Лексические функции и возможности оптимизации поиска информации в интернете (на материале параметрических слов)

Lexical functions and search engine optimization (based on words with numeric values).

Тимошенко С. П. (timoshenko@iitp.ru), **Цинман Л. Л.** (cinman@iitp.ru)

Институт проблем передачи информации им. А. А. Харкевича, Москва, Россия

На базе лингвистического процессора ЭТАП-3 была разработана опция экспериментального перифразирования. Она дополняет двух- или трехсловный поисковый запрос о числовом значении параметра до неполного предложения. Эксперимент доказал, что показатель точности поиска повышается в среднем на 24 %.

1. Вводные замечания

Задача повышения точности поиска в Интернете не всегда хорошо решается чисто математическими методами. Для лингвиста это предсказуемо, поскольку в таких ситуациях мы имеем дело с асимметрией между означающим и означаемым. Применительно к поиску можно сказать, что означаемым является искомый смысл, а означающим — вся совокупность выражающих этот смысл предложений. Предложения языка L, выражающие один и тот же смысл, могут очень сильно отличаться друг от друга. Задача поисковой машины в таком случае — распознать смысл, игнорируя формальные различия. Соблазнительно попробовать решить эту задачу с помощью лингвистических инструментов, ориентированных на описание разнообразных способов выражения смысла — с помощью средств лексической семантики. Одним из таких инструментов является аппарат лексических функций, разработанный И. А. Мельчуком и А. К. Жолковским и значительно дополненный и усовершенствованный Ю. Д. Апресяном. См. [4–9; 1, 3].

2. Роль лексических функций в перифразировании

В самом общем виде можно сказать, что лексические функции — это тривиальные смыслы, словесное выражение которых в тексте зависит от того, при каком конкретном слове этот смысл выражается. Для некоторых фрагментов лексической системы языка разработанные лексической семантикой правила вида: «При слове X смысл f_1 выражается словом X′, при слове Y смысл f_1 выражается словом Y′» обладают большой предсказательной силой. Одним из таких фрагментов является класс параметрических слов. Под параметрическими словами мы понимаем имена существительные со значением параметра, допускающего числовое значение, например: *высота, вместимость, объем; продолжительность, возраст; мощность, сила, масса, давление, магнитуда; рождаемость, смертность; цена, стоимость, зарплата, выручка; энтропия, уровень, коэффициент, индекс* и т. д.

Правила лексической семантики, описывающие функционирование слов этого класса, в русском языке носят особенно строгий характер. Это связано с характерной чертой языка: хотя параметры являются универсальным типом предикатов, прототипическими представителями данного класса в русском языке являются не глаголы, а существительные ([2]:С. 74). Глаголов с соответствующими значениями, таких как *стоить, весить, длиться, вмещать*, очень мало. Даже для выражения такого тривиального параметра, как 'высота', специализированного глагола нет (в отличие, например, от английского: *The Pisa tower*

rises 56 meters).[1] Поэтому естественным способом приписывания какому-либо объекту определенного параметра является конструкция из параметрического существительного и глагола (*Пизанская башня* ***достигает в высоту*** *55 метров*).

Отношения между существительным и глаголом в огрубленом виде предстают как отношения названия ситуации и вспомогательного глагола, служащего для выражения категориальных значений — вида, времени и т.д. При более внимательном рассмотрении оказывается, что несвободные сочетания вида «глагол-связка + существительное» делятся на группы, объединенные общим элементом смысла, который и есть значение лексической функции.

Аппарат лексических функций лежит в основе опции перифразирования, реализованной в системе ЭТАП-3. Перифразирование позволяет, опираясь на записанные в словарных статьях слов значения лексических функций, построить некоторое количество предложений, синонимичных или квазисинонимичных заданному, то есть предложений, имеющих с ним одинаковую структуру на глубинном семантическом уровне. В системе ЭТАП для описания несвободных словосочетаний, вообще говоря, используется несколько десятков лексических функций. Они позволяют из предложения *Чистка матрицы стоит тысячу рублей.* получить следующий набор предложений (1):

Чистка матрицы имеет стоимость тысяча рублей.

Чистка матрицы достигает стоимости тысяча рублей.

Чистка матрицы имеет цену на тысячу рублей.

Стоимость чистки матрицы составляет тысячу рублей.

Стоимость чистки матрицы достигает тысячи рублей.

