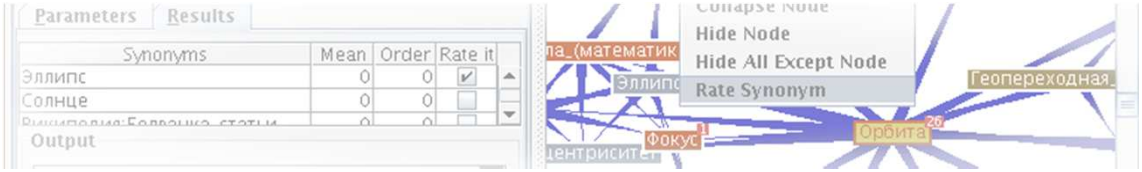
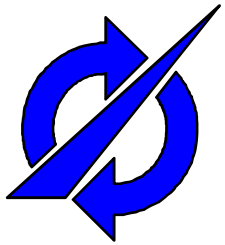
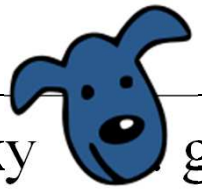


Оценка использования корпусов и электронных библиотек в Русском Викисловаре

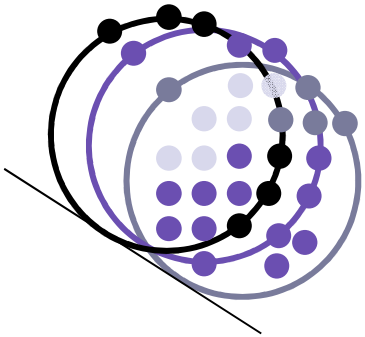
Санкт-Петербургский институт информатики и автоматизации РАН





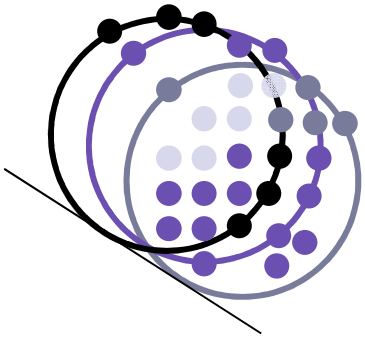
Крижановский Андрей (andrew.krizhanovsky gmail.com)

Содержание



- Введение
- Данные
  - Цитаты словарных статей Викисловаря
- Эксперимент
  - Корпуса текстов, электронные библиотеки
  - Авторы произведений
- Заключение

