Построение правил для автоматического извлечения словосочетаний из текста

Загорулько Максим Юрьевич
Научный руководитель н.с. ИСИ СО РАН, к.ф.-м.н. Е.А.Сидорова

Основная цель

При построении терминологических словарей важную роль играют многословные термины. Они составляют 80% всех терминов предметной области.

- Разработка алгоритмов извлечения из текста синтаксически связанных словосочетаний.

Постановка задачи

- Разработать формальное представление словосочетаний текста в виде последовательности слов, а также дерева зависимостей между словами.
- Разработать представление правил, предназначенных для автоматического извлечения словосочетаний из текста.
- Разработать словарь словосочетаний, поддерживающий эффективное извлечение словосочетаний из текста и обеспечивающий удобный доступ к его элементам.
- Разработать алгоритмы автоматического извлечения словосочетаний из текста по заданным правилам.
- Разработать пользовательский интерфейс, позволяющий лингвисту управлять процессом извлечения словосочетаний.

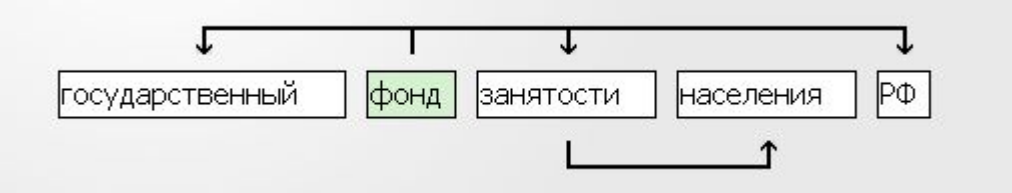
Структура словосочетания

Словосочетание – *Phrase* состоит из 4 элементов:

Phrase = <*Parts* , *Relations* , *root, title* >

Parts – Упорядоченная последовательность слов в словосочетании, где каждому ее элементу соответствует слово словосочетания в нормальной форме.

Пример 1: для словосочетания Государственный фонд занятости населения РФ



Структура словосочетания

Phrase = <Parts , Relations , root, title >

Relations – Набор пар: позиции главного и подчиненного слова, и набор морфологический признаков, по которым согласовываются подчиненное слово с главным

Пример 2: для словосочетания Институт Ядерной Физики:

Каждый из элементов Relations будет выглядеть так:

<2 (физика),1 (институт), (род, число, падеж)>

<3 (ядерный),2 (физика), (падеж - родительный)>

Структура словосочетания

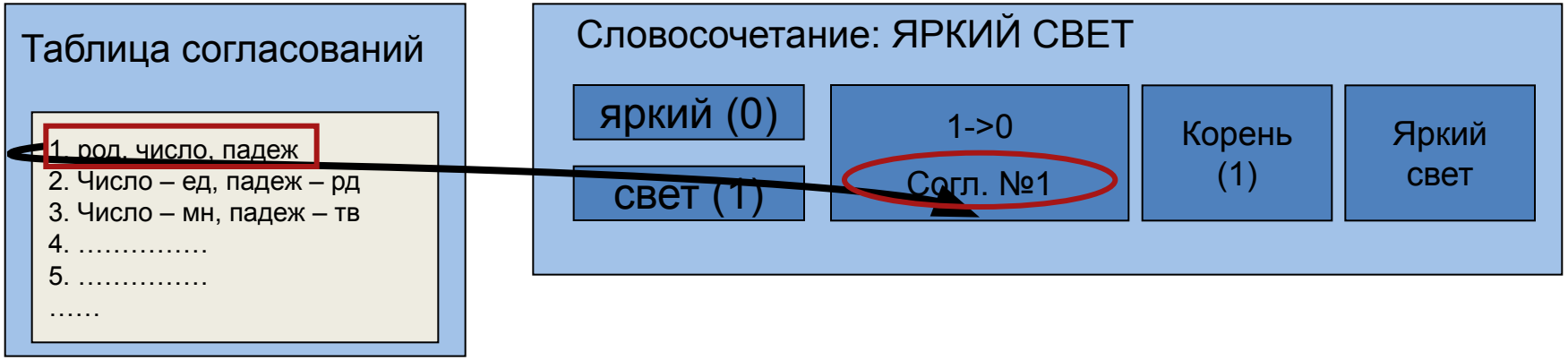
Pattern = <Words , Relations, root, title >

root - позиция корневого слова в словосочетании, то есть, является корневым, главным опорным словом.

title - наименование словосочетания.

Таблица согласований

Зачастую согласования между некоторыми частями различных правил или словосочетаний совпадают. Например очень часто встречаются такие согласования как (род, число падеж) или (падеж – родительный, число единственное). Поэтому целесообразно ввести единую таблицу согласований для всей системы.



