АВТОМАТИЗАЦИЯ ПОСТРОЕНИЯ АНГЛО-РУССКОГО WORDNET

А.М. Сухоногов
Петербургский Университет путей сообщения, кафедра ИВС
ASukhonogov@rambler.ru;

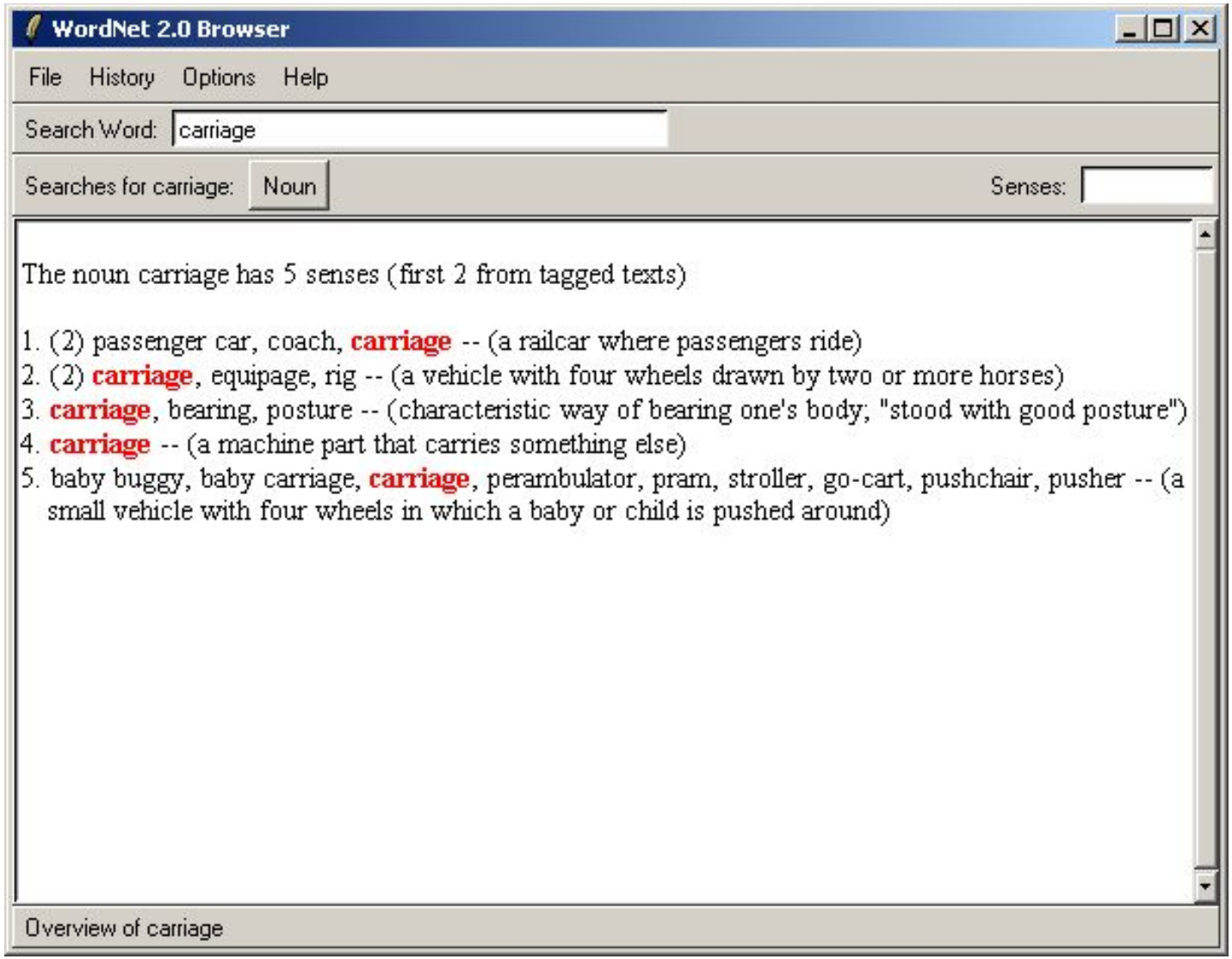
С.А. Яблонский
Петербургский Университет путей сообщения, кафедра ИВС
ЗАО “Руссикон”
serge_yablonsky@hotmail.com;
info@russicon.ru

Организация WordNet

WordNet – лексико-семантическая база данных, включающая:

- основную лексику языка (существительные, глаголы, прилагательные и наречия - более 100 тыс. словарных статей), организованную в виде синсетов.
  - ***Synset*** (синсет) – основная структура, представляющая словарную статью в WordNet. Синсет представляет множество лексем с одинаковым значением.
- таксономию отношений между синсетами (например, гипонимия, меронимия) и между лексемами (например, антонимия).
- определение семантических классов – ***TopOntology***

Princeton WordNet 2.0.



Почему WordNet ?

1. Наиболее полно отражает лексику английского и др. языков. Число входов (синсетов/слов) > 180 000.
2. Постоянное развитие PWN – версия 2.1.
3. ”Параллельный” перевод на >17 языков. (EuroWordNet, Balkanet, Корейский и др.)
4. Встроенные морфологические анализаторы, “привязанные” к национальным языкам.
5. PWN как межъязыковой индекс.
6. Разработка онтологий на базе WordNet. SUMO mappings to WordNet 2.0.
7. Разрабатывается RDF/OWL форматы WN для Semantic Web.