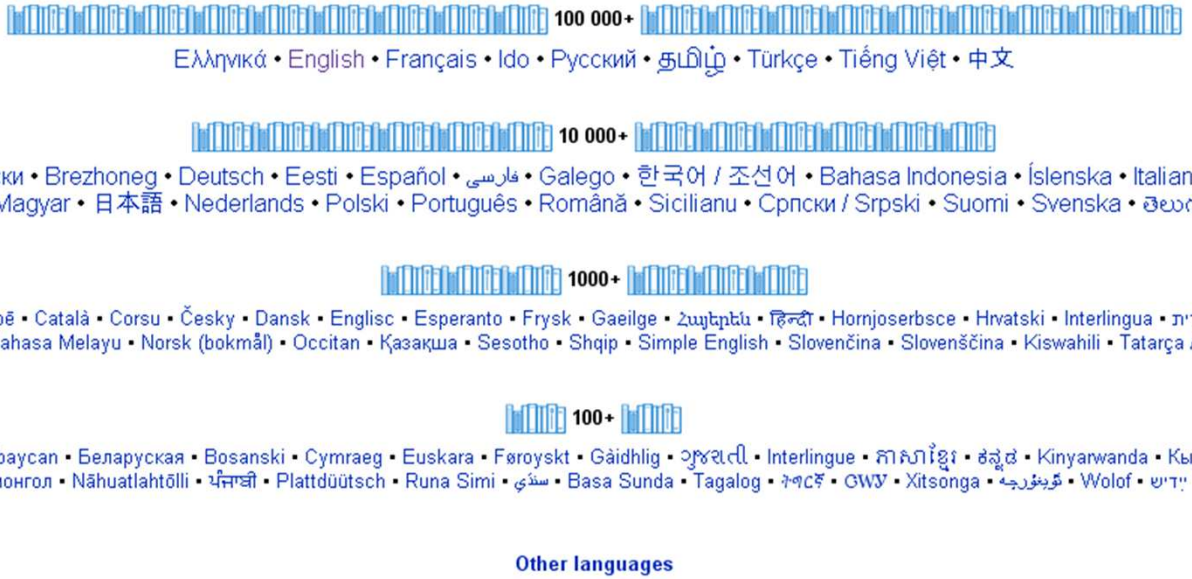
Цель



Анализ и оценка использования в Викисловаре данных корпусов и электронных библиотек

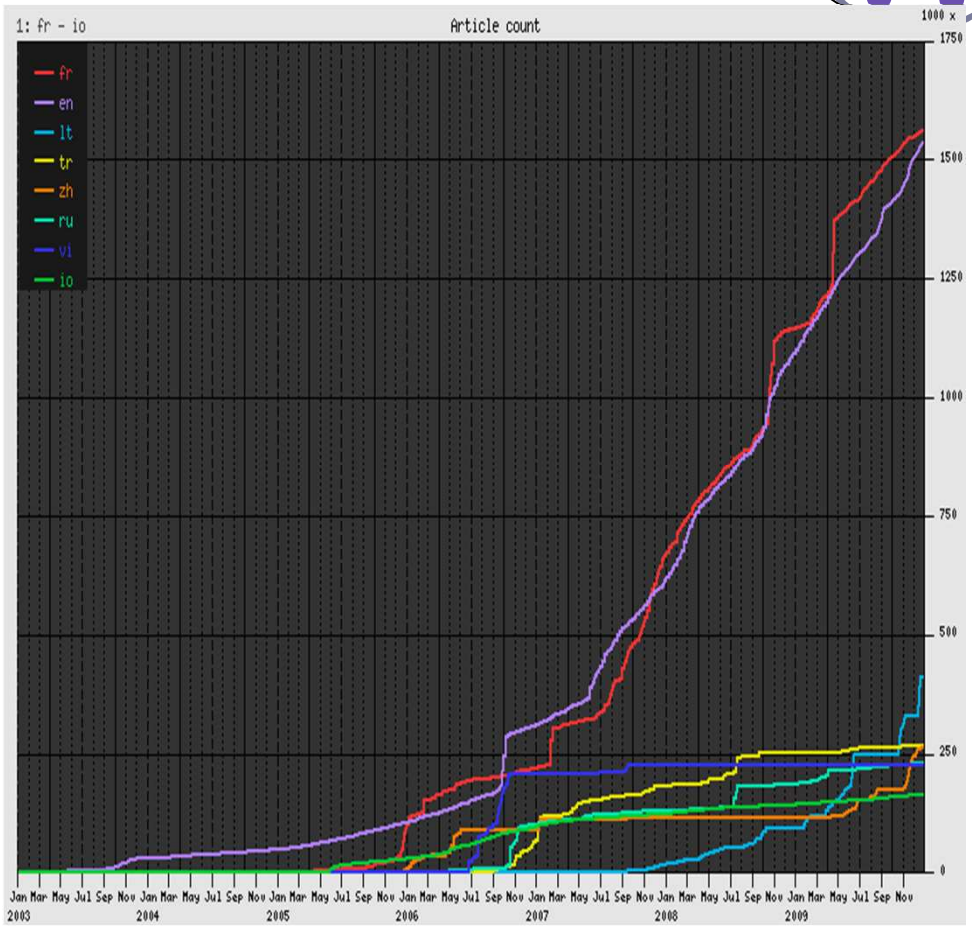
Викисловарь – много-функциональный многоязычный словарь и тезаурус

Грамматический
Толковый
Этимологический
Переводной



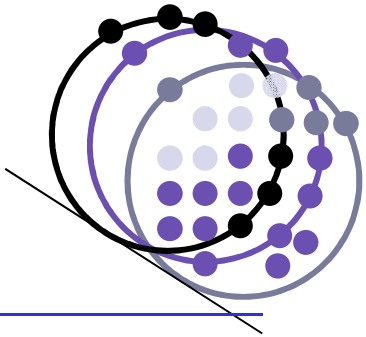
Развитие Викисловарей

+ Первым появился English Wiktionary в декабре 2002 г.
+ Проект Русский Викисловарь был запущен в мае 2004 г.



