Открытый лингвопроцессор и его применения в прикладных системах обработки текстов

А.В. Добров,
СПбГУ,
ООО «Гелайн»

Функциональность

Лингвопроцессор – это система, осуществляющая автоматическую обработку текстов на естественном языке

Открытый лингвопроцессор – это лингвопроцессор с открытым исходным кодом, опубликованным и доступным для загрузки под свободной лицензией

Функциональность

Лингвопроцессор – это система, осуществляющая автоматическую обработку текстов на естественном языке

Лингвопроцессоры используются в задачах информационного поиска, машинного перевода, классификации текстов и т.д.

Открытый лингвопроцессор – это лингвопроцессор с открытым исходным кодом, опубликованным и доступным для загрузки под свободной лицензией

Открытые лингвопроцессоры, как правило, не являются самостоятельными продуктами и ограничены по функциональности морфологическим и, в редких случаях, поверхностно-синтаксическим анализом

Функциональность

Лингвопроцессор – это система, осуществляющая автоматическую обработку текстов на естественном языке

AIIRE – это система, способная выполнять лингвистический анализ электронных текстов на всех языковых уровнях

Открытый лингвопроцессор – это лингвопроцессор с открытым исходным кодом, опубликованным и доступным для загрузки под свободной лицензией

AIIRE опубликован под лицензией GNU GPL, включен в некоторые репозитории и дистрибутивы ОС GNU/Linux, доступен для загрузки с веб-ресурсов дистрибутивов НауЛинукс и CentOS

Функциональность

AIIRE – это система, способная выполнять лингвистический анализ текстов на всех уровнях языка

- ✓ *Графематический анализ* (обработка символов и их сочетаний)
- ✓ *Морфологический анализ* (обработка словоформ)
- ✓ *Синтаксический анализ* (построение комбинированных структур составляющих и зависимостей)

А.С. Герд: Семантика пронизывает все уровни языка и потому не представляет собой отдельного уровня

Функциональность

А.С. Герд: Семантика пронизывает все уровни языка и потому не представляет собой отдельного уровня

- *Графематический уровень*: семантическая обработка небуквенных символов – цифр, знаков пунктуации и т.д.
- *Морфологический уровень*: загрузка лексических значений лемм из онтологии
- *Синтаксический уровень*: вычисление семантики предложений и высказываний (прагматика)

Функциональность

На входе – текст, на выходе – гипотезы синтаксического разбора и семантического представления содержания этого текста

Принцип работы

На входе – текст:

Он увидел их семью своими глазами.

Принцип работы

Распознаются атомарные единицы (в данном случае – словоформы, хотя это и необязательно)

Он увидел их семью своими глазами.

Принцип работы

Распознаются атомарные единицы (в данном случае – словоформы, хотя это и необязательно)

Он

Он увидел их семью своими глазами.

Принцип работы

Распознаются атомарные единицы (в данном случае – словоформы, хотя это и необязательно)

Он увидел

Он увидел их семью своими глазами.

Принцип работы

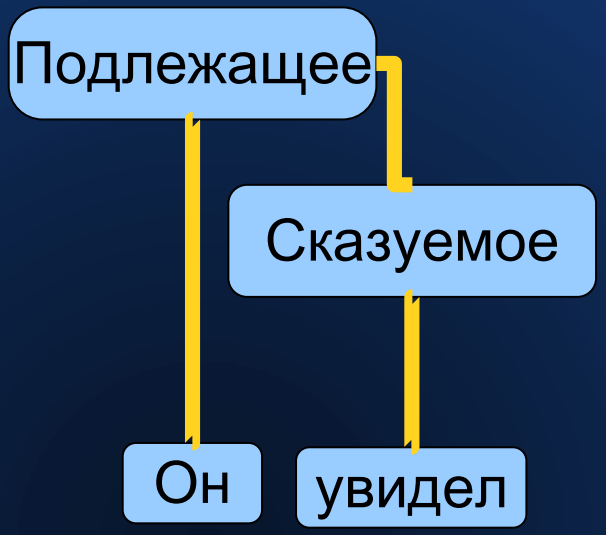
Каждую распознанную единицу система пытается связать с линейно предшествующими ей единицами:

Он увидел

Он увидел их семью своими глазами.

Принцип работы

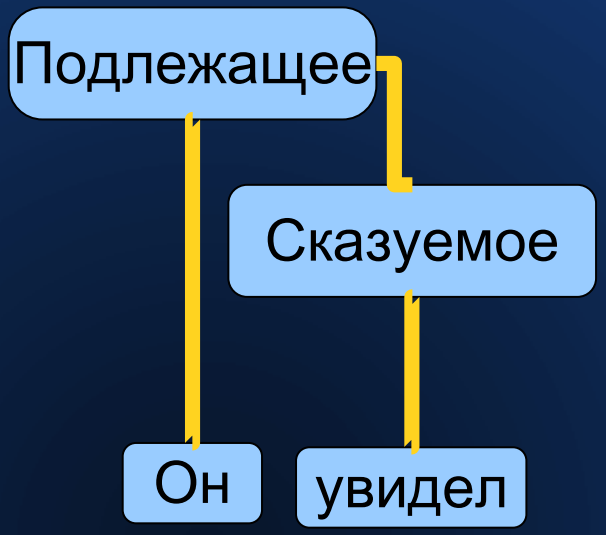
Каждую распознанную единицу система пытается связать с линейно предшествующими ей единицами:



Он увидел их семью своими глазами.

Принцип работы

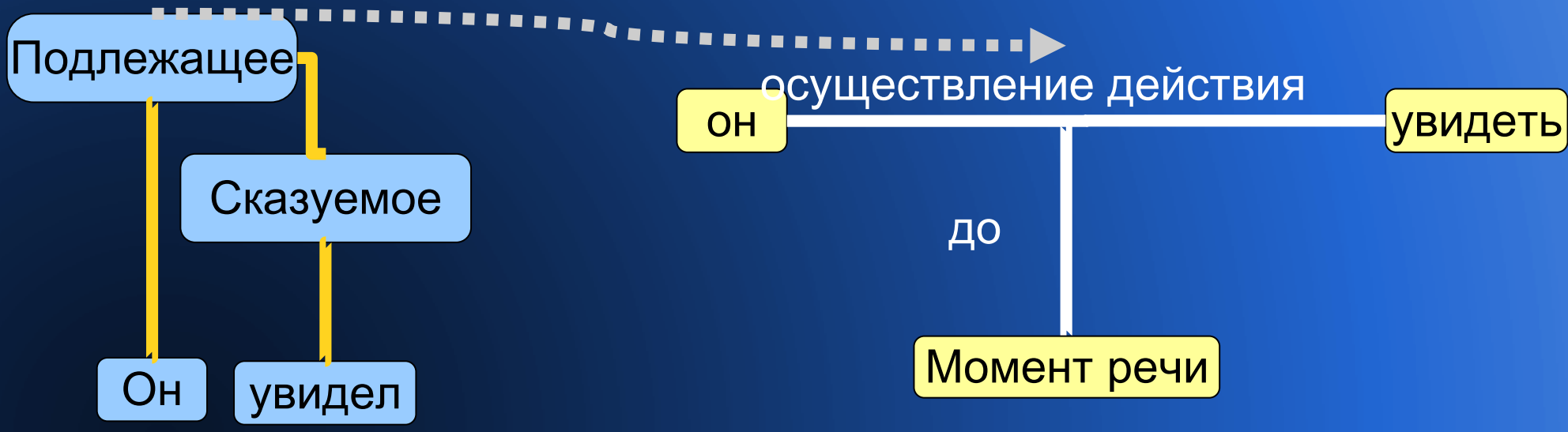
Каждая сложная единица сразу подвергается семантическому анализу:



Он увидел их семью своими глазами.

