

# АВТОМАТИЗАЦИЯ ПОСТРОЕНИЯ АНГЛО-РУССКОГО WORDNET

*А.М. Сухоногов*  
*Петтебургский Университет путей*  
*сообщения, кафедра ИВС*  
[ASukhonogov@rambler.ru](mailto:ASukhonogov@rambler.ru);

*С.А. Яблонский*  
*Петтебургский Университет путей*  
*сообщения, кафедра ИВС,*  
*ЗАО "Руссикон"*  
[serge\\_yablonsky@hotmail.com](mailto:serge_yablonsky@hotmail.com); [info@russicon.ru](mailto:info@russicon.ru)

Целью настоящей работы является описание построения русской версии WordNet и методов автоматизированного создания межъязыкового индекса для англо-русской версии WordNet.

## 1. Введение

Работа над словарем Princeton WordNet (PWN) [1,2] английского языка начата в Принстонском университете (США) в начале 80-х годов и продолжается сегодня. Сейчас доступна версия 2.0. этого словаря. Существующая версия WordNet охватывает более 120 тысяч слов общеупотребительной лексики современного английского языка. За период с марта 1996 по сентябрь 1999 года при финансировании Европейской комиссии был создан многоязычный вариант WordNet – EuroWordNet [3]. Эта лексическая система объединила в себе WordNet английского, датского, испанского, итальянского, немецкого, французского, чешского и эстонского языков, а за основу был взят Принстонский WordNet версии 1.5. Национальные версии проекта связаны с исходным PWN 1.5 и между собой через специальный межъязыкового индекса (Interlingual Index – ILI(PWN 1.5)). В 2004 году завершена работа над проектом BalkaNet, объединяющим греческий, болгарский, турецкий, чешский, французский, румынский и сербский языки. Поскольку в проекте за основу была взята версия 2.0 PWN, в BalkaNet межъязыковой индекс базируется на PWN 2.0 – ILI (PWN 2.0).

В настоящее время известно о нескольких реализациях WordNet подобных лексических баз данных для русского языка:

- Проект RussNet, разрабатывается с 1999 года на филологическом факультете СПбГУ [4]. Методика и принципы построения словаря проекта RussNet ориентированы на длительный процесс разработки ресурса группой лингвистов без автоматизации процесса построения и связи с исходным PWN.
- Проект тезауруса RuThes, используемого в УИС РОССИЯ [5]. Закрытый коммерческий ресурс.

Рассматриваемая в работе реализация русской версии WordNet (Russian WordNet - RWN) [6-10] ориентирована на формирование ядра RWN (более 100 тыс. слов) за счет

- привязки RWN к PWN;
- использования доступных русских, англо-русских и русско-английских словарей,

- автоматизации процедур построения и редактирования RWN.

Целью настоящей работы является описание

- построения русской версии WordNet,
- методов автоматизированного построения межъязыкового индекса для англо-русской версии WordNet<sup>1</sup>.

## 2. Краткое описание проекта RWN

Разработка RWN включает (рис.1):

- построение русской версии WordNet, достаточно полно (100-120 тыс. лексических единиц) описывающей лексику русского языка и сопоставимой по числу лексических единиц с английской версией. Для этого используются морфологический анализатор и лексические ресурсы [11] ЗАО "РУССИКОН"<sup>2</sup>, словари, свободно распространяемые в Интернете, (<http://www.slovarik.ru/>, <http://www.artint.ru/projects/frqulist.asp>), и ряд печатных изданий;
- интеграцию с другими лексическими системами на основе использования технологии Semantic Web.
- программное построение межъязыкового индекса, определяющего соответствие между синсетам PWN и RWN, на основе использования электронных версий словарей издательства Oxford Press<sup>3</sup>, ряда доступных в Интернете англо-русских и русско-английских словарей, WordNet-Domains [12].

<sup>1</sup> Для выполнения части данной работы (перевод 20000 синсетов PWN) авторским коллективом в 2005 году получен грант ЗАО "Яндекс" № 103003 - «Построения межъязыкового индекса...».

<sup>2</sup> ЗАО "Руссикон" ([www.russicon.ru](http://www.russicon.ru)) предоставило для создания версии WordNet для русского языка лингвистические ресурсы и программы.

