

# 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* =  $\langle \text{Parts}, \text{Relations}, \text{root}, \text{title} \rangle$

*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* =  $\langle$ Parts , *Relations* , root, title  $\rangle$

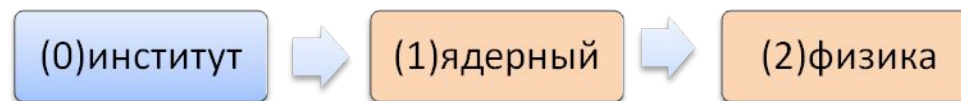
*Relations* – Набор пар: позиции главного и подчиненного слова, и набор морфологических признаков, по которым согласовываются подчиненное слово с главным

*Пример 2: для словосочетания Институт Ядерной Физики:*

*Каждый из элементов *Relations* будет выглядеть так:*

$\langle$ 2 (физика),1 (институт), (род, число, падеж) $\rangle$

$\langle$ 3 (ядерный),2 (физика), (падеж - родительный) $\rangle$



# Структура словосочетания

*Pattern* =  $\langle \text{Words} , \text{Relations} , \text{root} , \text{title} \rangle$

*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: Новосибирский Государственный Университет.*

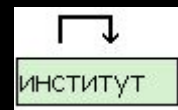
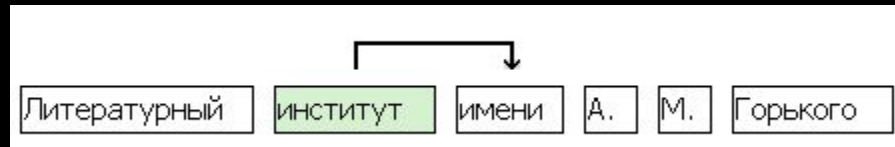
*Новосибирскому Государственному Университету*

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

*Пример 5: Институт гидродинамики.*

*Институту гидродинамики*





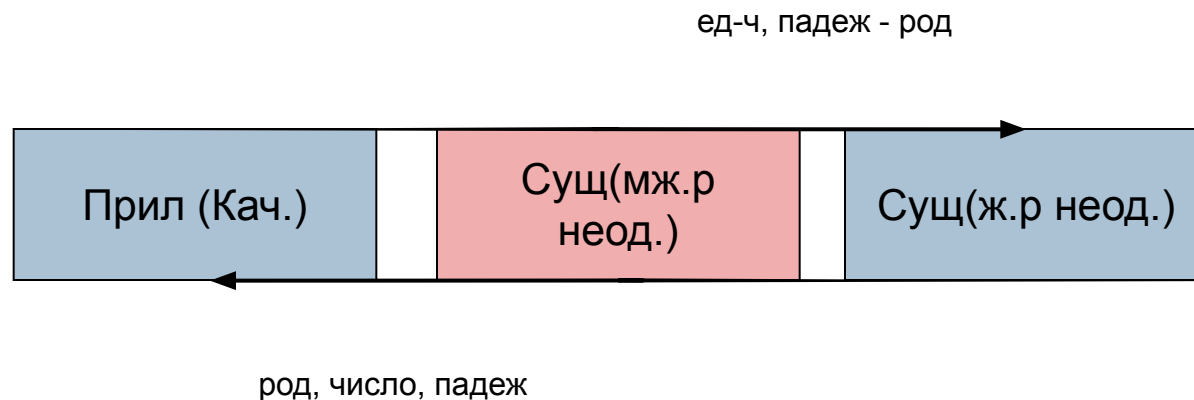
# Структура правил

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.

