УДК 004.9

МЕТОДЫ ИНФОРМАЦИОННОЙ КОГНИТИВНОЙ СЕМАНТИКИ

Цветков В.Я., д.т.н., профессор, E-mail:cvj2@mail.ru
Чехарин Е.Е., ст. преподаватель,
МГТУ МИРЭА, Москва, Россия,

Аннотация. Статья анализирует методы информационной когнитивной семантики в области анализа и интерпретации компьютерных текстов. Вводятся и обосновываются новые понятия: информационная конструкция, сигнификативные семантические информационные единицы, предикативные семантические информационные единицы, ассоциативные семантические информационные единицы. Показано, что информационная когнитивная семантика развивает идеи семантической теории информации. Раскрываются технологии корпусной кластеризации и автоматизированного построения справочников.
Ключевые слова. информация, информационные модели, семантика, когнитивная семантика, компьютерная лингвистика, информационные конструкции, семантические информационные единицы.

METHODS OF INFORMATION COGNITIVE SEMANTICS

Tsvetkov V.Y., DtechSci. Prof., E:mail:cvj2@mail.ru
Chekharin E.E., MSTU MIREA, Moscow, Russia

Abstract. The Article analyzes the methods of information cognitive semantics. The region of analysis and interpretation of computer texts. Article introduces and justifies the new concepts: information design, signifying semantic information units predicative semantic information units, associative semantic information units. Article shows that cognitive semantics information develops the ideas of semantic information theory. The article describes the technology of case clustering and automated construction manuals
Keywords: information, information models, semantics, cognitive semantics, computational linguistics, information design, semantic information units

Введение. Информационная когнитивная семантика является междисциплинарным направлением, объединяющим семантическую теорию информации и когнитивную семантику. Когнитивная семантика [1, 2] является частью когнитивной лингвистики [3, 4]. Основными принципами когнитивной семантики в аспекте семантической теории информации являются: концептуализация; фиксация концептуальной структуры в информационном потоке (речи); когнитивные ресурсы [5]. В рамках области когнитивной лингвистики, метод когнитивной семантики делит семантику (смысл) на смысловое построение и выражение знания. Таким образом, когнитивная семантика изучает многое из того, что традиционно входило в сферу прагматики, а также семантики. Информационная когнитивная семантика заменяет понятие «лексема» [6] на понятие «семантическая информационная единица» [7].

Соответственно анализ информационных единиц выполняется не только человеком, но и в компьютерных программах, включая интеллектуальную обработку информации. В то же время в [8] показано, что семантическая обработка информации позволяет создать новое качество.

Основная часть. Как научную область информационную когнитивную семантику интересуют ряд вопросов: что такое сигнификативный «смысл» для семантических информационных единиц? Что такое предикативный «смысл» для предложений? Что такое ассоциативный «смысл»для фраз? Каким образом осуществляется композиция семантических информационных единиц в предложение? С чем связана интерпретируемость информационных единиц? Как в рамках информационных моделей различать содержательность модели от ее формального представления? Чем семантическая информационная модель отличается от прочих информационных моделей? Какие вспомогательные средства анализа можно использовать для выявления смысла в информационных конструкциях? Во что трансформируется понятие концепта в информационной когнитивной семантике? Что общего и в чем различия между информационной когнитивной семантикой и лексической семантикой, структурной семантикой и теорией композиции соответственно?

Классические теории семантики, как правило, объясняют смысл языковых единиц (информационных единиц) с точки зрения необходимых и достаточных условий [9], смысл предложения с точки зрения истинных условий [10], а композицию с точки зрения пропозициональной функций.

В информационной когнитивной семантике смысл отдельных семантических информационных единиц также определяется полной, идентифицируемостью, интерпретируемостью и различимостью.

Идентифицируемость означает, что информационная единица имеет обозначение, присущее только ей.

Интерпретируемость означает, что существует механизм, позволяющий осуществлять интерпретацию данной информационной единицы.

Различимость означает, что описание информационной единицы позволяет морфологически отличать данную единицу от других.