Восемь самых больших Викисловарей (2003-2010)

10 (из 170) крупнейших Викисловарей

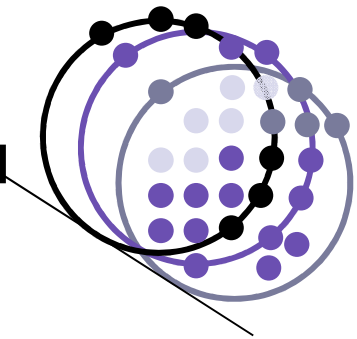


N	Языковая версия	Словарных статей	Администраторов	Активных редакторов
1	Английский	2 533 756	89	1044
2	Французский	2 031 596	23	312
3	Малагасийский	1 194 740	2	9
4	Китайский	1 116 479	8	44
5	Литовский	557 900	4	12
6	**Русский Викисловарь**	**291 187**	**6**	**141**
7	Турецкий	277 451	6	39
8	Вьетнамский	229 094	5	25
9	Польский	224 289	26	95
10	Тамильский	205 803	12	29

По данным на июнь 2011. Везде далее по данным на май 2011

Русский Викисловарь:

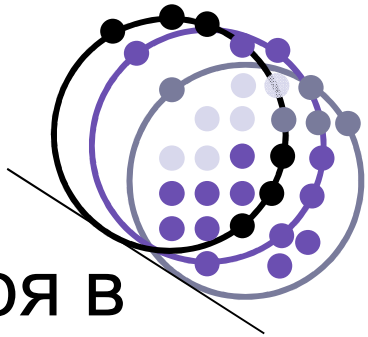
Число словарных статей по языкам (Многоязычность)



Раздел Русского Викисловаря	Число словарных статей
Русский	135 396
Украинский	89 712
Английский	35 993
Французский	20 854
Интерлингва	19 194
и так далее…	

- Словарные статьи о словах 385 языков.
- Цитаты представлены на 109 языках.
- Переводы русских слов – на 370 языков.

Задачи данной работы



- Преобразование текстов Викисловаря в машинно-читаемый словарь

  т.е. создание базы данных на основе Викисловаря

- Автоматическое извлечение из словаря:
  - Цитат, иллюстрирующих значение слова;
  - Сопутствующей информации (автор, название произведения, год, название корпуса или электронной библиотеки).

Фрагмент словарной статьи (раздаточный материал, стр. 1 и 3).

разг. завзятый игрок в карты ◆ Но этот чиновник был **картёжник** и проиграл её капитал, дачу, даже заложил все её бриллианты, да еще растратил казённые деньги.
А. Я. Панаева, «Воспоминания», 1889—1890 г. (цитата из Национального корпуса русского языка, см. Список литературы)

