

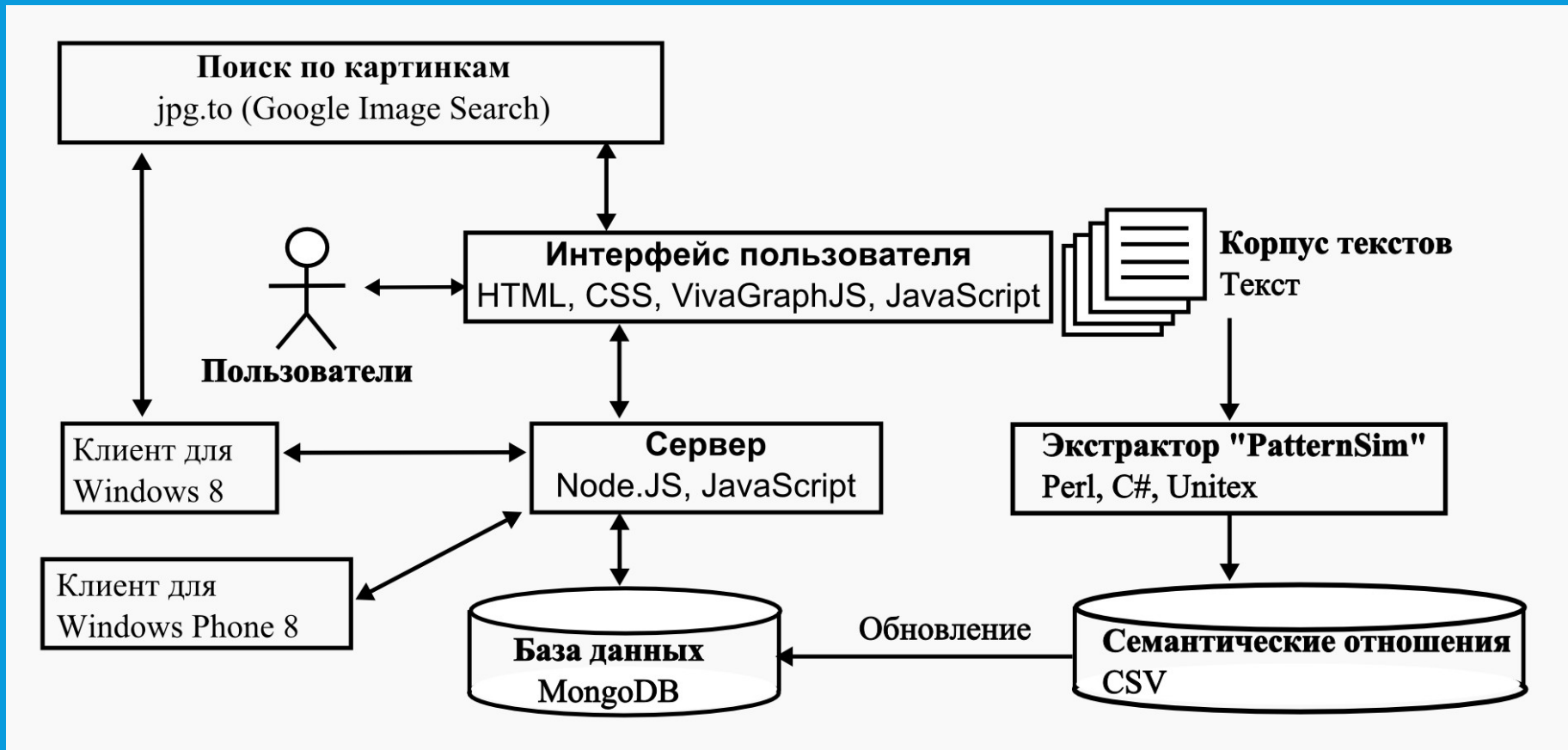
Поиск и визуализация семантически связанных слов

