

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

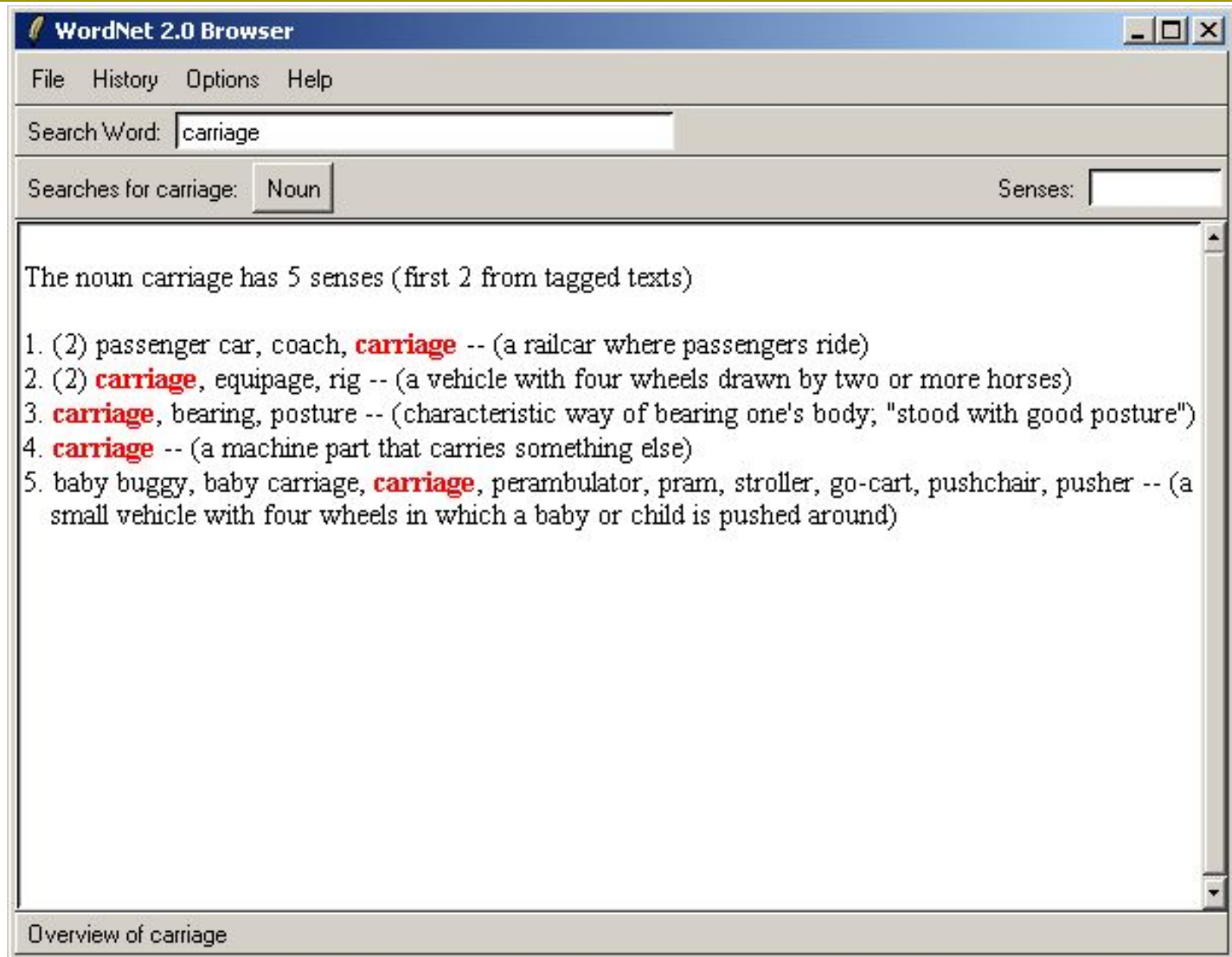
WordNet – лексико-семантическая база данных, включающая:

- 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 ?

