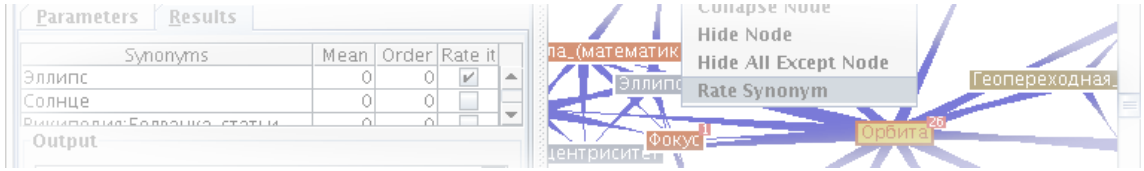
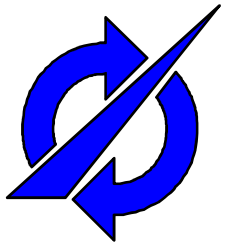
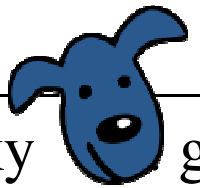


Машинная обработка Русского Викисловаря

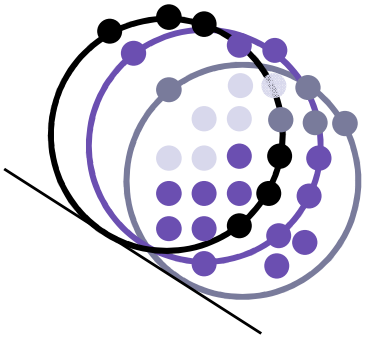
Санкт-Петербургский институт информатики и автоматизации РАН





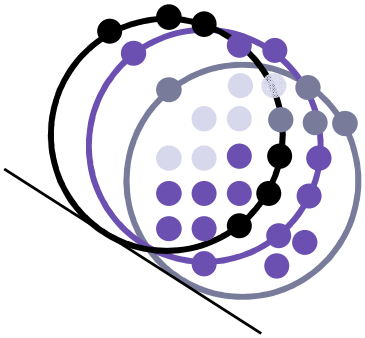
Крижановский Андрей (andrew.krizhanovsky gmail.com)

Содержание



- Викисловарь
  - применение
  - достоинства и трудности обработки
- MRD, парсер и сравнение Викисловарей
- Эксперимент
  - Корреляция мер семантической близости
- Результаты

Цель



Сделать возможным

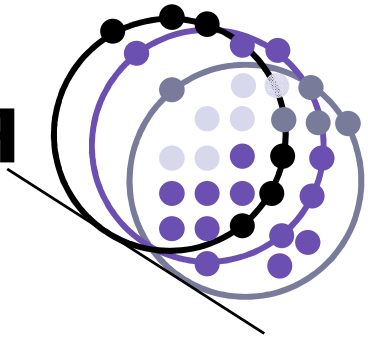
применение данных Викисловаря

(как лингвистического ресурса)

в различных компьютерных программах,

в задачах, связанных с обработкой текста.

Применение Викисловаря



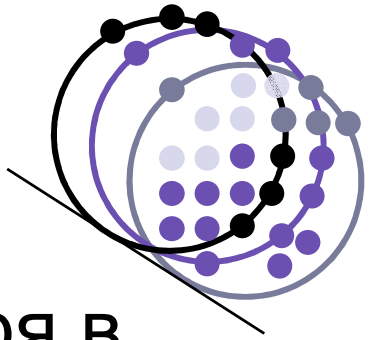
- в компьютерных программах:
  - текстовые поисковые системы

    расширение / переформулировка запросов с помощью тезаурусов
  - запросно-ответные системы

    распознавание запроса
- в задачах:
  - определение значения многозначного слова
  - сравнение онтологий (ontology matching)
  - автоматическое создание тезаурусов
  - машинный перевод
  - компьютерные игры для изучения языков

Медиа данные (звук + иллюстрации)