Стоимость чистки матрицы равняется тысяче рублей.

Цена чистки матрицы составляет тысячу рублей.

С помощью тех же лексических функций можно, подав на вход любое из получившихся предложений, снова получить весь набор перифраз.

Опция перифразирования была приспособлена для решения поисковых задач следующим образом: из именного словосочетания, в вершине которого находится параметрическое слово, можно получить ряд неполных предложений, содержащих все элементы, кроме численного значения параметра.

Ввод: *глубина Марианской впадины*

Перифразы, генерируемые системой перифразирования ЭТАП-3 в экспериментальном режиме (2):

Глубина Марианской впадины равна.

Глубина Марианской впадины составляет.

Глубина Марианской впадины достигает.

Глубина Марианской впадины равняется.

Марианская впадина имеет в глубину.

Марианская впадина достигает в глубину.

Марианская впадина имеет глубину.

Марианская впадина достигает глубины.

3. Коротко об алгоритмической организации экспериментального перифразирования

Коснемся лишь того фрагмента перифразирования, который относится к нашей задаче. Подробно перифразирование в системе ЭТАП описано в[3]. Приведенный выше куст перифраз (2) строится на основе информации о лексических функциях, содержащейся в словарной статье параметрического слова *глубина*, и некоторых универсальных правил перифразирования.

Для нашей задачи в слове *глубина* интерес представляют записи, относящиеся только к трем функциям:

OPER1:ИМЕТЬ/ДОСТИГАТЬ
FUNC2:СОСТАВЛЯТЬ1/ДОСТИГАТЬ/РАВНЯТЬСЯ1
LABOR1–2:ИМЕТЬ<В1>/ДОСТИГАТЬ<В1>

Из всего многообразия лексических функций, обслуживающих слово *глубина* эти три соединяют имя параметра с глаголом таким образом, чтобы получилось описание ситуации, когда что-либо имеет определенную глубину. Функция OPER1 позволяет обозначить ситуацию так, что параметр выступает

[1] Русские глаголы *выситься* и *возвышаться*, хотя и допускают конструкции со значением параметра (*Скала возвышается на 140 метров над уровнем Атлантического океана*.), в подавляющем числе случаев имеют не значение ***'X имеет высоту Y'***, а значение ***'X, имеющий большую высоту, располагается в Z'*** (или 'имея большую высоту'?)(Для глагола возвышаться это значение еще уточняется. ***'X, имеет большую высоту, и эта высота существенно больше высоты Z, рядом с которым располагается X'*** Типичные предложения с этими глаголами: *На ее могиле у Заонежского села Кузаранда высится стела. …над нами возвышается гигантский пресс.* (Примеры взяты из Национального корпуса русского языка и из СинТагРуса).

при соответствующем глаголе первым дополнением, например: *Гора Котопакси имеет **высоту** почти 6 км.* Функция FUNC2 позволяет обозначить ситуацию так, чтобы параметр был подлежащим при функциональном глаголе: ***Высота** горы Котопакси составляет почти 6 км.* Функция LABOR1–2 позволяет так обозначить ситуацию, что параметр занимает место второго дополнения при функциональном глаголе, а места подлежащего и сказуемого занимают первый и второй актанты соответственно: *Гора Котопакси достигает 5 870 м **в высоту**.*

Из универсальных правил перифразирования (их в системе ЭТАП несколько десятков) задействованы лишь три двусторонних правила

OPER1 + X <--> FUNC2 + X (*иметь глубину <--> глубина составляет*)

OPER1 + X <--> LABOR1–2 + X (*иметь глубину <--> иметь в глубину*)

FUNC2 + X <--> LABOR1–2 + X (*глубина составляет <--> иметь в глубину*)[2]

Если какая-либо лексическая функция имеет несколько значений (в нашем примере все три функции представлены альтернативными значениями), то система перифразирования строит предложения поочередно со всеми значениями. Один и тот же глагол может быть значением различных лексических функций (в нашем примере: *достигать высоты, высота достигает, достигать в высоту*). У различных параметрических слов значения этих трех лексических функций часто совпадают, но встречаются и заметные различия. Например, для слова *мощность* функция OPER1 имеет три значения

OPER1:ИМЕТЬ/ДОСТИГАТЬ/РАЗВИВАТЬ,

а функции LABOR1–2 для этого слова не существует.