Принцип работы

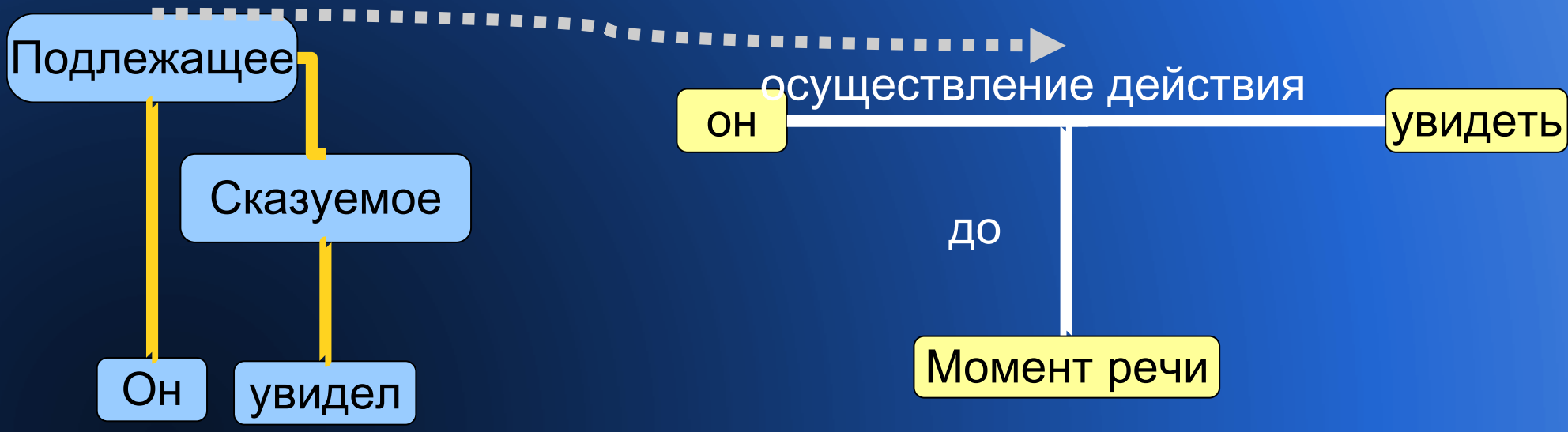
Каждая сложная единица сразу подвергается семантическому анализу:



Он увидел их семью своими глазами.

Принцип работы

Единицы, получившие хотя бы одну семантическую интерпретацию, учитываются при дальнейшем анализе:



Он увидел их семью своими глазами.

Принцип работы

Единицы, получившие хотя бы одну семантическую интерпретацию, учитываются при дальнейшем анализе:

Он+увидел

Он | увидел

Он увидел их семью своими глазами.

Принцип работы

После того, как все попытки связывания текущей единицы с предыдущими выполнены, осуществляется переход к следующей единице:

Он+увидел

Он | увидел

Он увидел их семью своими глазами.

Принцип работы

После того, как все попытки связывания текущей единицы с предыдущими выполнены, осуществляется переход к следующей единице:

Он+увидел | их (чей)

Он | увидел | их (кого)

Он увидел их семью своими глазами.

Принцип работы

Следующая единица также подвергается связыванию с предшествующими ей соседями:

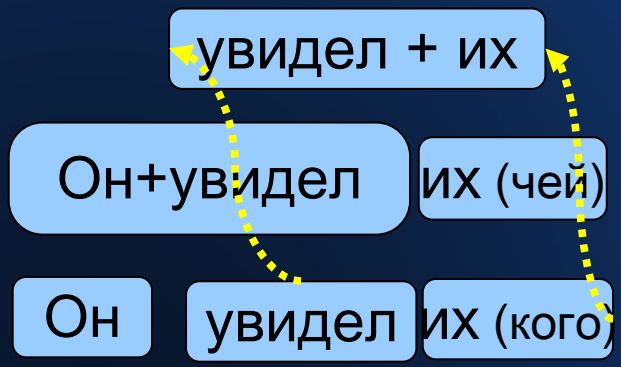
Он+увидел их (чей)

Он увидел их (кого)

Он увидел их семью своими глазами.

Принцип работы

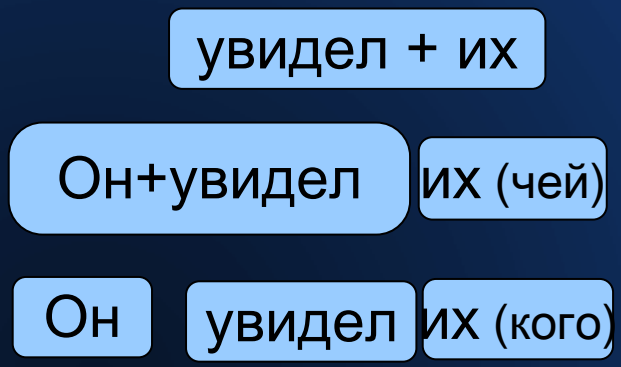
Следующая единица также подвергается связыванию с предшествующими ей соседями:



Он увидел их семью своими глазами.

Принцип работы

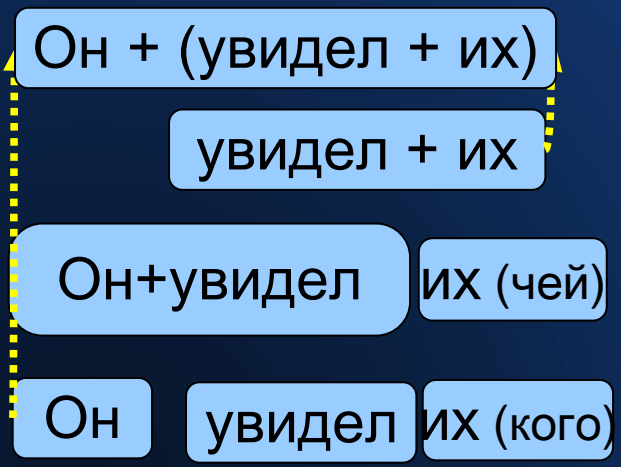
Единицы, полученные в результате связывания и имеющие семантическую интерпретации, связываются с предшествующими им соседями:



Он увидел их семью своими глазами.

Принцип работы

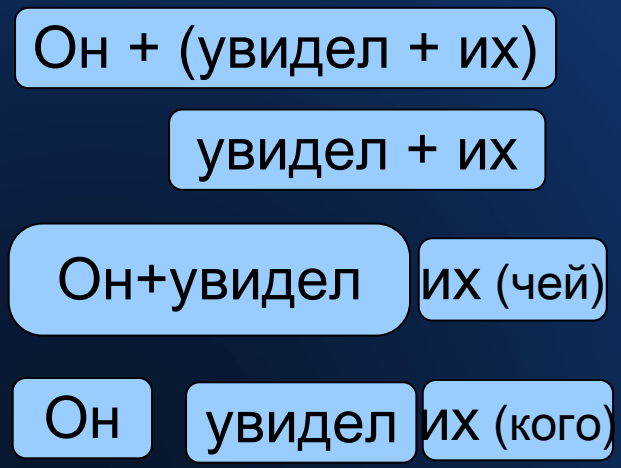
Единицы, полученные в результате связывания и имеющие семантическую интерпретации, связываются с предшествующими им соседями:



Он увидел их семью своими глазами.

Принцип работы

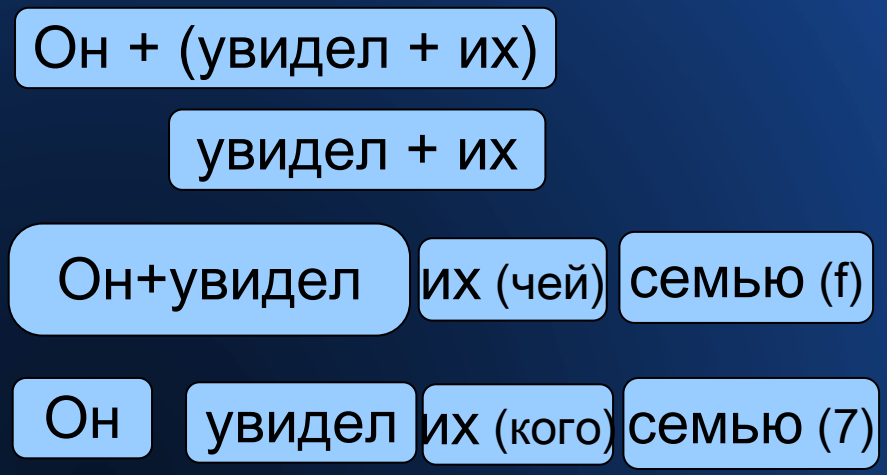
При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