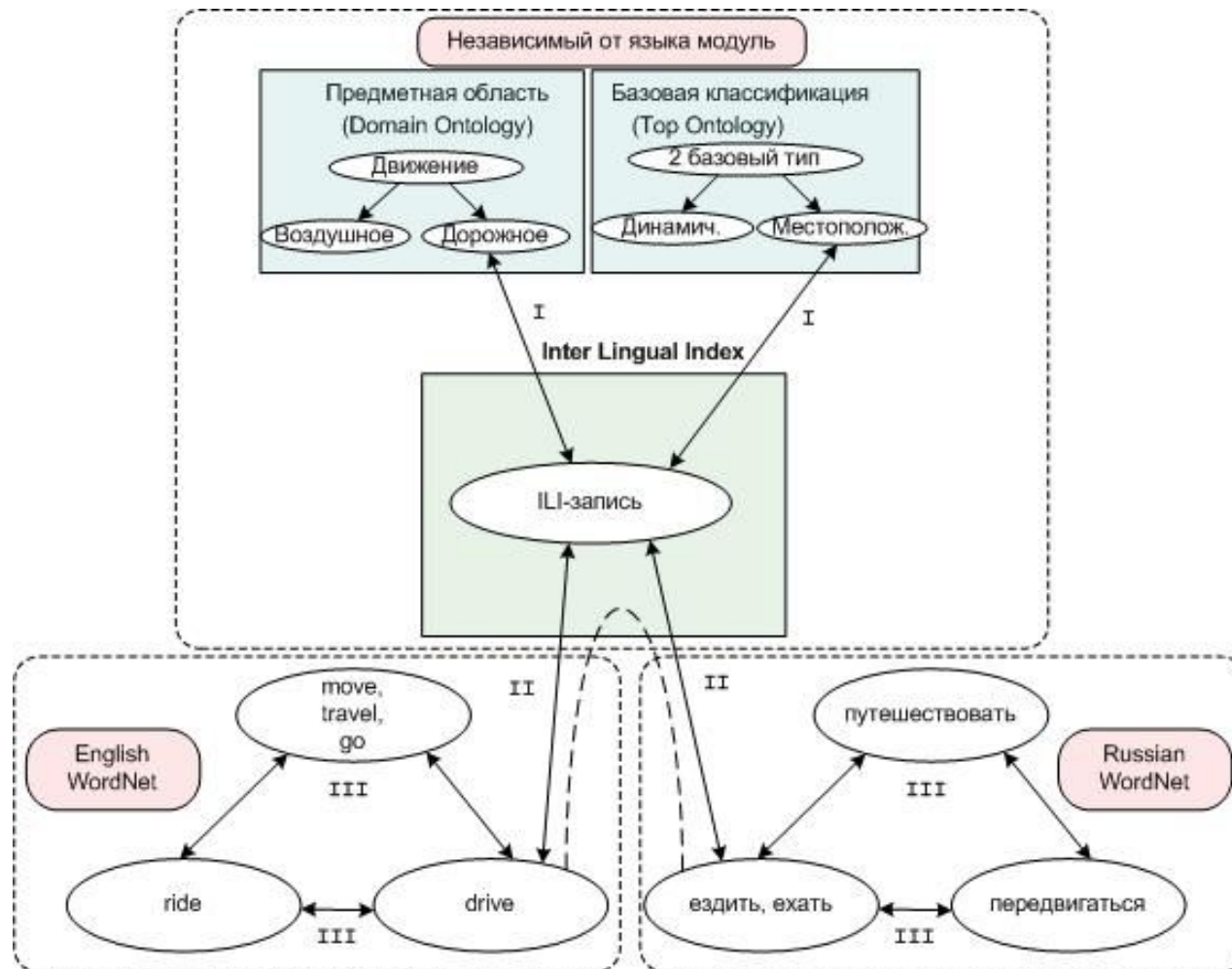
1. Наиболее полно отражает лексику английского и др. языков. Число входов (синсетов/слов) > 180 000.
2. Постоянное развитие PWN – версия 2.1.
3. “Параллельный” перевод на >17 языков. (EuroWordNet, Balkanet, Корейский и др.)
4. Встроенные морфологические анализаторы, “привязанные” к национальным языкам.
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



I - Независимые от языка отношения

II - Отношения между ILI и зависимыми от языка ресурсами

III - Отношения между зависимыми от языка ресурсами

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”

The image displays the Russian WordNet project interface, consisting of a web browser window and a desktop application window.