Задача

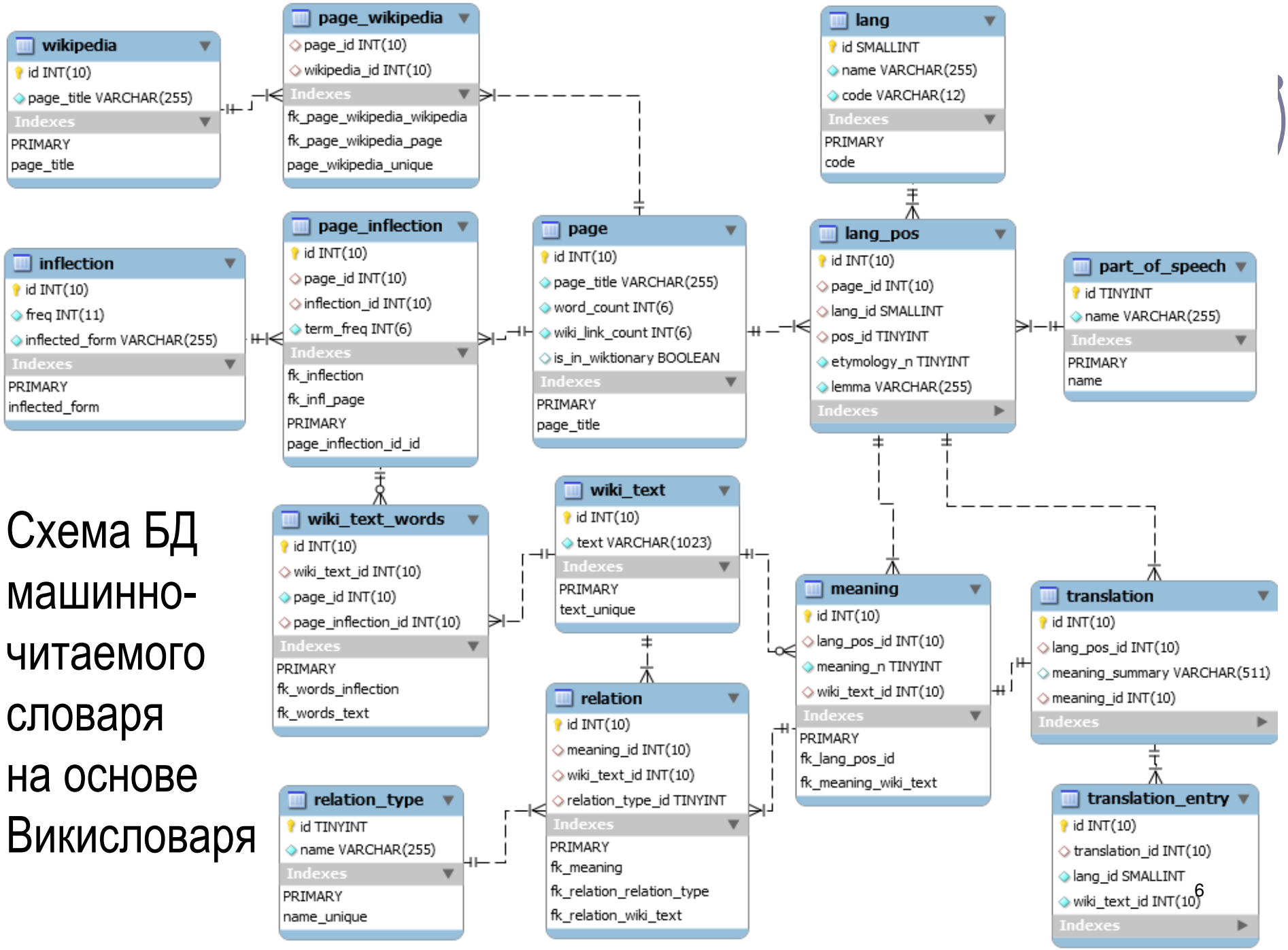


- Преобразования данных Викисловаря в машинную форму, а именно: машинно-читаемый словарь (MRD).

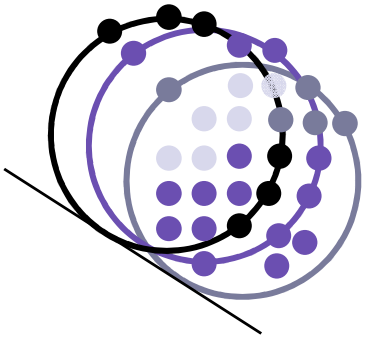
MRD включает:

- Данные (база данных),
- Алгоритмы и функции (API)

Схема БД машинно-читаемого словаря на основе Викисловаря



Викисловарь = вики + ?