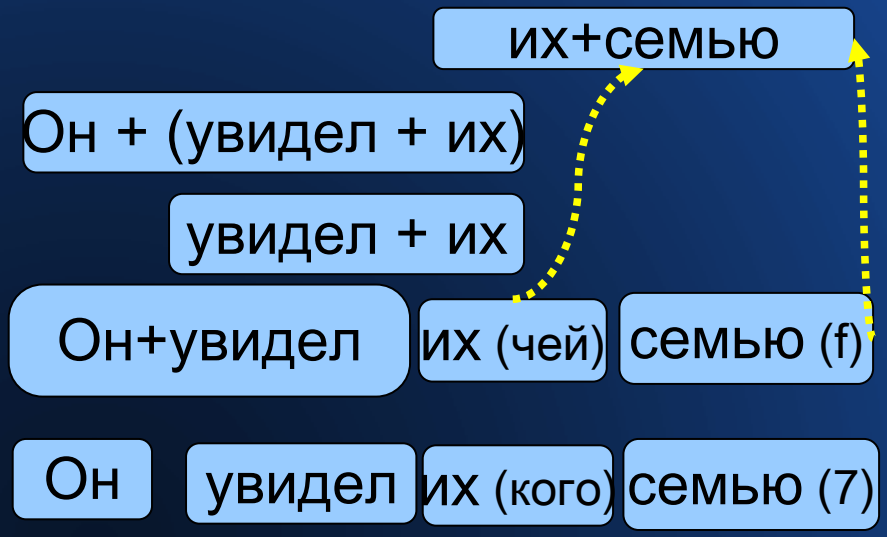
При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

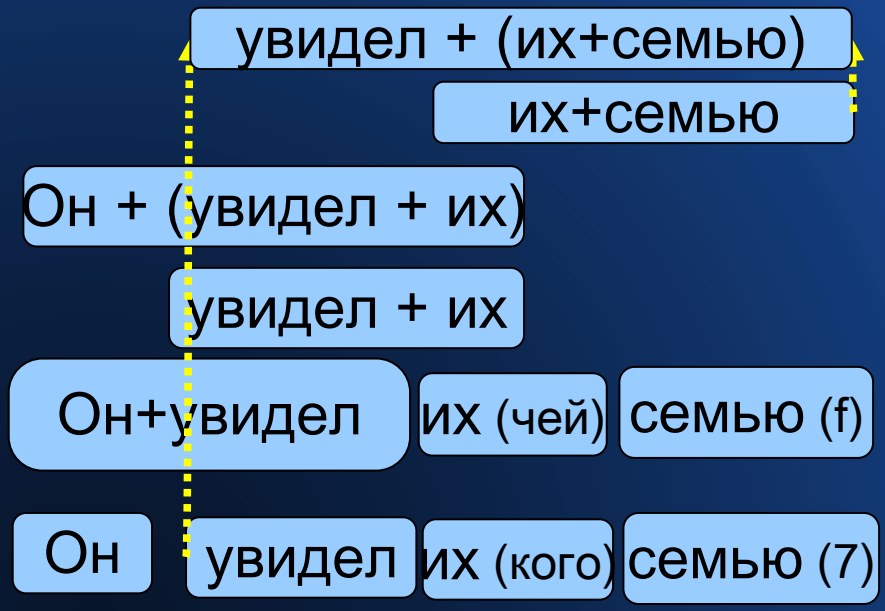
При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

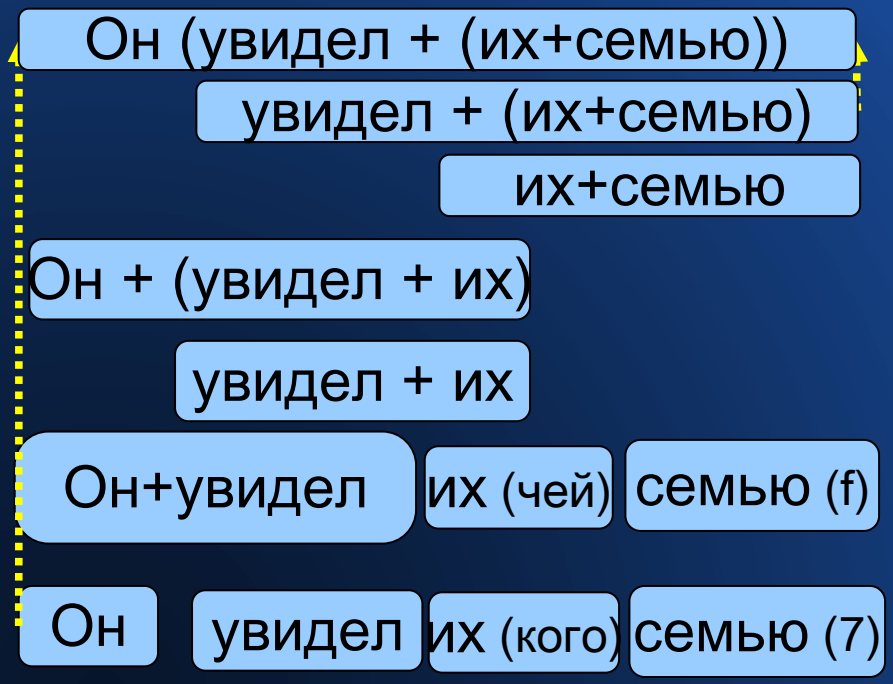
При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

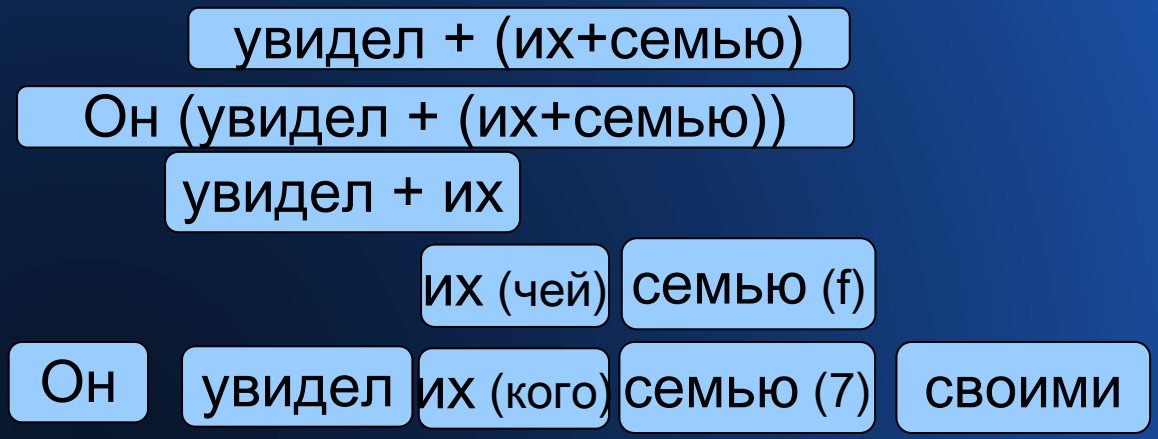
При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

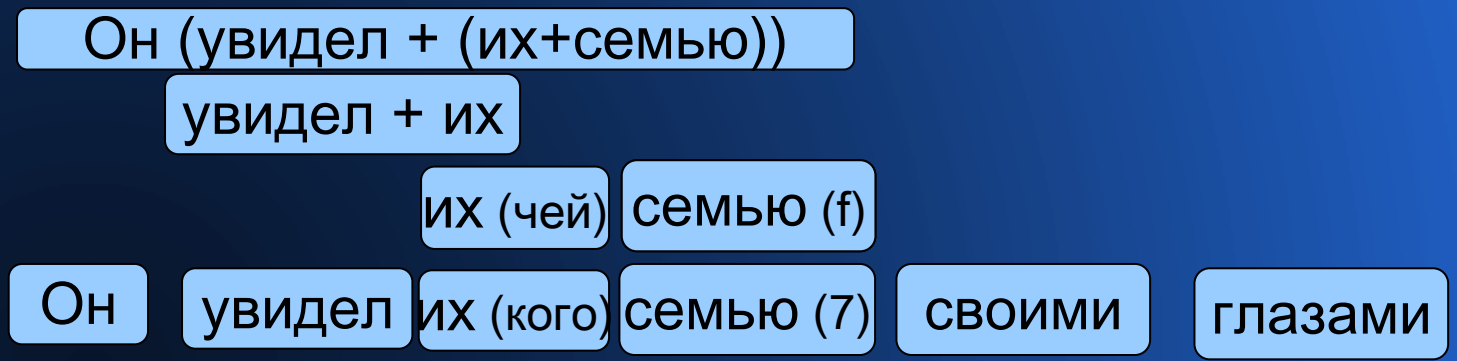
При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