Web Browser Window (Microsoft Internet Explorer):

- Address bar: <http://www.pgups.ru/WebWN/view.uk?sWord=%D0%B2%D0%B5%D1%82%D0%BA%D0%B0&event=findA>
- Page Title: Слово - Microsoft Internet Explorer
- Page Content: Russian WordNet logo, search bar with "ветка" entered, search results table, and navigation links.

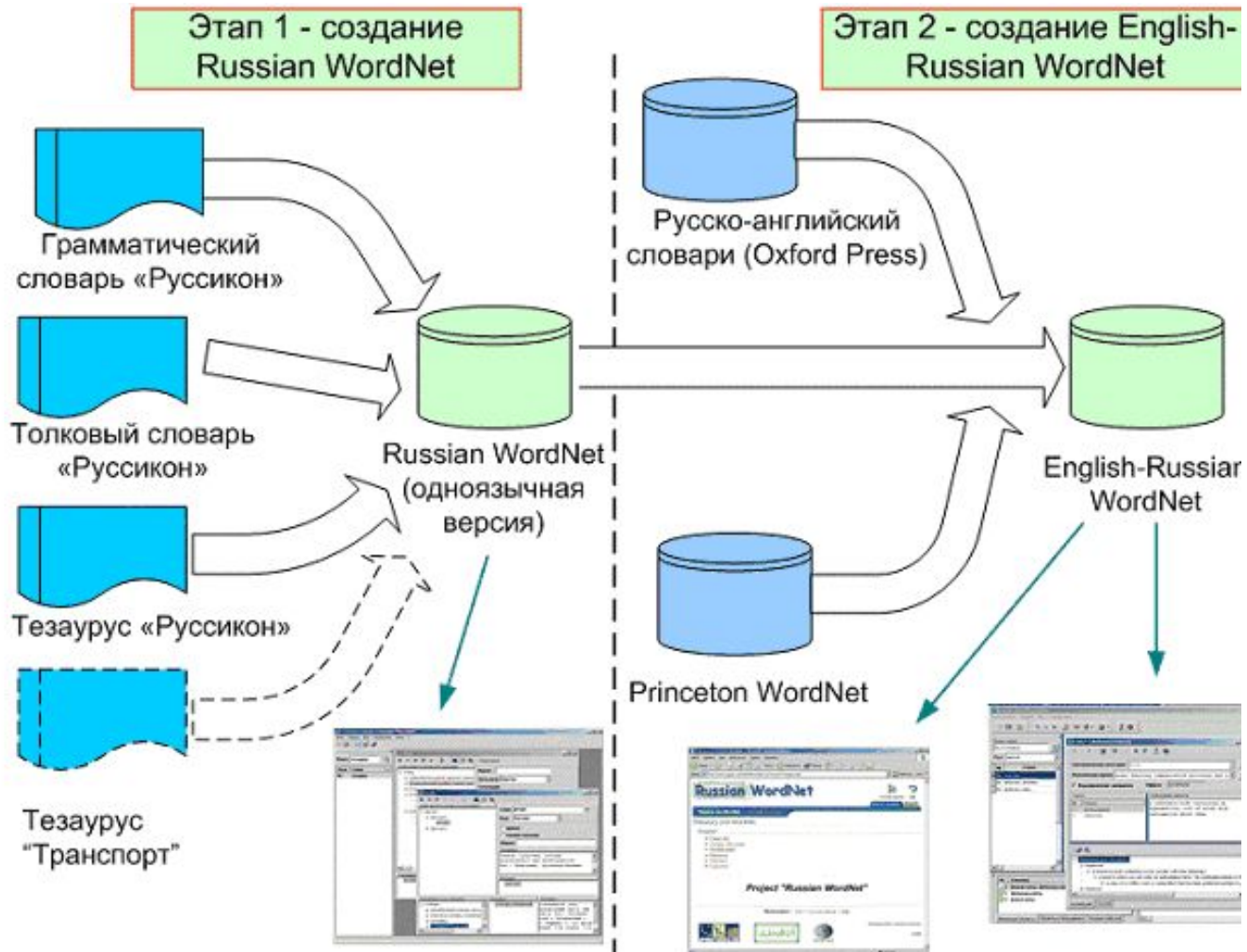
№ значения	Часть речи	Состав синсета	Толкование словарной статьи
1	Существительное	ветка	Линия железнодорожной дороги, отходящая в сторону от "Узкоколейная ветка Заводская железнодорожная ветка"
2	Существительное	ветка	Боковой отросток, побег, идущий от ствола дерева куста травянистого растения, небольшая ветвь. "Цветущая ветвь винограда (гроздь, кисть винограда)".