Полнота, хотя и является условной характеристикой, означает достаточность описания данной информационной единицы для ее применения в практической деятельности.

В информационной когнитивной семантике предложение рассматривается как совокупность связанных семантических информационных единиц (слов) [11].

Предложение как составная информационная единица обладает смысловой эмерджентностью, то есть не сводимостью смысла сложной информационной единицы – предложения к составляющим его простым информационным единицам. Предикативный смысл предложения рассматривается как отношение к действительности, что проверяется логическими средствами [12].

В информационной когнитивной семантике выделяют три группы простых информационных единиц: носители смысла основные; носители смысла вспомогательные; не имеющие самостоятельного смысла, но служащие для конструирования других единиц и связей между ними.

Если использовать понятие смысла языковых единиц (информационных единиц), то это дает основание разделить семантические информационные единицы по типам их смысла [13]. Слова обладают сигнификативным смыслом, следовательно их можно обозначить как сигнификативные семантические информационные единицы.

Сигнификативная семантическая информационная единица – это семантическая информационная единица, смысл которой по ее нормализованной форме определяется по словарям и тезаурусам. Примером такой единицы является слово.

Предикативная семантическая информационная единица – это семантическая сложная информационная единица, смысл которой определяется на основе предикации. Примером такой единицы является предложение как совокупность связанных слов.

Ассоциативная, или когнитивная, семантическая информационная единица – это семантическая сложная информационная единица, полный смысл которой определяется на основе ассоциаций. Примером такой единицы является фраза как совокупность связанных предложений. При этом предложения связаны между собой. Смысл одних предложений вытекает из смысла других предложений этой фразы.

Важным понятием и обобщением информационной когнитивной семантики является понятие «информационная конструкция» [14], которая подразумевает объединение информационных единиц, имеющее структуру . Информационная конструкция может создавать новый смысл всей конструкции и новые связи между составляющими ее элементами. Информационные конструкции позволяют вводить дополнительные информационные отношения между элементами конструкции. Этим повышается интерпретируемостью Информационные конструкции являются средством коммуникации, передачи знаний и средством моделирования картины окружающего мира [15].

Информационные единицы образуют свой язык [16]. Наглядный пример - кодирование и криптография, которые имеют свои языковые средства. Каждая

информационная единица имеет информационное окружение [17], которое определяется связями и отношениями в информационном поле.

Необходимо различать понятия информационного поля и информационного пространства [18]. Информационное пространство является пассивной или констатирующей субстанцией. Информационное поле [19] является активной субстанцией, которая содержит связи, отношения и «полевые переменные», характеризующие точки информационного поля. В отличие от физических непрерывных полей информационное поле может быть дискретным: топологическое поле, семантическая сеть. Интерпретируемость информационных единиц определяется их семантическим окружением [20]. Следовательно, необходимо различать семантическое окружение и информационное окружение.

В когнитивной информационной семантике существует разделение понятий информационное пространство, информационное поле, информационная среда и семантическое окружение [21, 22]. Информационная среда с одной стороны является характеристикой объекта. С другой стороны она содержит признаки, характерные для разных объектов класса или вида. Это делает информационную среду основным инструментом при кластеризации текстовых и информационных массивов.

Сложность информационных ресурсов приводит к необходимости редуцирования моделей и упрощения условий решаемых задач. Одно из таких упрощений применяется при анализе контентов или текстов. В задаче кластеризации текстов или информационных коллекций единиц предполагается, что информационные семантические единицы первоначально независимы друг от друга, а «значение» каждой информационной единицы задается семантической метрикой, определяющей ее статистический «вес» .

Один из вариантов анализа информационных единиц в компьютерных массивах основан на использовании процедуры лемматизации (англ. lemmatization). Эта процедура морфологического анализа, которая приводит информационную единицу (словоформу) к ее словарной форме (лемме). В этом случае слово приобретает сигнификативный смысл и проверяется по словарю. Соответственно по словарю определяется «вес» слова. Метод лемматизации применяется в поисковых алгоритмах в процессе схематизации веб-документов при их индексировании.