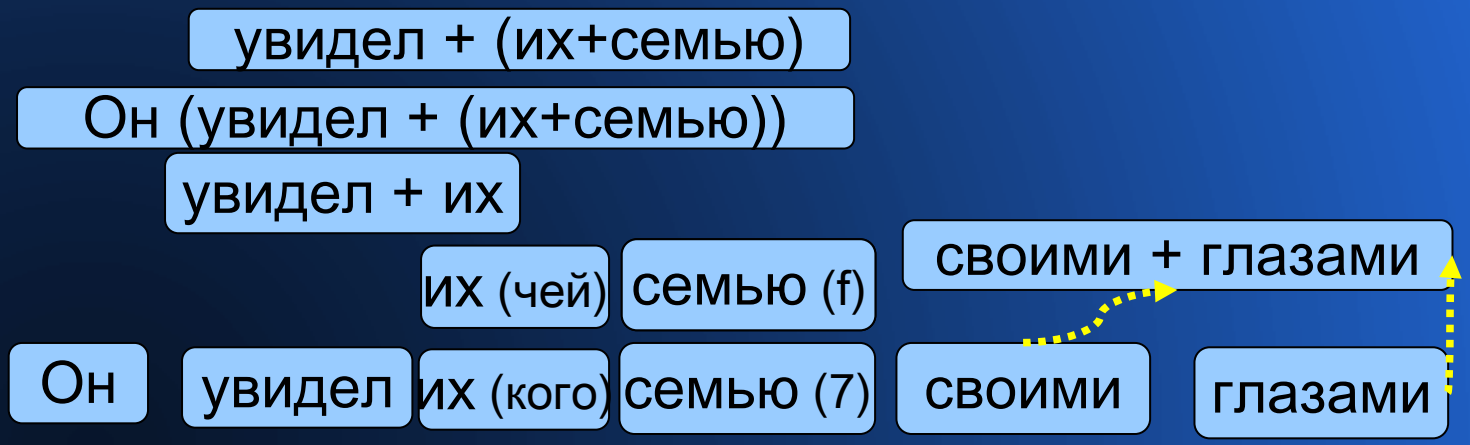
При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

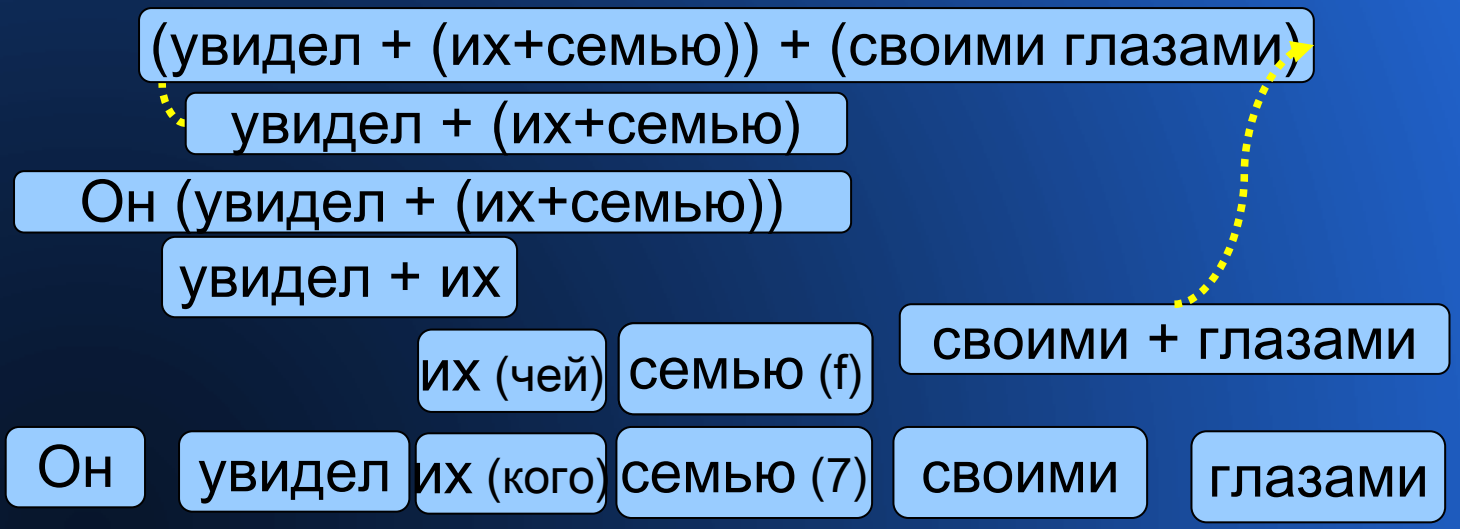
При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

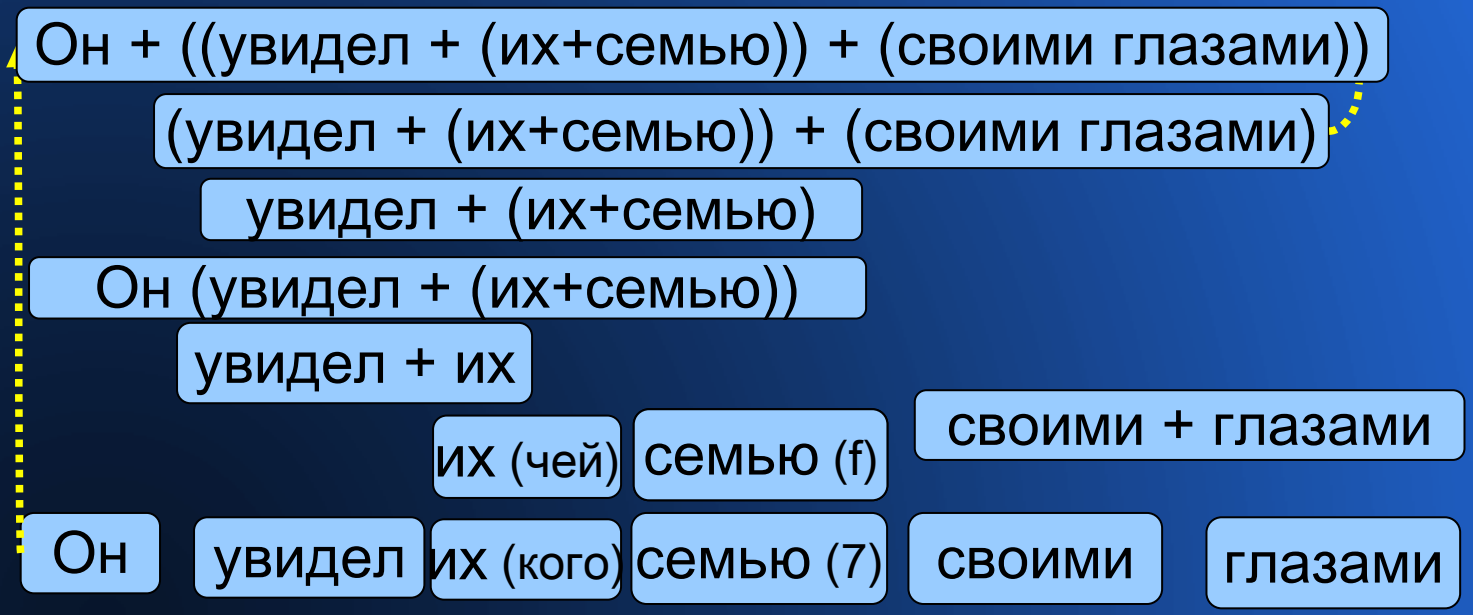
При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

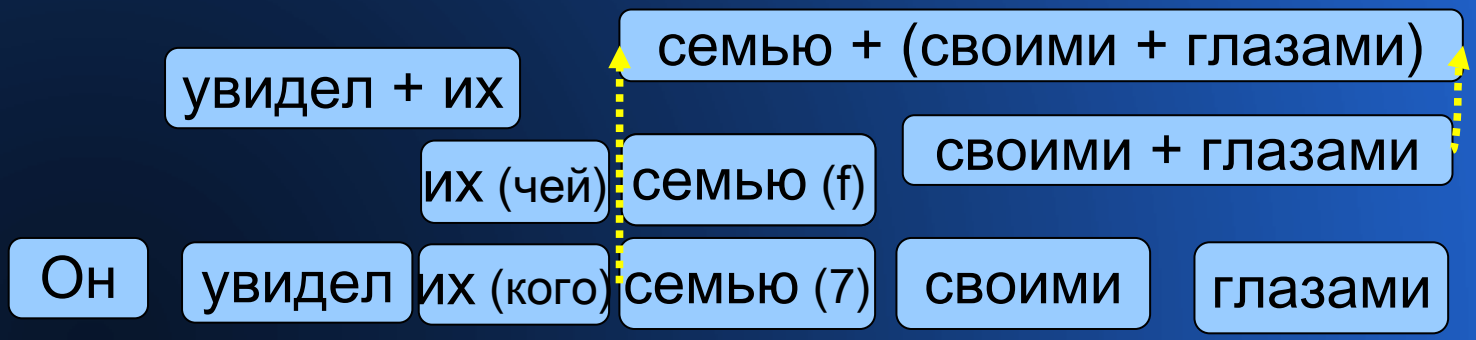
При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