- Структура статьи = f (язык, ~часть речи)
  - Определена последовательность частей статьи
  - Шаблоны:
    - структурные шаблоны ({{пример|}}, {{морфо|}})
    - словоизменений, этимологии, родств. слова, пометы…

Т.о. жёсткая схема даёт:

- \+ единообразие, системность
- \+ возможность автоматически анализировать текст

Викисловарь – много-функциональный многоязычный словарь и тезаурус

---

100 000+

Ελληνικά • English • Français • Ido • Русский • தமிழ் • Türkçe • Tiếng Việt • 中文

10 000+

Afrikaans • العربية • Български • Brezhoneg • Deutsch • Eesti • Español • فارسی • Galego • 한국어 / 조선어 • Bahasa Indonesia • Íslenska • Italiano • Kurdî / كوردی • Lietuvių • Limburgs • Magyar • 日本語 • Nederlands • Polski • Português • Română • Sicilianu • Српски / Srpski • Suomi • Svenska • తెలుగు • Volapük

1000+

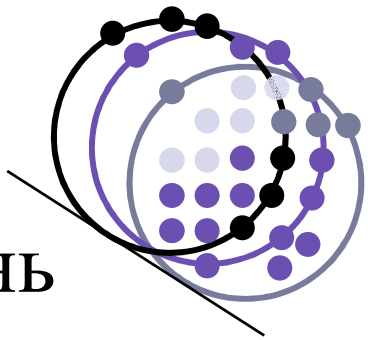
Asturianu • Bân-lâm-gú / Hō-ló-oē • Català • Corsu • Česky • Dansk • Englisc • Esperanto • Frysk • Gaeilge • Հայերեն • हिन्दी • Hornjoserbsce • Hrvatski • Interlingua • עברית • Kalaallisut • Kaszëbsczi • ພາສາລາວ • Latina • മലയാളം • Bahasa Melayu • Norsk (bokmål) • Occitan • Қазақша • Sesotho • Shqip • Simple English • Slovenčina • Slovenščina • Kiswahili • Tatarça / татарча • ไทย • Українська • اردو

100+

አማርኛ • Aragonés • Avañe'ẽ • Azərbaycan • Беларуская • Bosanski • Cymraeg • Euskara • Føroyskt • Gàidhlig • ગુજરાતી • Interlingue • ភាសាខ្មែរ • ಕನ್ನಡ • Kinyarwanda • Кыргызча • Latviešu • Македонски • मराठी • монгол • Nāhuatlahtōlli • ਪੰਜਾਬੀ • Plattdüütsch • Runa Simi • سنڌي • Basa Sunda • Tagalog • ትግርኛ • GW̃ • Xitsonga • ئۇيغۇرچە • Wolof • יידיש • isiZulu

Other languages

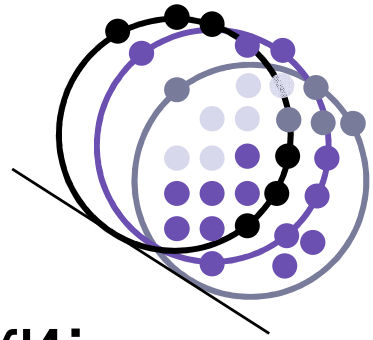
Данные Викисловаря: плюсы и трудности



+ Богатство
  + тезаурус (синонимы, антонимы…)
  + словосочетания
  + этимология
  + произношение
  + примеры употр-ий
  + переводы
  + …
+ Быстрый рост
+ Интервики (доп. д.)
+ Свободная лицензия