Здесь следует отметить преимущество искусственных языков, поскольку в них сигнификативные информационные единицы (слова) меньше изменяют форму (некоторые ее вообще не меняют). Это позволяет проводить более скоростной и более точный анализ, поскольку исключаются процедуры преобразования слов.

В естественных языках опорные или словарные формы различны для разных языков. Например, в русском языке словарной формой считается: существительные - именительный падеж, единственное число (выборы – выбор); глаголы - инфинитивная форма (искали – искать); прилагательные - единственное число, именительный падеж, мужской род (далекими - далекий). Однако в любом естественном языке существует некоторый процент слов, которые могут давать неоднозначные результаты в процессе лемматизации и приводить к двум леммам. В русском языке такой двойственностью обладают отглагольные существительные

Другой процедурой, применяемой при анализе информационных текстовых коллекций, является стемминг [23]. Он представляет собой процесс нахождения основы слова при заданной не словарной форме слова. При этом основа слова необязательно совпадает с морфологическим корнем слова, что является дополнением к морфологическому анализу.

Стемминг применяется в поисковых системах и является частью процесса нормализации текста. В результате процедуры стемминга образуется неструктурированная совокупность - «мешок слов» (bag-of-words) [24], с которой производят процедуры преобразования и кластеризации.

Исключение структуры и связей упрощает анализ и является аналогом изучения поверхности объекта без исследования его глубинного содержания. При таком «поверхностном» (shallow) [25] представлении текстовых данных «поверхностная» кластеризация методом «мешка слов» не учитывает структуру текста, не учитывает связи между единицами в тексте. Все это приносится в жертву скоростной обработке

Простой стеммер (программа) ищет внутреннюю флективную форму в таблице поиска. Преимущества этого подхода заключается в его простоте, скорости, а также легкости обработки исключений. К недостаткам можно отнести то, что все флективные формы должны быть явно перечислены в таблице: новые или незнакомые слова не будут обрабатываться, даже если они являются правильными. Поэтому таблица поиска может быть очень большой. В свою очередь это приводит к необходимости применения значительных вычислительных ресурсов, таких как кластеры [26]. Таблицы поиска, используемые в стеммерах, как правило, генерируются в полуавтоматическом режиме. Это требует их коррекции со стороны человека.

В последнее время получают развитие автоматические системы обработки текстов, основанные на методах и алгоритмах компьютерной лингвистики. Они выполняют глубокий лингвистический анализ текстов [27] на естественном языке. Классический лингвистический подход к анализу текста предполагает существование

относительно независимых уровней анализа таких как: морфологического, синтаксического и семантического. При этом задается последовательность анализа, в начале - морфологического, затем синтаксического и, наконец, семантического. По существу леманизация и стемминг относятся к первому этапу, потому что анализируют форму. В тоже время в литературе отмечают, что стемминг отличается от морфологического анализа.

Лингвистические методы анализа текстов основываются на правилах, разработанных экспертами-лингвистами. Для создания автоматических систем на основе этих правил требуется разработка модели естественного языка, что в каждом отдельном случае требует больших трудозатрат высококвалифицированных лингвистов и системных операторов. По существу эта задача переносится в область информационной когнитивной лингвистики [28], поскольку требует создания информационных моделей и информационных отношений отражающих лингвистические отношения.

Сложность этой задачи мотивировала развитие метода стемминга, как альтернативы. Одним из главных недостатков классических стеммеров является то, что они могут не различать слова со схожим синтаксисом, но с разными смысловыми значениями. Это приводит к проблеме снятия морфологической омонимии. Существует алгоритм снятия морфологической омонимии [29, 30], который построен на предположении, что морфологическая омонимия некоторого слова (сигнификативная информационная единица) может быть снята на основе соседних сигнификативных информационных единиц в предложении (предикативная информационная единица), которые составляют контекст этого слова.