Приведенный выше куст перифраз состоит из 8 предложений. Система перифразирования построит этот куст целиком, если ей на вход подадут либо именную группу (как в приведенном выше примере), либо любое предложение из этого куста.

Отметим, что для решения нашей задачи потребовалось установление отдельного режима работы системы ЭТАП. Дело в том, что все правила перифразирования в ЭТАПе настроены на работу с полными фразами, а фразы приведенного выше куста полными не являются. Поэтому при этом режиме работы системы после получения синтаксической структуры входного (неполного) предложения производится достройка этого предложения до полного. Например, если на входе была именная группа *глубина Марианской впадины*, то достраиваем это предложение до *глубина Марианской впадины равняется чему-то.* Глагол *равняется* — дежурное значение лексической функции FUNC2 для абсолютного большинства параметрических слов, а существительное *что-то* выполняет роль временного дополнения, необходимого при перефразировании. После синтеза очередной перифразы это временное существительное стирается. Если на вход ЭТАПа подается одно из предложений куста, где глагол присутствует, то синтаксическая структура пополняется только временным дополнением.

Упомянем еще одну проблему, которая возникает, когда на вход подается именная группа. Например, для запроса *водоизмещение 'Титаника'* наша система в настоящий момент построит запросы

водоизмещение 'Титаника' составляет,
'Титаник' имеет водоизмещение,
…,

в которых глаголы стоят в настоящем времени. Такие запросы не могут улучшить поиск, поскольку форма точного запроса не предполагает в качестве результата словоформы, отличающиеся от заданных наборами грамматических характеристик, а в текстах о 'Титанике' эти глаголы, скорее всего, употреблены в прошедшем времени. Подобная ситуация может возникнуть и с характеристикой вида глагола (сов/несов). Решение этой проблемы могло бы заключаться в порождении всех перифраз, где глаголы имеют различные характеристики времени и вида. Алгоритмически это сделать не сложно, но число перифраз при этом заметно возрастет.

В то же время упомянутая проблема исчезает, если на вход подается запрос, содержащий глагол в требуемой для данного запроса форме

водоизмещение 'Титаника' составляло
смертность в России к 2010 г. достигнет
…

В этом случае система перифразирования сохранит для всех перифраз характеристики глагола, поступившего на вход.

4. Эксперимент, определяющий эффективность поиска по перифразам

На основе исходного списка из 100 параметрических слов было составлено около 120 осмысленных коротких запросов (параметр + носитель пара-

[2] Особняком стоит правило OPER1 + X <--> X + *(быть) равным* (*иметь глубину <--> глубина равна*), поскольку, с одной стороны, это правило не универсальное (оно справедливо только для параметрических слов, у которых есть OPER1), а, с другой стороны, выражение *(быть) равным*, строго говоря, не является значением функции FUNC2.

метра). Именно к таким запросам, насколько можно судить по статистике Яндекса, чаще всего прибегают пользователи. С помощью системы ЭТАП-3 каждый запрос преобразовывался в блок перифраз. В ходе эксперимента перифразы вводились по одной в поисковые системы «Яндекс» и «Google». Поскольку нас интересовала принципиальная возможность улучшить поиск, автоматически расширяя запрос, время выполнения запроса и время обработки запроса системой ЭТАП-3 не учитывалось, равно как и дата проведения эксперимента и загрузка поисковых серверов. Поскольку система перифразирования выдает целостные структуры, не выходящие за пределы одного предложения и не предусматривающие разрывов и пропусков, для тестирования была выбрана форма точного запроса. Применение логического оператора «ИЛИ» внутри точного запроса языками поисковых запросов не поддерживается. Употребление дизъюнкции при неточном запросе приводит к тому, что находятся не сайты с целостной структурой, а сайты, просто содержащие заданные слова, возможно, не связанные синтаксически. Поэтому группа слов, выражающая численное значение и зависящая от глагола, может относиться к своершенно другому объекту или к другому параметру. Частично снимает эту проблему опция поиска с пропуском заданного количества слов. Можно себе представить неточный запрос вида *водоизмещение Титаника /+1 составлять*. Однако и он не дает стопроцентной точности. На первую же страницу почему-то попадает такой фрагмент:

Через два часа тридцать пять минут после катастрофы крен «**Титаника**» **составлял** почти 90 градусов. … Длина парома "Геральд оф Фри Энтерпрай» **составляла** 132 метра, **водоизмещение** — 7951 тонна. Он являлся составной частью флота, управляемого компанией…

Очевидно, эти опции пока еще не всегда корректно обрабатываются поисковыми машинами. Если усложнить запрос, насытив его «лингвистическими» маркерами (например «Гора Котопакси» /+1 достигает /+3 «в высоту»), доля неподходящих ответов еще возрастет.