Проекты WordNet

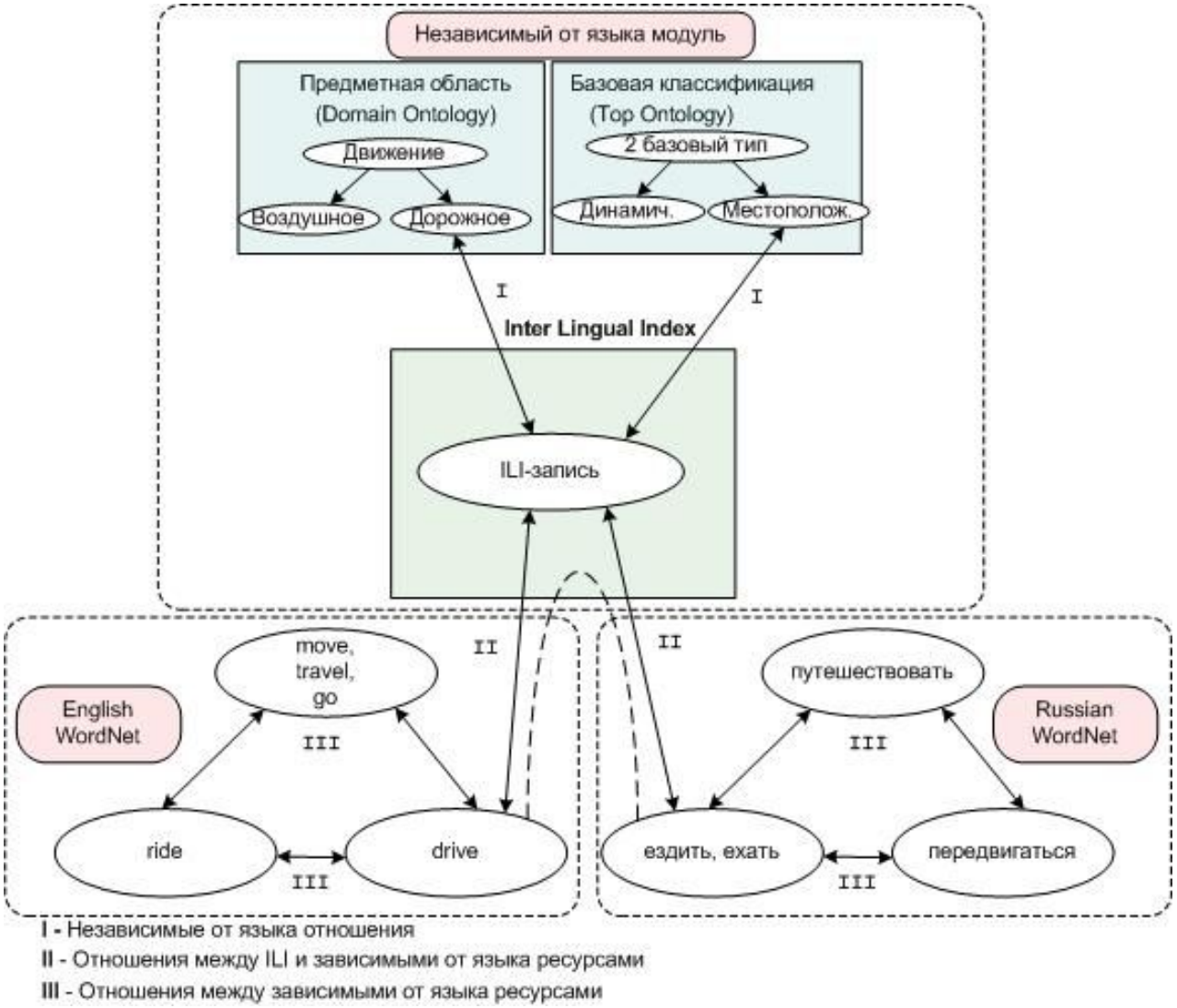
- Английский
- Датский
- Испанский
- Итальянский
- Немецкий
- Французский
- Чешский
- Эстонский

- Греческий
- Болгарский
- Турецкий
- Румынский
- Сербский

- Индийский
- Китайский
- Японский

GWA – Global WordNet Association (2001 г.)