Desktop Application Window (TenDraw):

- Search bar: ветка
- Search results list: ветеранка, ветеринар, ветерок, ветерочек, ветка, ветла, ветловый, ветнадзор, вето
- Tree view: ветка (expanded) showing sub-items like ветка, ветка, ветка
- Word details for "ветка":
 - Слово: ветка
 - Толкование: Линия железнодорожной дороги, отходящая в сторону от главного пути.
 - Значение: Узкоколейная ветка Заво...
- Word forms: ветка, ветке, ветку, веткой

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

Основные этапы «Russian WordNet»



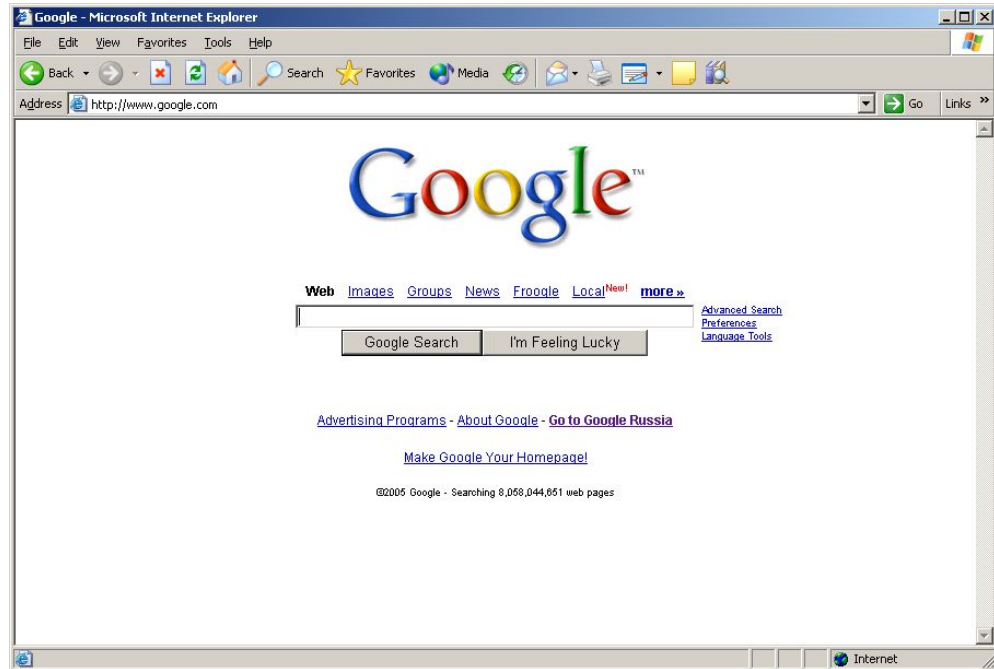
WordNet

В общем случае отображение $L1 \rightarrow L2$ невыполнимо, поскольку

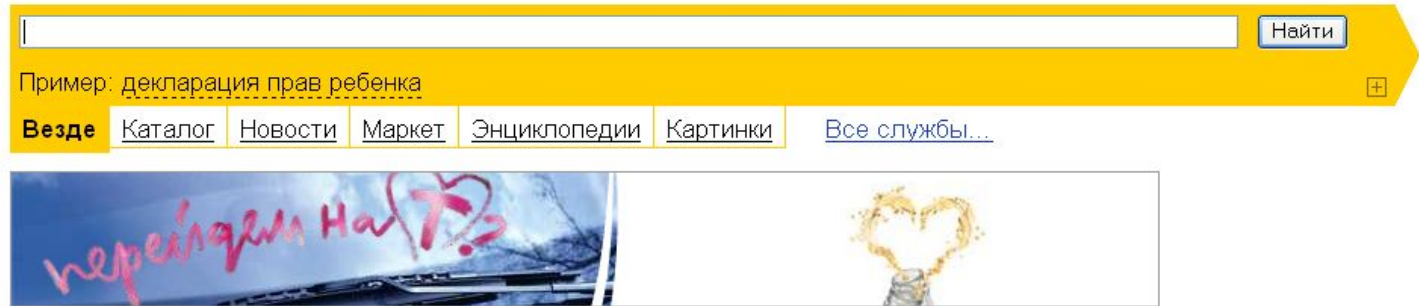
- для некоторого слова W_{L1} может не существовать соответствующего слова W_{L2} , т.е. перевод может отсутствовать,
- число значений $\text{lemmat}(W_{L1})$ может быть не равно числу значений $\text{lemmat}(W_{L2})$ и/или значения могут не совпадать
- некоторое слово W_{L1} может переводиться не одним 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



Яндекс 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.

$$p(x) = \frac{(w: x \in w)}{M} = \frac{f(x)}{M} \quad - \text{вероятность появления слова } x \text{ в коллекции из } M \text{ документов}$$

$$p(x, y) = \frac{(w: x, y \in w)}{M} = \frac{f(x, y)}{M} \quad - \text{вероятность совместного появления слова } x \text{ и } y \text{ в одном и том же документе}$$

$M=8\ 058\ 044\ 651$ (~8 млрд.)

[Google]

Определение «семантического расстояния» между словами

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

Условные вероятности появления слов в коллекции документов.