Фрагмент базы данных машинно-читаемого словаря для хранения цитат

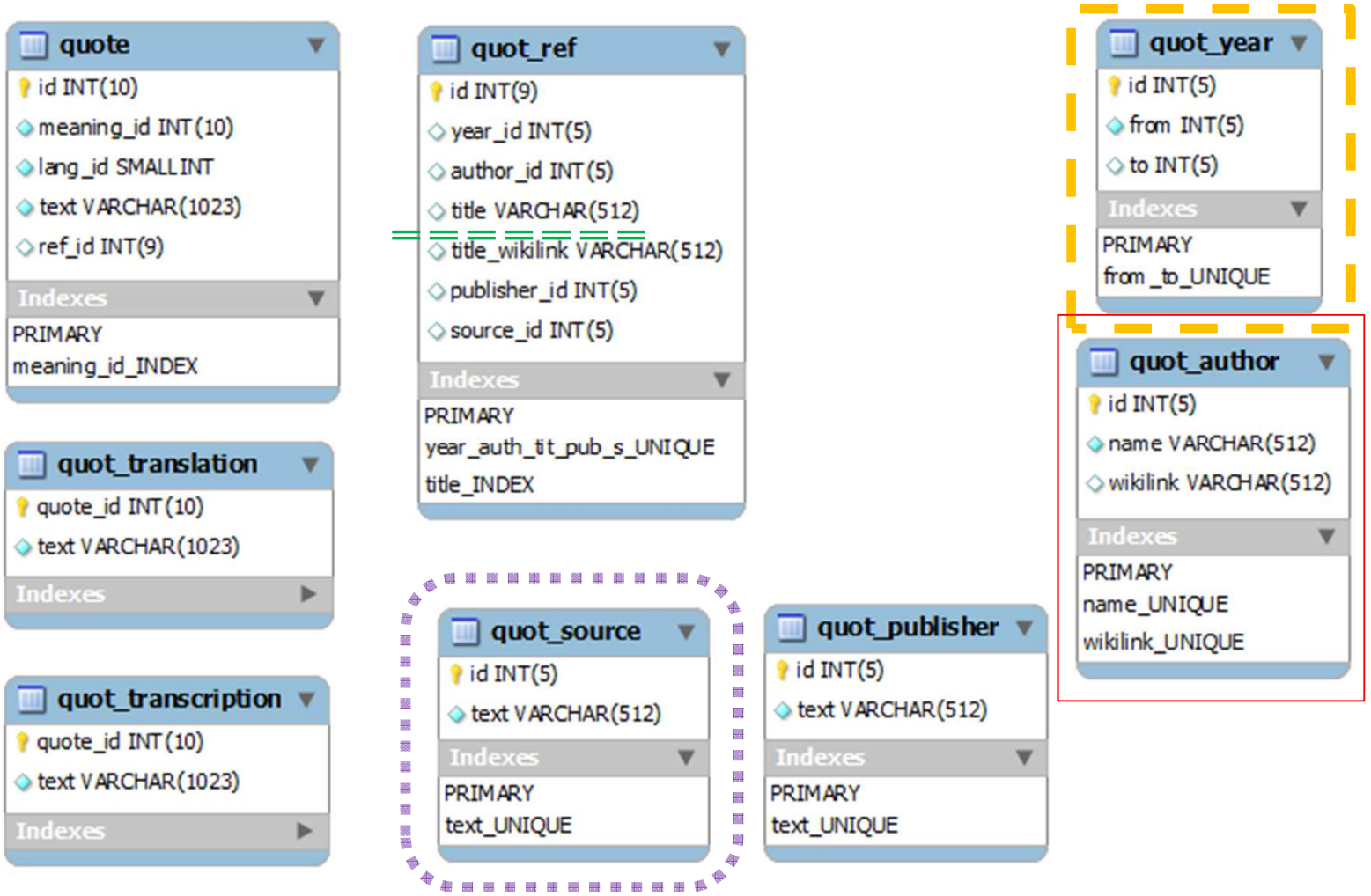