Употребление точной формы запроса привело к тому, что различия в работе двух упомянутых поисковых систем практически нивелировались, за исключением различий в составе баз документов (например, «Яндекс» индексирует больше личных блогов, чем Google). Эти различия оказались настолько незначительными, что мы сочли возможным ими пренебречь.

Цель эксперимента заключалась в том, чтобы определить, насколько использование системы экспериментального перифразирования повышает эффективноость поиска. Для этого был разработан следующий протокол оценки эффективности. Релевантным признается такой результат, когда искомая численная информация содержится непосредственно в сниппете, предлагаемом поисковой машиной. Показателем эффективности поиска считается количество релевантных результатов на первой странице (т. е. количество релевантных результатов среди первых десяти), выраженное в процентах. Если количество найденных документов меньше 10, то оно и принимается за 100%. В качестве контрольного уровня эффективности поиска был принят уровень эффективности поиска по исходным запросам — именным группам, подаваемым на вход системы перифразирования ЭТАПа-3. Эти запросы также оформлялись как точные — это ограничение дает возможность оценить чистый эффект использования перифраз.

В результате этих операций получаем данные по каждому запросу:

Запрос	**Найденные страницы**	**Найденные сайты**	**Кол-во релевантных документов на 1 странице**
«твердость алмаза»	2044	791	3/10 (30%)
«твердость алмаза равна»	42	18	10/10 (100%)
«твердость алмаза составляет»	18	6	3/6 (50%)
«твердость алмаза достигает»	0	0	0
«твердость алмаза равняется»	0	0	0
«алмаз имеет твердость»	79	35	9/10 (90%)
«алмаз достигает твердости»	0	0	0

Первая строка в данных по каждому запросу будет представлять эффективность поиска по необработанным словосочетаниям, а все остальные — по обработанным перифразам. Чтобы оценить среднее изменение результата, можно взять среднее значение эффективности необработанных запросов и среднее значение по всем перифразам вместе. Однако такой подход кажется нам неверным. Представим себе, как могла бы работать поисковая машина, умеющая генерировать перифразы. Она получает на входе запрос, создает перифразы, а затем обрабатывает каждую из них как точный запрос. Какие-то перифразы приносят результат, какие-то — нет. Очевидно, что на экран в таком случае выводится сумма найденных по всем перифразам документов, а несработавшие перифразы, конечно, ничего не добавляют, но и не ухуждшают картины. Поэтому мы рассчитывали повышение эффективности поиска для каждого запроса отдельно, и уже для этих результатов затем рассчитывался средний показатель, характеризующий общее изменение эффективно-

сти. Так, для приведенного выше запроса средний показатель увеличения эффективности составит

$$((100\% - 30\%) + (50\% - 30\%) + (90\% - 30\%)) / 3 = 50\% \text{ или } 0{,}5.$$

Бывают случаи, когда необработанный запрос приносит какое-то количество релевантных ответов, а перифразы не приносят ничего. Например, *«абсолютный минимум температуры на Земле»*. Это связано с тем, что информация по этому вопросу очень часто представлена в Рунете перепечаткой статьи из Большой советской энциклопедии, где использована конструкция с нулевой связкой вместо глагола. В этом и других похожих случаях мы признавали результат использования системы перифразирования отрицательным: если результаты поиска по необработанному запросу были релевантны в 10 случаях из 10, то эффективность поиска с перифразированием равна −100% или −1, если в 6 случаях из 10 — −60% или −0,6 и т.д.

Распределение показателей изменения эффективности поиска при использовании системы перифразирования представлено на гистограмме (рис. 1).