<sup>3</sup> Научный коллектив из сотрудников ПГУПС (кафедра ИВС) и ЗАО «Руссикон» выиграл в 2003 году конкурс издательства Oxford Press на лучший исследовательский проект по использованию словарей Oxford Press. В настоящее время издательства Oxford Press предоставило

## 2.1. Построение RWN

В WordNet словарными статьями являются синсеты – множества слов-синонимов, обозначающих один и тот же концепт в заданном контексте. Для синсета явно указывается часть речи и толкование. Каждое слово, входящее в состав синсета, может дополнительно иметь ряд атрибутов, например, признак доминантности, пометы типа «идиома», «близкое значение» и т.д. Для каждого слова может быть приведен пример его употребления в заданном контексте – определяется набор речений и фразеологизмов, также определяются толкования.

В настоящее время (по состоянию на май 2005г.) RWN включает:

- 55397 существительных, образующих 71729 синсет;
- 34400 глаголов, образующих 44998 синсет;
- 25315 прилагательных, образующих 33571 синсет;
- 10071 наречий, образующих 9716 синсетов;
- парадигмы для всех лемм словника;
- словообразовательные отношения;
- выделенную из EuroWordNet Top Ontology, с возможностью расширения;
- WordNet Domains с возможностью определения новых доменов.

Для проекта RWN разработаны методы и программные средства, позволяющие значительно сократить время разработки (рис. 1).

Так разработан редактор TenDrow для редактирования WordNet и пакет специальных утилит построения WordNet и ILI-индекса.

В настоящее время редактор TenDrow (рис. 2)

- работает с СУБД Oracle9i/10g и Interbase/Firebird;
- осуществляет обмен данными между БД и OWL-представлением WordNet (экспорт / импорт данных);
- поддерживает форматы лексических файлов Princeton WordNet 2.0 и VisDic1.3.36 (для загрузки в БД).

Редактор TenDrow предназначен для создания и редактирования широкого класса тезаурусов и близких к ним структур.

Схема данных Russian WordNet, позволяющая представить любую структуру WordNet-подобного словаря рассматривается в работе [8].

для создания русской версии WordNet XML версии следующих словарей: Oxford Russian Dictionary, 3<sup>d</sup> Edition, New Oxford Dictionary of English, 2<sup>nd</sup> Edition, New Oxford Thesaurus of English.

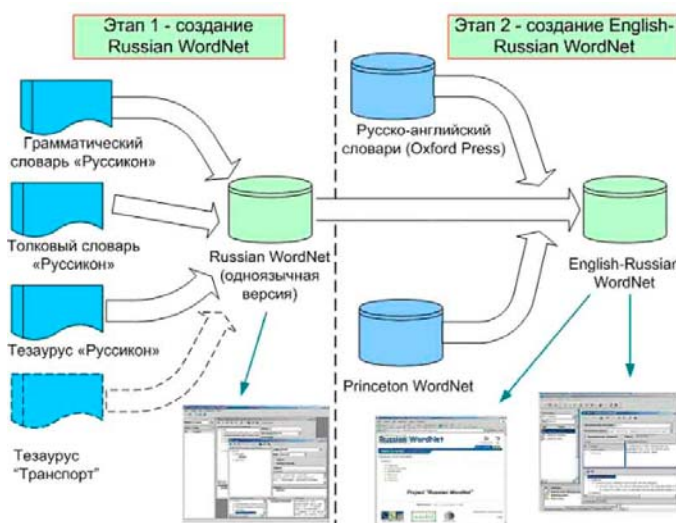


Рис. 1. Основные этапы создания Russian WordNet

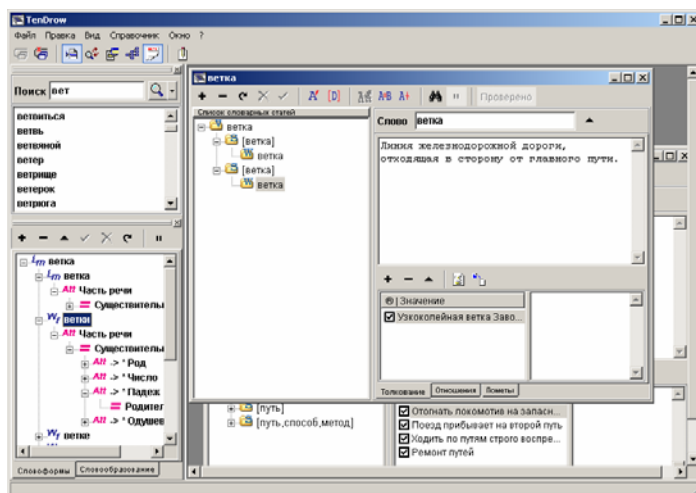


Рис. 2. Редактор TenDrow. Основное рабочее окно

Данные проекта хранятся и обрабатываются в СУБД Oracle9i/10g, что позволяет:

- организовать многопользовательский доступ к данным;
- открыть доступ к текущему состоянию проекта всем членам коллектива разработчиков;
- осуществлять резервирование данных средствами СУБД;
- достаточно просто реализовывать загрузку / выгрузку данных в различные форматы (TXT, XML, OWL, лексические файлы Princeton WordNet и т.д.).

Вид интерфейса для просмотра базы данных WordNet в Интернете приведен на рис. 3.

## 2.2. Представление WordNet в Semantic Web

В рамках технологии Semantic Web консорциум W3C разрабатывает стандарт представления WordNet, в котором WordNet представляется как фрагмент описания онтологии. Для этого словарь достаточно представить в формате описания RDF и обеспечить совместный доступ к этому ресурсу для распределенных программ – агентов. Формат OWL принят в качестве основного для экспорта и импорта данных в/из базы данных в проекте Russian WordNet. OWL/RDFS-схема Russian WordNet соответствует основным рекомендациям W3C-консорциума. OWL-схема проверена на корректность с помощью RDF-анализатора Jena (<http://www.w3.org/TR/2003/PR-rdf-schema-200331215>).

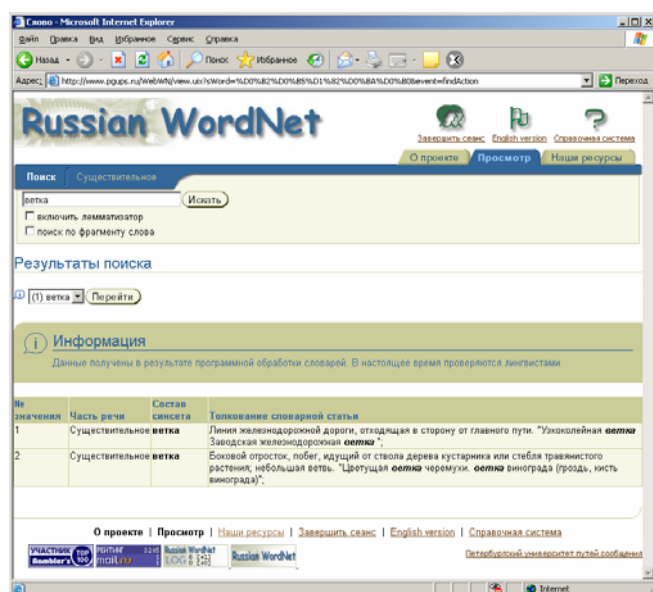


Рис. 3. Сайт проекта Russian WordNet (<http://www.pgups.ru/WebWN/wordnet.uix>)

Пример экспортированного из системы варианта WordNet (Princeton WordNet 2.0) в формате OWL доступен на сайте проекта. Таким образом, русский WordNet может использоваться и как один из компонентов технологии W3C/SemanticWeb.

## 3. Автоматизированное построение межязыкового индекса

В проекте EuroWordNet впервые была осуществлена попытка объединения WordNet для нескольких европейских языков вместе с Princeton WordNet с помощью межязыкового индекса ILI (PWN1.5). Индекс позволяет находить аналоги английских синсетов на других языках, а также обеспечивает доступ к 63 сущностям верхнего уровня онтологии (top-ontology) PWN. Эта top-ontology определяет общую для всех языков

семантическую сеть, что обеспечивает связь между версиями WordNet для различных языков.

В проекте BalkaNet наряду с рассмотрением (PWN1.5) в качестве ILI (PWN 2.0) используется сама версия PWN 2.0.

В проекте RWN решаются две задачи англо-русского и русско-английского перевода WordNet:

- перевод 63 сущностей top-ontology на русский и обратно;
- автоматизированный перевод ILI (PWN 2.0) с помощью электронных версий словарей и разработанных программных средств.