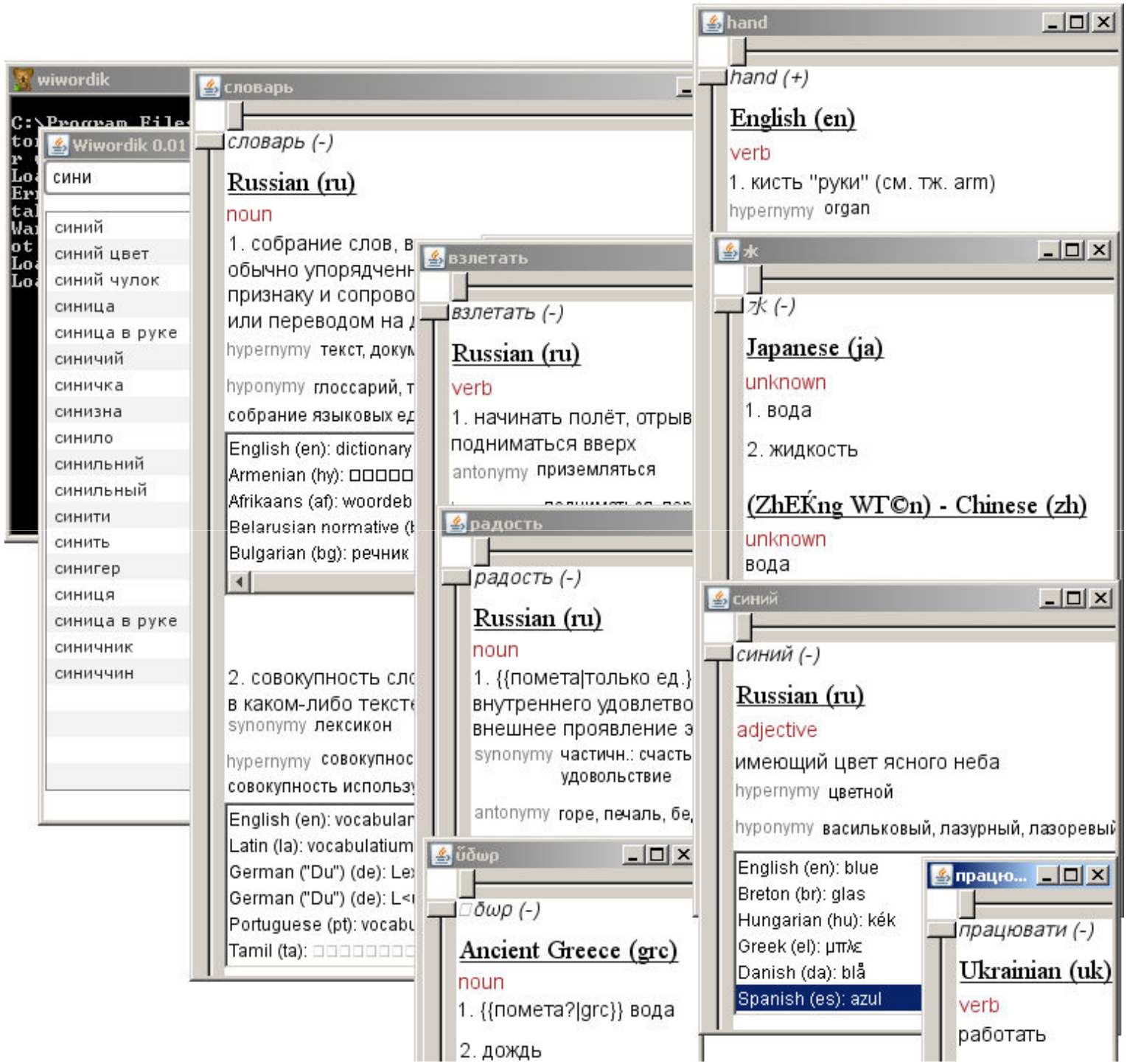
– Разная степень стандартизации и формализации (структура статьи) в разных Викисловарях
– Быстрый рост данных, но *толпа*:
  – Ручной ввод данных =>
  – Ошибки => Парсер д.б. устойчив!
– Омонимия вне страницы (см. дальше)

Реализация 1



- Программный код включает наработки:
  - synarcher – поиск синонимов в Википедии
  - wikidf – индексирование текстов Википедии
- Java
- База данных:
  - MySQL - для разработки и тестирования
  - SQLite – в скачиваемом приложении
- JUnit тестирование