Межъязыковой индекс ILI – Inter-lingual-index



WordNet русского языка

1. Проект филологического факультета, кафедра компьютерной лингвистики СПбГУ

   http://www.phil.pu.ru/depts/12/RN/bibliography_ru.shtml

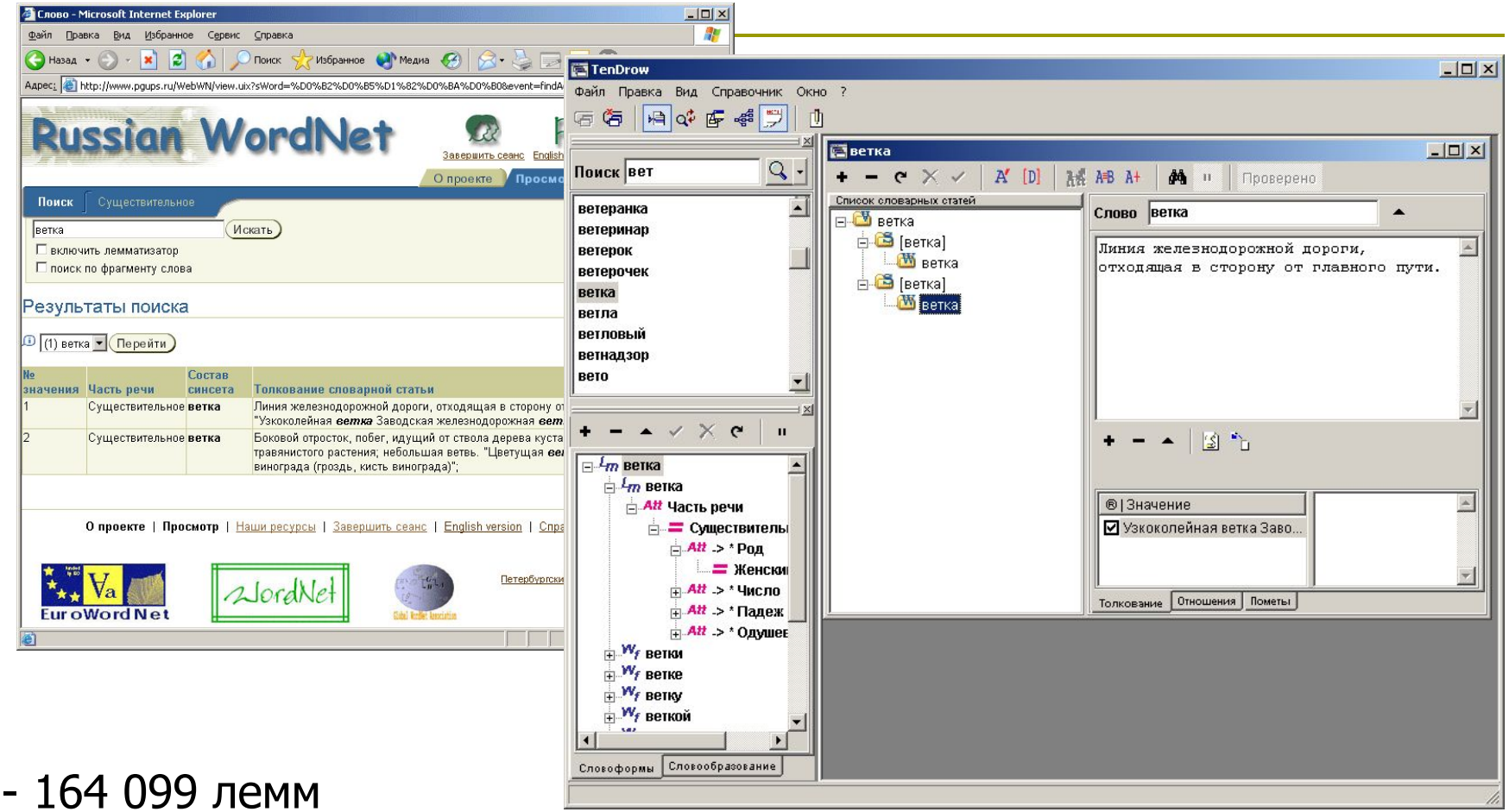
   http://www.kiberry.ru:8085/index.jsp

2. Проект “УИС Россия”

   http://www.cir.ru/

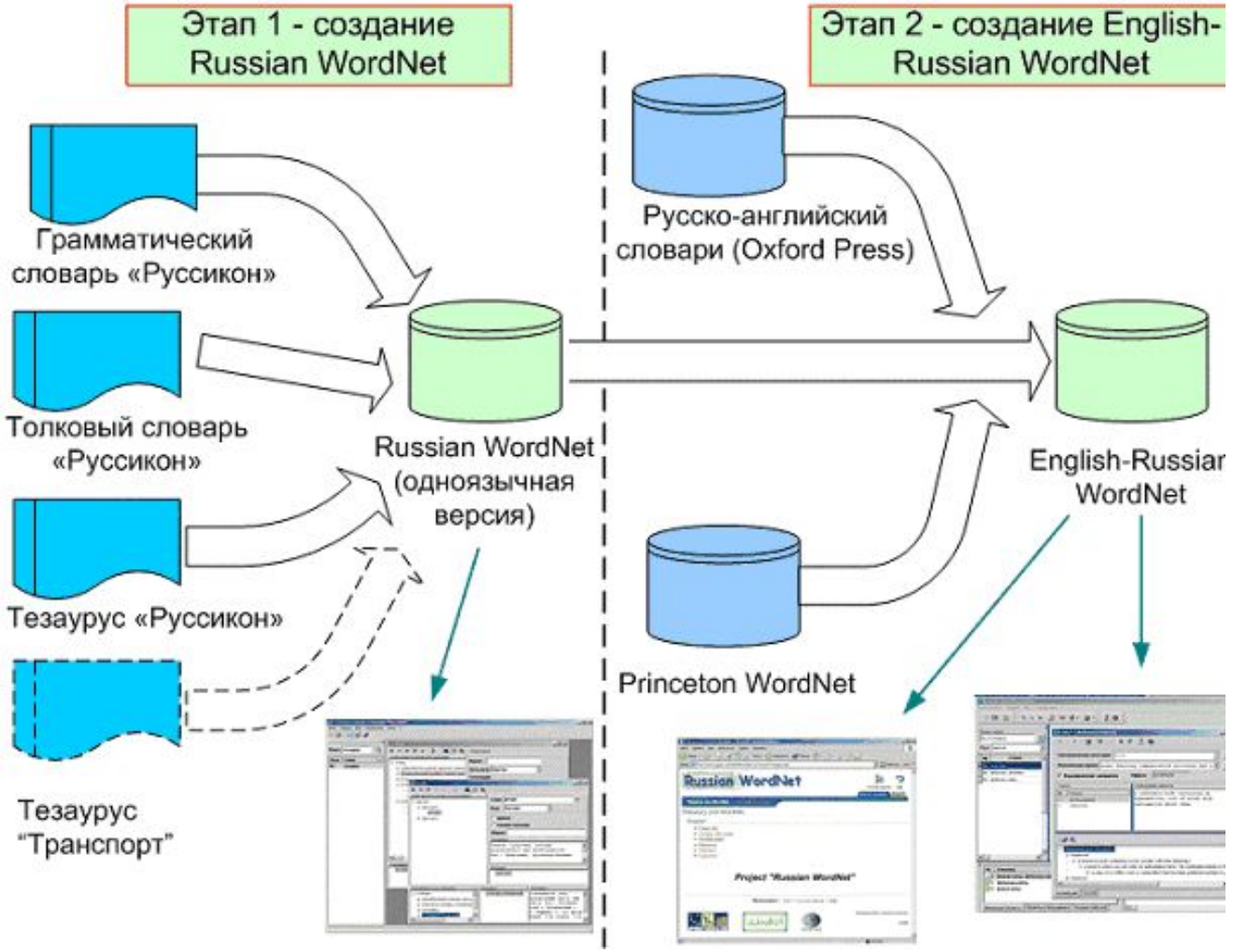
3. Проект “Russian WordNet”