<http://serelex.cental.be>

Введение

- Запрос => список **семантически связанных** слов
- **Интерактивный поиск** связанных лексических единиц
- Информация, **извлеченная** из корпуса ЕЯ текстов
- Большое покрытие
- Достаточная точность
- Три способа визуализации результатов
- Открытый исходный код

Архитектура системы



Экстрактор: метрика близости

1. Лексико-синтаксические шаблоны

such NP as NP, NP[,] and/or NP;

NP, NP [,] and other NP;

2. Извлечение конкордансов из корпуса

such diverse {[occupations]} as {[engineers]} and {[scientists]}

{[mango]}, {[pineapple]}, {[jackfruit]} and other {[fruits]}

3. Лемматизация конкордансов


4. Нормализация

419,751 лемм – 11,251,240 отношений

Сервер

- Node.js
- В ответ на запрос возвращается множество слов, отсортированных по семантической близости
- Лемматизация запросов
- Приблизительный поиск с помощью расстояния Левенштейна
- REST API

Пользовательский интерфейс

 **serelex**
Finds semantically related words

Like 58 Tweet 12 +1 6

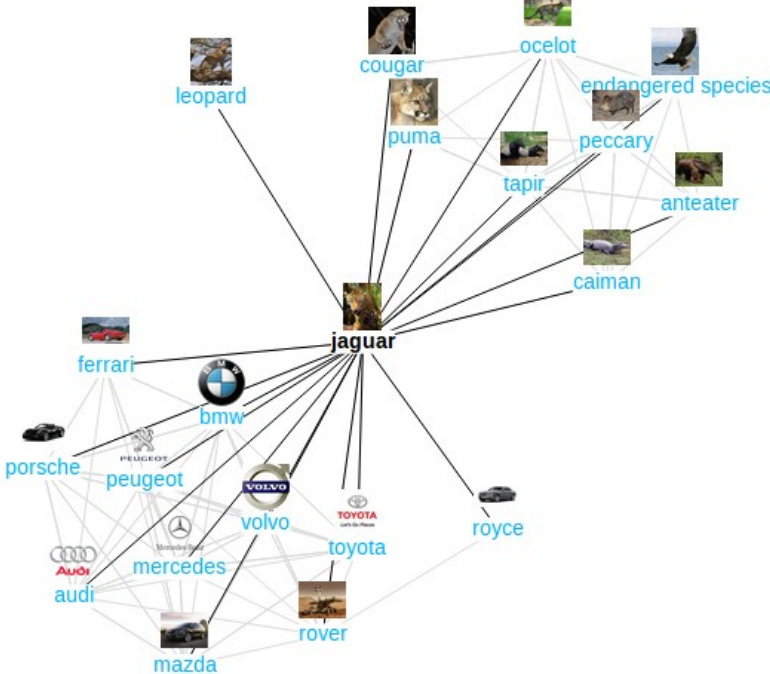
Show as **Graph** Images Secondary links **On** Off Icons **On** Off Info panel **On** Off Search Language **English** French

For example, [strawberry](#)


Results count: 685

- 1 [tapir](#)
- 2 [ocelot](#)
- 3 [puma](#)
- 4 [porsche](#)
- 5 [anteater](#)
- 6 [audi](#)
- 7 [cougar](#)
- 8 [mazda](#)
- 9 [rover](#)
- 10 [bmw](#)
- 11 [volvo](#)
- 12 [caiman](#)
- 13 [endangered species](#)
- 14 [ferrari](#)
- 15 [peugeot](#)
- 16 [toyota](#)
- 17 [leopard](#)
- 18 [mercedes](#)
- 19 [peccary](#)
- 20 [royce](#)

[Show next 20 results](#)



The diagram is a semantic network centered on the word 'jaguar'. It features two main clusters of related terms. The first cluster, on the right, includes animal-related terms: 'leopard', 'cougar', 'puma', 'ocelot', 'endangered species', 'peccary', 'tapir', 'anteater', and 'caiman'. Each term is accompanied by a small representative image. The second cluster, on the left, includes car-related terms: 'ferrari', 'bmw', 'volvo', 'toyota', 'rover', 'mazda', 'mercedes', 'audi', and 'porsche', also with small representative images. Lines of varying thickness connect the central 'jaguar' node to these peripheral nodes, indicating the strength or type of semantic relationship.