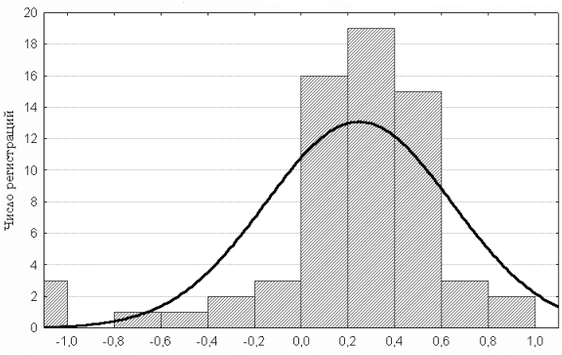

Рис. 1. Изменение эффективности поиска при использовании запросов на основе перифраз аппроксимация: 65 × 0,2 × normal (x; 0,2395; 0,3966)

На графике видно, что в среднем точность поиска повышается на 24%. Любопытно, что если разделить параметрические слова по тематическим группам, то этот средний показатель будет варьироваться. Если смотреть данные по «географическим» запросам (параметры географических объектов, как природных, так и культурных, например, *ширина реки Амазонки в среднем течении* или *численность населения Киева*), то повышение эффективности поиска составит в среднем 18,5%, а в группе данных по физическими запросам (примеры запросов: *молярная масса меди, сила тяжести на Марсе*) этот показатель будет равен 27%. Результаты поиска по геометрическим объектам (примеры запросов: *площадь поверхности сферы, объем конуса*) оказались неожиданными. Реализующий значение лексической функции FUNC2 глагол *равняться,* а также добавленная в поисковое перифразирование конструкция с кратким прилагательным *равен,* отличаются от других средств тем, что регулярно присоединяют в качестве первого дополнения формулу или её словесное описание: *Объём конуса равен одной трети произведения основания на высоту; Объем конуса равняется одной трети объема цилиндра с теми же основанием и высотой*. Показатель эффективности использования перифраз для поиска информации такого рода составляет 13%. Однако ЭТАП-3 можно настроить на поиск подобной информации, добавив в систему перифразирования глаголы *характеризовать*, *измеряться* и, возможно, некоторые другие. Отдельно подчеркнем, что эти глаголы не являются значениями лексических функций параметрических слов. Точно так же не описываются лексическими функциями конструкции типа *площадь восьмиугольника может быть вычислена как…*. Информацию о том, что формулы могут вводиться подобными предложениями мы черпаем из экстралингвистических фактов. Соответственно, в рамках системы ЭТАП можно построить ряд правил, позволяющих получать подобные предложения, однако эти правила, скорее всего, будут крайне неполными: при ничем не ограниченной сочетаемости слов практически невозможно описать все возможные предложения и тем боле угадать предложения, реально присутствующие в интернет-документах.

5. Переводы запросов на английский язык

Поскольку одной из опций процессора ЭТАП является автоматический перевод текста, то для системы не составляет труда перевести все полученные путем перифразирования словосочетания на английский. Из запроса *высота Пизанской башни* с помощью ЭТАПа легко получить набор словосочетаний (3):

The height of the Pisa tower equals

The height of the Pisa tower reaches

The height of the Pisa tower is reaching

The height of the Pisa tower amounts to

The height of the Pisa tower attains

The height of the Pisa tower is attaining

Однако эффективность поиска в данном случае практически не повышается. И это вновь связано с особенностями строя языка, в частности, с тем хорошо известным фактом, что в английском язы-

ке параметры задаются прилагательными со значением верхнего полюса шкалы (30 feet high, 6 feet tall, 25 years old), что русскому языку практически не свойственно (за исключением известных редких примеров типа *как велика вероятность n*). В случае параметров для носителей языка естественнее употреблять прилагательные: *The Cupola is 55 meters high and 16 meters wide.* Поэтому перифразы, построенные ЭТАПом с помощью существующих на сегодняшний день правил, несмотря на то, что грамматически они абсолютно правильны, могут вообще не встречаться среди источников. Именно так дело обстоит с Пизанской башней.