Проект “Russian WordNet”



- 164 099 лемм
  и их парадигмы, более 3,5 млн. словоформ
- 202 866 синсетов (значений)

Основные этапы «Russian WordNet»

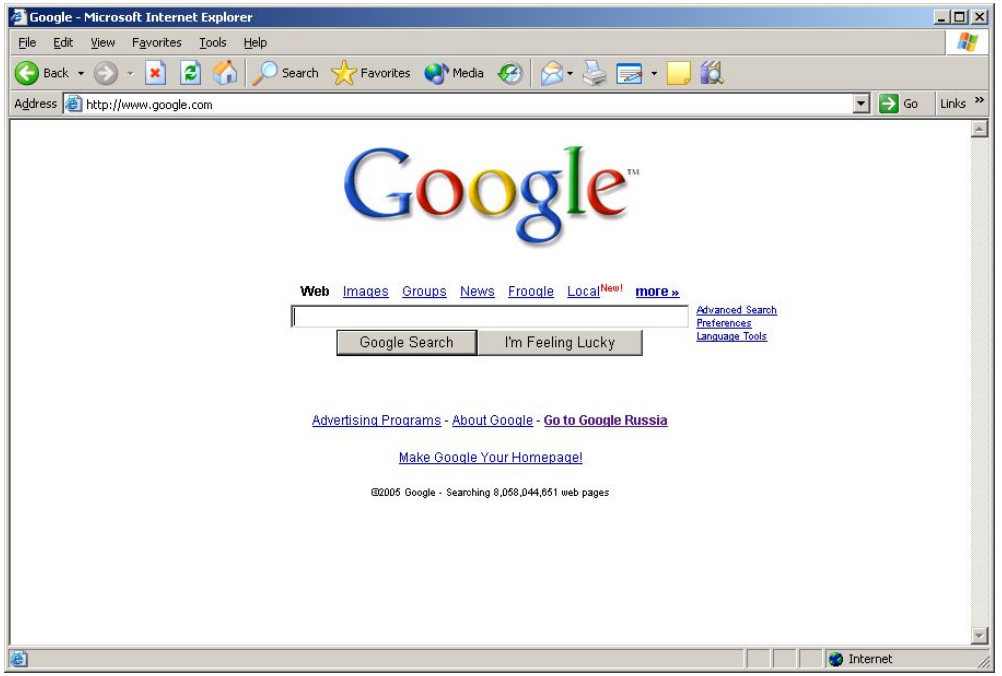


WordNet

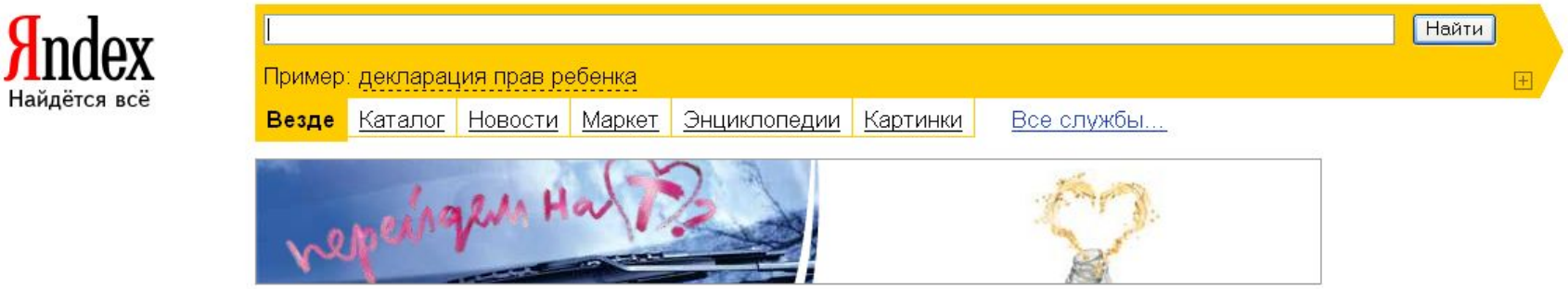
В общем случае отображение L1->L2 невыполнимо, поскольку
- для некоторого слова W_{L1} может не существовать соответствующего слова W_{L2}, т.е. перевод может отсутствовать,
- число значений lemmat (W_{L1}) может быть не равно числу значений lemmat (W_{L2}) и/или значения могут не совпадать
- некоторое слово W_{L1} может переводиться не одним словом а некоторым словосочетанием, не являющимся в общем случае фразеологизмом или устойчивым словосочетанием языке L2.