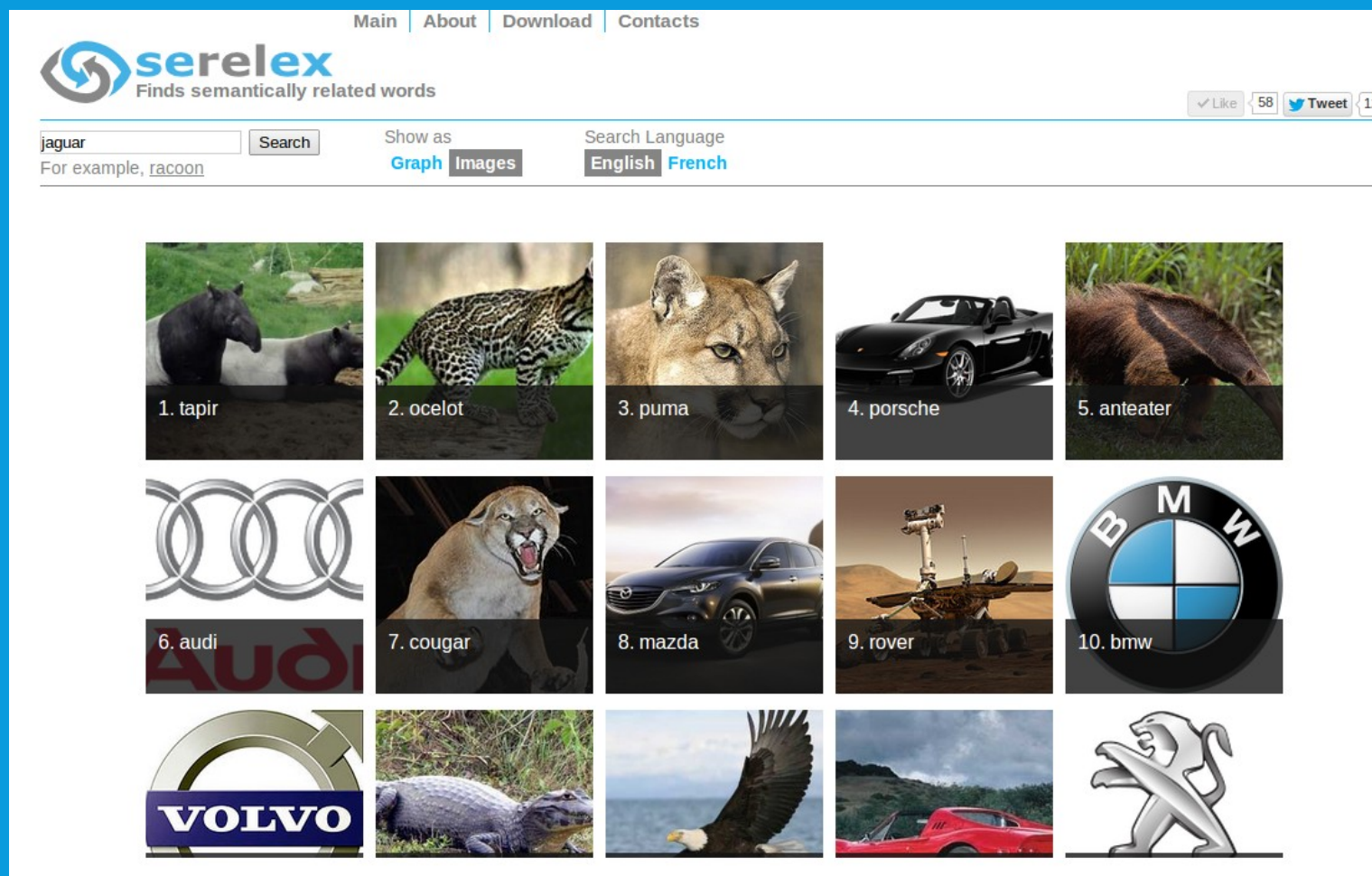
Jaguar — The jaguar is a big cat, a feline in the Panthera genus, and is the only Panthera species found in the Americas. The jaguar is the third-largest feline after the tiger and the lion, and the largest in the Western Hemisphere. The jaguar's present range extends from Southern United States ...
[Wikipedia](#)

