Филиппович Ю.Н., Окунев Ю.А.

Программный комплекс для анализа текста Словаря Академии Российской 1789–1794 гг.

Введение

Цель проекта электронного издания Словаря Академии Российской 1789–1794 гг. существенно отличается от цели простой репликации его печатного переиздания в электронной форме. Она состоит в создании интегрированной инструментальной информационно-программной среды для автоматизации исследований, как самого Словаря, так и материалов его сопровождения, которая позволит представлять процесс формирования и развития русского литературного языка и его тенденций. Реализация проекта направлена на решение следующих основных научных задач: разработку модели построения несуществующих до настоящего момента словников двух изданий Словаря Академии Российской близких по времени; проведение лексико-семантического сопоставления, как самих словников, так и помет внутри словарей, что позволит определить динамику развития взглядов на проблему формы русского литературного языка; создание частных словников для решения конкретных исследовательских задач, одной из которых является выявление смены подхода к отбору терминологической, специальной и заимствованной лексики при создании словарей; выявление источниковой базы Словаря Академии Российской; введение в научный оборот электронной версии Словаря.

Актуальность проекта обосновывается следующими положениями:

Во-первых, инструментальные средства, разрабатываемые в проекте, поддерживают актуальное научное направление и конкретные исследовательские задачи, впервые направленные на информационное моделирование и лексико-семантическое сопоставление уникальных памятников русской лексикографии — первых толковых словарей русского языка — гнездового и алфавитного изданий Словаря Академии Российской.

Во-вторых, разработка современных инструментальных средств исследования словарного материала XVIII века позволит привлечь внимание к данному направлению многих ученых, в том числе и молодых, а создание словарных баз данных и информационной системы обеспечит доступность «первого детища Академии Российской» для отечественных и зарубежных ученых, студентов и аспирантов. Одно дело держать книгу в руках, другое — кроме самой книги еще иметь набор инструментов препарирования ее содержания для научных целей, иметь быстрый доступ к любому из фрагментов ее материала, решая задачи поиска нужного слова, цитаты или получения их количественной оценки.

В-третьих, компьютерная оснащенность области исследований истории языка существенно отстает от других областей лингвистики. Проект способствует сокращению существующего разрыва между количеством специальных информационных и программных средств современной и исторической компьютерной лингвистики.

Сегодня уже не требует доказательства факт, что проблема построения словарей выходит из числа чисто лингвистических, а именно лексикографических проблем. Это связано с факторами двух типов:

Во-первых, с интенсификацией процессов информатизации и компьютеризации различных отраслей гуманитарного знания и, в том числе, лексикографии. Более того, экспансия современных компьютерных технологий стимулирует и обратный процесс — все большее стремление к формализации лексикографических объектов для последующей их реализации в форме информационно-программных изделий, а именно электронных словарей и энциклопедий. Использование вычислительной техники в лексикологии и лексикографии становится не вспомогательным средством для обработки отдельных

параметров (собирание и обобщение данных), а инструментом для конструирования лексикографических объектов. Другими словами, уже не идет речь об использовании отдельных формальных методов в лексикографии, а рассматриваются возможности использования целых компьютерных технологий.

Во-вторых, с явлением так называемого «ословаривания» любого знания, представлением его в лексикографической форме, т.е. в виде словарей различного типа. Сегодня широко распространяются электронные версии самых различных словарей. Это профессионально выполненные программные продукты, которые предоставляют потенциальному потребителю максимум преимуществ, которые может дать современная компьютерная техника перед традиционным печатным изданием. Названные электронные словари и энциклопедии являются копиями классических лексикографических продуктов со встроенными мощными информационно-справочными системами. Создаются и «уникальные», узкоспециализированные собственные разработки, чаще всего имеющие форму хорошо оформленных тематических баз данных.

Из вышесказанного следует, что проблема использования современных компьютерных технологий в лексикографии остается по-прежнему актуальной и требует детального изучения.

1. Общая характеристика системы «Интерлекс» 3.0 beta

Система автоматизированного анализа естественно-языкового описания предметной области «Интерлекс» необходима для реализации возможностей вычислительной техники в рамках автоматизированной технологии построения новых лексикографических объектов в заданной предметной области. Она является основой для построения различных версий адаптированных систем, в том числе и программного комплекса анализа текста гнездового Словаря Академии Российской 1789–1794 гг., который предполагается использовать при его реконструкции в алфавитную версию — Словарь Академии Российской 1806–1822 гг. Основной особенностью разрабатываемой версии системы «Интерлекс» в этом случае будет являться ориентация на тексты XVIII – начала XIX вв., а также на существовавшие в тот период времени лексикографические источники.

Разработанная и реализованная как действующий макет системы «Интерлекс» 3.0 beta представляет собой некоторое функциональное ядро инструментальных средств обработки текстовой информации, на основе которой возможно построение различных предметно-проблемных лексикографических информационно-программных комплексов.

Функциональные характеристики системы. Данная система предоставляет следующие функциональные возможности: построение словников на основании исходных текстов и/или ранее созданных словников; частотный и динамический анализ словников; формирование словаря на основании извлеченных словоформ; создание списков определений и эксцерпций базовых слов; связывание словоформ с базовыми словами.

Требования к входным и выходным данным. Ввод данных может производиться с клавиатуры и из текстовых файлов кодировок ASCII и UNICODE. Выходные данные отображаются в виде экранных форм, а некоторые из них также могут быть сохранены на жестком диске в формате Microsoft Excel.