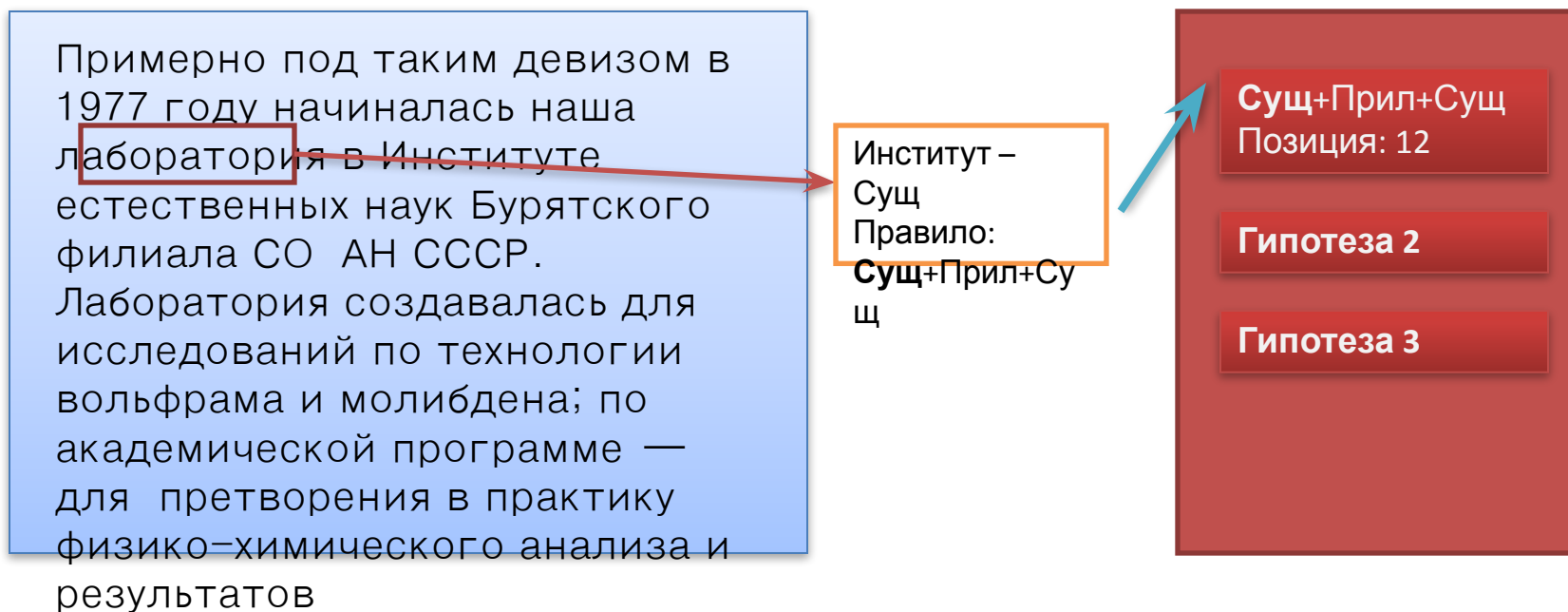
Пример 1: для правила [Сущ] + [Прил] + [Сущ] (Центральный Банк России):



# 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.

Примерно под таким девизом в 1977 году начиналась наша лаборатория в Институте естественных наук Бурятского филиала СО АН СССР. Лаборатория создавалась для исследований по технологии