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

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 $f(x)=f(y)=0$
- $NGD=\infty$, 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 ИЛ-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 ИЛ-индекса. Основные этапы.

Подготовительный этап

The screenshot shows the TenDraw application window titled "TenDraw - [Отношения синсетов]". The interface includes a menu bar (Файл, Правка, Вид, Справочник, Окно, ?), a toolbar with various icons, and a main workspace. The workspace is divided into several sections:

- WordNet Section:** Shows "English WordNet 2.0" selected in the "WordNet" dropdown, "Существительное" (Noun) in the "Часть речи" (Part of speech) dropdown, and "Нуропуту : Нуропут" in the "Дерево" (Tree) dropdown.
- Search Section:** Includes a search box labeled "Искать" and a checkbox for "поиск по подстроке" (search by substring).
- Main List:** A scrollable list of WordNet synsets. The synsets are displayed in a tree-like structure. The selected synset is "[allotropy, allotropism] the phenomenon of an element existing in two or more physical forms".

The selected synset is highlighted in blue. Below it, the expanded tree structure shows the following hierarchy:

- [allotropy, allotropism] the phenomenon of an element existing in two or more physical forms
 - [chemical phenomenon] any natural phenomenon involving chemistry (as changes to atoms or molecules)
 - [natural phenomenon] all non-artificial phenomena
 - [phenomenon] any state or process known through the senses rather than by intuition or reasoning

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.