Google сегодня

- Поисковый индекс, включающий порядка ~10 миллиардов документов, в т.ч. на русском языке (сколько?)
- Свободно распространяемый (с ограничениями) Java API для доступа к поисковому индексу



Яndex сегодня



- В поиске Яндекса сегодня:
  - уникальных серверов: 2 100 646,
  - уникальных документов: 727 070 847,
  - объем проиндексированной информации: 20 228 ГБ.
- Свободно распространяемый XML API для доступа к поисковому индексу

Определение «семантического расстояния» между словами

Пусть x – слово, w – страница (документ), проиндексированный поисковой машиной Google.

$$p(x) = \frac{(w : x \in w)}{M} = \frac{f(x)}{M}$$

- вероятность появления слова x в коллекции из M документов

$$p(x, y) = \frac{(w : x, y \in w)}{M} = \frac{f(x, y)}{M}$$

- вероятность совместного появления слова x и y в одном и том же документе

M=8 058 044 651 (~8 млрд.)

[Google]

Определение «семантического расстояния» между словами

$$p_y(x) = \frac{p(x,y)}{p(y)},$$

$$p_x(y) = \frac{p(x,y)}{p(x)}$$

Условные вероятности появления слов в коллекции документов.

Эти вероятности характеризуют зависимость, существующую между словами x и y, позволяют определять ассоциативные связи между словами.

Определение «семантического расстояния» между словами

Normalized Google distance (NGD):

$$NGD(x,y) = \frac{\max\{\log(f(x), \log f(y)\} - \log f(x,y)}{\log M - \max\{\log f(x), \log f(y)\}}$$

- Функция не определена для f(x)=f(y)=0
- NGD=∞, при f(x,y)=0, f(x)>0, f(y)>0
- NGD>0 в других случаях.
  Значения NGD(x,y) лежат в диапазоне от 0 до ∞,
  D(x,x)=0 для любого x.
- Функция симметрична, NGD(x,y)=NGD(y,x)

* Paul Vitanyi, Rudi Cilibrasi "Normalised Google Distance"

Наши ресурсы

- New Oxford Dictionary (SGML-формат, по лицензии на использование в исследовательских целях)

  Более 180 тыс. слов, 290 тыс. примеров употребления
- Доступ к ресурсам Яндекса, грант #103003 “Построения межъязыкового индекса для русской и английской версий WordNet”

Автоматизированное построение ILI-индекса. Основные этапы.