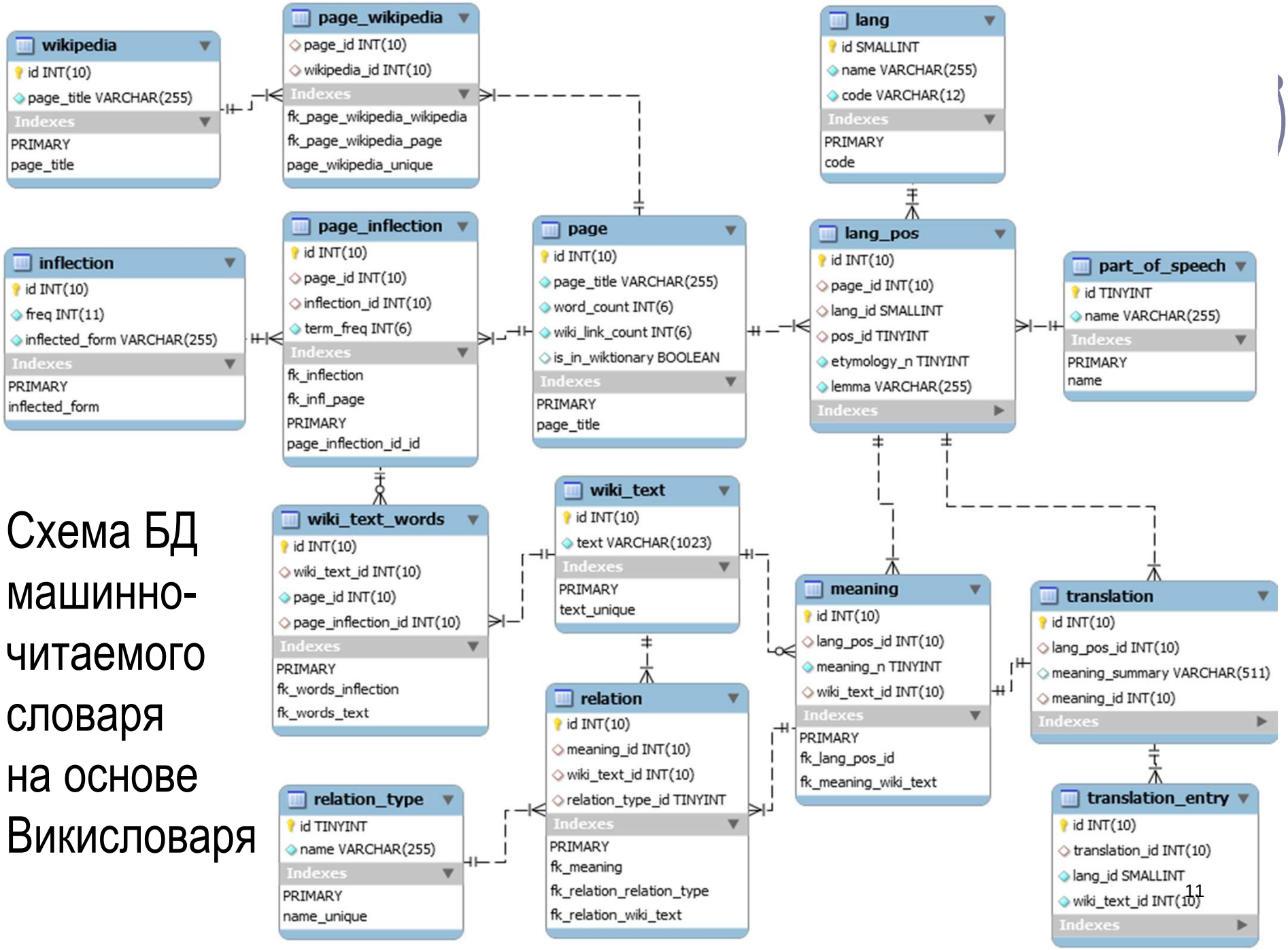
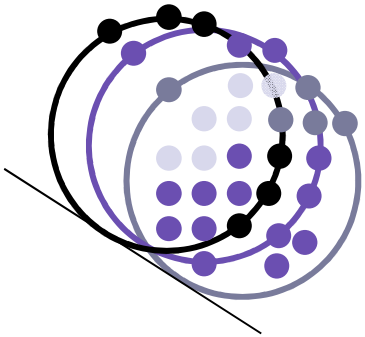