## 3.1. Исходные данные

Используются

- лексические БД Princeton WordNet 2.0 (PWN) и Russian WordNet (RWN).
- англо-русский и русско-английский Oxford Russian Dictionary, 3<sup>d</sup> Edition,
- New Oxford Dictionary of English, 2<sup>nd</sup> Edition,
- New Oxford Thesaurus of English.
- англо-русские и русско-английские словари, распространяемые по лицензии GPL в составе сервера словарей DICT Server (реализация RFC 2229 [14]) и ряд других словарей, свободно распространяемых в Интернете.
- WordNet Domains [12].
- Англо-русский словарь под ред. Ю.Д. Апресяна.

Словари конвертируются в отдельную схему СУБД. Для словарей определяется API, позволяющее

- для заданного слова находить множество слов – переводов и словарных статей в New Oxford Dictionary of English, 2<sup>nd</sup> Edition и New Oxford Thesaurus of English,
- по слову-переводу находить множество словарных статей русско-английского словаря (инвертированный словарь).

## 3.2. Особенности перевода WordNet

Пусть заданы два языка L1 и L2 и слово  $W_{L2}$  является переводом слова  $W_{L1}$ . При этом будем считать, что выполняется условие совпадения частей речи слов пары  $\langle W_{L2}, W_{L1} \rangle$ .

Если для каждой пары  $\langle W_{L2}, W_{L1} \rangle$  выполняются следующие условия:

- WordNet языка L1 ( $WN_{L1}$ ) содержит lemmat ( $W_{L1}$ ) и WordNet языка L2 ( $WN_{L2}$ ) содержит lemmat ( $W_{L2}$ ), где функция lemmat (W) преобразует словоформу W в лемму,
- все возможные значения lemmat ( $W_{L1}$ ) присутствуют в  $WN_{L1}$  и все возможные значения lemmat ( $W_{L2}$ ) присутствуют в  $WN_{L2}$ ,

- $WN_{L1}$  и  $WN_{L2}$  связаны через ILI-подобный механизм,

то возможно связать пару WordNet  $\langle WN_{L2}, WN_{L1} \rangle$  для языков L1 и L2 так, что алгоритм связи будет сопоставлять все синсеты  $WN_{L1}$  всем синсетам  $WN_{L2}$  через ILI.

В общем случае такое отображение невыполнимо, поскольку:

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

Поэтому идеального перевода WordNet с одного языка на другой (в нашем случае с английского на русский) не существует, и алгоритм построения связи PWN и RWN через ILI (PWN 2.0) должен разрешать противоречия 1)-3).

В настоящей работе основное внимание уделяется методу автоматического построения перевода-соответствия PWN и RWN. Этап “ручной” проверки и устранения неточностей и ошибок не рассматривается.

### 3.3. Порядок построения межъязыкового индекса.

При построении межъязыкового индекса (ILI) ставится задача воспроизведения основных деревьев отношений (гипонимии и меронимии) PWN в RWN. Порядок построения ILI индекса определяется структурой этих деревьев и начинается с корня каждого дерева. При построении ILI осуществляется последовательный обход деревьев гипонимии и меронимии PWN, затем - синсетов, которые не попали ни в одно из этих деревьев. По каждому дереву осуществляется обход в ширину: синсеты - корни дерева, затем - все синсеты 1-го уровня, 2-го уровня и т.д. Обход прекращается при достижении уровня листьев дерева для текущего узла дерева. Такой подход позволяет воспроизвести основные отношения, определенные в PWN.

Согласно принятому порядку построения ILI-индекса, исходным является синсет PWN -  $s_{PWN}$ . На первом этапе для этого синсета строится список синсетов-альтернатив из RWN -  $S_{RWN} = \{s_{RWN1}, s_{RWN2}, \dots, s_{RWNn}\}$ . Если  $S_{RWN} = \{\emptyset\}$ , построение ILI для синсета  $s_{PWN}$  невозможно, поскольку используемые для перевода словари не содержат ни одного слова или словосочетания из множества слов синсета  $s_{PWN}$ . Такие синсеты выделяются и должны быть обработаны вручную. Если  $S_{RWN} \neq \{\emptyset\}$ , для каждой пары синсетов вычисляется значение оценочной функции  $R = R(s_{PWN}, s_{RWNi})$ . На втором этапе для

каждого синсета  $s_{RWNi}$  из  $S_{RWN}$  строится свой список синсетов-альтернатив PWN -  $S_{PWN} = \{s_{PWN1}, s_{PWN2}, \dots, s_{PWNm}\}$  и также вычисляется значение оценочной функции  $R'_i = R'(s_{RWNi}, s_{PWN})$ . В список  $S_{PWNi}$  каждого синсета-альтернативы из  $S_{RWN}$  гарантированно попадает и исходный синсет -  $s_{PWN}$ , поскольку для перевода используется единый англо-русский и русско-английский словарь.