Требования к характеристикам технических средств и программному обеспечению. Процессор Intel Pentium III и выше. Оперативная память 256 Мб и выше. Объем свободного места на жестком диске 512 Мб и выше. Операционная система Microsoft Windows XP и выше. Виртуальная среда Microsoft Framework версия 2.0 и выше. СУБД Microsoft Access, Microsoft Excel XP.

Архитектура системы. Развертывание системы должно соответствовать структуре, представленной на рис.1.

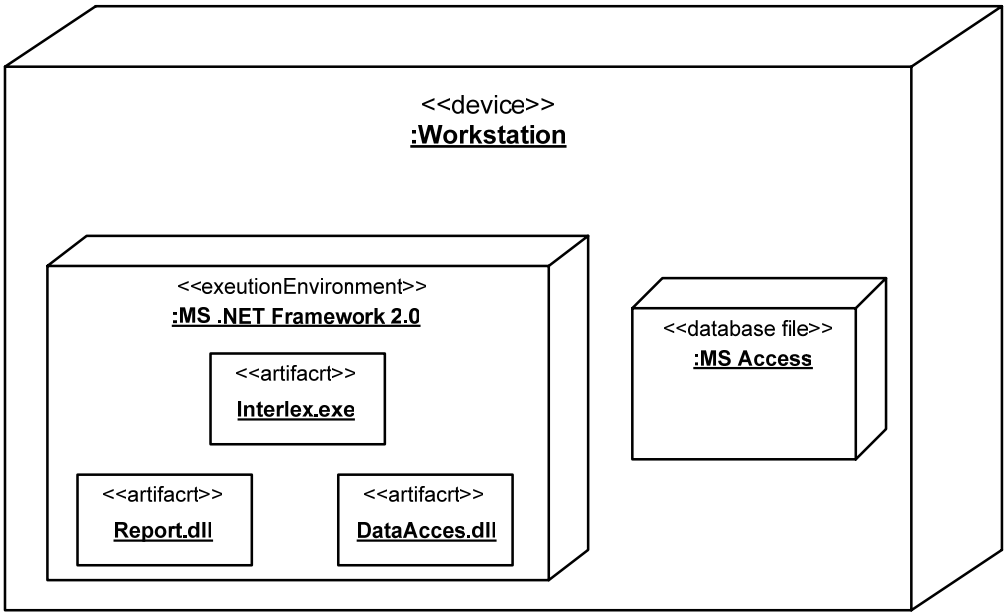


Рисунок 1. Диаграмма развертывания системы

2. Интерфейс пользователя

Интерфейс пользователя представляет собой многооконную трехуровневую систему: главное диалоговое окно системы, панели словаря базовых слов и режимов работы. Основных режимов работы три: «Словарь», «Словоформы» и «Словник». В режиме «Словник» может быть реализовано решение конкретных задач (методика или последовательность действий) построения простого и группового словников, а также проведение статистического анализа (квантитативного исследования) характеристик словников и динамики их изменения (пополнения).

Главное диалоговое окно системы. После запуска исполняемого файла interlex.exe из директории программы на экран выводится главное диалоговое окно программы.