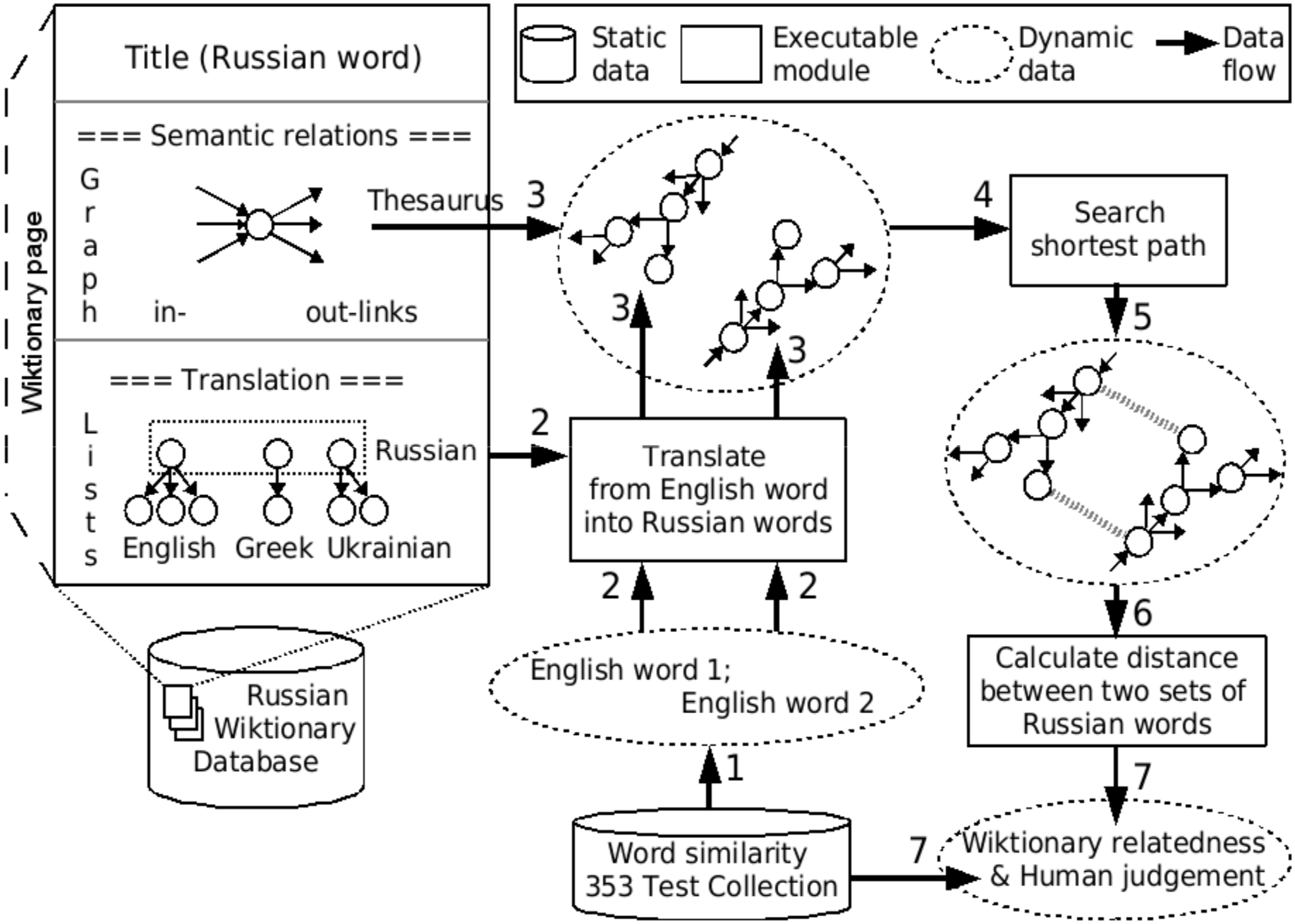
Реализация


wiwordik
Wiwordik 0.01
сини
синий
синий цвет
синий чулок
синица
синица в руке
синичий
синичка
синизна
синило
синильний
синильный
синити
синить
синигер
синиця
синица в руке
синичник
синиччин
словарь
словарь (-)
Russian (ru)
noun
1. собрание слов, в
обычно упорядченн
признаку и сопрово
или переводом на д
hypernymy текст, докум
hyponymy глоссарий, т
собрание языковых ед
English (en): dictionary
Armenian (hy): □□□□□
Afrikaans (af): woordeb
Belarusian normative (
Bulgarian (bg): речник
2. совокупность сло
в каком-либо тексте
synonymy лексикон
hypernymy совокупнос
совокупность использу
English (en): vocabular
Latin (la): vocabulatium
German ("Du") (de): Le
German ("Du") (de): L<
Portuguese (pt): vocabu
Tamil (ta): □□□□□□□□
взлетать
взлетать (-)
Russian (ru)
verb
1. начинать полёт, отрыв
подниматься вверх
antonymy приземляться
радость
радость (-)
Russian (ru)
noun
1. {{помета|только ед.}}
внутреннего удовлетво
внешнее проявление э
synonymy частичн.: счасть
удовольствие
antonymy горе, печаль, бе
ὕδωρ
δωρ (-)
Ancient Greece (grc)
noun
1. {{помета?|grc}} вода
2. дождь
hand
hand (+)
English (en)
verb
1. кисть "руки" (см. тж. arm)
hypernymy organ
ж
ж (-)
Japanese (ja)
unknown
1. вода
2. жидкость
(ZhEḱng WT©n) - Chinese (zh)
unknown
вода
синий
синий (-)
Russian (ru)
adjective
имеющий цвет ясного неба
hypernymy цветной
hyponymy васильковый, лазурный, лазоревый
English (en): blue
Breton (br): glas
Hungarian (hu): kék
Greek (el): μπλε
Danish (da): blå
Spanish (es): azul
працю...
працювати (-)
Ukrainian (uk)
verb
работать