Схема БД машинно-читаемого словаря на основе Викисловаря



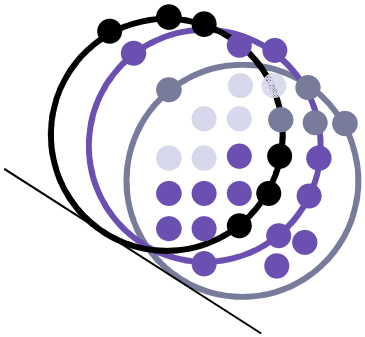
Эксперименты:
Автоматическое извлечение цитат из Викисловаря



- всего 50.5 тыс. цитат
- 17 тыс. цитат с указанием источников
- 23.5 тыс. с указанием авторов

Язык	Число цитат
Русский	42 262
Английский	1202
Татарский	952
Латинский	873
Немецкий	845
Сербский	606
Французский	463
Украинский	437

Эксперименты:
Корпуса текстов, электронные библиотеки в Викисловаре



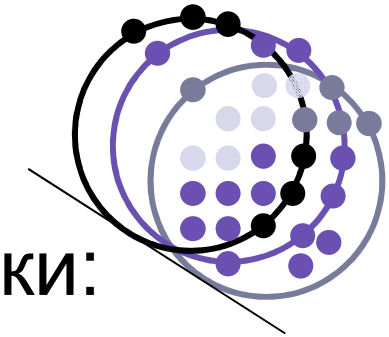
N	Шаблон в Викисловаре	Описание	Число цитирований
1	НКРЯ	**Национальный корпус русского языка**	16479
2	Lib	**Библиотека Максима Мошкова**	230
3	source	Викитека (электронная библиотека текстов)	90
4	OLD	Oxford Latin Dictionary, изд. OUP, 1968-82	74
5	Даль	Толковый словарь живого великорусского языка В.И. Даля	56
6	БП	Интернет-библиотека Беларуская Палічка	50
7	Ушаков	Толковый словарь русского языка: В 4-х т. / Под ред. Д. Н. Ушакова. — М.: Сов. энцикл.: ОГИЗ, 1935—1940.	48
8	BYU	BYU Corpus of American English	33
9	Википедия или ВП	Википедия	27
10	МАС	Малый академический словарь русского языка в 4-х т. (МАС)	21
11	CREA	Corpus de referencia del español actual	20
12	ЯРГ	Материалы проекта «Языки русских городов»	9
13	IS	Украино-русский параллельный корпус компании ElVisti	7
14	COCA	Корпус современного американского английского языка	6

Эксперименты:

Популярные в Викисловаре авторы произведений

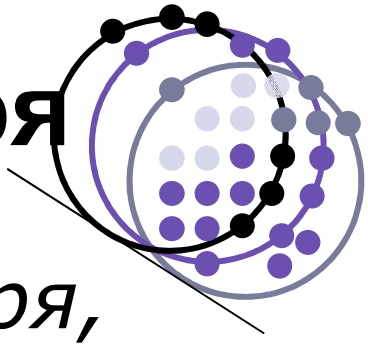
Автор	Число цитат
Чехов	716
Л. Н. Толстой	529
Пушкин	520
Достоевский	500
Тургенев	457
Гоголь	321
Лесков	245
Булгаков	207
Стругацкие	171
Виктор Астафьев	142

Реализация



- Программный код включает наработки:
  - synarcher – поиск синонимов в Википедии
  - wikidf – индексирование текстов Википедии
- Базы данных:
  - MySQL - для разработки и тестирования
  - SQLite – в скачиваемом приложении
- Надёжность и устойчивость (>300 Unit-тестов)
- Визуализация (Wiwordik, JavaFX)
- Открытая лицензия

{{Шаблоны}} Викисловаря

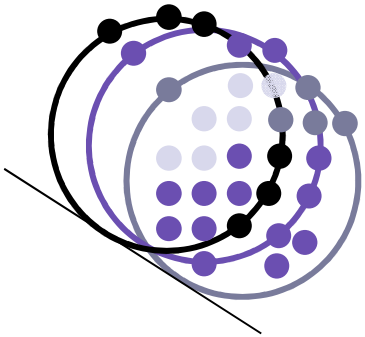


- это особые страницы Викисловаря, которые могут содержать программный код и позволяют:

+ централизованно сменить внешний вид сразу многих словарных статей;
+ делать массовые изменения статей в автоматическом режиме (например, указать род, указать перевод и т.д., при наличии соответствующих баз данных).