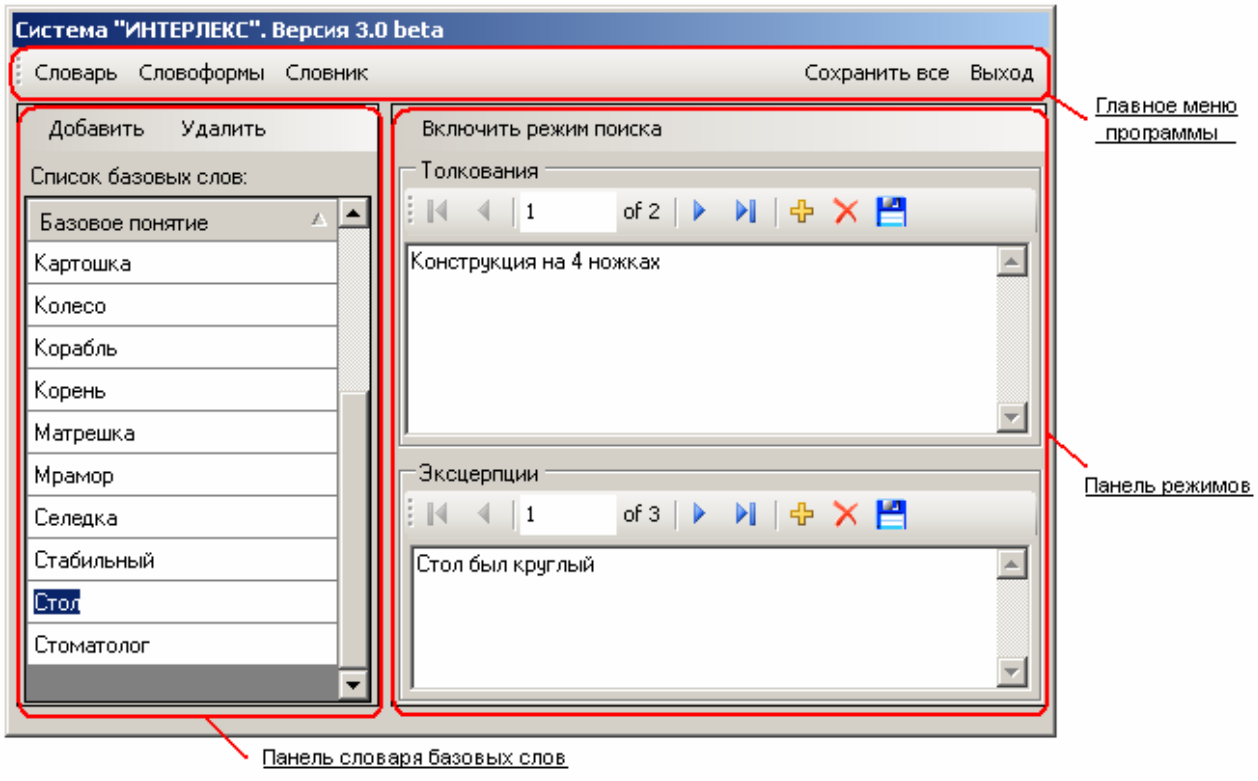


Рисунок 2. Главное диалоговое окно системы

Данное окно состоит из следующих элементов: главное меню программы, панель словаря базовых слов, панель режимов.
Главное меню программы представляет собой набор следующих кнопок:

1. Словарь	Переключение «Панели режимов» в режим «Словарь»
2. Словоформы	Переключение «Панели режимов» в режим «Словоформы»
3. Словник	Переключение «Панели режимов» в режим «Словник»
4. Сохранить все	Сохранение изменений внесенных после предыдущего нажатия данной кнопки или после запуска программы
5. Выход	Завершение работы с программой

Панель словаря базовых слов представляет собой список базовых слов, хранящихся в канонической форме (Рис.2).

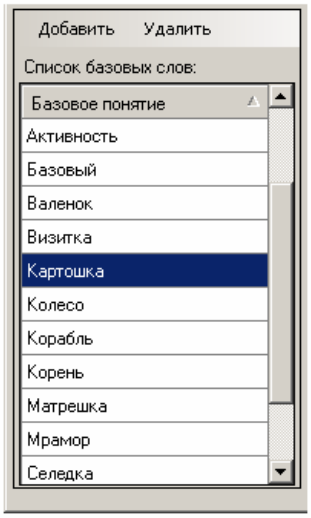


Рисунок 3. Панель словаря базовых слов.

Для добавления нового слова необходимо нажать на кнопку «Добавить». В результате чего появится следующее окно ввода (Рис.4):

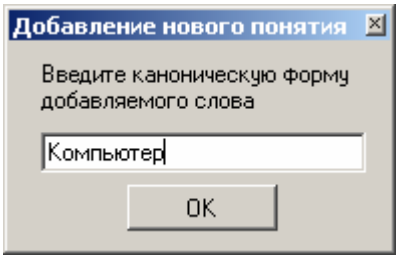


Рисунок 4. Окно добавления базового слова

Для удаления слова из словаря необходимо нажать на кнопку «Удалить».

Для изменения написания слова необходимо установить курсор в поле списка, соответствующе данному слову, и произвести ввод с клавиатуры.

Панель режимов. На данной панели возможно отображение следующих панелей: «Словарь», «Словоформы», «Словник». Переключение между панелями происходит при нажатии, соответствующей кнопки на главном меню программы. Рассмотрим каждый из режимов в отдельности.

Режим «Словарь». Окно режима «Словарь» представлено на рис. 5.