Для снятия морфологической омонимии алгоритм последовательно«обходит» предложение справа налево, рассматривая каждое слово только один раз. Следовательно, алгоритм обладает линейной сложностью относительно размера предложения. Другая проблема заключается в том, что некоторые алгоритмы стемминга могут быть пригодны для одного набора терминов, но вызывать много ошибок в другом.

Поэтому следующим шагом развития алгоритма является метод на основе размеченных лингвистических «корпусов текстов» [31]. Основная идея стемминга на основе корпуса текстов состоит в создании классов эквивалентности для слов классических стеммеров, и последующем разбиении некоторых слова, объединенных на основе их встречаемости в корпусе. При использовании этого метода производится дополнение массивов текстов на естественном языке соответствующей

лингвистической информацией, включая разметку именованных сущностей. Разработка таких лингвистических ресурсов менее трудоемка, чем разработка модели ЕЯ.

При использовании «корпусного метода» автоматические лингвистические анализаторы конструируются с использованием хорошо апробированных методов машинного обучения. В результате машинного обучения обобщают примеры, представленные в лингвистическом корпусе текста. В процессе обобщения конструируют процедуры обработки и анализа текстов. В частности, этот метод успешно применяется для создания синтаксических анализаторов. Проблемой на этом пути является отсутствие априорной информации о возможных синтаксических связях предложения, которая может быть получена с помощью лингвистических правил.

Основной проблемой при использовании систем такого анализа является ресурсоёмкая процедура «перебора альтернатив», которая во времени обладает экспоненциальной сложностью. Соответственно это опять требует значительных вычислительных ресурсов.

Следует отметить, что термин «корпус текстов» означает совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту [30]. Соотвественно «корпусная лингвистика» - раздел языкознания, занимающийся разработкой, созданием и использованием текстовых (лингвистических) корпусов.

Можно констатировать наличие двух подходов к анализу компьютерной текстовой информации: «поверхностные», упрощающие структуру текста , но обладающие высокой скоростью анализа; «глубинные», основанные на учете структуры и связей в тексте, но сложные для алгоритмизации и требующие большого времени для анализа

Поэтому возникло третье направление включающее «гибридные» методы обработки и анализа компьютерных текстов, сочетающие в себе как «поверхностные», так и глубинные методы лингвистического анализа [32], Эти методы в некоторых случаях сравнимы по скорости работы с «поверхностными» методами и позволяют получать высокое качество анализа, свойственное лингвистическим методам.

Часто анализ текстов начинается с простой процедуры информационного поиска. В аспекте информационной когнитивной семантики представляет интерес не статистический поиск и не формальный поиск по образцу, а поиск с включением элементов анализа по лингвистическим признакам и принципам.

Достаточно применяемая модель информационного поиска использует булеву алгебру [33, 34]. Модель предполагает, что документ представляет собой не

структурированный набор слов, а запрос — выражение булевой алгебры над отдельными информационными (не семантическими) единицами. При выполнении поискового запроса вначале происходит поиск всех информационных коллекций, содержащих все шаблоны запроса. Затем, каждый документ из информационной коллекции оценивается на соответствие полному выражению запроса. В результате пользователю возвращаются документы, которые удовлетворяют условию, сформулированному методами булевой алгебры

Результатом поиска на основе булевой алгебры является набор документов, в котором все термины обладают равными весами. Это приводит к тому, что одинаковый вес получают как значимые слова, так и малозначимые. Возникает проблема определения вес информационной единицы для повышения эффективности поиска. Рассмотрим решение этой проблемы на примере относительно простой задачи построение словаря.