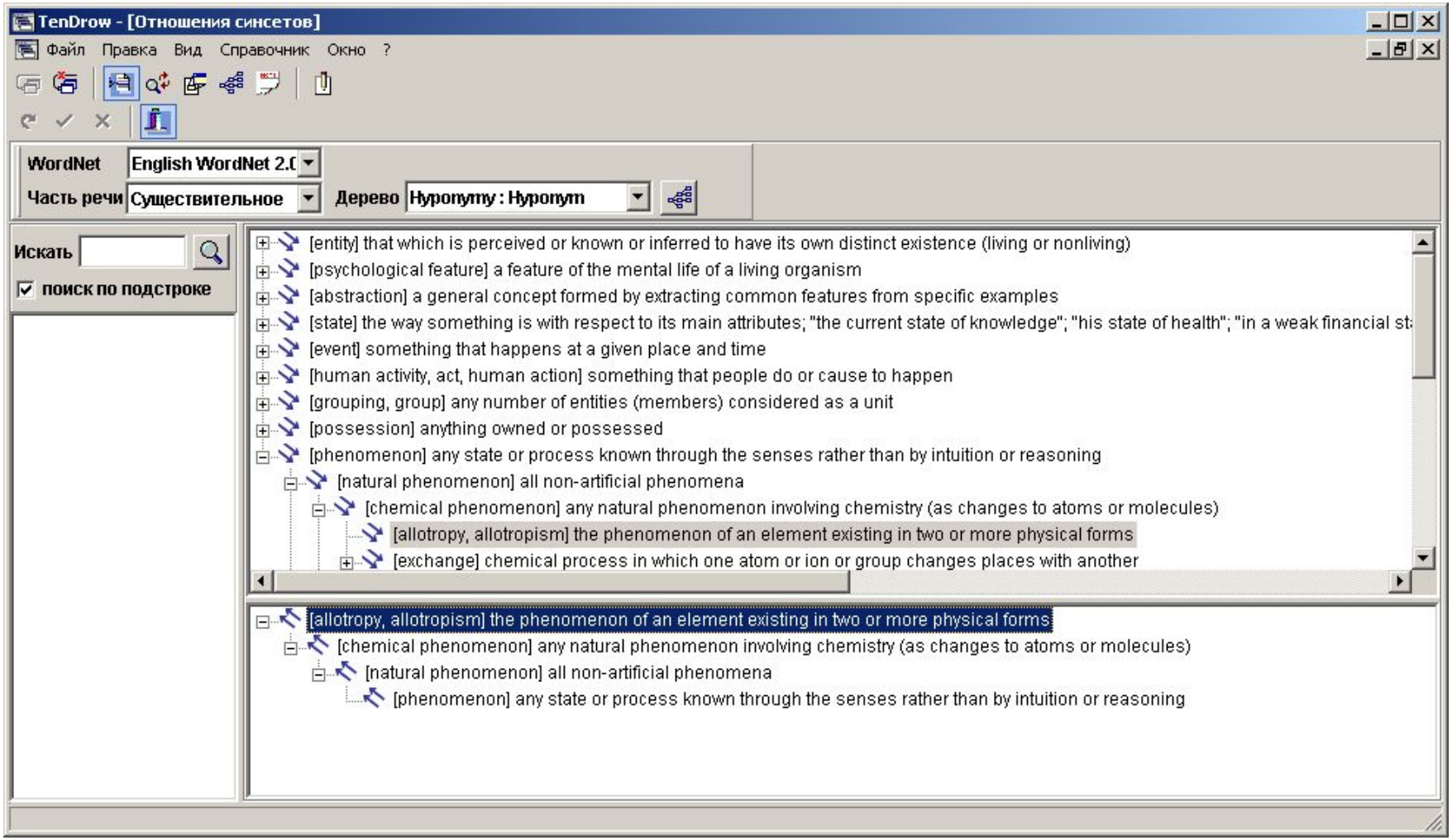
Подготовительный этап

- Построение частотных словарей для:
  - 153 235 лемм Princeton WordNet (PWN)
  - 164 099 лемм Russian WordNet (RWN)
  - ~2,5 млн. сочетаний (пар) лемм PWN
  - ~2,5 млн. сочетаний (пар) лемм RWN
- Ручной перевод и определение соответствия синсетов PWN и RWN для наиболее общих, философских значений. Синсеты – корневые элементы деревьев гипонимии (род/вид) и меронимии (часть/целое).

  Например: {**entity**}, {**psychological feature**}, {**abstraction**}, {**state**}, {**event**}, {**human activity, act, human action**}, {**grouping, group**}, {**possession**}, {**phenomenon**}

Автоматизированное построение ILI-индекса. Основные этапы.

Подготовительный этап



Автоматизированное построение ILI-индекса. Основные этапы.

Построение ILI-индекса

- Обход дерева гипонимии (затем – меронимии) PWN «в ширину» начиная от корня к листьям.
- Для каждого синсета PWN - подбор эквивалентного или наиболее близкого синсета/значения в RWN, формирование записи ILI-индекса.

Автоматизированное построение ILI-индекса. Перевод синсетов PWN.

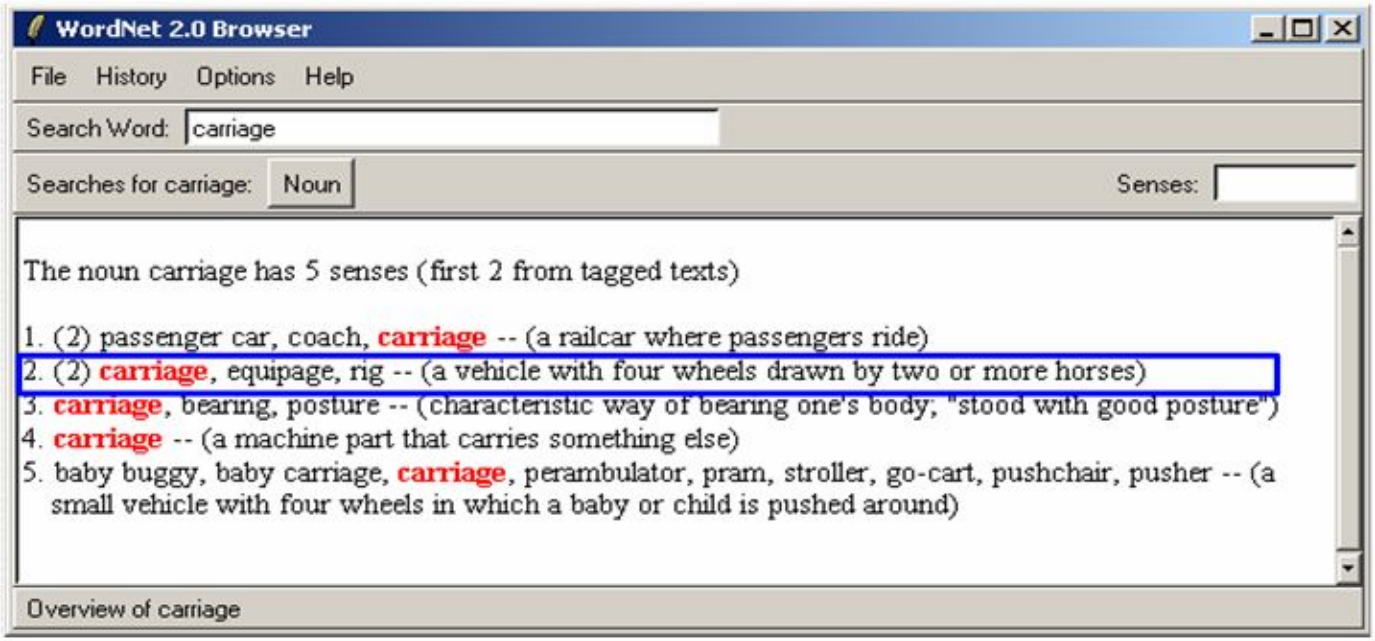
Вариант 1

Синсет PWN состоит более чем из 1 слова, (для 2х слов существуют переводы в англо-русском словаре). Переводы слов PWN присутствуют в словнике RWN.

Вариант 2

Синсет PWN состоит из 1 леммы или англо-русский словарь содержит перевод только одной леммы.