Основным критерием для определения соответствия исходного синсета  $s_{PWN}$  и синсета-альтернативы  $s_{RWNi}$  мы считаем минимальное значение суммы приведенных значений оценочной функции:

$$ILI(s_{PWN}, s_{RWN}) = \min(R_P(s_{PWN}, s_{RWNi}) + R'_P(s_{RWNi}, s_{PWN})) \quad (1), \text{ где}$$

$$R_P = \frac{R(s_{PWN}, s_{RWNi})}{\sum_{i=1}^n R(s_{PWN}, s_{RWNi})},$$

$$R'_P = \frac{R'(s_{RWNi}, s_{PWN})}{\sum_{i=1}^m R'(s_{RWNi}, s_{PWN})} \quad (2)$$

Для получения корректного соответствия синсетов PWN и RWN необходимо, чтобы значения, для которых строится межъязыковой индекс, были определены как в PWN, так и в RWN. Используемые для перевода словари содержали, по крайней мере, корректный перевод одного из слов синсетов. Совпадение переводов словосочетаний в составе синсетов практически гарантирует их соответствие. В настоящей статье не рассматриваются методики для оценки достоверности полученного результата.

### 3.4. Порядок построения списка синсетов-альтернатив

Пусть  $s = \{w_1, w_2, \dots, w_n\}$  - синсет, для которого требуется построить список  $S$  синсетов-альтернатив. Англо-русский и русско-английский словари позволяют получить множества переводов слов синсета:  $w_i = \{t_{i1}, t_{i2}, \dots, t_{in}\}$ . Для каждого перевода  $t_{ik}$  слова  $w_i$  исходного синсета  $s$  можно построить множество синсетов-альтернатив. Однако, поскольку синсет по определению представляет только одно из значений, только один перевод слова  $w_i$  будет представлять значение исходного синсета. Дополнение же списка  $S$  синсетами, содержащими другие слова из множества  $\{t_{i1}, t_{i2}, \dots, t_{in}\}$  приведет к появлению альтернатив с заведомо иным значением, значительно расширит область поиска, затруднит выбор эквивалентного значения из  $S$  с использованием оценочной функции.

Актуальной является задача определения пары  $\{w_i, t_k\}$ , такой, что  $t_k$  представляет наиболее

Таблица 1

вероятный перевод значения синсета  $s$ . В случае, если из-за неполноты англо-русского и русско-английского словаря (если отсутствует перевод слова  $w_i$  со значением синсета  $s$ ), пара  $\{w_i, t_k\}$  будет определена некорректно. Даже такое соответствие  $\{w_i, t_k\}$  приведет лишь к незначительному расширению списка  $S$ , эквивалентный исходному синсет попадет в список  $S$  при переводе других слов синсета  $s$ .

Для определения пары  $\{w_i, t_k\}$  используется методика расчета семантического расстояния, основанная на выявлении характерных зависимостей между словами по значениям их условных вероятностей в достаточно большом и представительном корпусе текстов, таком, каким сейчас является Internet. За основу взята формула [13] для определения нормализованного поискового расстояния (также известного как *NGD – normalized Google distance*):

$$NGD(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log M - \max\{\log f(x), \log f(y)\}} \quad (3)$$

где  $M$  – число web-документов, проиндексированных поисковой системой; для Google на момент написания статьи  $M = 8\,058\,044\,651$  страниц;  $F(x) = Mp(x)$  – число web-страниц, на которых встречается слово  $x$ ;  $f(y) = Mp(y)$  – число web-страниц, на которых встречается слово  $y$ ;  $f(x, y) = Mp(x, y)$  – число web-страниц, на которых одновременно встречаются  $x$  и  $y$ .