Для подготовки словаря терминов выбирается информационная коллекция, которая представляет собой один файл, содержащий необходимый текст смешанный со служебной информацией языков SQL и HTML. Для выделения тематического словаря из файла материалов базы данных над ним проводят следующие мероприятия:

- очистка от служебных тегов и спецсимволов языков HTML и SQL;
- проверка и удаление элементов кода PHP;
- очистка от предлогов и местоимений, а также малозначимых слов, содержащих менее четырех букв;
- удаление всех дополнительных символов ‘=’, ‘+’, ‘-‘, ‘\’ и т.п., очистка от английских букв;
- составление списка слов и подсчет частоты вхождений повторяющихся слов;
- синтаксический анализ полученного списка и проверка орфографии;
- морфологический анализ списка и преобразование всех слов к именительному падежу и единственному числу;
- пересчет частоты вхождений повторяющихся слов.

Для каждого термина из полученного списка программно рассчитана частота его вхождения в файл то есть общая вероятность возможности встретить. Д.К.Зипфом (G.K. Zipf) установлено, что зависимость частоты вхождения слова от его ранга – равносторонняя гипербола, и эта зависимость одинакова для всех текстов [35]. Причем наиболее значимые слова текста располагаются в средней части диаграммы, так как слова с максимальной частотой, как правило, являются предлогами, частицами, местоимениями, а редко встречающиеся слова в большинстве случаев не имеют

решающего значения.

Полученные термины были расположены в порядке убывания их частот и пронумерованы. Порядковый номер частоты есть *ранг частоты*. Если взять широкий диапазон, ключевые термины могут смешаться со вспомогательными словами (шумом); если взять узкий диапазон, можно потерять смысловые термины. Чтобы избавиться от лишних слов и в тоже время поднять рейтинг значимых слов, вводят инверсную частоту термина [36]. Значение этого параметра тем меньше, чем чаще слово встречается в материалах базы данных. Инверсную частоту принято вычислять по следующей формуле:

$$inv(t_i) = \log \frac{N_{DB}}{N_{t_i}}, \quad (1)$$

где $inv(t_i)$ – инверсная частота i-го термина из всего набора терминов;

N_{DB} – общее количество материалов в базе данных;

Nt_i – количество материалов базы данных с термином i.

Теперь каждому i-му термину присвоим весовой коэффициент, отражающий его значимость следующим образом:

$$weight(M_j(t_i)) = freq(M_j(t_i)) \cdot inv(t_i), \quad (2)$$

где $weight(M_j(t_i))$ – вес i-го термина в j-м материале;

$freq(M_j(t_i))$ – частота i-го термина в j-м материале.

Общий вес i-го термина в базе данных ВОП нами предлагается вычислять как произведение его инверсной частоты на общую частоту вхождения термина в дамп по следующей формуле:

$$weight(t_i) = \sum_j weight(M_j(t_i)) = inv(t_i) \cdot \sum_j freq(M_j(t_i)), \quad (3)$$

где $weight(t_i)$ – общий вес i-го термина в базе данных ВОП;

$inv(t_i)$ – инверсная частота i-го термина;

$\sum_j freq(M_j(t_i))$ – частота вхождения i-го термина в информационную коллекцию.

Для взвешивания терминов в документе наиболее часто используются следующие подходы.

Первый подход основан на оценке количества вхождений слова в данном документе, что является достаточно простой и почти очевидной характеристикой. Если некоторое слово часто встречается в тексте документа, то, скорее всего, этот документ некоторым образом связан по смыслу с этим словом. Недостаток этого

подхода к оценке веса слова проявляется в случаях, когда анализируемая коллекция содержит документы различной длины. При этом больший вес будут получать более длинные документы, так как они содержат больше слов.

Второй подход использует скорректированную нормализованную частоту появления слова в документе [37] (TF-IDF). Нормализованная частота определяется как отношение числа вхождений слова к общему количеству слов документа. При относительной простоте эта характеристика обеспечивает неплохое качество поиска [38]. Недостатком метода является то, что недооцениваются длинные документы, так как в них больше слов и средняя частота слов в тексте ниже.

В информационной когнитивной семантике, как и в когнитивной семантике существует признание того, что смысл не является чем-то постоянным, а зависит способов понимания и интерпретации. Наиболее ярким примером является гештальт [39] и его интерпретация.