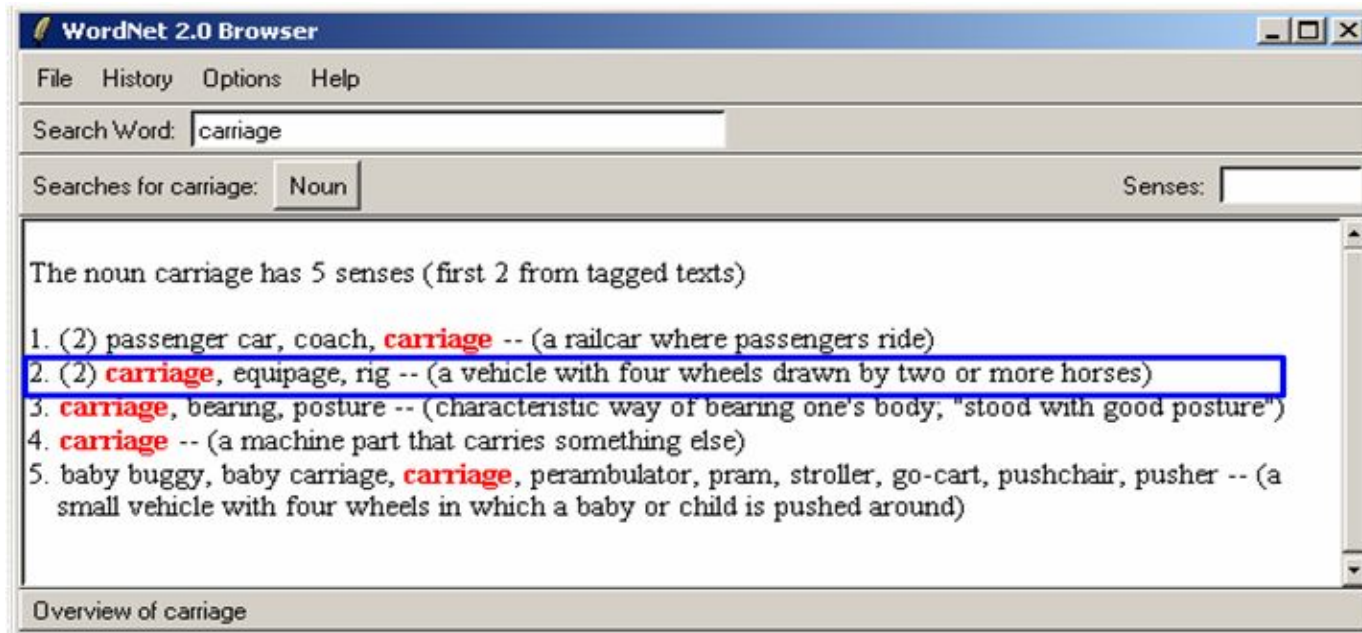
Вариант 1

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 ИЛ 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 ИЛ с использованием Google API

F(x)		приспособление	устройство	агрегат	оборудование	костюм	упряжка	карга
Количество ресурсов в Google		137000	883000	191000	1910000	742000	9290	67700
экипаж	594000	5970	39500	6170	49900	25400	771	7450
карга	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 ІІІ с использованием Google API

NGD(x,y) * 100	приспособление	устройство	агрегат	оборудование	костюм	упряжка	карга
экипаж	48,34	34,07	48,00	43,66	36,31	69,86	46,02
карга	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 ИЛ с использованием Google API

NGD(x,y) * 100	приспособление, устройство	агрегат, оборудование	костюм	упряжка, каjeta
экипаж, каjeta	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, equirage, rig] => [экипаж, карета, упряжка]