Демонстрация алгоритма построения ILI с использованием Google API. Вариант 1



carriage – {экипаж, карета}, {пассажирский вагон}, {тележка, вагонетка}, {гондола} и т.д. (20 переводов)

equipage –{экипаж}

rig – {приспособление, устройство}, {агрегат, оборудование}, {костюм},{упряжка, карета} и т.д. (35 переводов)

Демонстрация алгоритма построения ILI с использованием Google API

F(x) Количество ресурсов в Google		приспособление	устройство	агрегат	оборудование	костюм	упряжка	карета
		137000	883000	191000	1910000	742000	9290	67700
экипаж	594000	5970	39500	6170	49900	25400	771	7450
карета	67700	576	5110	347	4850	10200	490	67700
пассажирский вагон	784	35	184	51	332	156	0	44
тележка	96900	807	13500	3410	31700	5650	137	631
вагонетка	4880	137	629	256	875	476	6	44
гондола	8530	10	554	126	833	600	9	143

Демонстрация алгоритма построения ILI с использованием Google API

NGD(x,y) * 100	приспособление	устройство	агрегат	оборудование	костюм	упряжка	карета
экипаж	48,34	34,07	48,00	43,66	36,31	**69,86**	**46,02**
карета	49,82	56,50	37,63	71,59	46,13	**42,17**	**0,00**
пассажирский вагон	75,33	92,95	77,26	103,72	91,12	-	62,79
тележка	46,75	45,85	37,80	49,10	52,49	57,92	44,44
вагонетка	62,90	79,47	62,11	92,11	79,11	53,72	62,79
гондола	76,73	80,86	68,77	92,70	76,62	50,75	52,71

Демонстрация алгоритма построения ILI с использованием Google API

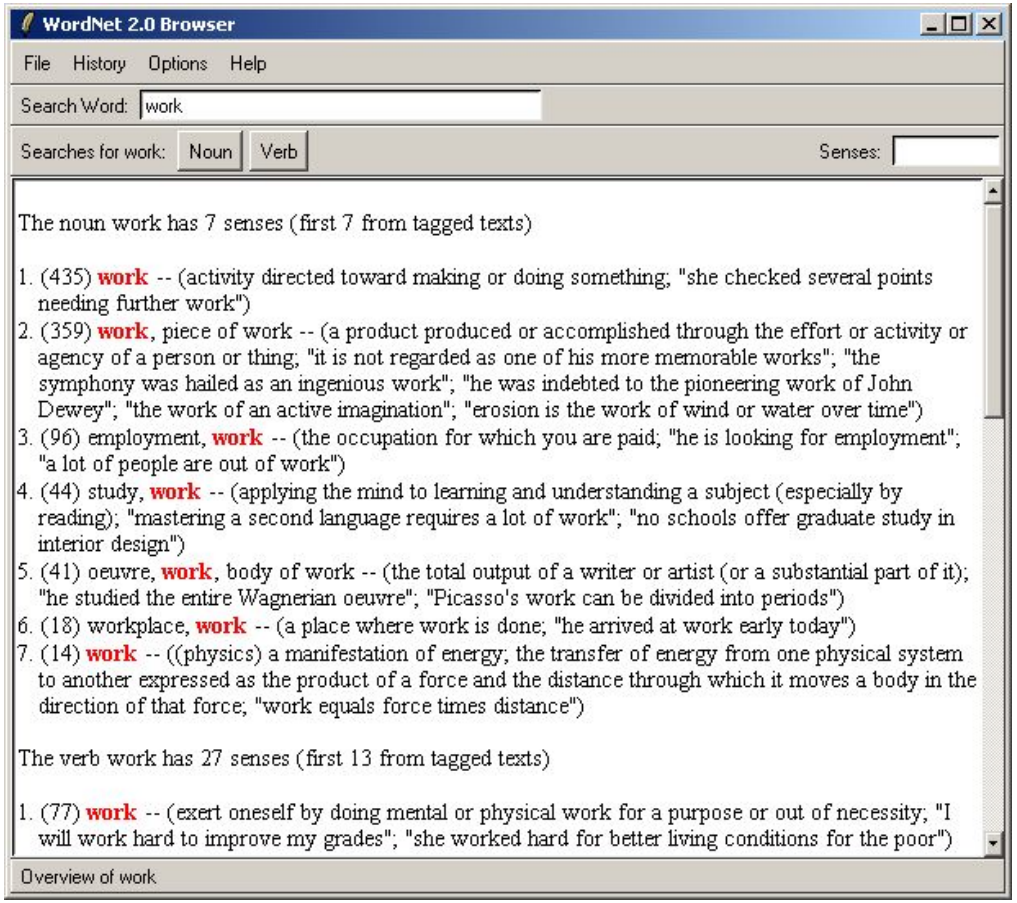
NGD(x,y) * 100	приспособление, устройство	агрегат, оборудование	костюм	упряжка, карета
экипаж, карета	34,07	37,63	36,31	**0,00**
пассажирский вагон	75,33	77,26	91,12	62,79
тележка, вагонетка	45,85	37,80	52,49	44,44
гондола	76,73	62,11	76,62	50,75

[carriage, equipage, rig] => [экипаж, карета, упряжка]

Демонстрация алгоритма построения ILI с использованием Google API. Вариант 2

Синсет PWN состоит из 1 леммы или англо-русский словарь содержит перевод только одной леммы.