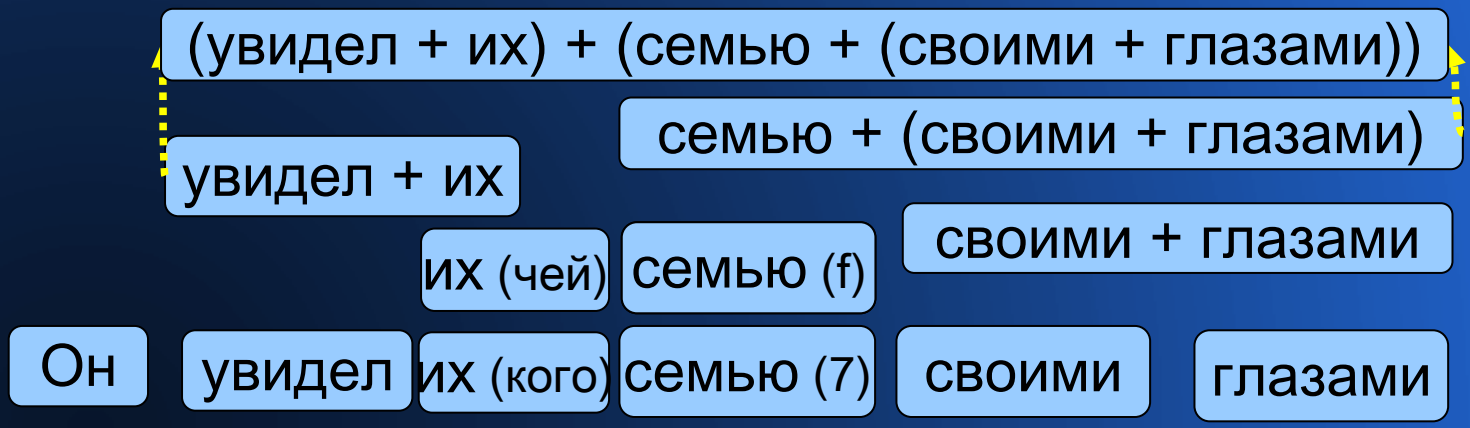
При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

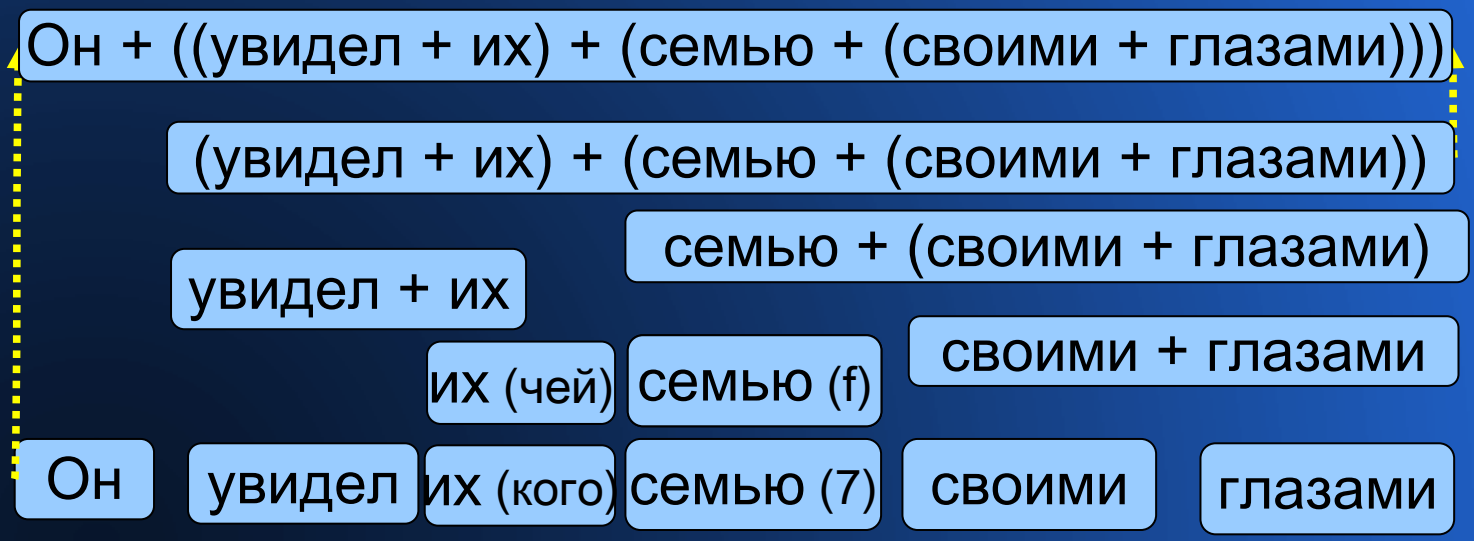
При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

При появлении новых атомарных единиц могут начать действовать ранее неактивные единицы:



Он увидел их семью своими глазами.

Принцип работы

Среди всех выделенных единиц выбираются гипотезы с максимальным покрытием:

Он + ((увидел + (их+семью)) + (своими глазами))

Он + ((увидел + их) + (семью + (своими + глазами)))

Он увидел их семью своими глазами.

Принцип работы

У каждой гипотезы может быть несколько семантических интерпретаций:

Он + ((увидел + (их+семью)) + (своими глазами))

Он сам увидел их семью

Он понял, что их семья – это его глаза

Он + ((увидел + их) + (семью + (своими + глазами)))

Он увидел их 7 своими глазами

Он понял, что они – это его 7 глаз

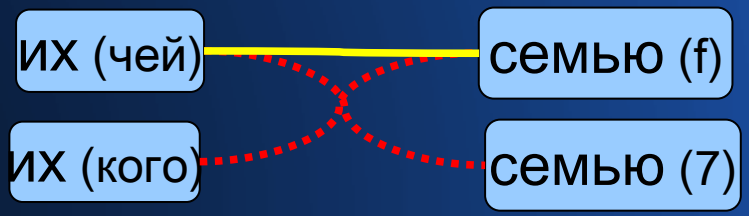
Он увидел их семью своими глазами.

Что это дает?

- ✓ Неоднозначность на более низких уровнях снимается за счет того, что анализ на более высоком уровне производится сразу.

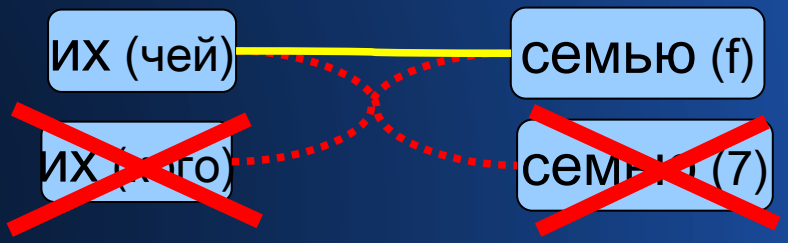
Что это дает?

- Неоднозначность на более низких уровнях снимается за счет того, что анализ на более высоком уровне производится сразу.
  - Морфологическая неоднозначность частично снимается благодаря невозможности синтаксической связи:



Что это дает?

- ✓ Неоднозначность на более низких уровнях снимается за счет того, что анализ на более высоком уровне производится сразу.
  - ✓ Морфологическая неоднозначность частично снимается благодаря невозможности синтаксической связи:



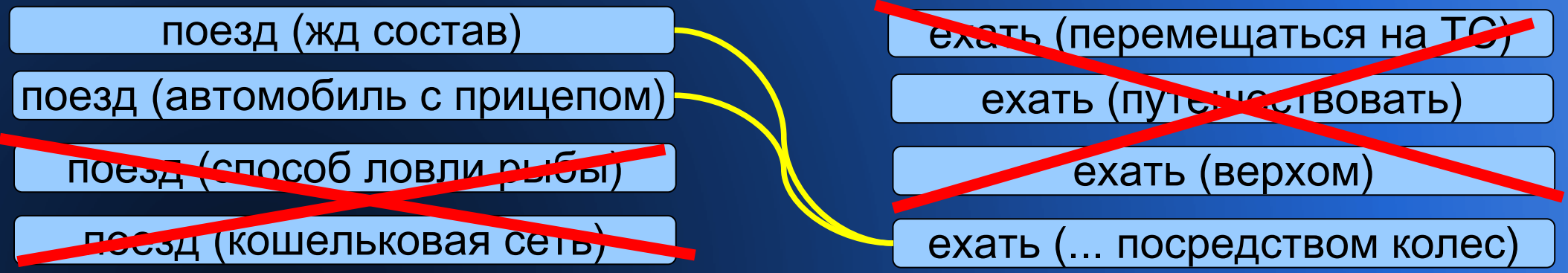
Что это дает?