В качестве примера приведем значительно упрощенный, в демонстрационных целях, вариант определения списка  $S$  синсетов-альтернатив для следующего синсета PWN:

{carriage, equipage, rig} – a vehicle with four wheels drawn by two or more horses

По слову «carriage» в PWN находится 5 синсетов, в том числе и вышеприведенный. Русско-английский и англо-русский словари дают следующие варианты переводов (упрощено):

carriage – {экипаж, карета}, {пассажирский вагон}, {тележка, вагонетка}, {гондола} и т.д. (всего 20 вариантов перевода);

equipage – предположим, нет перевода;

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

Используются доступные интерфейсы поисковых систем, например, Google (<http://www.google.com/apis/>), получаем следующие значения частот (по числу найденных web-страниц) (таблица 1).

Результатом перевода, как видно, не всегда является одно слово. В большинстве случаев, перевод представлен группой слов. Для таких групп, определяется минимальное значение NGD (таблица 3).

| $f(x)$             | приспособление | устройство | агрегат | оборудование | костюм | упряжка | карета |
|--------------------|----------------|------------|---------|--------------|--------|---------|--------|
|                    | 147000         | 1100000    | 206000  | 3260000      | 561000 | 9420    | 69000  |
| экипаж             | 479000         | 995        | 50800   | 4570         | 81400  | 33200   | 600    |
| карета             | 69000          | 430        | 738     | 317          | 1510   | 7640    | 395    |
| пассажирский вагон | 507            | 42         | 187     | 53           | 326    | 138     | 0      |
| тележка            | 159000         | 546        | 17700   | 702          | 52300  | 849     | 112    |
| вагонетка          | 896            | 145        | 591     | 256          | 886    | 392     | 6      |
| гондола            | 195            | 1          | 11      | 0            | 13     | 2       | 0      |

В таблице первый столбец – частоты для переводов «carriage», первая строка – частоты для переводов «rig» по данным, полученным с использованием Google API. В соответствующих ячейках находятся частоты совместного появления одновременно двух слов на одной web-странице. Например, слова «экипаж» и «устройство» одновременно встречаются на 50800 – web-страницах, «упряжка» и «пассажирский вагон» вообще ни в одном документе вместе не встречаются.

Согласно (3) рассчитывается «семантическое расстояние» между словами-переводами, таблица 2.

Таблица 2

| $NGD(x, y) \cdot 10^3$ | приспособление | устройство | агрегат | оборудование | костюм | упряжка | карета |
|------------------------|----------------|------------|---------|--------------|--------|---------|--------|
| экипаж                 | 30.33          | 11.89      | 20.86   | 13.49        | 11.41  | 33.94   | 20.51  |
| карета                 | 31.84          | 35.46      | 35.65   | 34.26        | 18.63  | 29.31   | 0.06   |
| пассажирский вагон     | 54.7           | 46.55      | 53.24   | 45.48        | 46.92  | -       | 49.29  |
| тележка                | 30.34          | 16.76      | 29.53   | 15.39        | 32.02  | 44.02   | 31.11  |
| вагонетка              | 41.18          | 37.09      | 37.57   | 38.57        | 37.79  | 76.09   | 51.44  |
| гондола                | -              | 83.67      | -       | 84.35        | 140.05 | -       | 86.07  |

Таблица 3

| $NGD(x, y) \cdot 10^3$ | приспособление, устройство | агрегат, оборудование | костюм | упряжка, карета |
|------------------------|----------------------------|-----------------------|--------|-----------------|
| экипаж, карета         | 11,89                      | 13,49                 | 11,41  | 0,06            |
| пассажирский вагон     | 46,55                      | 45,88                 | 49,92  | 49,29           |
| тележка, вагонетка     | 16,76                      | 15,39                 | 32,02  | 31,11           |
| гондола                | 83,67                      | 84,35                 | 140,05 | 86,07           |

Здесь минимальное семантическое расстояние зафиксировано для переводов {упряжка, карета} и {экипаж, карета}. Исходя из этого, список  $S$  будет состоять из следующих синсетов RWN:

{упряжка} – несколько лошадей, собак, оленей и т.п., запряженных одной упряжью;

{упряжка, упряжь} – совокупность приспособлений, предметов для запряжки лошадей или другой живой тяги;

{карета} – закрытый со всех сторон четырехколесный конный экипаж на рессорах;

{повозка, экипаж} – колесный безрессорный экипаж простого устройства, а также общее название различных колесных экипажей;

{экипаж} – команда, личный состав корабля, самолета, танка а также состав специалистов, обслуживающих какой-либо движущийся механизм; {экипаж} – береговая воинская часть морской пехоты, часто служащая для пополнения флотских команд.