Disambiguates:

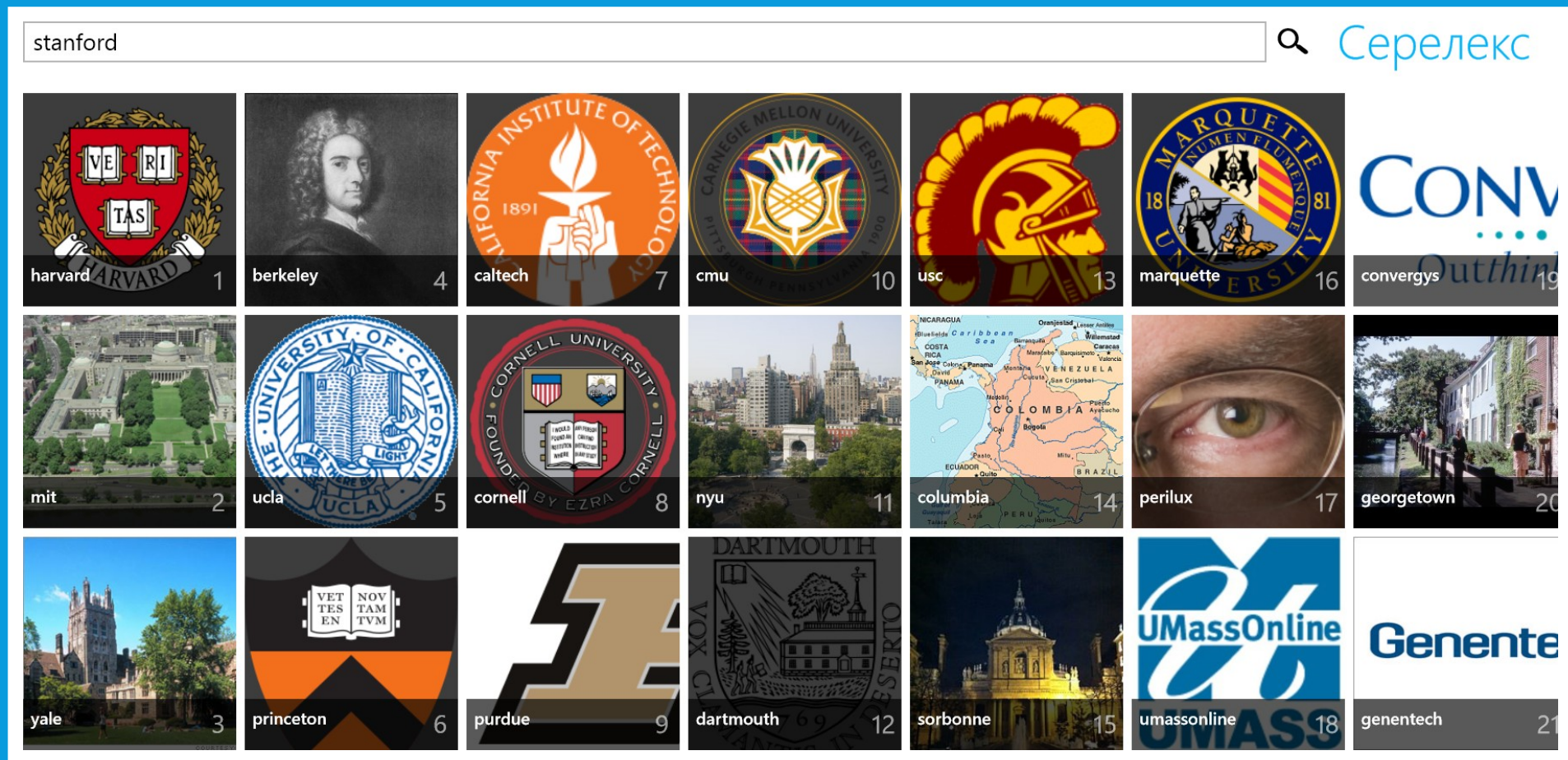
- ▶ [Jaguar](#)
- ▶ [Jacksonville Jaguars](#)
- ▶ [Jaguar Cars](#)
- ▶ [SEPECAT Jaguar](#)
- ▶ [Atari Jaguar](#)