Сущ+Прил+Сущ  
Позиция: 12

Гипотеза 2

Гипотеза 3

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

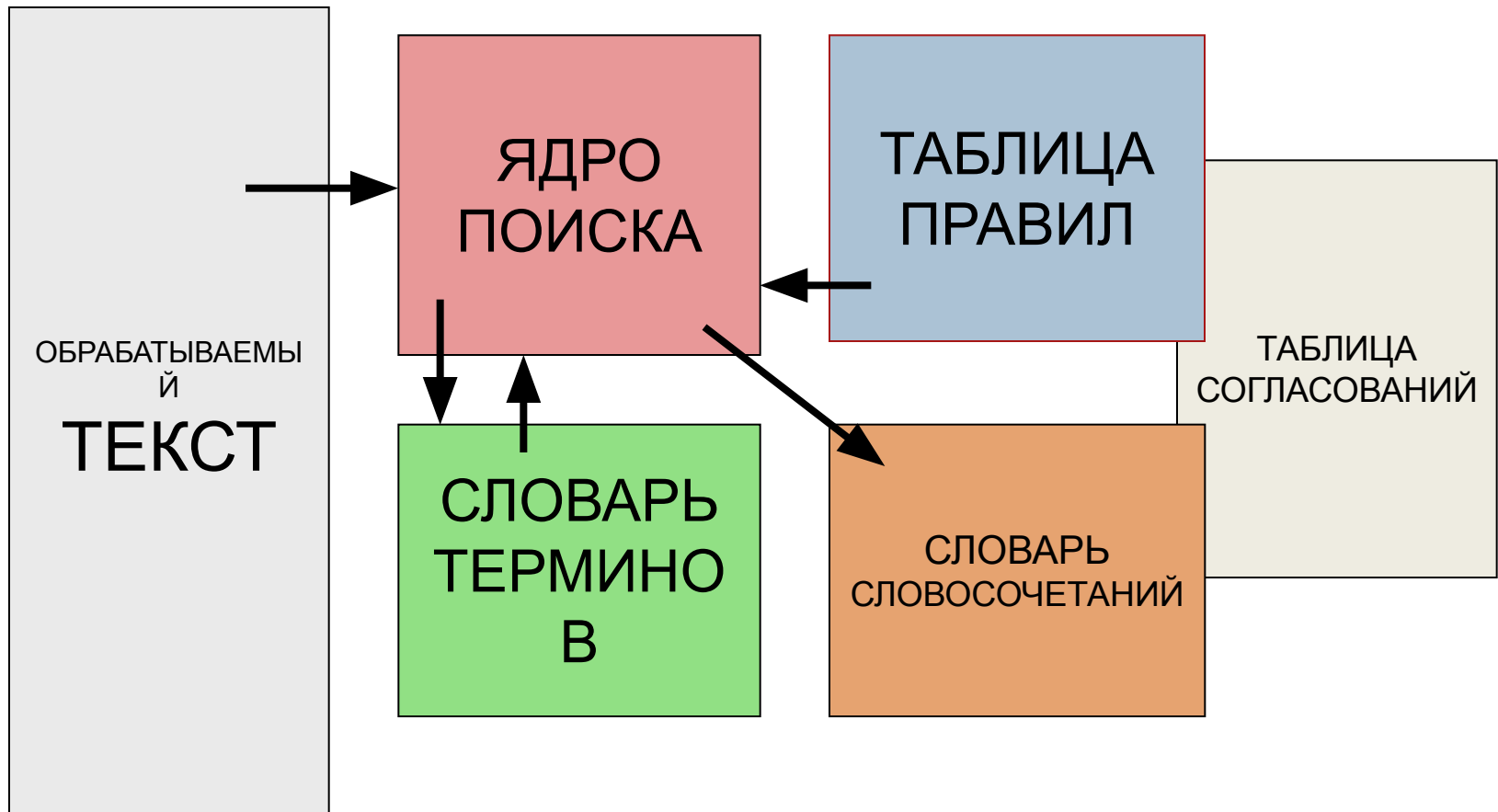
Институт естественных наук



Сущ+Прил+Сущ

**4 шаг:** На основе оставшихся гипотез формируем новое словосочетание

# Извлечение словосочетаний



# Словарь словосочетаний

Словарь предметной лексики - Новый словарь

Словарь Правка Сервис Вид Настройки О программе

Термины | Словокомплексы | Стоп-термины | Стоп-словокомплексы | Классы | Словарь СК | Таблица правил

Всего словокомплексов: 214

Название шаблона	Корень	Правило
абсолютный монархия	монархия	[Прил] + [Сущ]
азиатский часть	часть	[Прил] + [Сущ]
акция протест	акция	[Сущ] + [Сущ]
арбитражный суд	суд	[Прил] + [Сущ]
атлантический океан	океан	[Прил] + [Сущ]
балтийский море	море	[Прил] + [Сущ]
безграничный распространение	распространение	[Прил] + [Сущ]
битва под	битва	[Сущ] + [Сущ]
боевой действие	действие	[Прил] + [Сущ]
большой часть	часть	[Прил] + [Сущ]
<b>большой дефицит</b>	<b>дефицит</b>	<b>[Прил] + [Сущ]</b>
большой показатель	показатель	[Прил] + [Сущ]
большой успех	успех	[Прил] + [Сущ]
большой часть	часть	[Прил] + [Сущ]
большой число	число	[Прил] + [Сущ]
быстрый накопление	накопление	[Прил] + [Сущ]
быстрый развитие	развитие	[Прил] + [Сущ]
важный форпост	форпост	[Прил] + [Сущ]
ведущий мировой	мировой	[Прил] + [Сущ]
ведущий роль	роль	[Прил] + [Сущ]
великий князь	князь	[Прил] + [Сущ]
версия страница	версия	[Сущ] + [Сущ]
верховный правитель	правитель	[Прил] + [Сущ]
взаимный выражение	выражение	[Прил] + [Сущ]
власть быть	власть	[Сущ] + [Сущ]
власть династия	власть	[Сущ] + [Сущ]
власть субъект	власть	[Сущ] + [Сущ]
власть субъект	власть	[Сущ] + [Сущ]
внезапный нападение	нападение	[Прил] + [Сущ]
внешний долг	долг	[Прил] + [Сущ]
военный действие	действие	[Прил] + [Сущ]
военный положение	положение	[Прил] + [Сущ]

