоформы**, что позволяет видеть текст в его нормальном виде. Однако в исследовательских целях бывает полезно представить контексты в виде последовательности не реальных словоупотреблений, а лексем, частей речи и пр., т.е. в виде некоего "квазитекста". Настройка единиц представления производится в окне «Строить конкорданс» путем выбора нужного вида словника. После его выбора именно в виде единиц этого словника будут отображаться

все контексты и текст произведения. Так, выбрав «Части речи», Вы получите текст типа «*существительное прилагательное существительное, глагол…..*».

Если текст и контексты перестали отображаться нормально, нужно проверить, стоит ли в окне «Строить конкорданс» тип информации «Словоформы» и не поменялась ли эта установка в результате случайного выбора какого-то другого, неподходящего типа информации.

Получение адреса контекста

Адрес контекста – это название произведения. Он показывается внизу экрана контекстов, а также в верхнем поле этого экрана, сопровождая перемещение по контекстам. Этот адрес сохраняется по умолчанию при копировании в файл.

Если произведение большое и одного названия недостаточно, то можно уточнить адрес с помощью выбора такого типа информации, как «Части произведений».

Переход из контекста в нужное место текста

Установив курсор на нужном контексте, перейдите в закладку «Текст». Флаг «в начало» при этом **не должен** быть установлен. В этом случае выдается именно то место текста, из которого взят данный контекст. Если же флаг «в начало» установлен, то осуществится переход в начало текста произведения и место непосредственного вхождения выбранного контекста в текст будет „потеряно“.

Копирование контекстов в файл

Копирование контекстов в файл осуществляется из окна контекстов - кнопка «Конк>Буфер». Скопированный текст можно просмотреть - кнопка «Буфер» - и при необходимости сохранить в текстовый файл.

При копировании в текст записывается сопровождающая информация, которую можно настраивать (см. «Настройка сопроводительных данных»)

По умолчанию при каждом контексте записывается условие отбора, а также словоформа, по которой найден контекст, название произведения и дата. См. ниже результат поиска и контекстов с глаголом «воздвигнуть»:

```
Части речи : Глагол
Слова: воздвигнуть
**************************
Словоформы\Названия произведений\Дата
***************
```

Найдено 9 строк

****

воздвиг\КОЛЬНА (ПОДРАЖАНИЕ ОССИАНУ)\1814:

1 гробницы...||И грозным видом поражен,||Вопросит сын иноплеменный:||"Кто памятник воздвиг надменный||И старец, летами согбен,||Речет: "Тоскар наш незабвенный,||Герой

воздвиг\ПОЛТАВА\1828-1829\ :

2 В гражданстве северной державы,||В ее воинственной судьбе,||Лишь ты воздвиг, герой Полтавы,||Огромный памятник себе.||В стране - где мельниц ряд

воздвиг\Я памятник себе воздвиг нерукотворный...\1836\ :

3 Exegi monumentum. ||Я памятник себе воздвиг нерукотворный,||К нему не заростет народная тропа,||Вознесся выше он

воздвигла\КАМЕННЫЙ ГОСТЬ\1826-1830\ :

4 Г у а н . ||Так здесь похоронили командора? || М о н а х . ||Здесь; памятник жена ему воздвигла|| И приезжает каждый день сюда||За упокой души его молиться,||

Воздвигни\КОЛЬНА (ПОДРАЖАНИЕ ОССИАНУ)\1814\ :

5 черный вран стрежет.||Гряди - и там, где их не стало,|| Воздвигни памятник побед!"||Он рек, и в путь безвестный, дальный ||Пустился

воздвигну\ПОЛТАВА\1828-1829\ :

6 И скоро в смутах, в бранных спорах,||Быть может, трон воздвигну я.||Друзей надежных я имею:||Княгиня Дульская и с нею||

воздвигнув\БОРИС ГОДУНОВ\1824-1825\ :

7 его неистощимой||Проистекут источники на нас; ||И, царскую на то воздвигнув чашу,||Мы молимся тебе, царю небес. || Ш у й с к и й (пьет). ||Да здравствует

Воздвигнул\БАХЧИСАРАЙСКИЙ ФОНТАН\1821-1823\ :

8 России,||В Тавриду возвратился хан,||И в память горестной Марии || Воздвигнул мраморный фонтан, ||В углу дворца уединенный.||Над ним крестом осенена||

воздвигнутый\К ВЕЛЬМОЖЕ\1830\ :

9 вихорь бури,||Падение всего, союз ума и фурий,||Свободой грозною воздвигнутый закон,||Под гильотиною Версаль и Трианон||И мрачным ужасом смененные

II.3. Работа с текстами произведений

Переход в текст и перемещение по нему

Для перехода в текст произведения нужно перейти в закладку “Текст”. На возникшем экране появится полный текст произведения или же его фрагмент (для больших произведений). Если на экран выведен недостаточный фрагмент, нужно нажать Home для попадания в начало текста (его части или композиционного элемента) или PageUp/PageDown для перемещения выше или ниже.

Для того чтобы курсор сразу устанавливался в начало произведения (а также выбранной его части или композиционного элемента), а не на том контексте, из которого осуществлен выход в текст, поставьте флаг «в начало».

Для удобной работы в режиме просмотра текстов или их частей выполните следующие действия:

1) выберите нужный тип информации («Названия произведений», «Части произведений» или «композиционные элементы»),

2) установите флаг «в начало»;

3) откройте закладку «Текст».

Теперь по мере передвижения по названиям или другим единицам словника в нижнем окне будет появляться нужный текст с курсором, установленным в его начало. Шрифт текста можно поменять (см. ниже).