- ✓ Неоднозначность на более низких уровнях снимается за счет того, что анализ на более высоком уровне производится сразу.
  - ✓ Лексическая неоднозначность частично снимается за счет недопустимости того или иного отношения:

поезд (жд состав)	ехать (перемещаться на ТС)
поезд (автомобиль с прицепом)	ехать (путешествовать)
поезд (способ ловли рыбы)	ехать (верхом)
поезд (кошельковая сеть)	ехать (... посредством колес)

Что это дает?

- ✓ Неоднозначность на более низких уровнях снимается за счет того, что анализ на более высоком уровне производится сразу.
  - ✓ Лексическая неоднозначность частично снимается за счет недопустимости того или иного отношения:



Что это дает?

- Неоднозначность на более низких уровнях снимается за счет того, что анализ на более высоком уровне производится сразу.
  - Синтаксическая неоднозначность также частично снимается в ходе семантического анализа, ср:

1) *Президент встретил премьера в городе*

2) *Президент встретил премьера в пальто*

Что это дает?

- ✓ Неоднозначность на более низких уровнях снимается за счет того, что анализ на более высоком уровне производится сразу.
- ✓ Из тысяч возможных комбинаций лингвистического анализа исключаются те, которые не соответствуют правилам грамматики или противоречат семантическим ограничениям.

Что это дает?

- ✓ Неоднозначность на более низких уровнях снимается за счет того, что анализ на более высоком уровне производится сразу.
- ✓ Из тысяч возможных комбинаций лингвистического анализа исключаются те, которые не соответствуют правилам грамматики или противоречат семантическим ограничениям.
- ✓ В задачах информационного поиска (индексирования, автоматической рубрикации, data-mining, …) частично решается проблема информационного шума.

Что это дает?

- ✓ Неоднозначность на более низких уровнях снимается за счет того, что анализ на более высоком уровне производится сразу.
- ✓ Из тысяч возможных комбинаций лингвистического анализа исключаются те, которые не соответствуют правилам грамматики или противоречат семантическим ограничениям.
- ✓ В задачах информационного поиска (индексирования, автоматической рубрикации, data-mining, …) частично решается проблема информационного шума.
- ✓ В задачах МП (в перспективе) повышается качество перевода

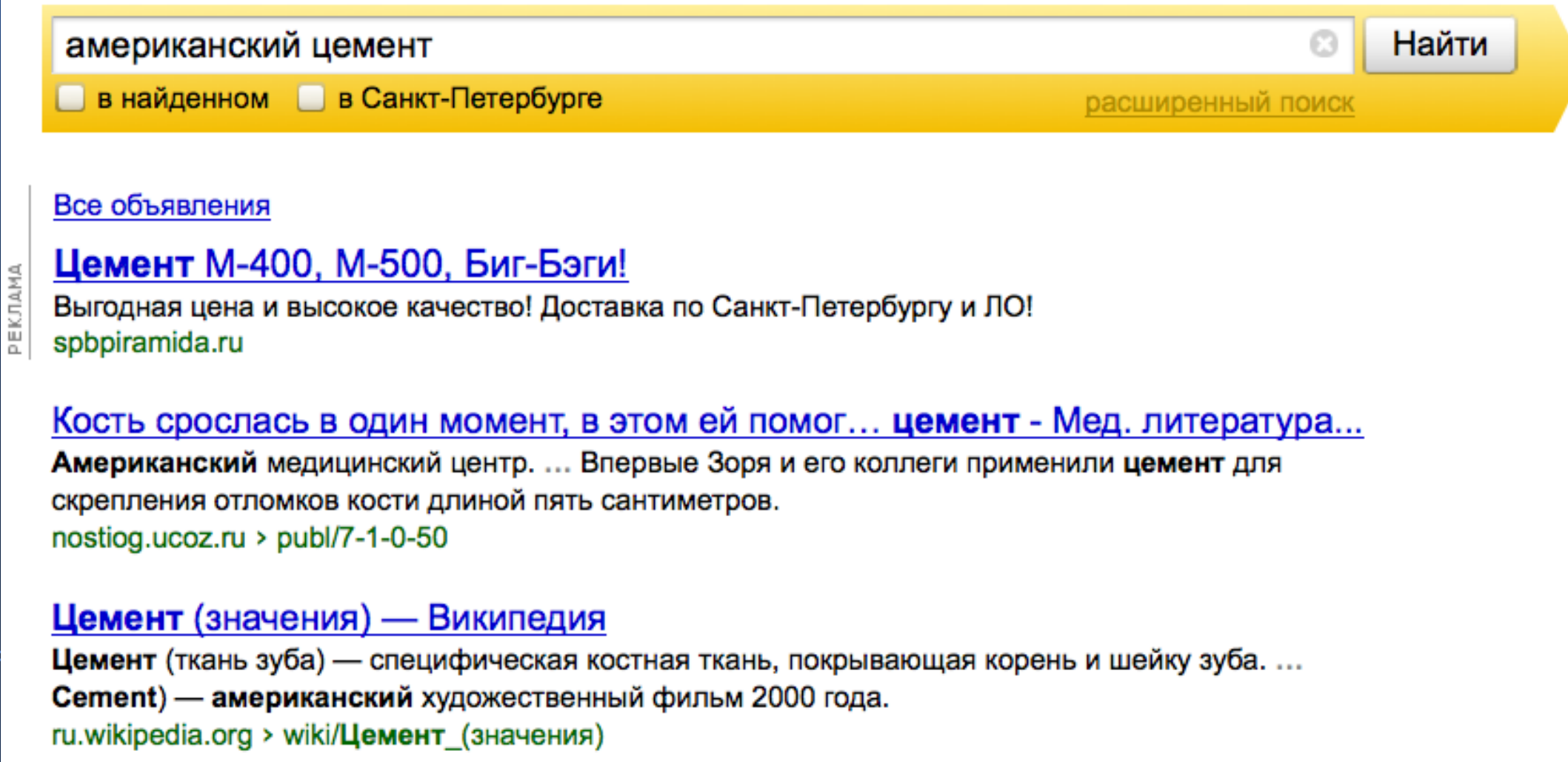
Информационный поиск

Задачи, решаемые при помощи AIIRE:

- ✓ Семантический поиск по запросу
- ✓ Автоматическая рубрикация с учетом семантики