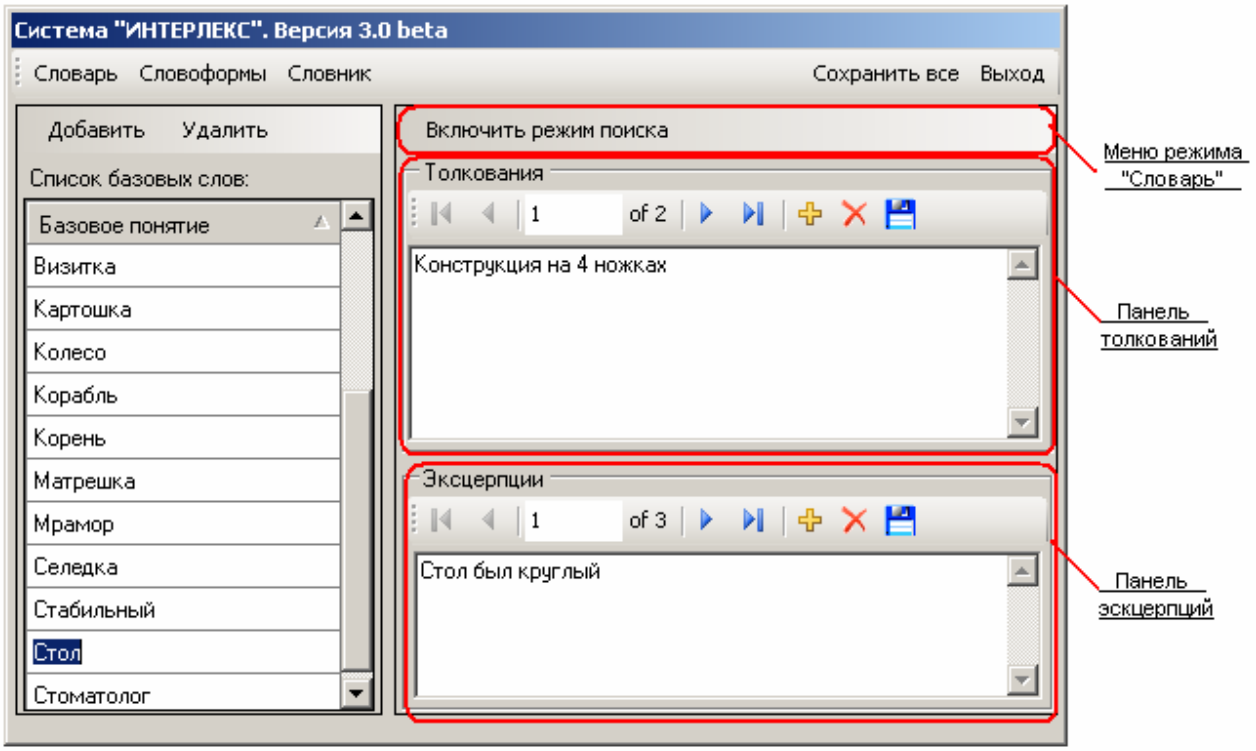


Рисунок 5. Режим «Словарь»

Данный режим состоит из трех компонент: Меню режима «Словарь», Панель толкований, Панель эксцерпций.

Меню режима «Словарь» состоит из кнопки «Включить режим поиска», при нажатии на которую выводится следующее диалоговое окно:

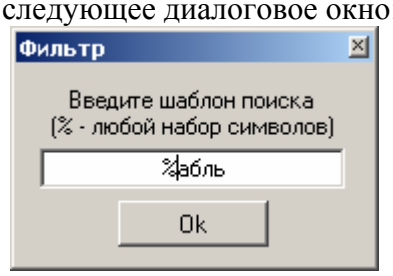


Рисунок 6. Указание фильтра в режиме поиска

В данном окне указывается фильтр, согласно которому будет произведена выборка из списка базовых слов.

Панели толкований и эскцерпций представляют собой списки толкований указанного базового понятия и примеров его использования соответственно. На каждой из этих панелей расположено следующее меню навигации, позволяющее производить перемещение по списку, в также вставку и удаление записей.

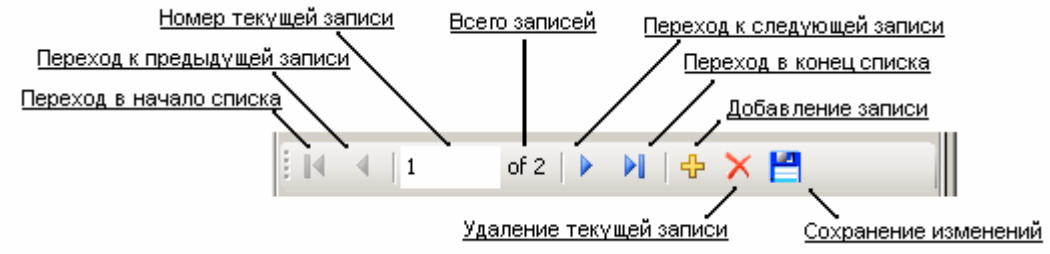


Рисунок 7. Меню навигации

Режим «Словоформы». Визуальное представление режима указано на рис.8. Режим «Словоформы» представляет собой перечень словоформ, имеющихся в базе данных системы, с указанием их соответствия канонической формы из списка базовых слов (при условии, что ранее была произведена привязка).

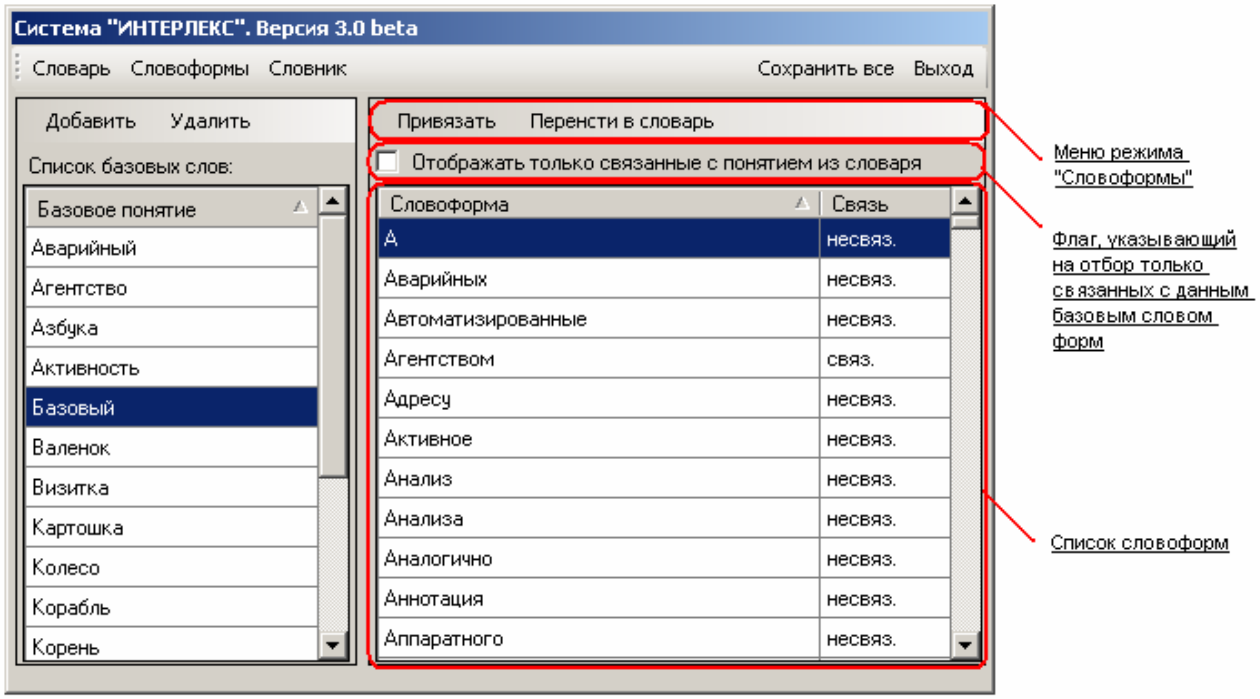


Рисунок 8. Режим «Словоформы»

Режим «Словоформы». Меню этого режима включает следующие кнопки:

«Привязать» – производит установление связи между выделенным базовым словом и выделенной словоформой.