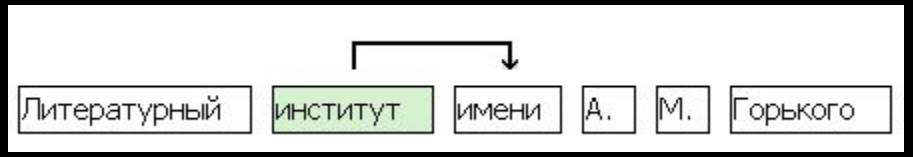
Согласование

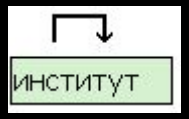
- *Морфологическое согласование* - набор параметров для согласования главного слова с подчиненным словом (падеж, род, число и пр.). Т.е. параметры, по которым необходимо осуществить согласование опорного слова данной части с зависимым словом при склонении словосочетания.

Пример 4: Новосибирск<u>ий</u> Государственный Университет.
Новосибирск<u>ому</u> Государственному Университет<u>у</u>

- *Морфологическое управление* - набор морфологических признаков и их значений, определяющих форму слову, например: «падеж=родительный», «род=мужской», «число=единственное».

Пример 5: Институт гидродинамики.
Институту гидродинамики


Литературный
институт
имени
А.
М.
Горького


институт


Литературный
институт
имени

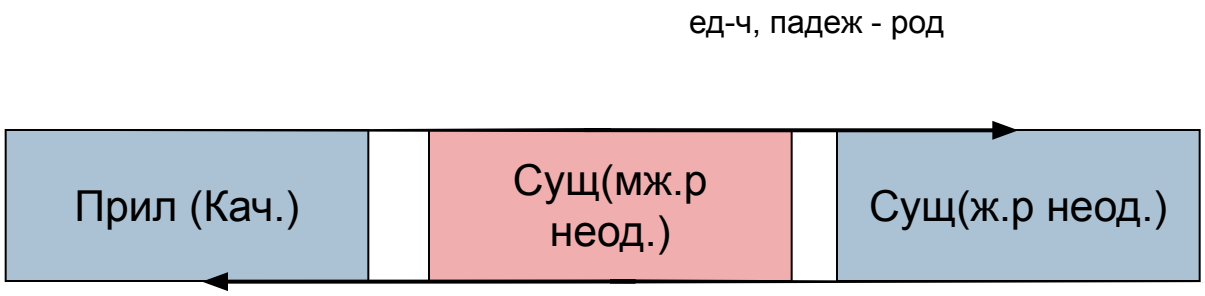
Структура правил

Правило – *Pattern* состоит из 4 элементов, по аналогии с тем как строится *Phrase, за исключением поля Parts*:

Pattern = <Parts , Relations , root, title >

Parts – Упорядоченная последовательность наборов морфологических классов.

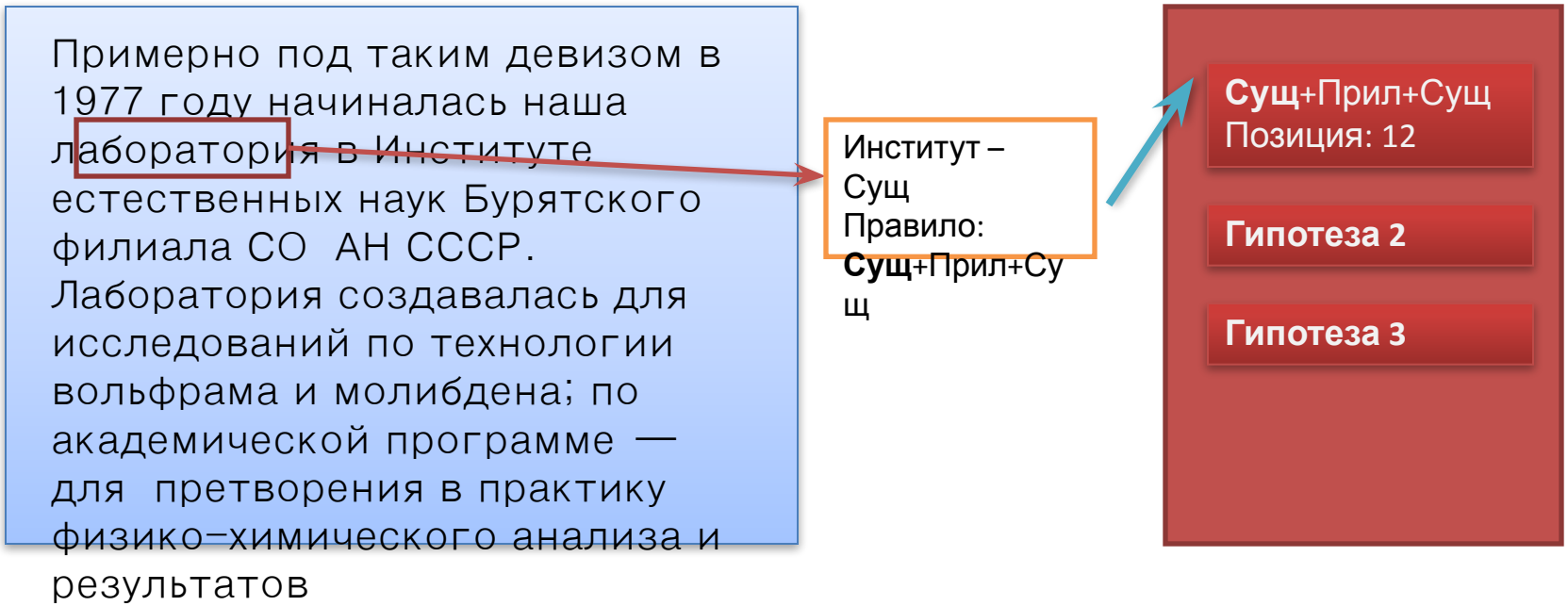
Пример 1: для правила [Сущ] + [Прил] + [Сущ] (Центральный Банк России):