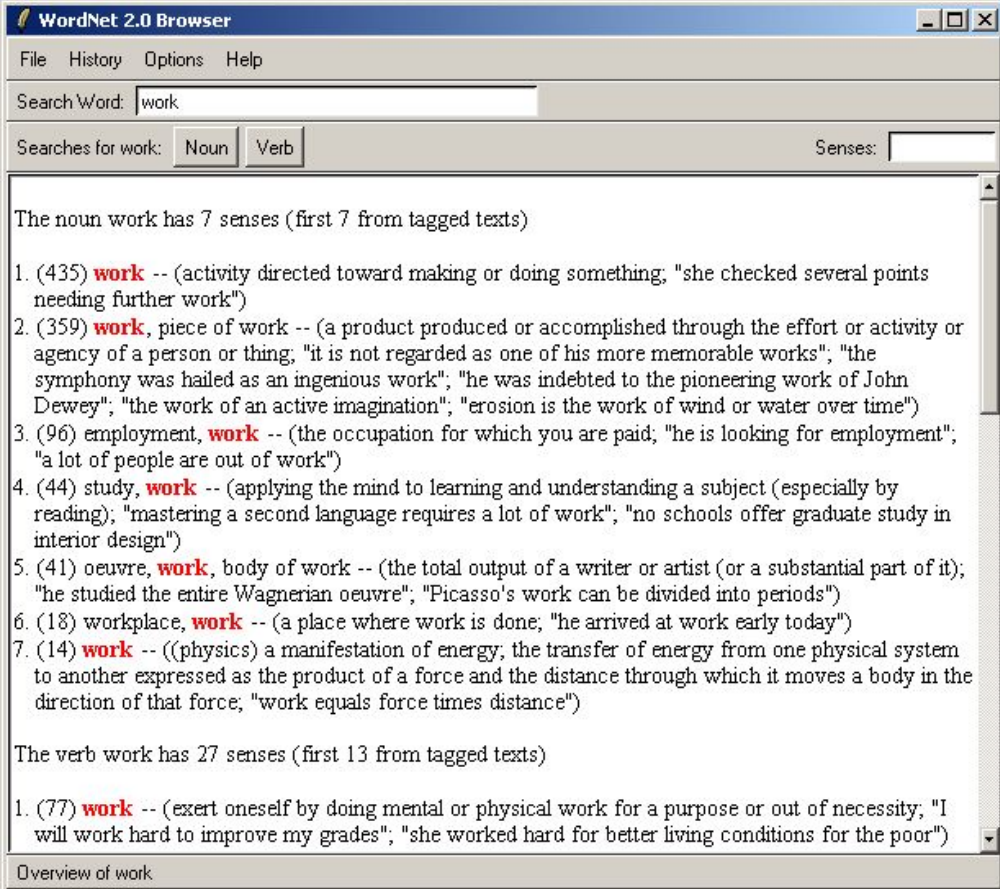
Демонстрация алгоритма построения ИЛ с использованием Google API. Вариант 2

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] => [???



WordNet 2.0 Browser

File History Options Help

Search Word: work

Searches for work: Noun Verb Senses:

The noun work has 7 senses (first 7 from tagged texts)

1. (435) **work** -- (activity directed toward making or doing something; "she checked several points needing further work")
2. (359) **work**, piece of work -- (a product produced or accomplished through the effort or activity or agency of a person or thing; "it is not regarded as one of his more memorable works"; "the symphony was hailed as an ingenious work"; "he was indebted to the pioneering work of John Dewey"; "the work of an active imagination"; "erosion is the work of wind or water over time")
3. (96) employment, **work** -- (the occupation for which you are paid; "he is looking for employment"; "a lot of people are out of work")
4. (44) study, **work** -- (applying the mind to learning and understanding a subject (especially by reading); "mastering a second language requires a lot of work"; "no schools offer graduate study in interior design")
5. (41) oeuvre, **work**, body of work -- (the total output of a writer or artist (or a substantial part of it); "he studied the entire Wagnerian oeuvre"; "Picasso's work can be divided into periods")
6. (18) workplace, **work** -- (a place where work is done; "he arrived at work early today")
7. (14) **work** -- ((physics) a manifestation of energy; the transfer of energy from one physical system to another expressed as the product of a force and the distance through which it moves a body in the direction of that force; "work equals force times distance")

The verb work has 27 senses (first 13 from tagged texts)

1. (77) **work** -- (exert oneself by doing mental or physical work for a purpose or out of necessity; "I will work hard to improve my grades"; "she worked hard for better living conditions for the poor")

Overview of work

Демонстрация алгоритма построения ИЛ с использованием Google API. Вариант 2

The screenshot displays the TenDraw application window titled "TenDraw - [Отношения синсетов]". The interface includes a menu bar (Файл, Правка, Вид, Справочник, Окно, ?), a toolbar, and a search section. The search section is set to "WordNet English WordNet 2.0" and "Часть речи: Все". The search term "work" is entered in the "Искать" field, and the "поиск по подстроке" checkbox is checked. A list of search results is shown on the left, with "[work] activity directed toward making..." selected. The main window displays a hierarchical tree of related terms and their definitions, such as "[human activity, human action, act] something that people do or cause to happen" and "[activity] any specific activity, 'they avoided all recreational activity'".