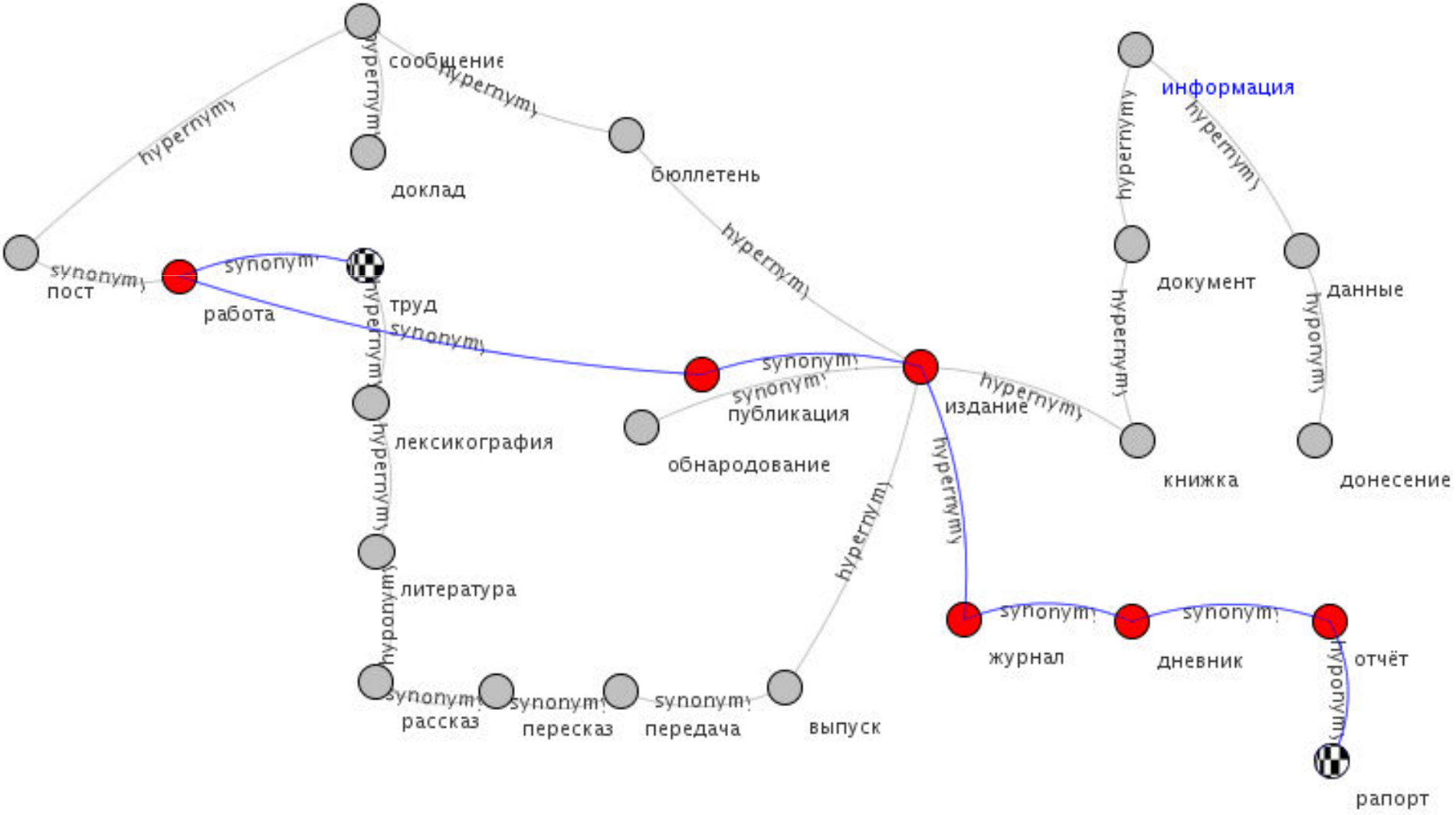
Размеры Викисловарей

	Wiktionary editions as of September 2007, from [24]				A part of Wiktionary extracted by the parser. Wiktionary edition as of January 2009.				
	English Wiktionary		German Wiktionary		Russian Wiktionary				
	English	German	English	German	Total[5]	English	German	Russian	Ukrainian
Entries	176,410	10,487	3,231	20,557	247,580	2,813[6]	13,072	124,301	88,575
Part of speech (POS)									
Nouns	99,456	6,759	2,116	13,977	108,448	935	336	58,843	40,607
Verbs	31,164	1,257	378	1,872	26,290	342	49	356[7]	24,096
Adjectives	23,041	1,117	357	2,261	26,864	184	18	2,168	23,536
Unknown	POS which were not recognized by the parser				80,293	1,321	12,648	57,573	331
Semantic relations									
Synonyms	29,703	1,916	2,651	34,488	**28,718**	1,345	665	24,338	310
Antonyms	4,305	238	283	10,902	10,480	238	234	9,062	54
Hypernyms	42	0	336	17,286	18,975	444	474	17,033	115
Hyponyms	94	0	390	17,103	8,585	176	473	7,574	12
Holonyms	–	–	–	–	216	1	0	215	0
Meronyms	–	–	–	–	322	8	2	306	0
Total	–	–	–	–	**67,296**	2,212	1,848	58,528	491