[Show all](#)

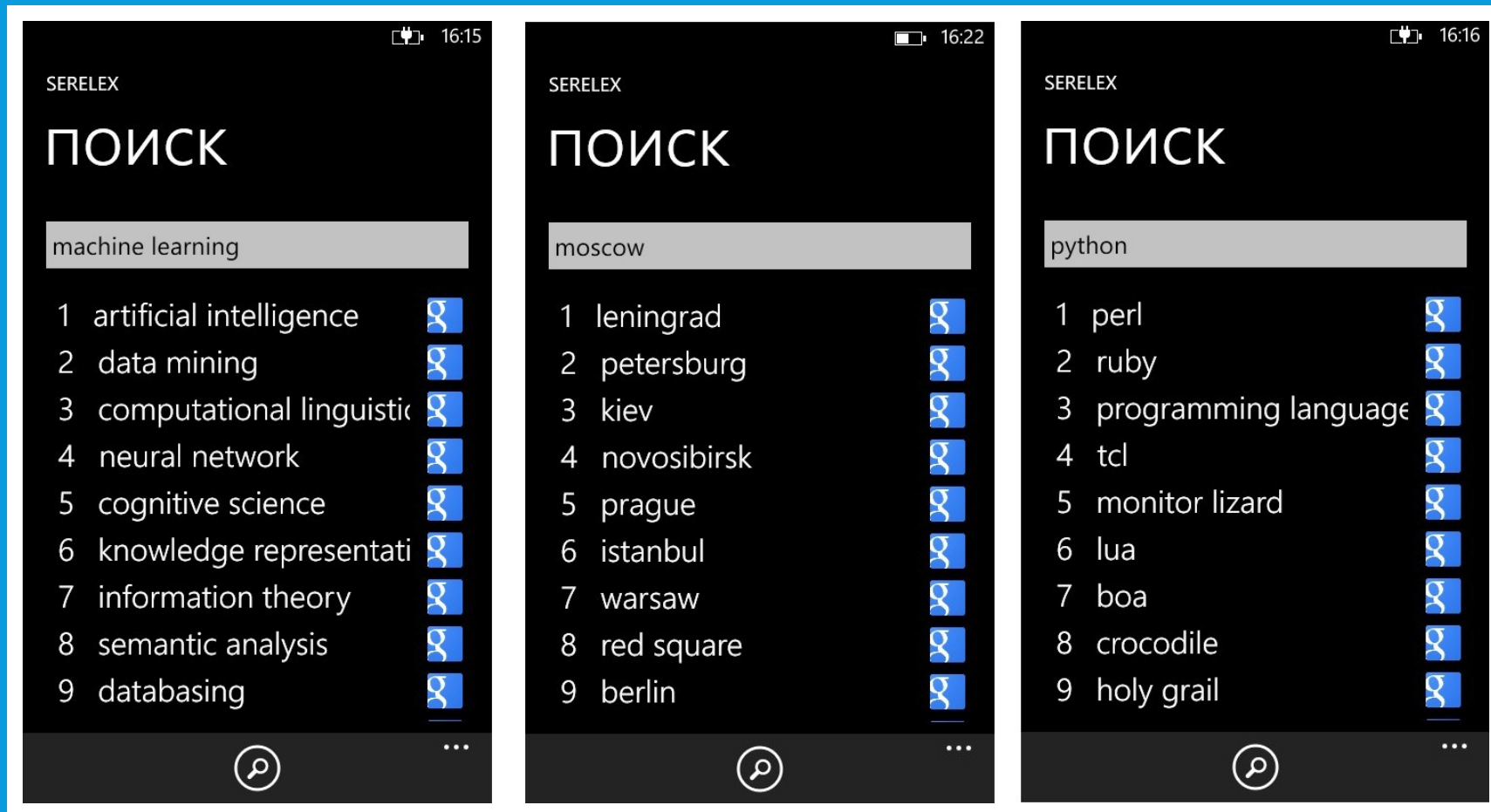
Пользовательский интерфейс



Пользовательский интерфейс



Пользовательский интерфейс



Оценка результатов

- **Корреляция** с суждениями о семантической близости
- **Ранжирование** семантических отношений
- **Извлечение** семантических отношений
- **Удовлетворенность** пользователей качеством поиска

Корреляция с суждениями о семантической близости