Семантические и когнитивные информационные модели могут быть разными для разных языков, но сопоставимыми по смыслу. Поэтому целесообразно использовать идеи когнитивной семантики для анализа искусственных языков

Заключение. Методы информационной когнитивной семантики развивают идеи семантической теории информации [40], когнитивной лингвистики [41] и принципов синхронизации исследований [42]. В целом это направление исследований направлено на построение целостной картины мира. Спецификой информационной когнитивной семантики является использование информационных единиц и информационных отношений. При этом вводятся такие новые понятия как сигнификативные, предикативные и ассоциативные - информационные единицы. Информационная когнитивная семантика направлена на сближение информационных методов и методов когнитивной семантики и когнитивной лингвистики. Дальнейшее развитие этого направления связано с применением моделей семантического окружения и информационных отношений в информационном поле. Возможно развитие понятий информационная конструкция в лингвистическом и семантическом аспектах. Пока это техническая обобщенная информационная модель предназначенная для структурного анализа в первую очередь и для семантического анализа во вторую. В тоже время информационная конструкция одинаково применима для ЕЯ и ЯИ. Это сближает исследования и позволяет вырабатывать общие модели и технологии анализа. Смысловой анализ информационных конструкций целесообразно выполнять с использованием семантического окружения информационных единиц. Методика описания информационного окружения дает возможность разрабатывать алгоритмы

поиска области истинности и релевантности при заданных информационных единицах. Представляет интерес разработка методов семантического анализа с использованием понятия и процедуры информационного морфизма. Остается перспективным исследование в области делимости информационных единиц.

Список литературы

1. Никитин М. В. Основания когнитивной семантики. – С-Пб.: Изд-во РГПУ им. АИ Герцена, 2003.-277с.

2. Алефиренко Н. Ф. Когнитивная семантика: миф или реальность? //Вестник Томского государственного педагогического университета. – 2006. – №. 5. – с.43-48

3. Кубрякова Е. С. О когнитивной лингвистике и семантике термина «когнитивный» //Вестник ВГУ. Серия «Лингвистика и межкультурная коммуникация. – 2001. – №. 1. – С. 4-10

4. Маслова В. Введение в когнитивную лингвистику – М. : Флинта : Наука, 2007. – 296 с

5. Хазова С. А. Когнитивные ресурсы совладающего поведения: Эмпирические исследования: Монография - Кострома: Костромской гос. ун-т. – 2010.

6. Кошленко М. М. Лексема и фразеосочетание //Их место в уровневой организации языка.-В сб. и. – Т. 23. – С. 213-216.

7. V. Ya. Tsvetkov. Semantic Information Units as L. Florodi’s Ideas Development // European Researcher, 2012, Vol.(25), № 7, p.1036- 1041.

8. Цветков В.Я. Семантика информации // Дистанционное и виртуальное обучение. 2012. № 10. С. 4-7.

9. Бондарко А. В. Грамматическое значение и смысл. – Наука, Ленингр. отд-ние, 1978.

10. Болдырев Н. Н. Концепт и значение слова //Методологические проблемы когнитивной лингвистики. – 2001. – С. 25-36.

11. Цветков В. Я. Информационные единицы сообщений // Фундаментальные исследования. – 2007. - №12. - с.123 – 124.

12. Сеченов И. М., Витгенштейн Л. Базовые понятия когнитивной лингвистики в их взаимосвязи. – 2005

13. Ахманова О.С. Словарь лингвистических терминов. – М.: КомКнига, 2007. – 576 с.