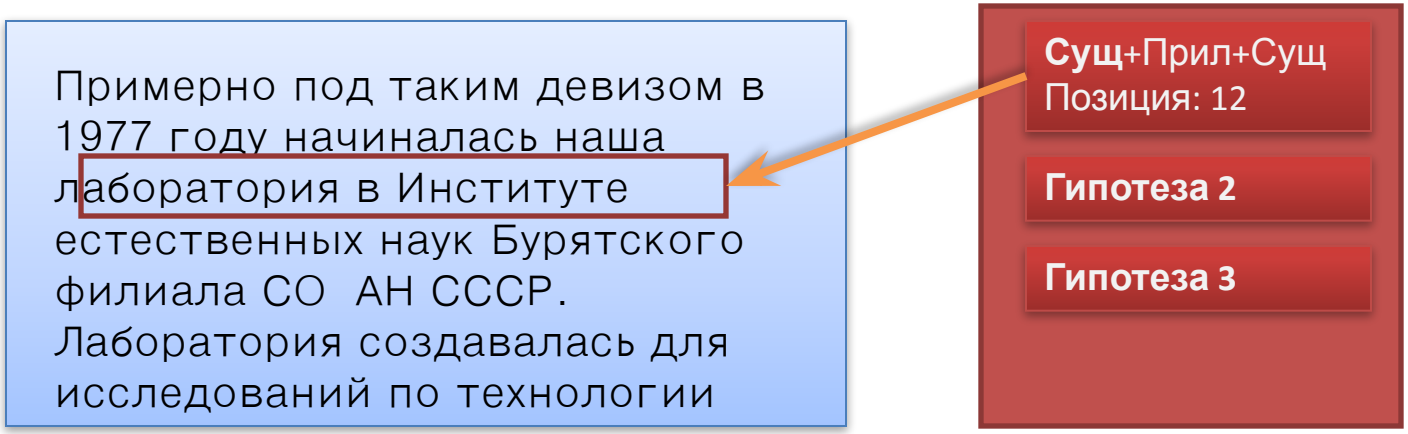
Алгоритм извлечения словосочетаний из текста

0 шаг: (1 обход текста) Составляется словарь терминов.

1 шаг: (2 обход текста) Для каждого слова текста ищем все правила, с таким же морфологическим классом корневого элемента. Запоминаем позиции данных правил и сами правила, составляем из них список гипотез.