«Перенести в словарь» – вызывает окно, указанное на рис. 6 с целью добавления канонической формы выделенной словоформы в список базовых слов.

Установление флага *«Отображать только связанные с понятием из словаря»* приводит к отображению в списке словоформ только связанных с выделенным базовым словом словоформ.

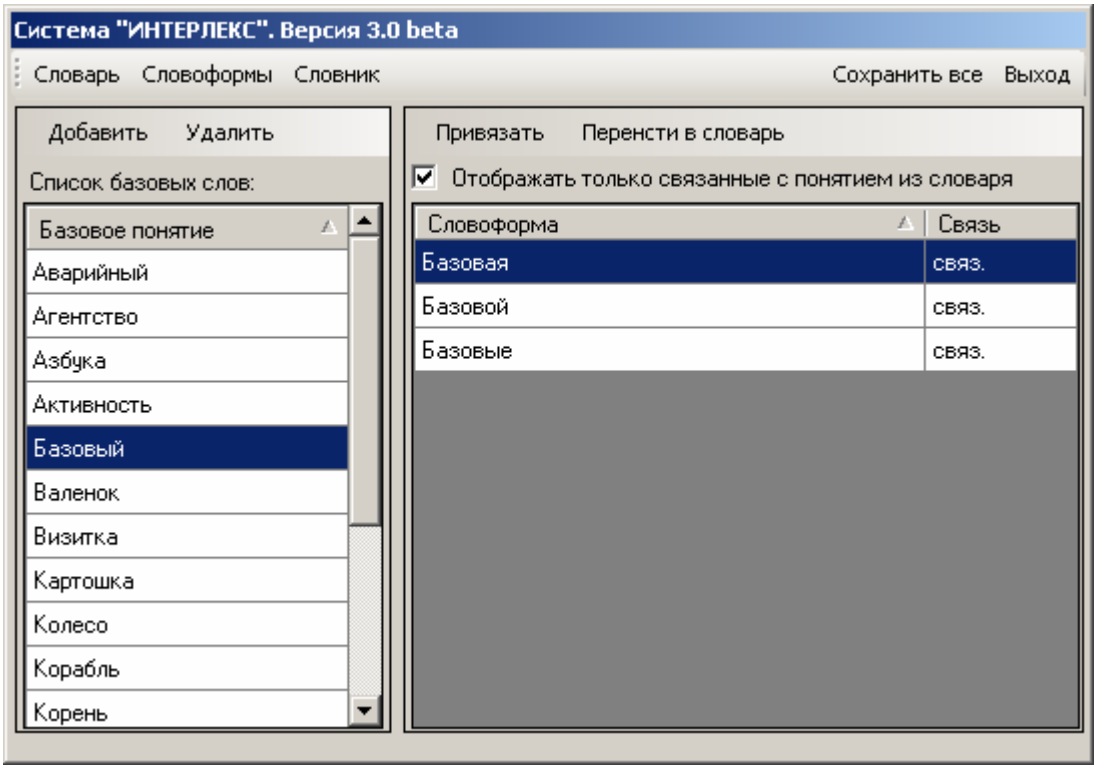


Рисунок 9. Действие флага «Отображать только связанные с понятием из словаря»

Режим «Словник». Визуальное представление режима указано на рис. 10.

Меню режима «Словник» включает следующие кнопки: «Построение» – построение словника; «Построение простого словника» – построение словника на основе текстового файла; «Построение группового словника» – построение словника на основе представленных в системе словников; «Анализ» – проведение анализа; «Динамический анализ» – анализ блока словников; «Частотный анализ» – частотный анализ выбранного словника; «Перенести в словарь» – вызывает окно, указанное на рис. 6 с целью добавления канонической формы выделенной словоформы в список базовых слов.

Список исходных текстов представляет собой псевдонимы текстов заданных при построении словника. Список словоформ и их частот представляет собой таблицу, в одном столбце которой указана словоформа, а в другом – количество данных словоформ в указанном тексте.