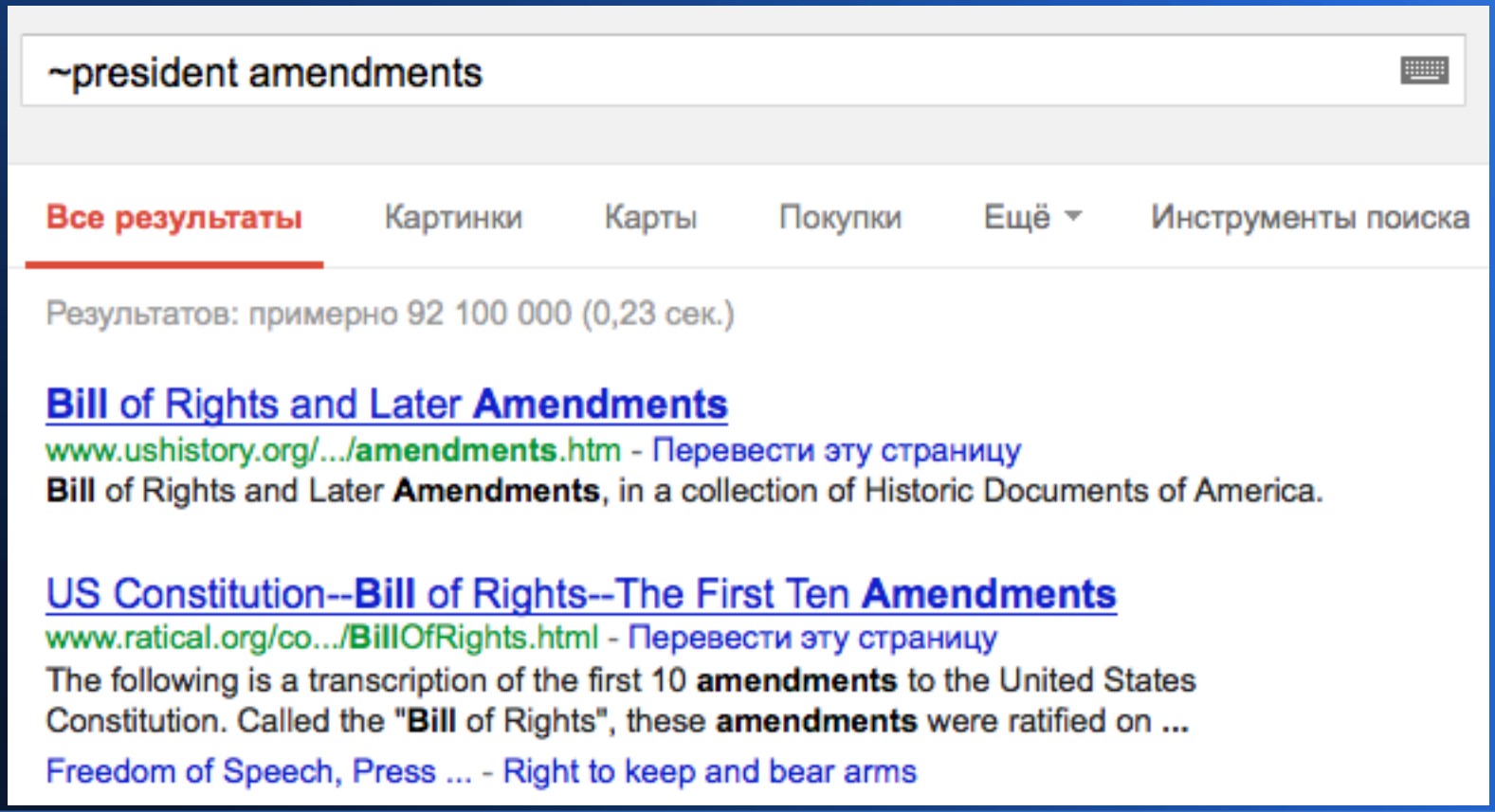
Семантический поиск

Фрагмент выдачи одной из систем адресного поиска:



Семантический поиск

Игнорирование синтаксических и семантических связей приводит к курьезам:



Семантический поиск

AIIRE: поиск информации должен производиться не по словам, и даже не по фразам; поиск информации должен производиться по *содержанию поискового запроса*.

Семантический поиск

AIIRE: поиск информации должен производиться не по словам, и даже не по фразам; поиск информации должен производиться по *содержанию поискового запроса*.

Для этого поисковый индекс должен содержать в качестве ключей не слова, а концептуальные графы.

Семантический поиск

AIIRE: поиск информации должен производиться не по словам, и даже не по фразам; поиск информации должен производиться по *содержанию поискового запроса*.

Для этого поисковый индекс должен содержать в качестве ключей не слова, а концептуальные графы.

Поиск по точному и синонимическому соответствию производится путем обнаружения графа в индексе

Семантический поиск

AIIRE: поиск информации должен производиться не по словам, и даже не по фразам; поиск информации должен производиться по *содержанию поискового запроса*.

Для этого поисковый индекс должен содержать в качестве ключей не слова, а концептуальные графы.

Поиск по точному и синонимическому соответствию производится путем обнаружения графа в индексе

Поиск с учетом родо-видовых отношений выполняется путем построения индекса по самим концептуальным графам

Семантический поиск

Поиск по точному и синонимическому соответствию производится путем обнаружения графа в индексе

- ✓ AIIRE строит гипотезы семантического анализа в виде *концептуальных графов*
- ✓ AIIRE производит *нормирование концептуальных графов*

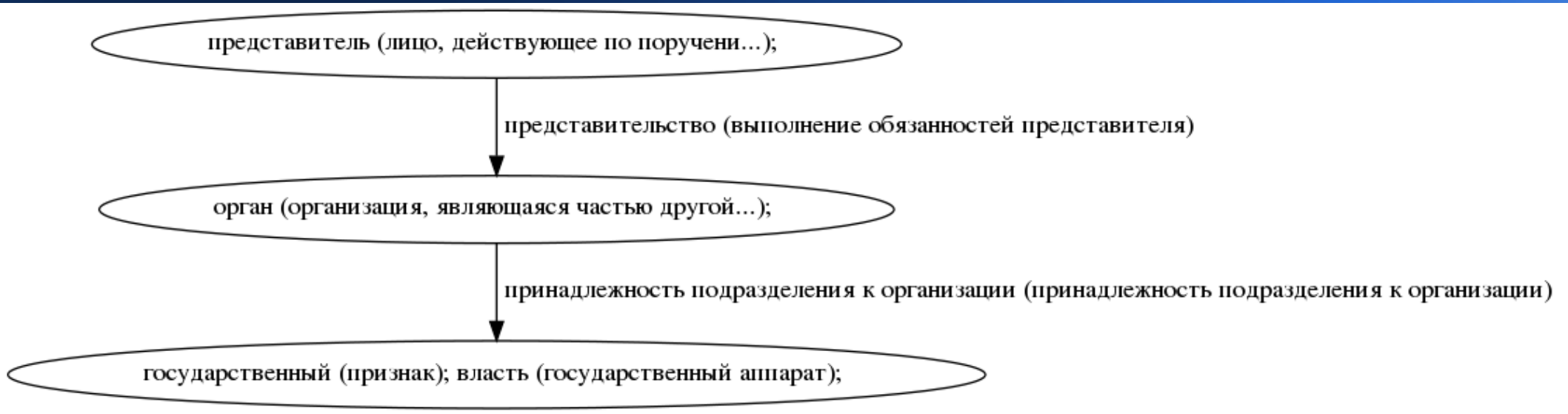
Поиск с учетом родо-видовых отношений выполняется путем построения индекса по самим концептуальным графам

- ✓ При нормировании концептуальный граф регистрируется в *дополнительном индексе* при всех содержащихся в нем концептах и их онтологических «предках»
- ✓ При поиске производится пересечение множеств концептуальных графов, соответствующих запросу

Семантический поиск

Концептуальные графы:

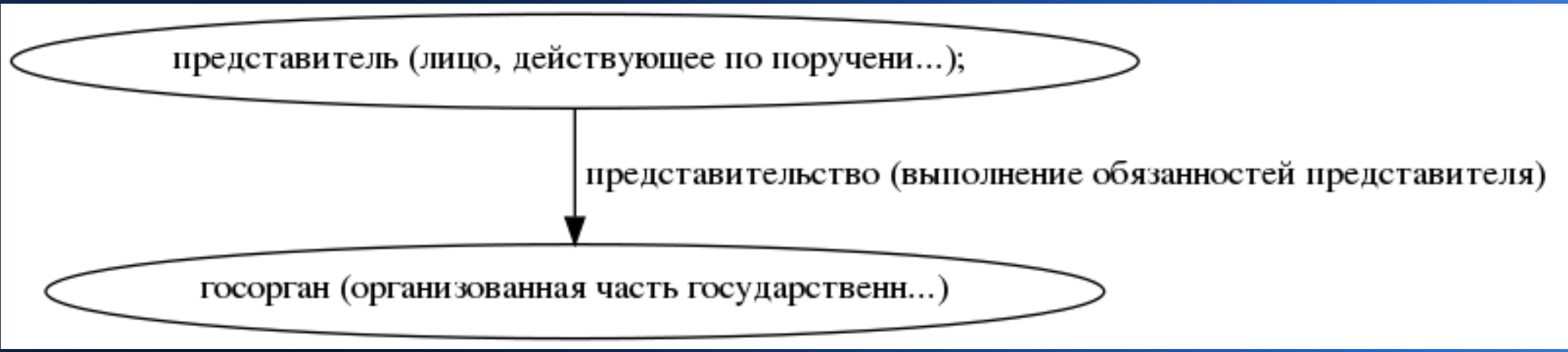
Содержание фразы «представитель органа государственной власти»:



Семантический поиск

Нормализация концептуальных графов:

Подграфы заменяются на более простые узлы, соответствующие представителям синонимических рядов