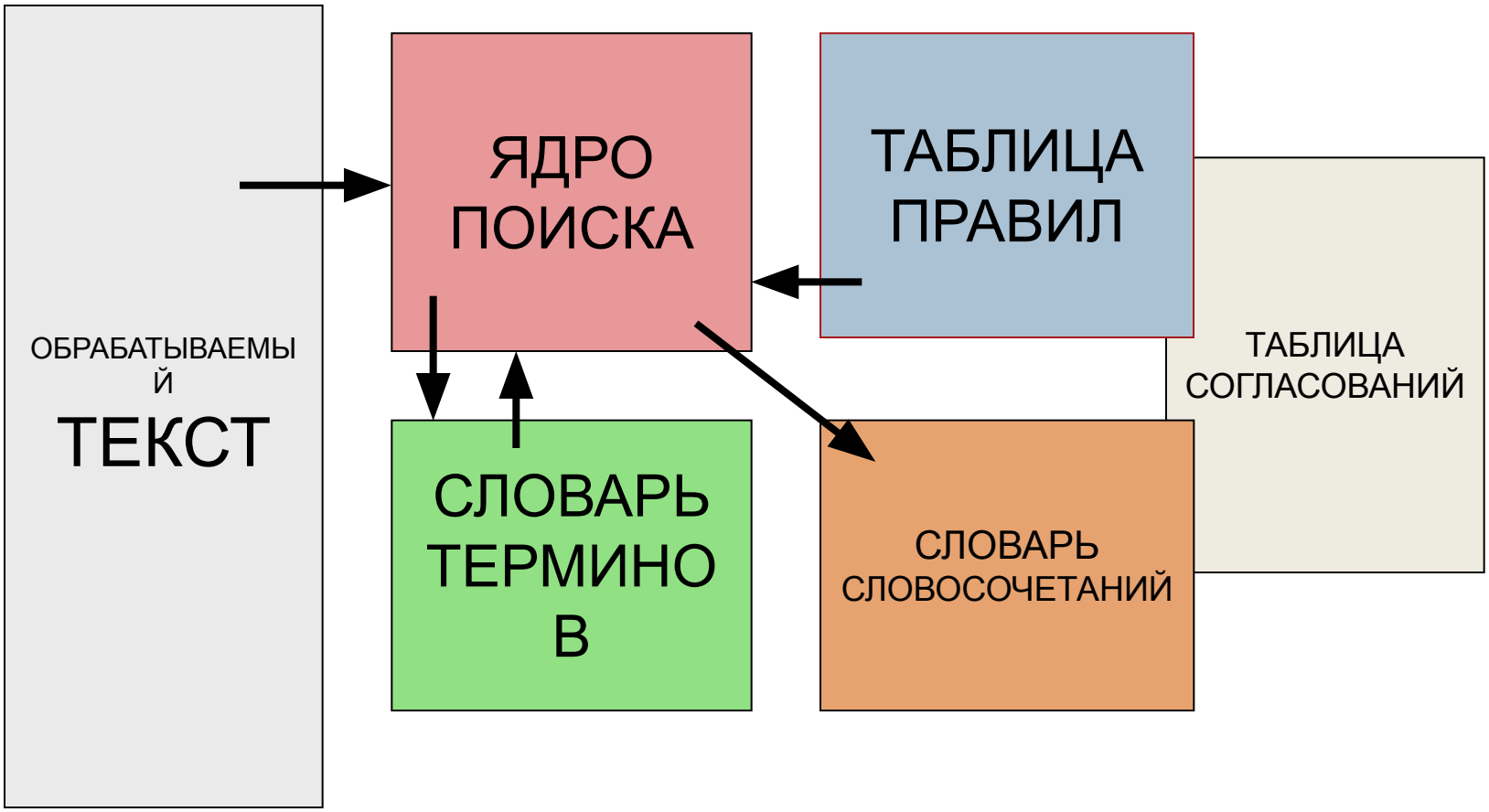
2 шаг: (3 обход текста) Для каждой гипотезы в соответствии с текущей позицией в тексте проверяем соответствие морфологического класса элемента правила и слова в тексте. Если соответствие отсутствует – удаляем гипотезу из списка.



3 шаг: Для каждой гипотезы проверяем согласование заданное в правилах. Если согласование не выполнено – удаляем гипотезу из списка.