Согласования СК:

1 -> 0 Род, Число, Падеж

Статистика:

Встречаемость в массиве:  
Частота в массиве (на 100000 понятий):  
Текстов с данным понятием:

Паран	Текстов	Слов	Частота	Вес	Эксперт

большой дефицит

Терминов: 2171 (2171) СК: 8 (8) Стоп-терминов: 0 (0) Стоп-СК: 0 (0)

# Таблица правил

Словарь предметной лексики - Новый словарь

Словарь Правка Сервис Вид Настройки О программе

Термины | Словокомплексы | Стоп-термины | Стоп-словокомплексы | Классы | Словарь СК | Таблица правил

Всего правил: 24

Сохранить таблицу правил

[Глаг] + [Глаг]	0
[Прил] + [Сущ]	39
[Прил] + [Сущ]	0
[Прил] + [Сущ]	86
[Сущ] + [Сущ]	46
[Сущ] + [Сущ]	2
[Сущ] + [Сущ]	31
[Сущ] + [Сущ]	48
[Прил] + [Сущ]	1
[Прил] + [Сущ]	0
[Прил] + [Сущ]	1
[Прил] + [Сущ]	28
[Прил] + [Сущ]	0
[Прил] + [Сущ]	75
[Прил] + [Сущ]	29
[Прил] + [Сущ]	0
[Прил] + [Сущ]	60
[Прил] + [Сущ]	40
[Прил] + [Сущ]	0
[Прил] + [Сущ]	88
[Сущ] + [Сущ]	50
[Прил] + [Сущ]	0
[Прил] + [Сущ]	0
<b>[Прил] + [Сущ]</b>	<b>0</b>

Терминов: 2171 (2171) СК: 8 (8) Стоп-терминов: 0 (0) Стоп-СК: 0 (0)

# Редактор словосочетаний

Редактор точных СК - признак недостаточный грамотность

Наименование:  
признак недостаточный грамотность Начать...

Собранно по правилу: [Сущ] + [Прил] + [Сущ]

признак недостаточный грамотность

Слово в словаре

Выбранное слово в нормальной форме:  
грамотность (Род: жр, Одушевленность: но)  
грамотность (Род: жр, Одушевленность: но)  
грамотность Поиск...