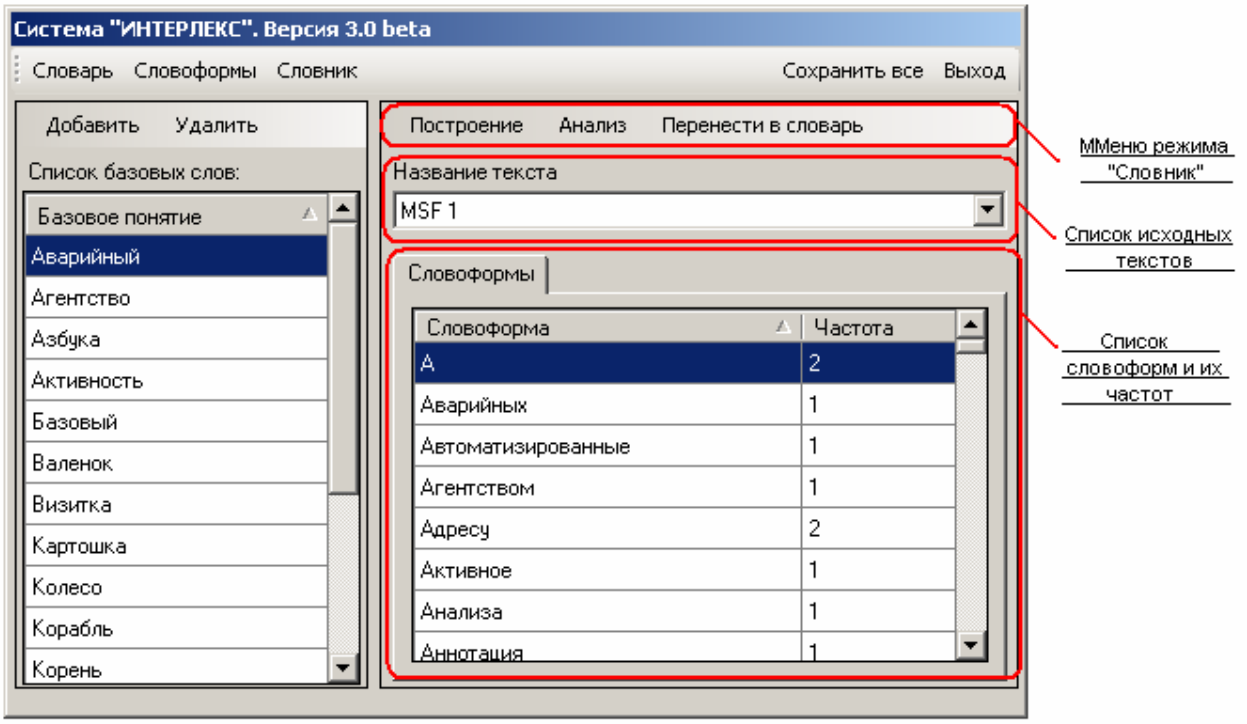


Рисунок 10. Режим «Словник»

Построение простого словника производится путем нажатия на соответствующую клавишу меню режима «Словник», которое приводит к отображению диалогового окна (рис.11). В данном диалоговом окне производится выбор текстового файла, по которому будет строиться словник, присвоение ему псевдонима, выбор кодировки исходного текстового файла, а также указание символов, которые могут являться частью словоформы.

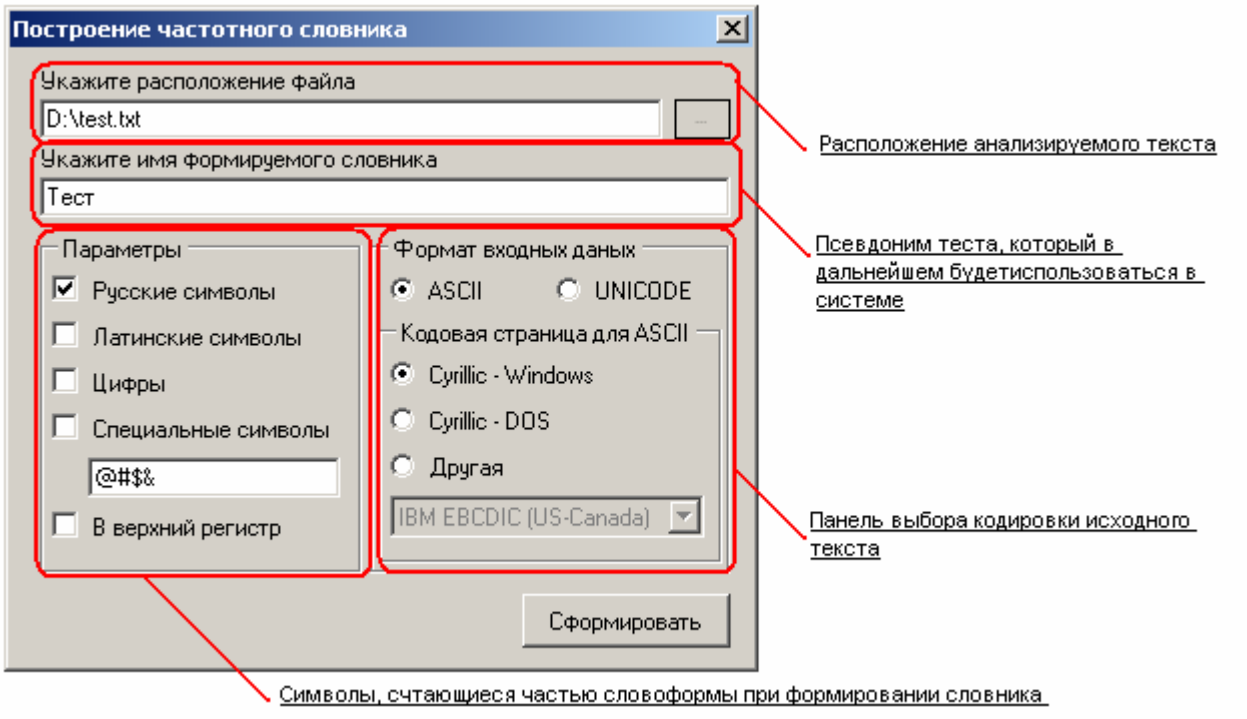


Рисунок 11. Диалоговое окно формирования словника

Построение группового словника производится путем нажатия на соответствующую клавишу меню режима «Словник», которое приводит к отображению следующего

диалогового окна, в котором путем удержания клавиши Control необходимы выделить словники, принимающие участие в построении.