- Три проверочных набора данных: MC, RG, WordSim

automobile; car; 3.92

brother; monk; 2.84

glass; magician; 0.11

- Корреляция Спирмена

mc	RG	WORDSIM
0.6	0.73	0.52

Ранжирование семантических отношений

- **Пример:**

1; alligator; animal (related)

. . .

25; alligator; lizard (related)

26; alligator; twin (random)

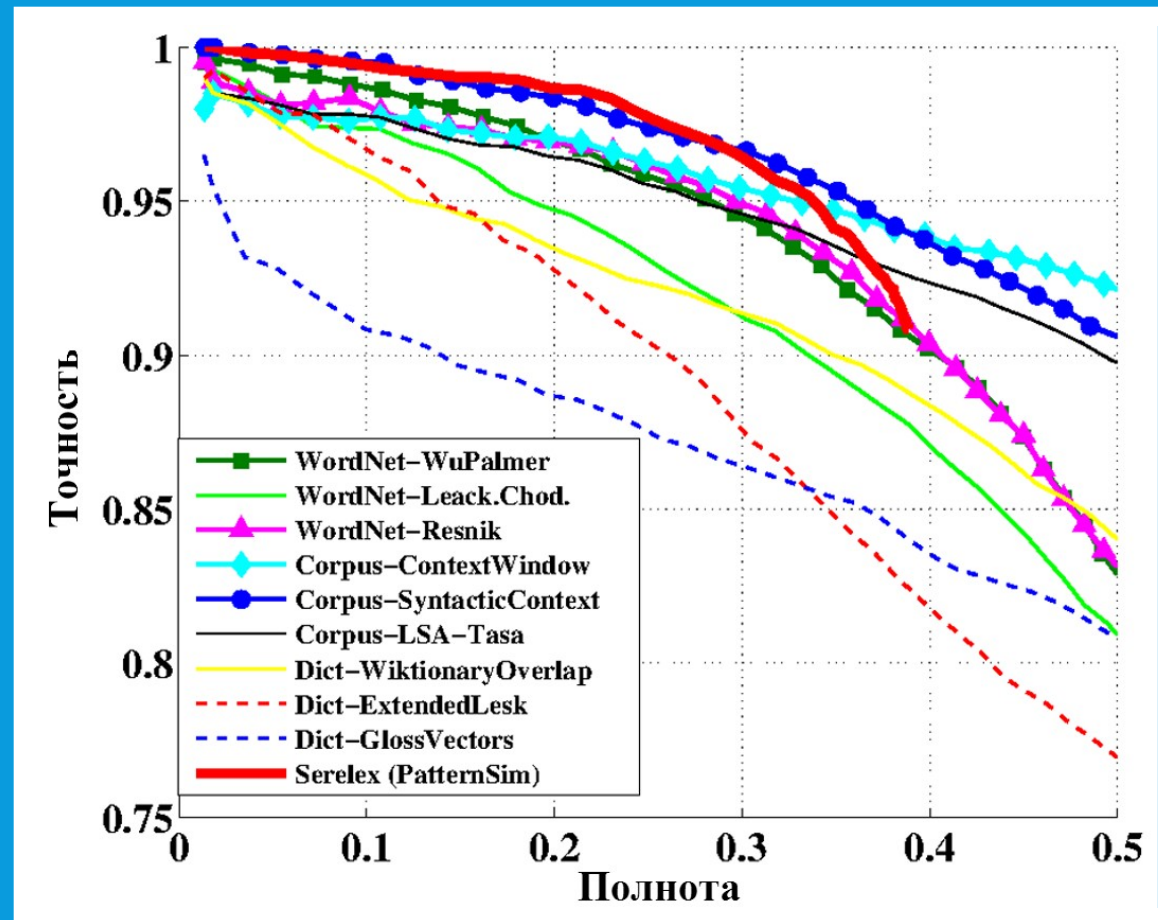
. . .