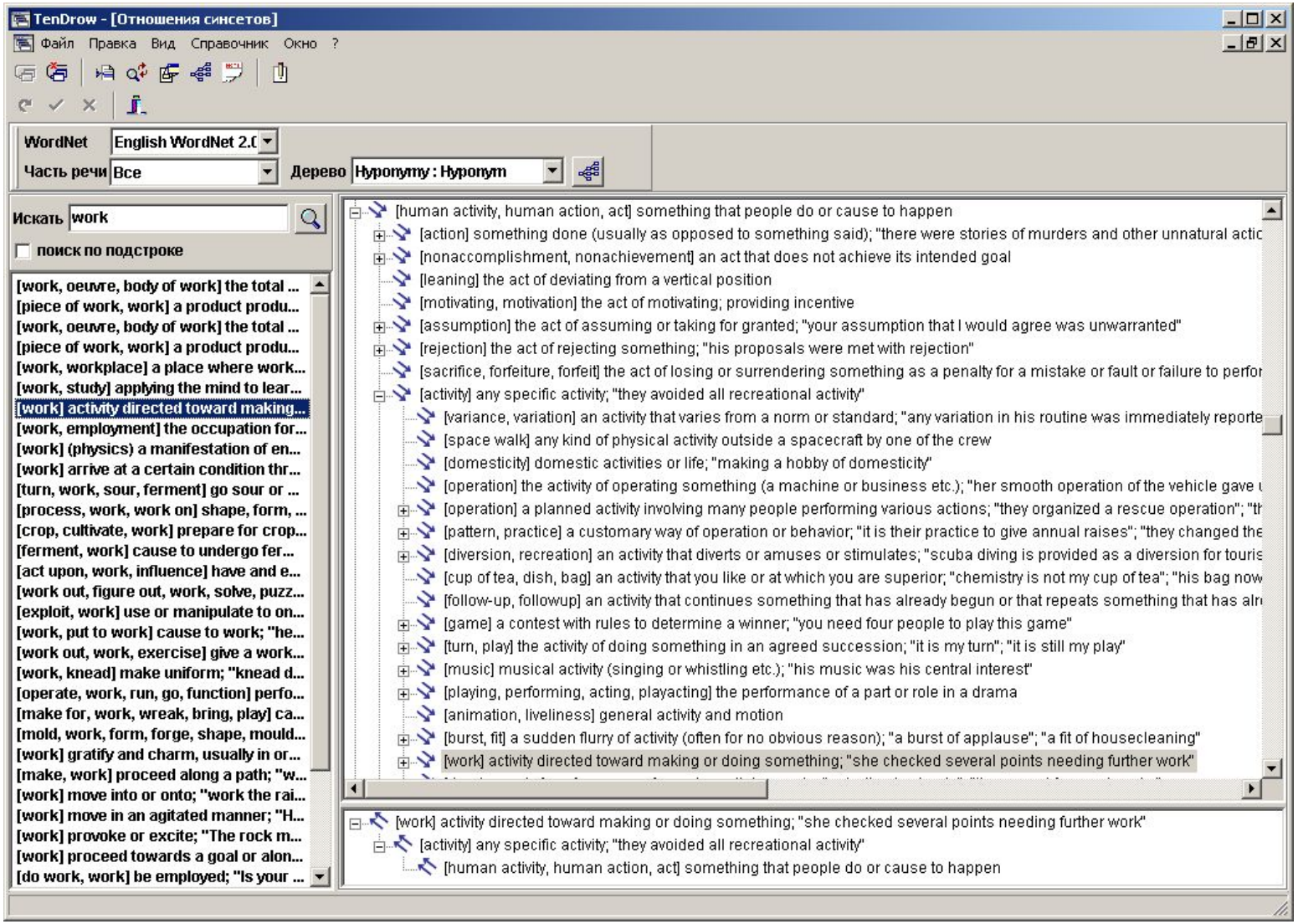
- **work**
- **love и др.**



Англо-русский словарь содержит более **20** вариантов перевода **work** !!!

[work] => [???]

Демонстрация алгоритма построения ILI с использованием Google API. Вариант 2



Демонстрация алгоритма построения ILI с использованием Google API. Вариант 2

- Определяется гипероним синсета PWN. Например, для синсета

  [work] - activity directed toward making or doing something; "she checked several points needing further work"

  гиперонимом (родительский узел в дереве род/вид) является синсет:

  [activity] - any specific activity; "they avoided all recreational activity«
- Для синсета [activity] на предыдущем шаге уже определен соответствующий синсет RWN – [дело, деятельность, занятие]
- Для всех переводов [work] вычисляется NGD=NGD(x,y) со словами синсета-гиперонима RWN (дело, деятельность, занятие)

Демонстрация алгоритма построения ILI с использованием Google API. Вариант 2

Для [**work**] в англо-русском словаре определены переводы:
work – {**служба, работа**}, {**произведение**}, {**изделие**}, {**исследование**}, {**труд**}, {**рабочий**} и т.д. (более 20 вариантов)

F(x)		дело	деятельность	занятие
		2 800 000	2 090 000	1 230 000
служба	2 550 000	623 000	959 000	89 400
работа	10 100 000	928 000	291 000	291 000
произведение	1 920 000	358 000	50 900	50 900
изделие	300 000	66 500	10 800	10 800
исследование	922 000	758 000	446 000	63 400
труд	1 310 000	1 770 000	101 000	101 000
рабочий	833 000	917 000	66 200	66 200

Демонстрация алгоритма построения ILI с использованием Google API. Вариант 2

NGD(x,y)*100	дело	деятельность	занятие
служба	18.87	**12.14**	41.58
работа	35.73	40.34	53.08
произведение	25.82	34.81	43.52
изделие	46.96	52.54	53.89
исследование	16.41	18.71	33.74
труд	**5.76**	18.07	**29.37**
рабочий	14.01	24.09	33.25

Из всех вариантов перевода
[work] - {служба, работа}, {произведение}, {изделие}, {исследование}, {труд}, {рабочий} и т.д. (более 20) выбирается:

[work] => {служба, работа}, {труд}

Статистика Russian WordNet

Лемм:

Существительных	Прилагательных	Глаголов	Наречий	Всего
74 886	32 593	41 487	13 137	164 099

Синсетов:

Слов в синсете	Существ.	Глагол.	Прилагат.	Наречий	Всего
1	84158	45221	38396	12925	182921
2	3270	7554	1689	269	12877
3	871	1512	423	162	2987
4	379	734	194	93	1414
5	212	388	119	66	795
6	135	273	73	57	545
7	84	195	45	32	357
…	…	…	…	…	…
Всего	89292	5635	1 41066	13716	202806

Спасибо за внимание