☻ Сложность освоения (новые редакторы)

{{Шаблоны}}. Примеры



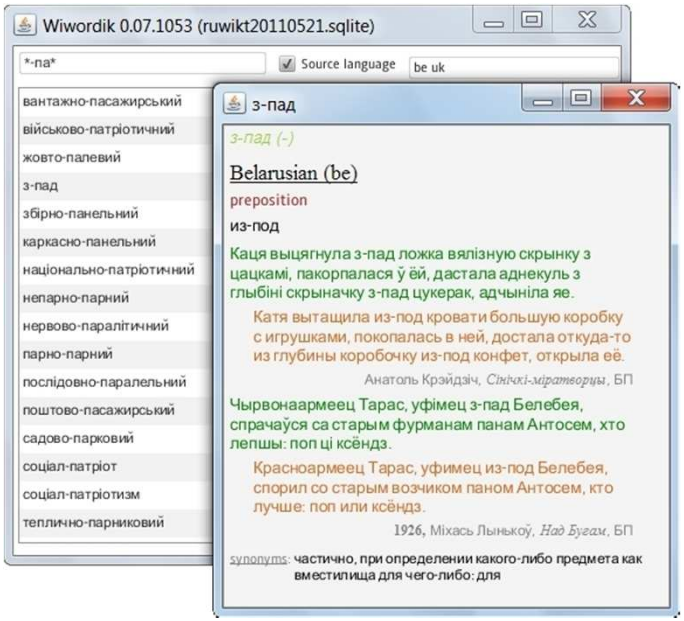
- {{-ru-}}, **{{пример|}}**
- Фонетические: {{transcriptions||}}, {{медиа}}
- {{морфо|под|вод|и|ть|ся}}
- Морфологические:
  - {{сущ ru}}, {{прил ru}}, {{сущ ru m ina}}, {{adv ru}}
  - {{сущ eo}}, {{прил eo}}, {{adv eo}}, {{гл eo}}
- **Шаблоны библиографии**
  - **{{НКРЯ}}, {{Ушаков1940}}, {{Эпитет1913}}**
- Технические (из Википедии):
  - {{За}}, {{wikify}}, «вавилонские шаблоны»

Результаты

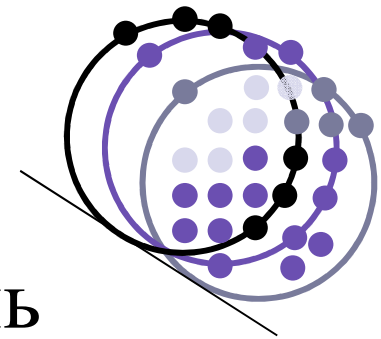
- Выполнен анализ и оценка использования в Русском Викисловаре данных корпусов и электронных библиотек

- Разработана графическая оболочка для данных Викисловаря

Сайт проекта:
http://code.google.com/p/wikokit/



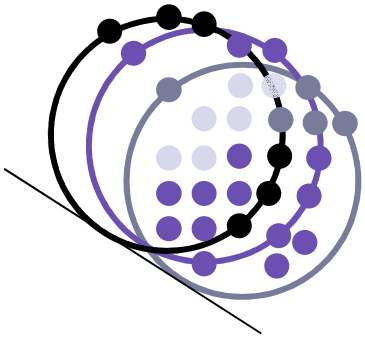
Данные Викисловаря: плюсы и трудности



+ Богатство
  + тезаурус (синонимы, антонимы…)
  + фразеологизмы
  + этимология
  + произношение
  + толкование, цитаты
  + переводы
  + …
+ Быстрый рост
+ Интервики (доп. д.)
+ Свободная лицензия

– Разная степень стандартизации и формализации в разных Викисловарях
– Быстрый рост данных, но ошибки, так как:
  – ручной ввод данных,
  – новичкам требуется время на усвоение правил Словаря
– Омонимия вне страницы

Будущая работа
(сделано и *ещё делать*)



- Извлечение из Русского Викисловаря
  - Толкование
    - определение
    - *помета*, **цитата**, *картинка*
  - Парадигматические отношения (синонимы, гиперонимы…, помета)
  - Перевод
  - *Фонетика*
    - *Транскрипция, Аудио*
  - *Этимология*
  - *Фразеологизмы, поговорки, пословицы*
  - …

Английский Викисловарь

Спасибо за внимание!

http://ru.wiktionary.org/