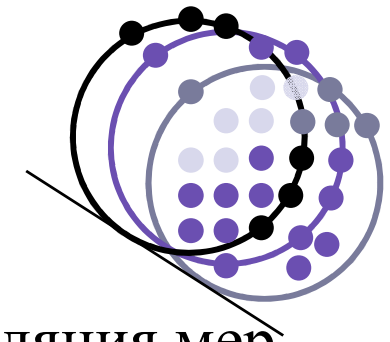
WordNet (2006): 150,000 слов, 115,000 синсетов (наборов синонимов)


Static data
Executable module
Dynamic data
Data flow
Wiktionary page
Title (Russian word)
=== Semantic relations ===
Graph
in-
out-links
Thesaurus
3
=== Translation ===
Lists
Russian
English
Greek
Ukrainian
2
Translate from English word into Russian words
Russian Wiktionary Database
English word 1; English word 2
1
Word similarity 353 Test Collection
4
Search shortest path
5
6
Calculate distance between two sets of Russian words
7
Wiktionary relatedness & Human judgement

Кратчайший путь в Русском Викисловаре



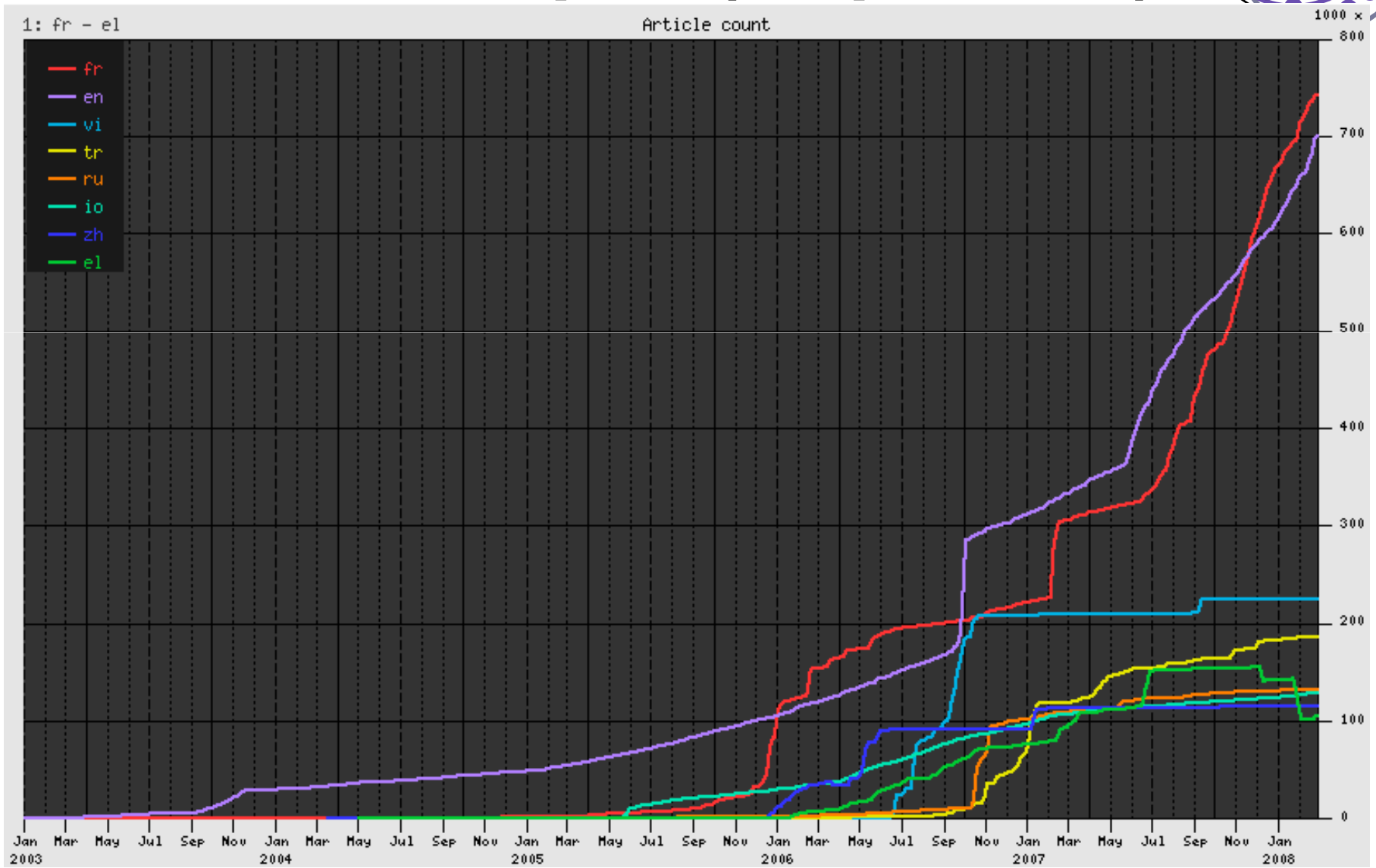
Корреляция мер семантической близости



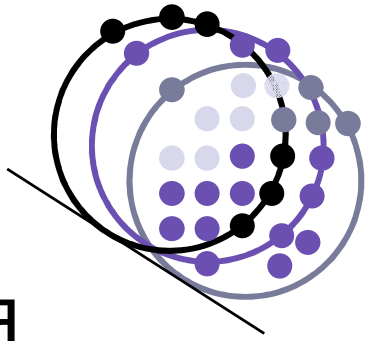
Dataset	WN	WP	WT	Others
Metric or Algorithm	I. Path based measures (in taxonomy)			
wup	0.3	0.47	–	–
lch	**0.34**	**0.48**	–	–
res $_{hypo}$	–	0.25-0.37[8]	–	–
jarmasz	–	–	–	**0.539** RT[9]
path $^{max}_{len}$	–	–	**0.24**	–
	II. Words frequency in corpus			
jaccard	–	–	–	Google 0.18
res	0.34	–	–	–
LSA	–	–	–	IntelliZap 0.56
ESA	–	**0.75**	–	–
	III. Text overlapping			
lesk	0.21	0.2	–	–
text	–	0.19	–	–

Корреляция мер семантической близости слов:
1) значения экспертов (набор 353-TC),
2) значения вычислены автоматически на основе WordNet, Английской Википедии, Русского Викисловаря

Восемь самых больших Викисловарей (Март 2008)

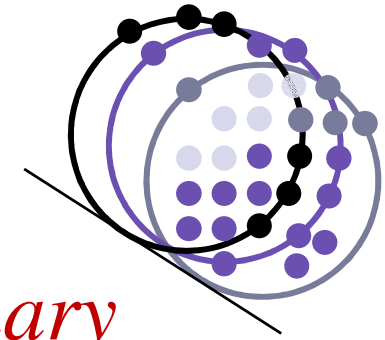


Результаты



- Создан парсер Русского Викисловаря
  - Спроектирована схема БД
  - Реализован доступ к БД (API, Java)
- Выполнено сравнение результатов поиска семантически близких слов на основе Викисловаря и тезауруса WordNet
- Сайт проекта (Wiki tool kit)
  - http://code.google.com/p/wikokit/

Сделано и *ещё делать*
(схема БД, парсер)



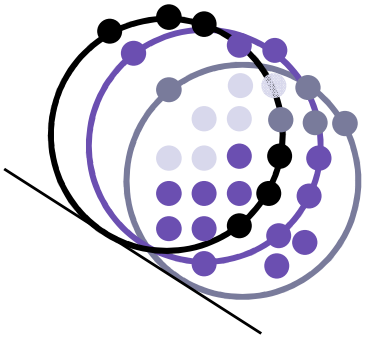
Русский Викисловарь

- Извлекаются (RE)
  - Толкование
    - определение
    - помета, цитата, картинка
  - Отношение
    (синонимы…, помета)
  - Перевод
  - *Фонетика*
    - *Транскрипция, Аудио*
  - *Этимология*
  - …
- *API Базы данных*

English Wiktionary

- Извлечь
  - Толкование
  - Отношение
    (синонимы…)
  - Перевод
  - …
- *API Базы данных*

Планы



- Продолжить разработку MRD
  - Наращивание функц-ти парсера, отладка
  - + English Wiktionary
- Визуализация (JavaFX)
  - MRD браузер
  - Игры и тесты (изучение иностранных языков)

Спасибо за внимание!