Согласование

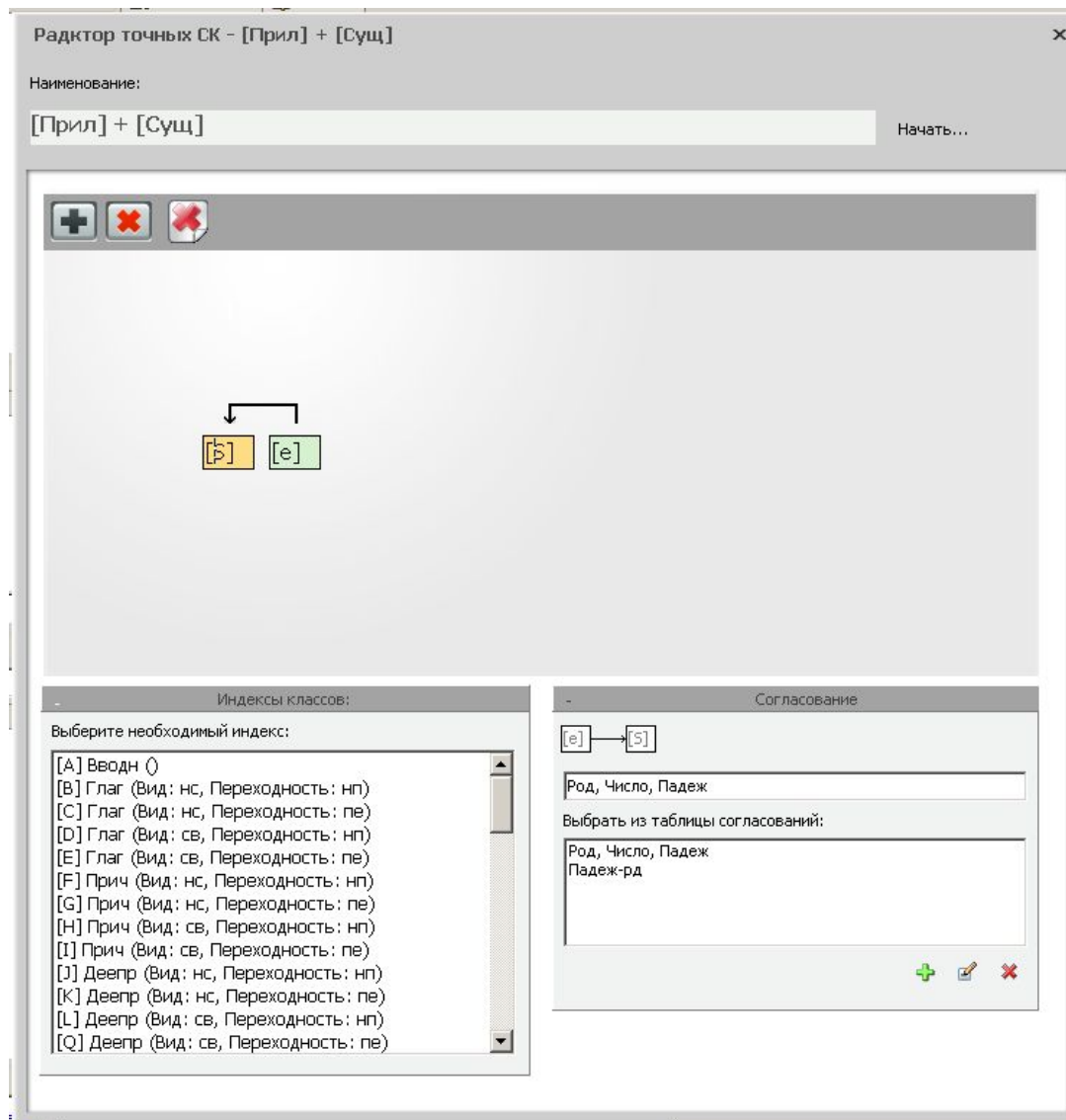
признак → грамотность

Падеж-рд

Выбрать из таблицы согласований:  
Род, Число, Падеж  
Падеж-рд

The image shows a software interface for editing word combinations. At the top, the title is 'Редактор точных СК - признак недостаточный грамотность'. Below it, there's a text field containing 'признак недостаточный грамотность' and a 'Начать...' button. A toolbar with three icons (plus, minus, and a red X) is visible. The main area displays the text 'Собранно по правилу: [Сущ] + [Прил] + [Сущ]' and a diagram of the word combination 'признак недостаточный грамотность'. The words are in separate boxes, with arrows indicating their relationship: a top arrow from 'признак' to 'грамотность', a bottom arrow from 'грамотность' to 'признак', and a middle arrow from 'недостаточный' to 'грамотность'. Below this are two panels: 'Слово в словаре' and 'Согласование'. The 'Слово в словаре' panel shows search results for 'грамотность' with grammatical information. The 'Согласование' panel shows a diagram of 'признак' and 'грамотность' with an arrow between them, and a table for selecting agreement options like 'Падеж-рд'.