### 3.4. Порядок вычисления оценочной функции

Оценочная функция  $R = R(s_{WN1}, s_{WN2})$  строится для определения оценки степени соответствия синсетов из словарей WordNet русского и английского языков ( $WN_{L2}, WN_{L1}$ ), алгоритм её расчета и формула, в общем, подобны (3) для расчета нормализованного поискового расстояния. Дополнительно оценивается соответствие:

- перевода словника синсета и словника синсета-альтернативы;
- перевода словника синсета и лемм толкования синсета-альтернативы;
- перевода лемм толкований синсета и словника синсета-альтернативы;
- перевода лемм толкований синсета и лемм толкования синсета-альтернативы;
- словника и толкования синсета-альтернативы и синсета, который был включен в состав ПИ и является родительским в дереве гипонимии/меронимии по отношению к рассматриваемому.
- учитываются стилистические пометы и соответствия, определенные для синсета в WordNet Domains[12].

Перед вычислением значения оценочной функции выполняется предварительная обработка всех синсетов:

- толкования синсетов разбиваются на слова. Для каждого из слов определяется лемма. Для толкований синсетов PWN используется лемматизатор, входящий в состав DICT Server, для толкований синсетов RWN используется лемматизатор ЗАО «Руссикон». Набор лемм заменяет исходные словоформы в толковании синсета на время вычисления индекса. Из набора лемм исключаются леммы, образующие перевод словника синсета. С помощью словарей переводятся все леммы, полученные из толкований синсетов.
- для каждого синсета определяется частота/вероятность. Эти значения определяются по статистике, полученной от поисковых систем (Yandex, Google) по методике, описанной выше. Для PWN

дополнительно используются частоты, приведенные в составе WordNet sense\_key.

### Список литературы:

1. Fellbaum C. WordNet: an Electronic Lexical Database. MIT Press, Cambridge, MA, 1998.
2. Miller G. et al. Five Papers on WordNet. CSL-Report, vol.43. Princeton University, 1990. <ftp://ftp.cogsci.priceton.edu/pub/wordnet/5papers.ps>
3. Vossen, P. EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dordrecht: Kluwer, 1998.
4. Сайт проекта RussNet [http://www.phil.pu.ru/depts/12/RN/index\\_ru.shtml](http://www.phil.pu.ru/depts/12/RN/index_ru.shtml).
5. Портал УИС «Россия» <http://www.cir.ru/>.
6. Balkova V., Suhonogov A., Yablonsky S.A. Russia WordNet. From UML-notation to Interne/Intranet Database Implementation, In: *Proceedings of the Second International WordNet Conference, GWC 2004*. – Brno, Czech Republic, 2004, pp. 31-38.
7. Yablonsky S.A. Integration of Russian Language Resources. In: *Proceedings 4th International Conference on Language Resources & Evaluation*. – Barcelona, Spain, 2004.
8. Сухоногов А.М., Яблонский С.А. Разработка русского WordNet. – *Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды шестой всероссийской научной конференции RDCL`2004*, Пушино, 2004. – стр. 113-117.
9. Сухоногов А.М., Яблонский С.А. Словари типа WordNet в технологиях Semantic Web. – *Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004*. Труды конференции. В 3-х т., Т.2. М.:Физматлит, 2004. – стр. 557-564.
10. Сухоногов А.М., Яблонский С.А. Использование WordNet при поиске и индексации текстов.- *Международная конференция «Корпусная лингвистика - 2004»*: Тезисы докладов. – СПб., 2004. – стр. 84-89.
11. Yablonsky S.A. Russicon Slavonic Language Resources and Software. In: A. Rubio, N. Gallardo, R. Castro & A. Tejada (eds.) *Proceedings First International Conference on Language Resources & Evaluation*, (pp. 1141–1147). – Granada, Spain. 1998.
12. Magnini B., Cavaglia B. Integrating Subject Field Codes into WordNet // *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation / M. Gavrilidou, G. Crayannis, S. Markantonatu, S. Piperidis, G. Stainhaouer (eds.)*. – Athens, Greece, 31 May – 2 June, 2000. – P. 1413–1418.
13. Rudi Cilibrasi, Paul Vitanyi Automatic Meaning Discovery Using Google, <http://www.arxiv.org/abs/cs.CL/0412098>.
14. RFC2229 – A Dictionary Server Protocol,
15. <http://www.faqs.org/rfcs/rfc2229.html>