50; alligator; electronic (random)

- **Оценка точности и полноты**

- Точность сопоставима с альтернативными метриками
- Полнота ниже в связи с разрежённостью подхода

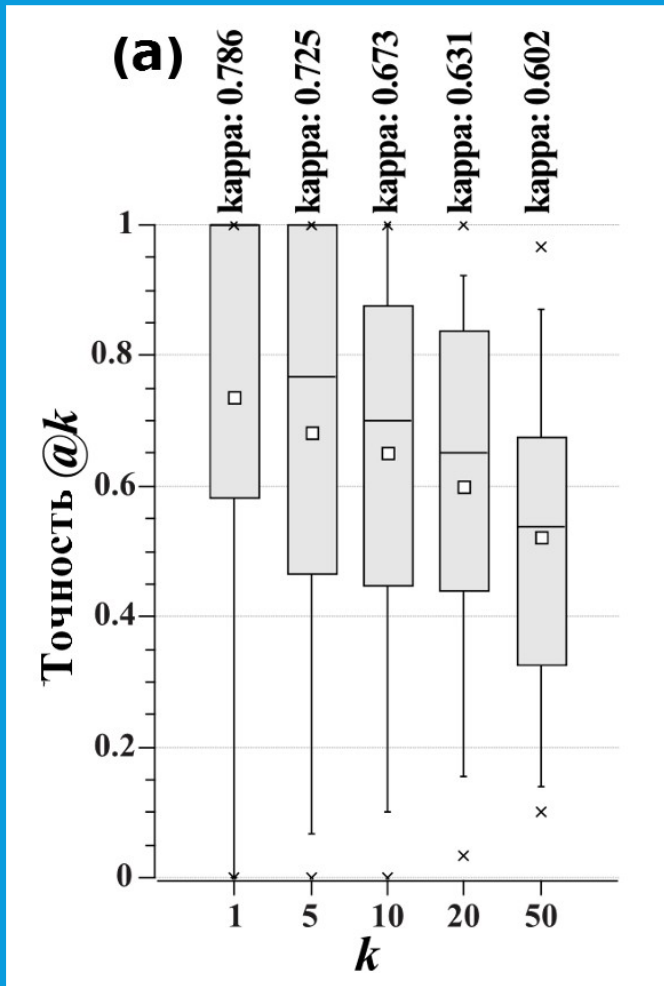
Ранжирование семантических отношений



Извлечение семантических отношений

- **Оценка точности извлечения для 49 слов**
 - Асессоры аннотировали первые 50 результатов
 - Запрос fruit
 - 1; vegetable (relevant)*
 - 2; mango (relevant) ...*
 - 50; house (non-relevant)*
- **Средняя точность**
 - 79% - для первого результата
 - 56% - 50 первых результатов
 - Значительное согласие (каппа Флейса: 0.61-0.80)