14. Tsvetkov V. Ya. Information Constructions // European Journal of Technology and Design, 2014, Vol.(5), № 3- p147-152

15. Тупик Н. В. Модель мира человека как элемент системы управления // Когнитивный анализ и управление развитием ситуации (CASC´2001). Материалы 1-й Международной конференции., М.: ИПУ РАН, 2001, Т.3, с.163 - 168

16. Цветков В. Я. Язык информатики // Успехи современного естествознания. – 2014 .- №7- с.129-133

17. Цветков В.Я., Чехарин Е.Е. Окружение информационных единиц // Вестник МГТУ МИРЭА «MSTU MIREA HERALD» 2014 - № 2 (3) - с.36- 42

18. Ожерельева Т.А. Об отношении понятий информационное пространство, информационное поле, информационная среда и семантическое окружение // Международный журнал прикладных и фундаментальных исследований. – 2014. – № 10 – с. 21-24

19. Tsvetkov, V.Ya. Information field**.** Life Science Journal 2014- 11(5). –pp.551-554

20. V. Ya. Tsvetkov. Semantic environment of information units // European Researcher, 2014, Vol.(76), № 6-1, p. 1059-1065

21. Ожерельева Т.А. Когнитивные особенности получения второго высшего образования // Перспективы науки и образования- 2013. -№3. – с106 -111.

22. Вагин В.Н., Фомина М.В. Аргументация в индуктивном формировании понятий// Образовательные ресурсы и технологии. 2014. № 2. С. 34-39.

23. Lovins J. B. Development of a stemming algorithm. – MIT Information Processing Group, Electronic Systems Laboratory, 1968.

24. Wallach H. M. Topic modeling: beyond bag-of-words //Proceedings of the 23rd international conference on Machine learning. – ACM, 2006. – С. 977-984.

25. Moschitti A. et al. Exploiting syntactic and shallow semantic kernels for question answer classification //ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. – 2007. – Т. 45. – №. 1. – С. 776.

26. Казенников А.О., Соловьев И.В. Извлечение структурированного новостного сообщения из веб-страниц при использовании дополнительной информации RSS. // Вестник МГТУ МИРЭА «MSTU MIREA HERALD» 2014 - № 2 (3) - с.276-288

27. Филлипс Л., Йоргенсен М. В. Дискурс-анализ //Теория и метод. Харьков: Гуманитарный центр. – 2004.

28. Майоров А.А. Лингвистический анализ термина георефернция // Перспективы науки и образования- 2013. -№4. – с214 -219.

29. Berry M.W. Survey of Text Mining: Clustering, Classification, and Retrieval. Springer, 2003

30. Jurafsky D., Martin M. Statistical Speech and Language Processing. Prentice Hall,

1999.

31. Gries S. T. Corpus-based methods and cognitive semantics: The many senses of to run // trends in linguistics studies and monographs. – 2006. – Т. 172. – p.57.

32. Большакова Е. И. и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика //Журнал исследований социальной политики. – 2013/

33. Manning C., Raghavan P., and Schütze H. , Introduction to Information Retrieval. Cambridge University Press, 2008

34. Apresyan Y., Boguslavsky I., and Iomdin L. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT // First International Conference on Meaning – Text Theory (MTT"2003). Москва. 2003. pp. 279-288

35. Zipf G. K. Human behavior and the principle of least effort. Cambridge: U. Press, 1949

36. Матвеев П.Н. Принципы построения поисковой системы для образовательного портала – Интернет-порталы: содержание и технологии. Сб. науч. ст. Вып.2 / Редкол.: Тихонов А.Н. (пред.) и др.; ГНИИ ИТТ «Информика». – М.: Просвещение, 2004. – 499 с.

37. Salton G., Buckley C., "Term-Weighting Approaches in Automatic Text Retrieval," *Journal of Information Process Management*, Vol. 24, No. 5, 1988. pp. 513-523

38. Croft W.B., Metzler D., and Strohman T. Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing, 2009.

39. Tsvetkov V. Ya., Maslov A. S. Informative Description of Gestalt // European Journal of Technology and Design, 2014, Vol.(5), № 3- p153-160

40. Carnap R. et al. An outline of a theory of semantic information. – Research Laboratory of Electronics, Technical Report №247, MIT, 1952. – 49p.

41. Croft & Cruse, William & D. Alan (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press. p. 3.

42. Афанасьев Ю.И. Теория взаимодействия анализ в условиях синхронизации процесса// Образовательные ресурсы и технологии. 2014. № 3. С. 47-52.