4 шаг: На основе оставшихся гипотез формируем новое словосочетание

Извлечение словосочетаний



Словарь словосочетаний

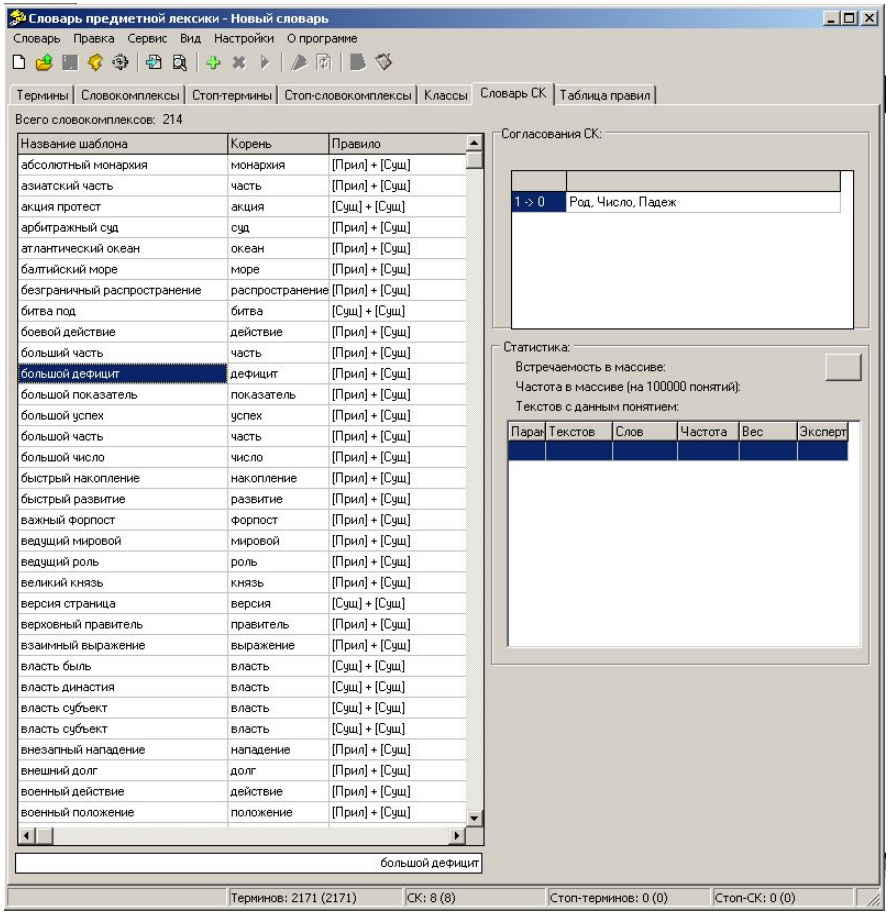
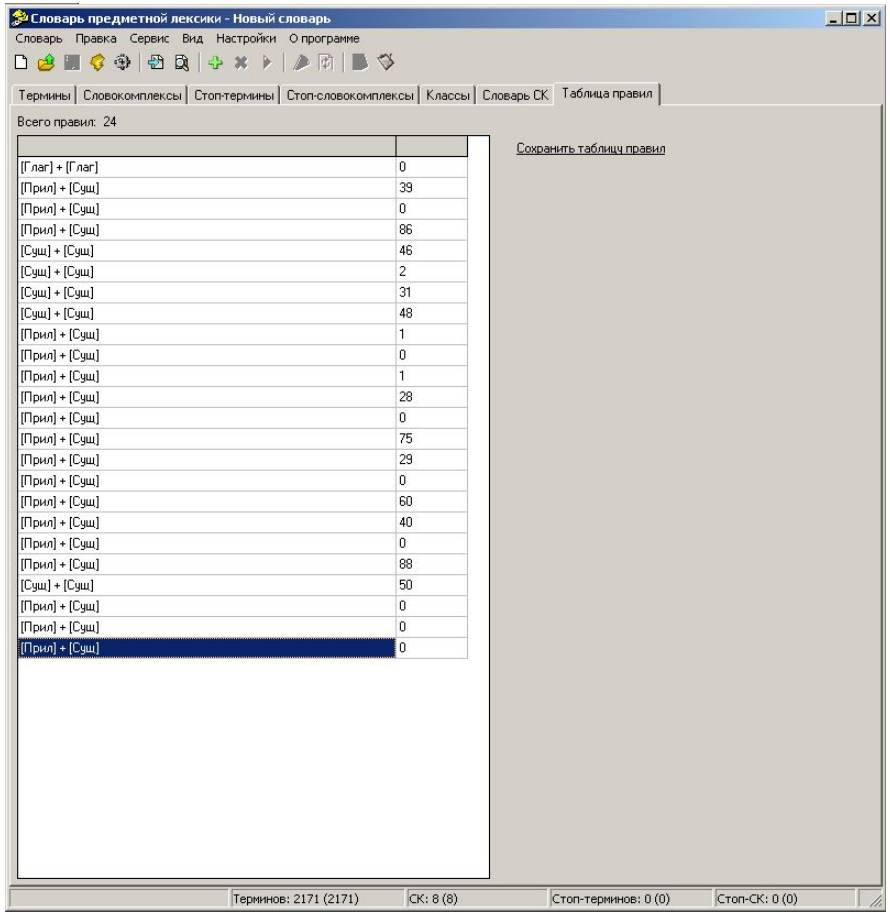
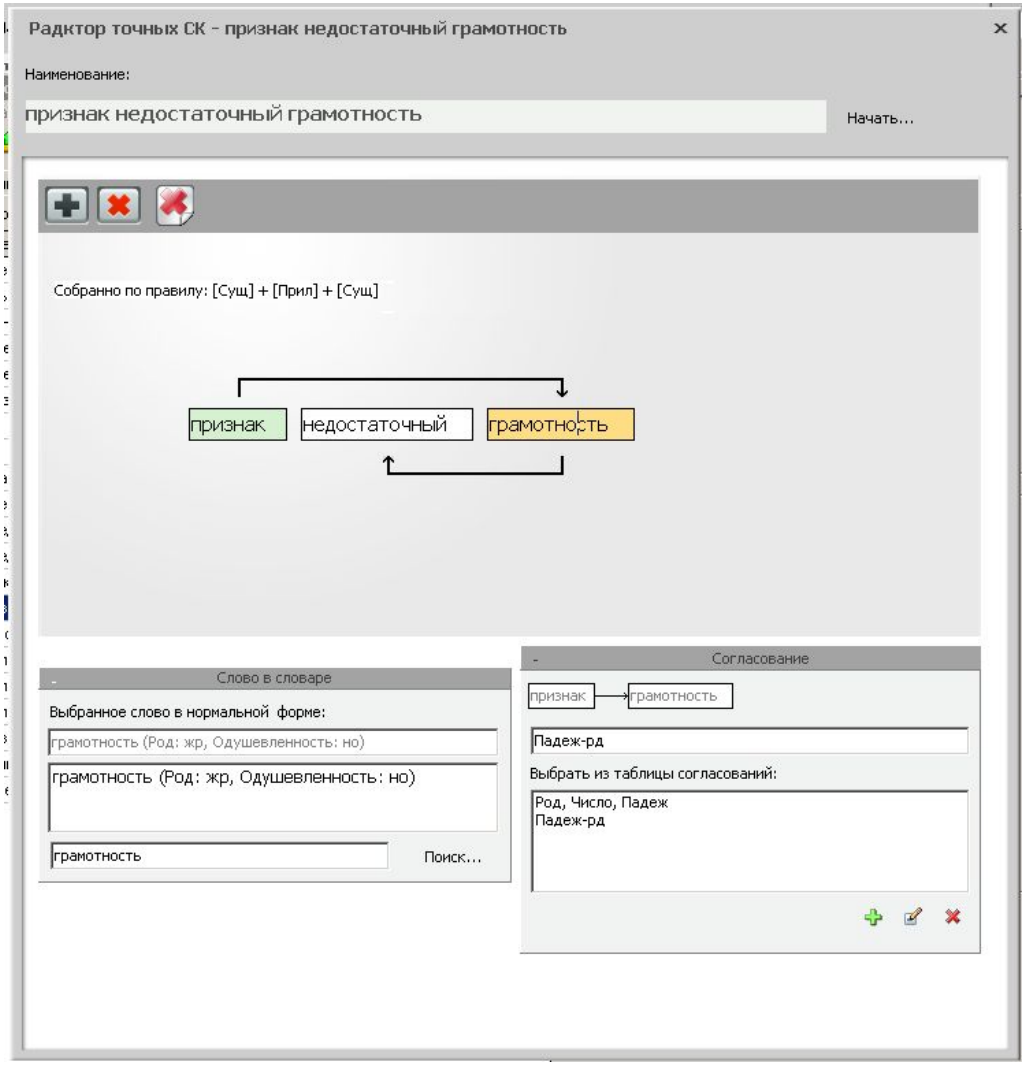