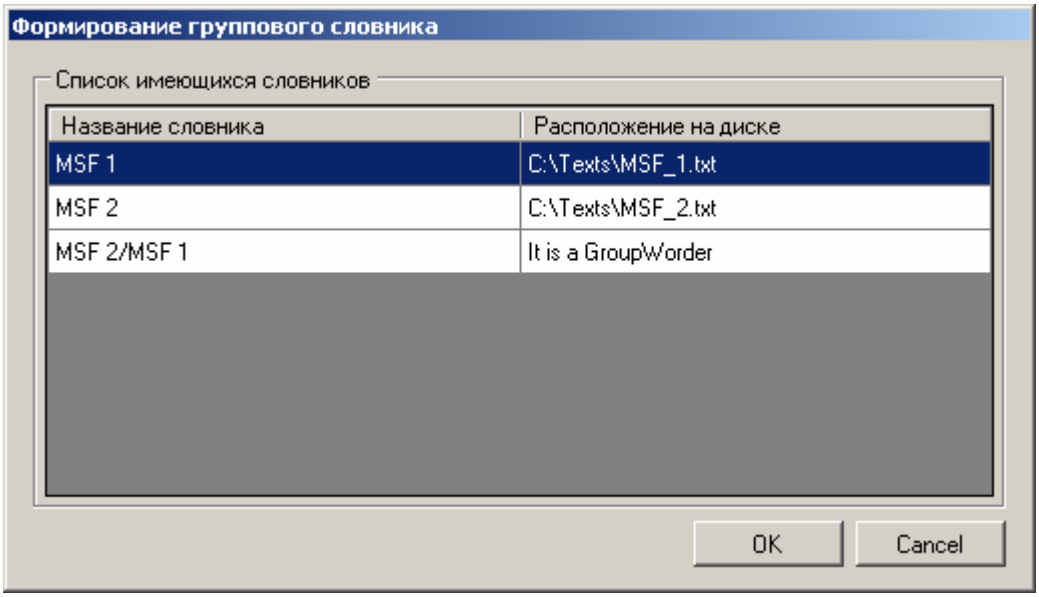


Рисунок 12. Диалоговое окно формирования группового словника

Проведение анализа. Система «Интерлекс» предусматривает проведение частотного и динамического анализов, выполнение которых производится по нажатию соответствующей кнопки. Независимо от вида анализа, пользователю необходимо указание частотных интервалов.

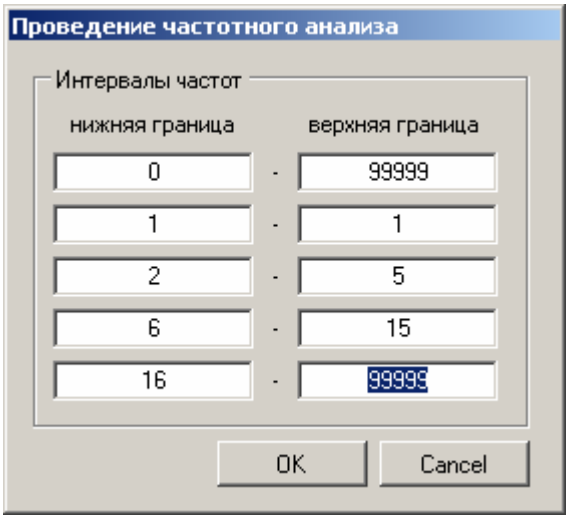
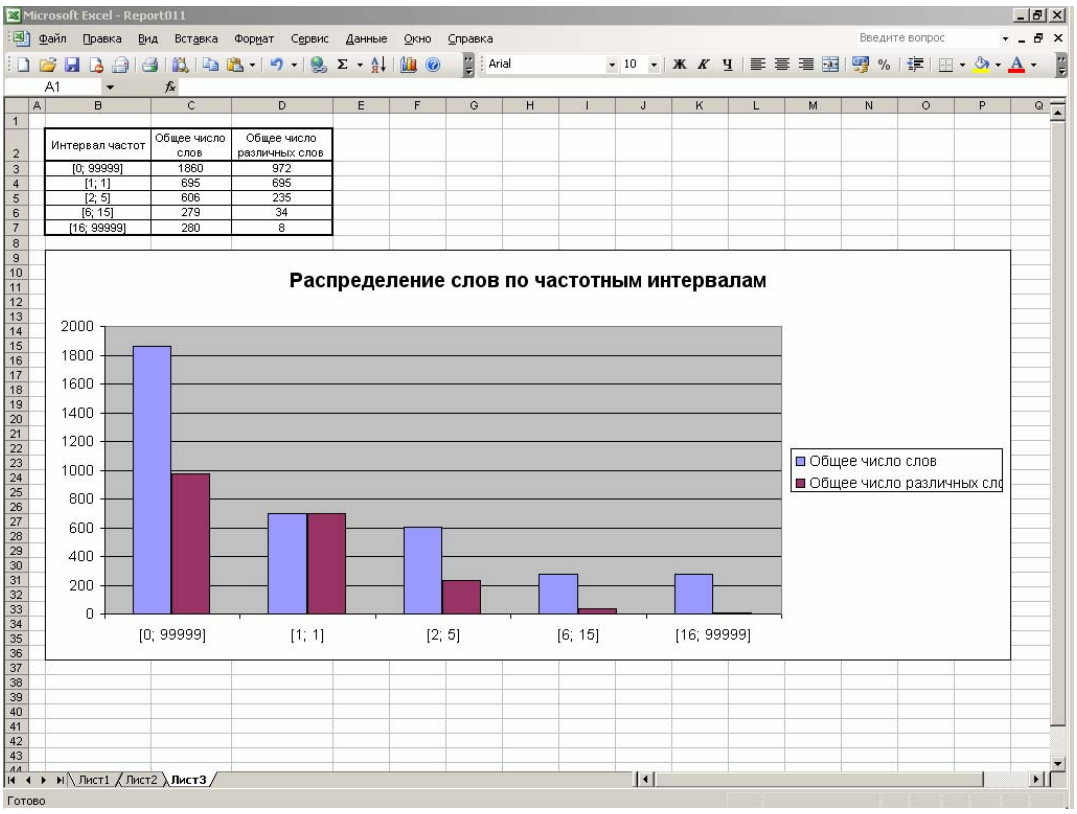