# Редактор правил





# Согласование морфологических признаков.

**Морфологическое согласование:**

Тип согласования:

Управление

Согласование

Морфологические признаки:

- Род
- Число
- Одушевленность
- Падеж
- Падеж\_расширенный
- Вид
- Переходность
- Залог
- Время
- Лицо
- Наклонение
- Краткость
- ТипПрил
- ТипНар
- СтепеньСравнения
- Неизменяемость
- PluraliaTantum
- ИмяСобственное
- ИмяТип

Добавить...

**Мульти-согласование**

<b>Род</b> <input type="checkbox"/> жр <input type="checkbox"/> мр <input type="checkbox"/> мр-жр <input type="checkbox"/> ср	<b>Число</b> <input type="checkbox"/> ед <input type="checkbox"/> мн	<b>Одушевленность</b> <input type="checkbox"/> но <input type="checkbox"/> од	<b>Падеж</b> <input type="checkbox"/> вн <input type="checkbox"/> дт <input type="checkbox"/> им <input type="checkbox"/> пр <input type="checkbox"/> рд <input type="checkbox"/> тв	<b>Падеж_расширенный</b> <input type="checkbox"/> 2пр <input type="checkbox"/> 2рд <input type="checkbox"/> зв	<b>Вид</b> <input type="checkbox"/> нс <input type="checkbox"/> св	<b>Переходность</b> <input type="checkbox"/> нп <input type="checkbox"/> пе
<b>Залог</b> <input type="checkbox"/> дст <input type="checkbox"/> стр	<b>Время</b> <input type="checkbox"/> буд <input type="checkbox"/> нст <input type="checkbox"/> прш	<b>Лицо</b> <input type="checkbox"/> 1л <input type="checkbox"/> 2л <input type="checkbox"/> 3л	<b>Наклонение</b> <input type="checkbox"/> пвл	<b>Краткость</b> <input type="checkbox"/> кр	<b>ТипПрил</b> <input type="checkbox"/> кач <input type="checkbox"/> притяж	<b>ТипНар</b> <input type="checkbox"/> вопр <input type="checkbox"/> относ <input type="checkbox"/> указат
<b>СтепеньСравнения</b> <input type="checkbox"/> прев <input type="checkbox"/> сравн	<b>Неизменяемость</b> <input type="checkbox"/> 0	<b>PluraliaTantum</b> <input type="checkbox"/> plur <input type="checkbox"/> sing <input type="checkbox"/> дфст	<b>ИмяСобственное</b> <input type="checkbox"/> имя_тип <input type="checkbox"/> лок <input type="checkbox"/> орг <input type="checkbox"/> фам	<b>ИмяТип</b> <input type="checkbox"/> имя <input type="checkbox"/> отч		

Очистить... Сохранить...

# Результаты обработки текстов

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

Тестирование

Текст обработан  Раскрыть текст

8—9 месяцев). В то же время в России находится 10 % всех пахотных земель мира. Флора и фауна

Основная статья: Природа России

Леса занимают свыше 40 % территории. На территории России находится пятая часть всех лесов мира и половина мировых хвойных лесов. Животный мир разнообразен — здесь обитают и белые медведи, и моржи, и тигры, и леопарды, и др.

В России расположены 35 национальных парков и 84 заповедника. Единственный в стране природный парк, находящийся в черте города — омская «Птичья гавань».

Население

Термины | Словосочетания | Классы | Tab\_ResultPhraseMorph

Всего найдено СК: 242 Процент: 74%

№	Название шаблона	Корнев	Правило	Начало	Конец
1	версия страница	0	[Суц] + [Суц]	0	1
2	российский федерация	1	[Прил] + [Суц]	0	1
3	русский столица	1	[Прил] + [Суц]	0	1
4	современный вид	1	[Прил] + [Суц]	0	1
5	оценка численность	0	[Суц] + [Суц]	0	1
6	постоянный население	1	[Прил] + [Суц]	0	1
7	российский федерация	1	[Прил] + [Суц]	0	1
8	северный часть	1	[Прил] + [Суц]	0	1

Обработка текста - поиск терминов Время: 48.547 сек. Слов: 6731 СК:8

# Результаты обработки текстов

Название текста	Слов в тексте	Гипотезы, прошедшие согласование	С+Срд	С+Ств	С+П	П+С	С+Прил+Срд
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.