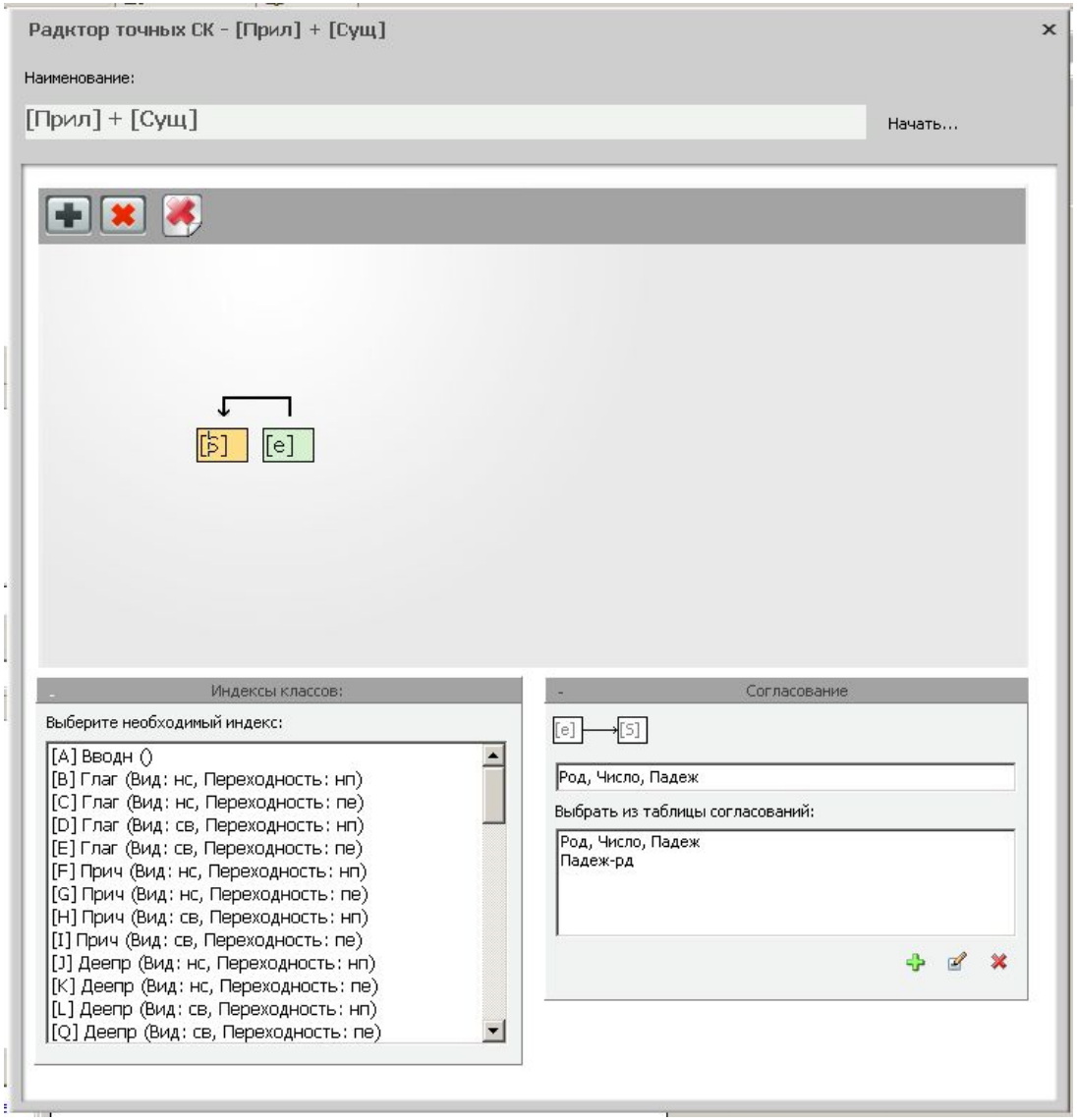
Таблица правил



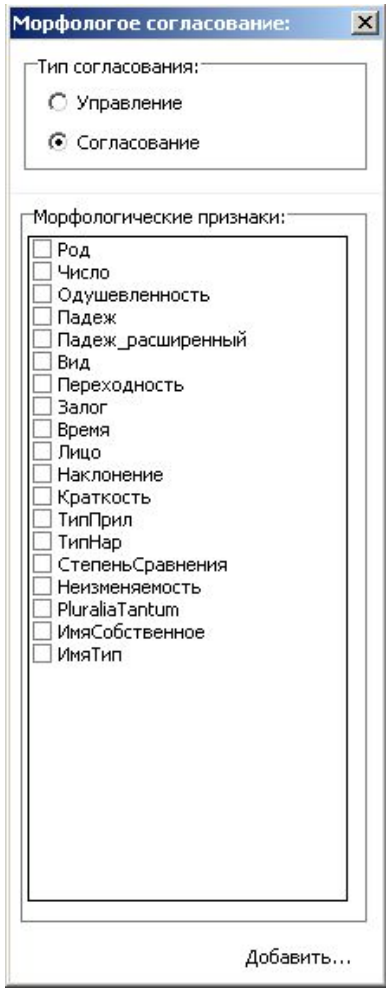
Редактор словосочетаний

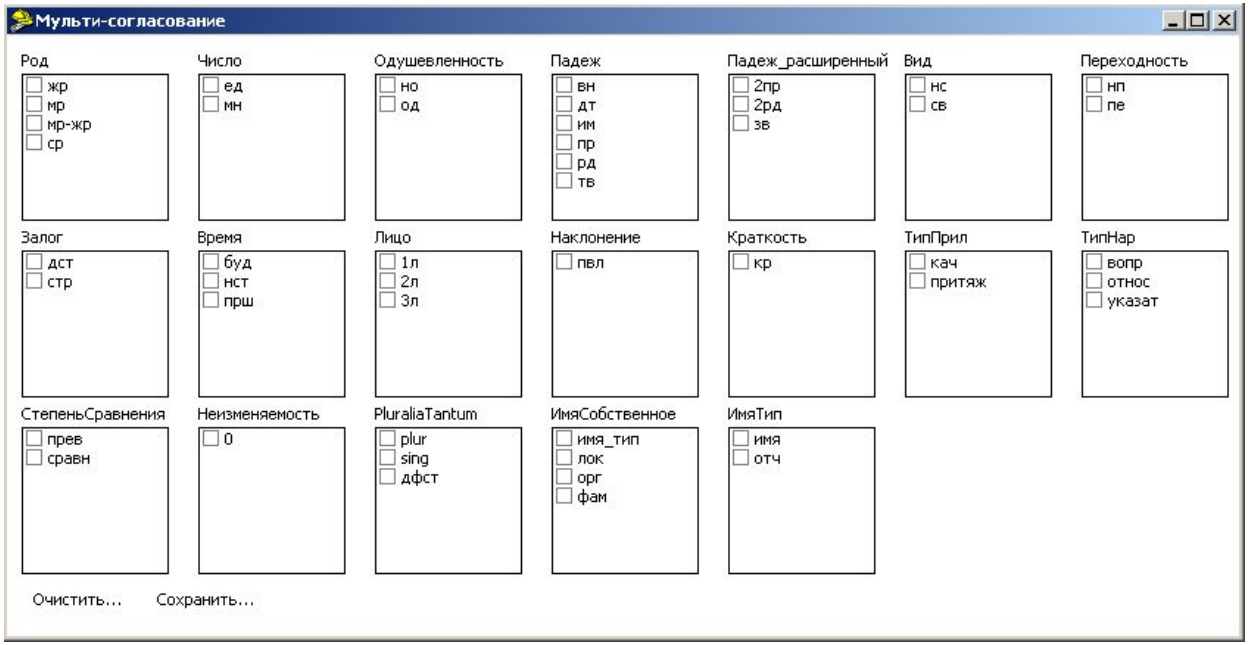


Редактор правил



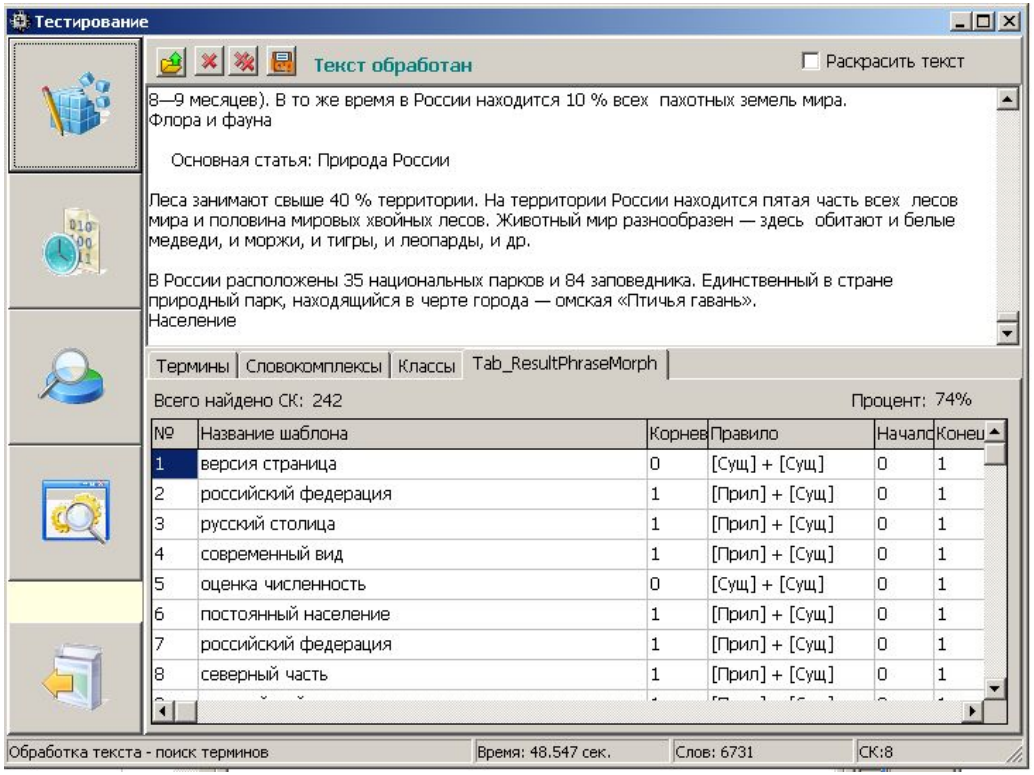
Согласование морфологических признаков.





Результаты обработки текстов

Было обработано 3 текста из разных предметных областей. Таблица правил содержала 5 основных правил



Результаты обработки текстов

Название текста	Слов в тексте	Гипотезы, прошедшие согласование	С+Ср д	С+Ст в	С+ П	П+С	С+Прил+Ср д
Отрывок из учебного пособия по гетерогенному катализу. №1	9 000	37%	539	42	12	357	69
Отрывок из учебного пособия по гетерогенному катализу. №2	19 000	38%	1167	99	39	660	84
Михаил Шолохов "Судьба человека"	7 000	40%	171	58	11	320	13

Перспективы развития

- Вложенность правил (рекурсия).
- Необязательные и альтернативные элементы.
- Синтез форм словосочетаний на основе нормальной формы.