Рисунок 13. Диалоговое окно указания частотных интервалов.

В дополнение к этому, при проведении динамического анализа, пользователь должен произвести выбор словников, участвующих в нем, при помощи диалогового окна, изображенного на рис.12. В зависимости от выбранного анализа на экран будет выведен один из файлов Excel, изображенных на рис.14 и 15.



Интервал частот	Общее число слов	Общее число различных слов
[0; 99999]	1860	972
[1; 1]	695	695
[2; 5]	606	235
[6; 15]	279	34
[16; 99999]	280	8

Рисунок 14. Частотный анализ словника.

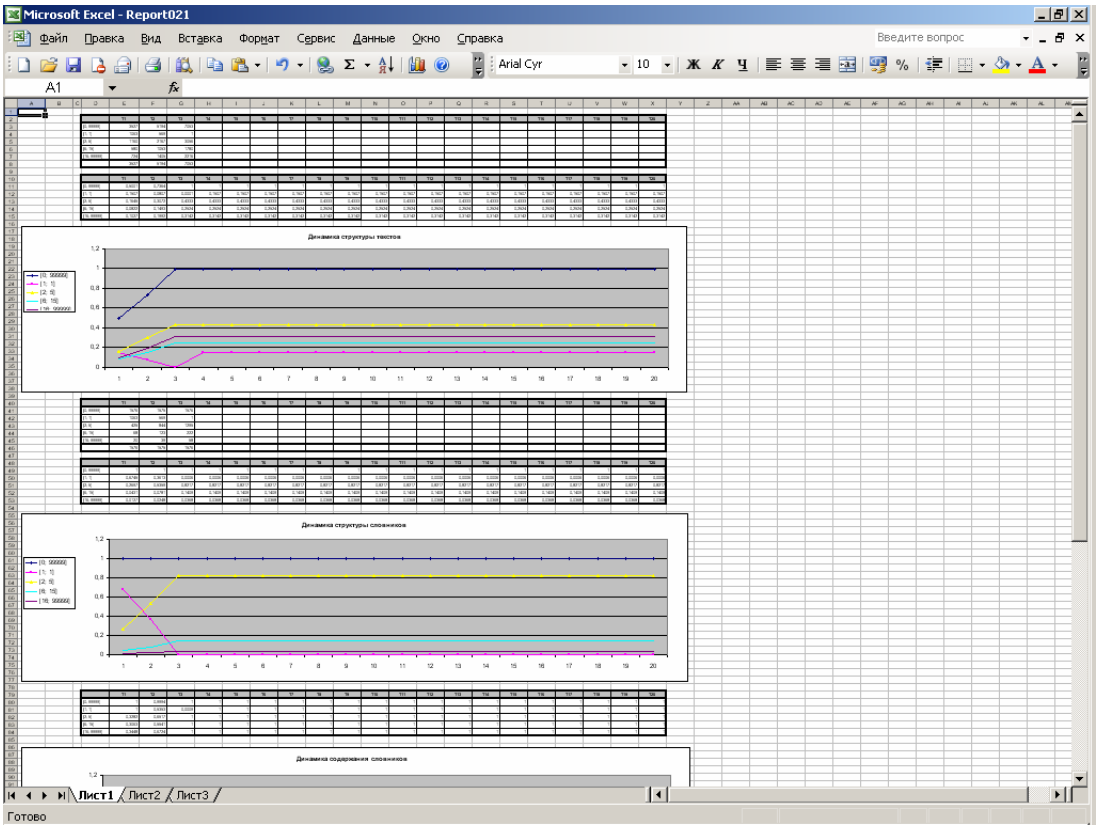


Рисунок 15. Динамический анализ словников.

Заключение

Система «Интерлекс» 3.0 beta предоставляет широкие возможности по анализу текстовых файлов различных кодировок и может послужить достойным инструментом при формировании словаря базовых слов с указанием для каждой канонической формы ее

определений, списка примеров использования, а также списка относящихся к ней словоформ. Дальнейшее развитие системы связано с расширением функциональности программного комплекса, реализация клиент-серверной архитектуры и разработка типовых исследовательских сценариев обработки данных САР.