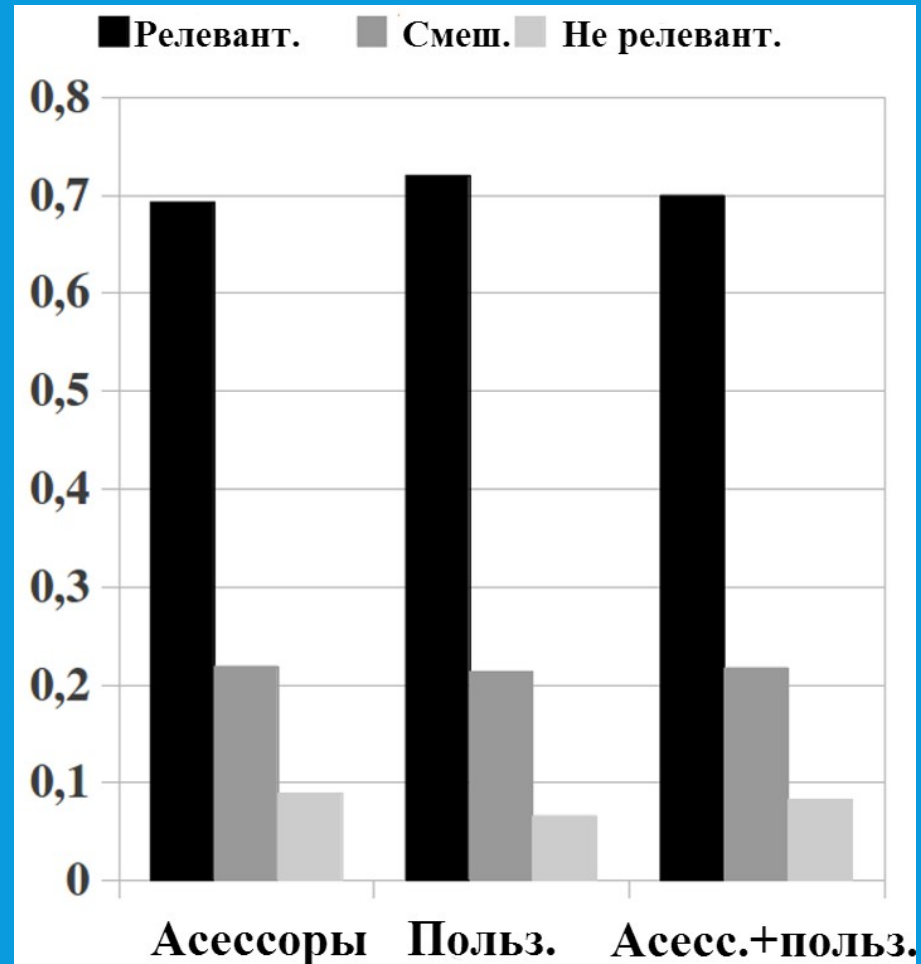
Извлечение семантических отношений



Удовлетворенность пользователей качеством поиска

- **23 асессора**
 - 20 запросов по своему усмотрению
 - Оценка первых 20 результатов – релевантные/нерелевантные/частично релевантные
- **460** суждений асессоров и **233** суждения анонимных пользователей
- **594 уникальных запроса**

Удовлетворенность пользователей качеством поиска



Выводы

- Разработана **система Серелекс**, позволяющая осуществлять поиск семантически связанных слов
- Точность системы сопоставима с **аналогичными разработками**
- **Не используются** составленные вручную словари
- Первый результат релевантен в **79% случаях**
- **70% пользователей** удовлетворены первыми 20 результатами