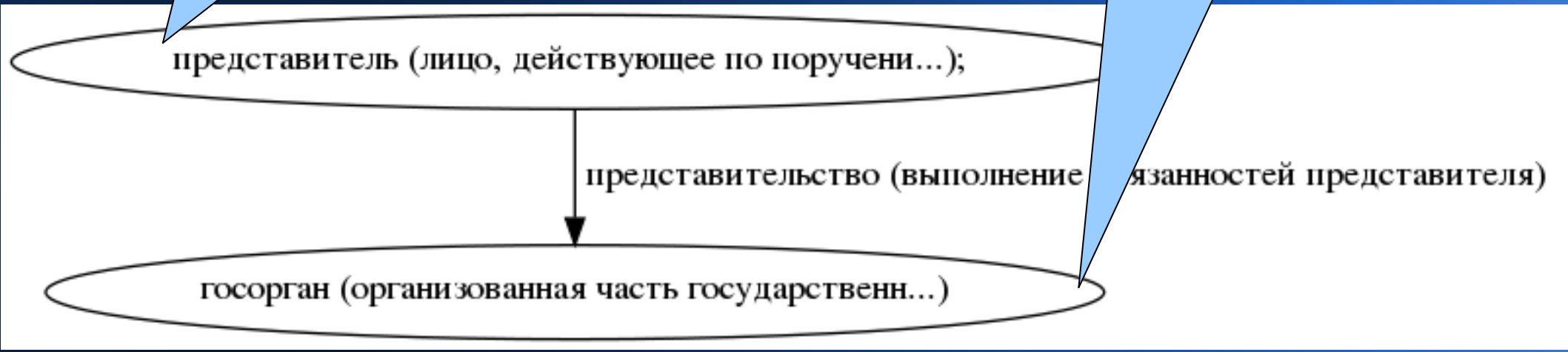


Семантический поиск

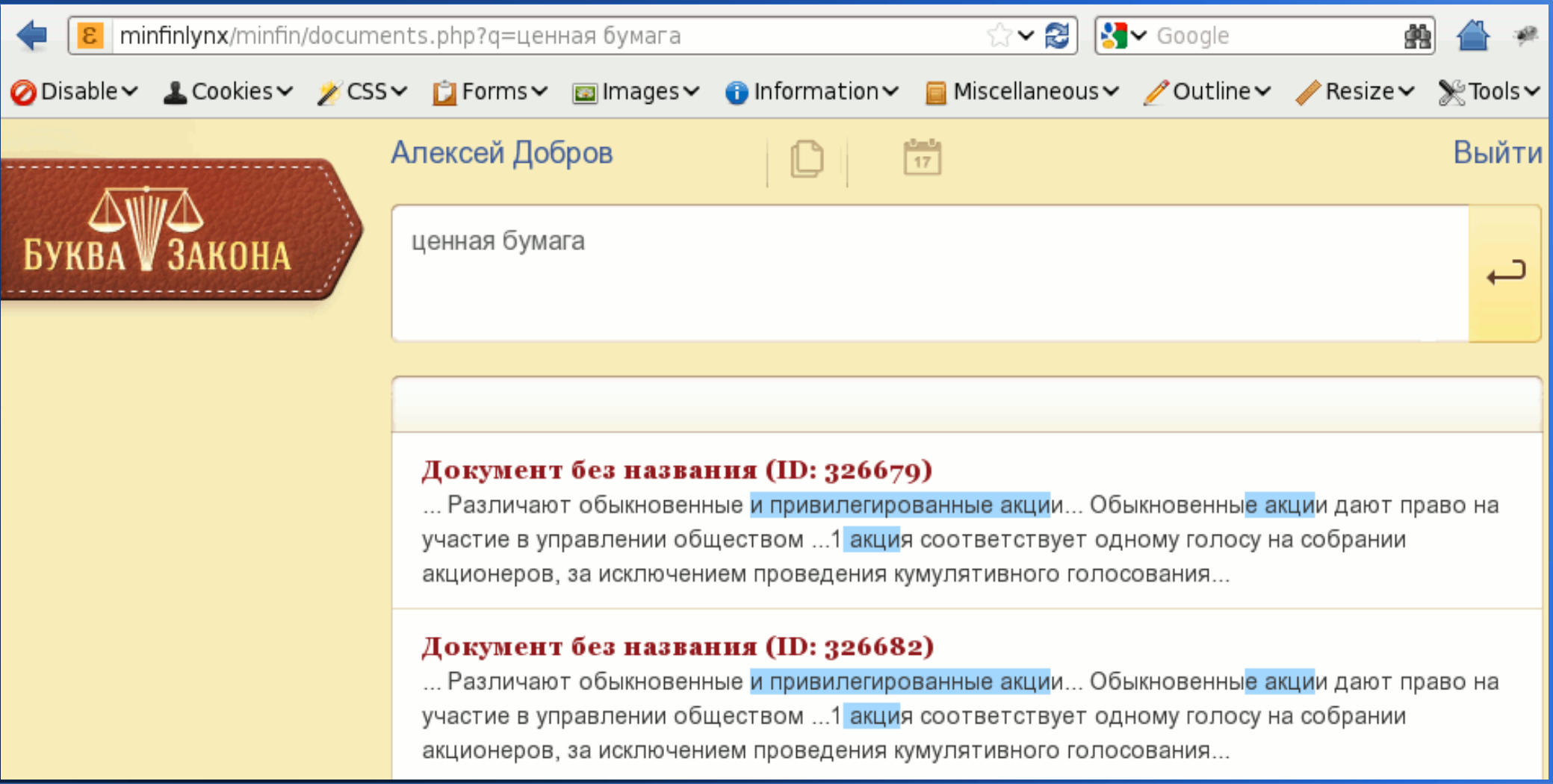
Построение индекса по концептуальным графам:

- лицо (человек)
- субъект права
- живое существо
- предмет
...

- подразделение
- организация
- совокупность людей
- совокупность живых существ
...



Семантический поиск



Автоматическая рубрикация

Автоматическая рубрикация (классификация) документов – это отнесение этих документов к заранее определенным рубрикам

Автоматическая рубрикация

Автоматическая рубрикация (классификация) документов – это отнесение этих документов к заранее определенным рубрикам

Инженерный подход: отнесение к рубрикам производится в соответствии с набором правил, определенных разработчиком

Автоматическая рубрикация

Автоматическая рубрикация (классификация) документов – это отнесение этих документов к заранее определенным рубрикам

Инженерный подход: отнесение к рубрикам производится в соответствии с набором правил, определенных разработчиком

Иными словами, автоматическая рубрикация – это *семантический поиск по запросу, эквивалентному образу рубрики*

Автоматическая рубрикация

Утилита «aire_classifier» входит в «aire»:

```
macintosh-3:Экономика distort$ cat news_13_02_27.11.2012
Официальный курс доллара на среду составляет 30,94 рубля
МОСКВА, 27 ноя – Прайм.  Официальный курс евро к рублю, установленный ЦБ РФ
на среду, понизился на 0,96 копейки – до 40,1893 рубля, курс доллара понизил
ся на 7,91 копейки – до 30,9410 рубля, следует из данных Банка России.
Все новости экономики и бизнеса на сайте агентства Прайм >>
Стоимость бивалютной корзины (0,55 доллара и 0,45 евро), рассчитанная по офи
циальным курсам на среду, сократилась  по сравнению с показателем на вторник
  на пять копеек, составив 35,1 рубля.
macintosh-3:Экономика distort$ aire_classifier < news_13_02_27.11.2012
Grammar size: 19250 signals.
Changing names to links...
Done.
происшествия (предметная область экстраординарных событий, значимых для неко
торого социума) 0,01
экономика (область деятельности, связанная с производством материальных благ
 и хозяйствованием) 0,85
в мире (совокупность событий, происходящих за рубежом) 0,09
социальная система (социальное устройство, характеризующееся определенными о
тношениями) 0,05
LSN reset in db /usr/share/aire/lang/ontology/routedb.db
LSN reset in db /usr/share/aire/lang/ontology/sigdb.db
```

Автоматическая рубрикация

Утилита «aire_classifier» входит в «aire»:

```
происшествия (предметная область экстраординарных событий, значимых для неко
торого социума) 0,01
экономика (область деятельности, связанная с производством материальных благ
 и хозяйствованием) 0,85
в мире (совокупность событий, происходящих за рубежом) 0,09
социальная система (социальное устройство, характеризующееся определенными о
тношениями) 0,05
```

AIRE-lang

Пакет «aire-lang» включает в себя языковые модули aire. На данный момент распространяется модуль русского языка, включающий в себя

- ✓ Грамматику (291 класс абстрактных конструкций) – полностью наша разработка
- ✓ Морфологический словарь (121636 лемм) – использовались материалы Викисловаря
- ✓ Онтологию (1919783 концептов, обработано – 24532, >350 типа отношения) – использовались материалы Википедии, Викисловаря, нескольких опубликованных тезаурусов и синонимических словарей; доля участия этих источников в разработанной онтологии минимальна.

Спасибо!