WordNet English WordNet 2.0
Часть речи: Все
Дерево: Нуропуты : Нуропут

Искать: work
 поиск по подстроке

[work, oeuvre, body of work] the total ...
[piece of work, work] a product produ...
[work, oeuvre, body of work] the total ...
[piece of work, work] a product produ...
[work, workplace] a place where work...
[work, study] applying the mind to lear...
[work] activity directed toward making...
[work, employment] the occupation for...
[work] (physics) a manifestation of en...
[work] arrive at a certain condition thr...
[turn, work, sour, ferment] go sour or ...
[process, work, work on] shape, form, ...
[crop, cultivate, work] prepare for crop...
[ferment, work] cause to undergo fer...
[act upon, work, influence] have and e...
[work out, figure out, work, solve, puzz...
[exploit, work] use or manipulate to on...
[work, put to work] cause to work; "he...
[work out, work, exercise] give a work...
[work, knead] make uniform; "knead d...
[operate, work, run, go, function] perfo...
[make for, work, wreak, bring, play] ca...
[mold, work, form, forge, shape, mould...
[work] gratify and charm, usually in or...
[make, work] proceed along a path; "w...
[work] move into or onto; "work the rai...
[work] move in an agitated manner; "H...
[work] provoke or excite; "The rock m...
[work] proceed towards a goal or alon...
[do work, work] be employed; "Is your ...

- [human activity, human action, act] something that people do or cause to happen
 - [action] something done (usually as opposed to something said); "there were stories of murders and other unnatural acti...
 - [nonaccomplishment, nonachievement] an act that does not achieve its intended goal
 - [leaning] the act of deviating from a vertical position
 - [motivating, motivation] the act of motivating; providing incentive
 - [assumption] the act of assuming or taking for granted; "your assumption that I would agree was unwarranted"
 - [rejection] the act of rejecting something; "his proposals were met with rejection"
 - [sacrifice, forfeiture, forfeit] the act of losing or surrendering something as a penalty for a mistake or fault or failure to perfor...
 - [activity] any specific activity; "they avoided all recreational activity"
 - [variance, variation] an activity that varies from a norm or standard; "any variation in his routine was immediately reporte...
 - [space walk] any kind of physical activity outside a spacecraft by one of the crew
 - [domesticity] domestic activities or life; "making a hobby of domesticity"
 - [operation] the activity of operating something (a machine or business etc.); "her smooth operation of the vehicle gave u...
 - [operation] a planned activity involving many people performing various actions; "they organized a rescue operation"; "tr...
 - [pattern, practice] a customary way of operation or behavior; "it is their practice to give annual raises"; "they changed the...
 - [diversion, recreation] an activity that diverts or amuses or stimulates; "scuba diving is provided as a diversion for touris...
 - [cup of tea, dish, bag] an activity that you like or at which you are superior; "chemistry is not my cup of tea"; "his bag now...
 - [follow-up, followup] an activity that continues something that has already begun or that repeats something that has alr...
 - [game] a contest with rules to determine a winner; "you need four people to play this game"
 - [turn, play] the activity of doing something in an agreed succession; "it is my turn"; "it is still my play"
 - [music] musical activity (singing or whistling etc.); "his music was his central interest"
 - [playing, performing, acting, playacting] the performance of a part or role in a drama
 - [animation, liveliness] general activity and motion
 - [burst, fit] a sudden flurry of activity (often for no obvious reason); "a burst of applause"; "a fit of housecleaning"
 - [work] activity directed toward making or doing something; "she checked several points needing further work"
- [work] activity directed toward making or doing something; "she checked several points needing further work"
 - [activity] any specific activity; "they avoided all recreational activity"
 - [human activity, human action, act] something that people do or cause to happen

Демонстрация алгоритма построения ИЛ 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 ИЛ 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 ИЛ 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