Кроме того, играет роль общеизвестный факт обязательности глагола-связки в английском языке. В тех случаях, когда параметр все-таки обозначается существительным, носителям английского языка не нужно подбирать глагол, который мог бы выразить грамматические категории — достаточно просто использовать глагол *be: The speed of light is 300 million metres per second*. На данном уровне экспериментальное перифразирование не может быть использовано для перевода запросов. Однако именно в силу того, что в английском разница между формой запроса и формой ответа столь заметна, это поле деятельности представляется довольно интригующим. Чтобы двигаться в этом направлении, необходимо расширять как правила перифразирования, так и инструменты, позволяющие переводить глубинные английские структуры предложений типа *The Cupola is 55 meters high в глубинные русские структуры предложений типа Высота купола составляет 55 метров. Это может стать шагом к построению полноценного глубинно-синтаксического представления в той действующей модели языка, которая лежит в основе лингвистического процессора ЭТАП-3.*

6. Неточные запросы

Данные эксперимента показывают, что использование перифраз повышает точность поиска за счет выбора только тех документов, которые содержат искомую информацию, наличие которой однозначно предсказывается глаголом, реализующим ту или иную лексическую функцию. Однако требование точного запроса приводит к тому, что отсеиваются и релевантные документы, в которых мысль выражена немного по-другому: не будет найден документ, содержащий предложение *Гора Котопакси, высота которой составляет почти 6 км…* Неточный поиск по перифразе приводит к тому, что конструкция разрывается и по запросу *продолжительность жизни в Голландии составляет,* находятся страницы новостей, где есть информация о продолжительности жизни в Китае и сообщение о политической ситуации в Голландии. Однако некоторые попытки такого поиска показывают, что и в этом случае точность поиска по перифразам выше точности поиска без них. Это происходит из-за того, что наличие глагола, пусть даже и оторванного от именного словосочетания, меняет общую направленность страницы. Например, по запросу *высота юкки* находятся страницы, содержащие объявления о продаже пальмы какой-либо высоты, а по запросу *высота юкки достигает* находятся статьи из разнообразных справочников по цветоводству, содержащих общую информацию о растении. Можно сказать, что в действие вступает стилистический фактор, так как глаголы, реализующие лексические функции OPER1, FUNC2, LABOR1–2 часто используются в научно-публицистических и научных текстах. Это открывает для системы перифразирования определенные перспективы, тем более что в системе ЭТАП-3 предусматривалось применение стилистических фильтров.

7. Выводы

Эксперимент показал, что точность результатов поиска по запросам, предполагающим численные ответы, может быть увеличена за счет использования системы экспериментального перифразирования. В случае точного запроса количество релевантных результатов увеличивается на 24%. Точный запрос означает, что к искомым документам предъявляются самые жесткие требования. В случае менее жестких требований, то есть при использовании нестрогого запроса, результаты непредсказуемы. Помимо варьирования форм запросов, зависящего от поисковой машины, пути улучшения работы экспериментального перифразирования лежат в расширении числа используемых лексических функций и более точной их настройки на разные типы информации. В частности, эксперимент показал, что, изменяя участвующие в перифразировании глаголы, можно получать либо конкретные числовые значения, либо формулы и их словесные описания.

Литература

1. *Апресян Ю. Д.* Избранные труды. Лексическая семантика. Синонимические средства языка. // М.: 1995. Т. 1, 2-е издание.
2. *Апресян Ю. Д.* Основания системной лексикографии // Языковая картина мира и системная лексикография. М.: Школа «Языки русской культуры», 2006.
3. *Апресян Ю. Д., Цинман Л. Л.* Формальная модель перифразирования предложений для систем переработки текстов на естественных языках // Русский язык в научном освещении. 2002. № 4. С. 102–146.
4. *Жолковский А. К., Мельчук И. А.* О возможном методе и инструментах семантического синтеза // Научно-техническая информация. 1965. № 6.
5. *Жолковский А. К., Мельчук И. А.* О семантическом синтезе // Проблемы кибернетики. — 1967. № 19. С. 177–238.
6. *Жолковский А. К., Мельчук И. А.* К построению действующей модели языка «Смысл <=> Текст» // Машинный перевод и прикладная лингвистика. 1969. №11. С. 5–35.
7. *Мельчук И. А.* Опыт лингвистических моделей «Смысл <=> Текст». Семантика, синтаксис // М.: Школа «Языки русской культуры», 1999.
8. *Мельчук И. А., Жолковский А. К.* Толково-комбинаторный словарь современного русского языка. Опыт семантико-синтаксического описания русской лексики // Wien: Wiener Slawistischer Almanach, 1984.
9. *Mel'čuk I.* Lexical Functions: A Tool for the Discription of Lexical Relations in a Lexicon // Lexical Functions in Lexicography and Natural Language Processing. Ed. By L. Wanner. Amsterdam (Philadelphia). 1996. P. 37–102.