Копирование текста в файл

Возможны два способа копирования текста в файл:

1) копируется весь текст

2) копируется пословным нажатием выделенный его фрагмент (в порядке выделения).

Для копирования текста нужно нажать кнопку «Сохранить текст». Для копирования фрагмента текста нужно сначала его пословно выделить с помощью Shift+левая клавиша мыши (повторное нажатие снимает выделение)

После просмотра (кнопка «Просмотр») записанное можно сохранить в файл.

III. Работа с “Путеводителем по Пушкину”

Перейдя в закладку «Путеводитель», Вы получаете доступ к данным справочной базы, созданной на основе «Путеводителя по Пушкину» (см. Академическое издание, т. 19).

«Путеводитель по Пушкину» - замечательный энциклопедический справочник, созданный ведущими пушкинистами XX-го века (М.К.Азадовским, М.П.Алексеевым, Н.С.Ашукиным, С.М.Бонди, Д.Д.Благим, Г.А.Гуковским, В.А.Мануйловым, Л.Б.Модзалевским, М.А.Цявловским, А.М.Эфросом, Д.Б.Якубовичем и др.). Впервые он был опубликован в приложении к журналу «Красная нива» за 1931 г. и затем без изменения включен в академическое собрание сочинений. Как пишут создатели путеводителя,

«…в него, в частности, входят:

- хронологическая канва биографии А.С.Пушкина;
- статьи, посвященные всем произведениям Пушкина;
- статьи, посвященные всем персонажам, упомянутым в произведениях;

- статьи, раскрывающие все географические названия, связанные с творчеством и биографией Пушкина;
- статьи, характеризующие литературные жанры, которыми пользовался Пушкин;
- статьи, толкующие вышедшие из употребления слова пушкинского словаря» (т. 19, стр. 865).

В «Путеводитель» нами добавлен ряд ссылок, а также несколько статей, составленных на основе примечаний и указателей, имеющихся в Академическом издании. Это, в первую очередь, примечания и предисловия Пушкина к собственным произведениям, информация об авторах коллективных произведений, а также переводы на русский язык стихотворений, написанных А.С. Пушкиным на франц. языке и переводы всех других иноязычных текстов и слов, входящих в корпус.

Заглавные слова во всех таких случаях заключены в <....>. После выбора алфавитного диапазона и нужного слова появляется статья «Путеводителя». Выйти на нее можно и в режимы работы с конкордансом (в закладке «Контексты») из типов информации «Слова», «Адресаты», «Гиперслова».

Шрифт текста статьи может быть изменен (см. ниже “Настройка шрифта”).

IV. Настройка данных

В правом верхнем углу находится закладка «Настройка списков».

В ней становится доступной кнопка «Выбор списка элементов» и меню типов элементов. При нажатии кнопки открывается окно, которое позволяет уменьшить (увеличить) число доступных типов информации и изменить порядок их расположения. Находящееся рядом меню позволяет выбрать, какой именно компонент конкорданса будет настраиваться. Если нужно изменить меню словника конкорданса, то следует выбрать в меню окна „Выбор списка элементов» значение «используемых в словнике». Если изменяется меню «строить конкорданс/текст на основе», то выбирается «видимых в конкордансе». Если нужно изменить набор данных, выдаваемых в текстовый файл, то выбирается «добавляемых к конкордансу»).

По умолчанию в файл, кроме единицы, на которую получены контексты, записываются ключевая словоформа контекста, название произведения и дата его написания.

V. Настройка шрифта.

До начала работы с системой нужно установить на Ваш компьютер шрифт Kiisarf.ttf. Если он не будет установлен, то по умолчанию загрузится

стандартный шрифт Arial и текст на французском языке будет отображаться с искажениями.

Размер шрифта может быть изменен. Для этого, установив курсор на какой-либо букве в любом окне (словника, контекстов, текста или путеводителя), нажмите правую клавишу мыши. Выберите нужные настройки из стандартного окна настройки шрифта системы Windows. Шрифт может отдельно настраиваться для каждого из окон системы.

Если вы установите шрифт Pushkin.ttf, находящийся на диске в директории Font и имитирующий почерк Пушкина, то у вас появится возможность посмотреть, как примерно мог выглядеть текст, написанный пушкинской рукой. Прилагаемый шрифт разработан компанией Paratype.

V. Помощь.

Работу облегчает подробная справочная система, вызываемая через кнопку «Помощь». В ней в гипертекстовом виде и с бóльшим количеством примеров изложено все вышесказанное.

В заключение

Выбор для первого выпуска именно пушкинских текстов не случаен. Они занимают столь большое место в нашей культуре и в нашем сознании, что с необходимостью найти стихотворение с нужной строчкой, с потребностью узнать, что поэт думал по тому или иному поводу, какие явления и понятия были для него важны, сталкиваются все, начиная от школьников и заканчивая специалистами. А.С.Пушкин во многом сформировал и наш словарь, и нашу картину мира. Изучив лексический состав текстов А.С.Пушкина, основную систему понятий, используемых им, его взгляды на мир и человека, мы лучше поймем себя, свое прошлое и, возможно, будущее.

И хотя излишнее увлечение количественными подсчетами может убить любую гармонию, но без кропотливого и полного анализа текста, в том числе и количественного, трудно сделать объективные и обоснованные выводы. Мы надеемся, что наша система поможет исследователю сказать:

Поверил я алгеброй гармонию. Тогда
Уже дерзнул, в науке искушенный,
Предаться неге творческой мечты.

А.С.Пушкин, «Моцарт и Сальери»