УДК 811.1
UDC 811.1

Петрова Инна Михайловна
Московский городской педагогический университет
г. Москва, Российская Федерация
Inna M. Petrova
Moscow City University
Moscow, Russian Federation
miinna@yandex.ru

**ПОТЕНЦИАЛ ПОИСКОВОЙ СИСТЕМЫ GOOGLE
ПРИ ПРОВЕДЕНИИ ИССЛЕДОВАНИЙ В РАМКАХ
КОГНИТИВНОЙ КОРПУСНОЙ ЛИНГВИСТИКИ
THE POTENTIAL OF GOOGLE SEARCH FOR STUDIES
IN COGNITIVE CORPUS LINGUISTICS**

Аннотация

Статья посвящена использованию поисковой системы Google в качестве аналога корпуса текстов при проведении когнитивных исследований языка. Цель статьи – определить значимость статистических данных, доступных в результате использования разных операторов и фильтров поисковой системы, при изучении когнитивных механизмов репрезентации языковой действительности в речи. В статье проводится сравнение результатов конкурирующих запросов в Google и Национальном корпусе русского языка (НКРЯ), сформированных на основе вариативности комбинаторики биномиальных фраз. Проведённый эксперимент показал, что объём и разнообразие языкового материала указывают на бóльшую эффективность использования Google для исследования комбинаторики и вариативности фразы, что позволяет сделать вывод относительно обоснованности применения этих данных в качестве лингвистического материала для дальнейшей интерпретации в когнитивных исследованиях языка.

Abstract

The paper investigates the possibility to employ the Google search system as an analogue of the corpus of texts for potential use in further cognitive research of a language. The purpose of the article is to elucidate the significance of statistical data available due to the use of different operators and filters of the search system in the study of cognitive mechanisms of representation of linguistic reality in speech. Experimental observations have been made to compare the results of competing queries in Google and Russian National Corpus based on the word order variability of binomial phrases. The results obtained showed that the volume and variety of language data justifies the employment of Google for these purposes. This leads to the conclusion that these data can be considered valid as linguistic material for further interpretation in cognitive language research.

Ключевые слова: корпусная когнитивная лингвистика, алгоритм поисковой системы, комбинаторика фразы, биномиальные конструкции, поисковый запрос.

Keywords: corpus cognitive linguistics, search engine algorithm, phrase combinatorics, binominal phrases, search query.

doi: 10.22250/2410-7190_2019_5_3_127_142

1. Введение

В современной лингвистике всё больше и больше исследований проводится с использованием корпусов текстов [Бурыкин, 2015 ; Манерко, 2018 ; Сулейманова, Демченко, 2018 и др.]. Это обусловлено тем, что ускоренная компьютерная обработка массивов текстов (корпусов), согласно заданным параметрам, предоставляет обширный лингвистический материал, позволяющий получать новые знания о языке. Перспективы использования этих технологий активно обсуждаются в научном сообществе [Голубкова, 2017 ; Рахилина, 2017 ; Чумарина, 2017 и др.]. Отмечая положительные и отрицательные стороны использования корпусных методик при изучении языковых явлений, практически все исследователи констатируют неизбежность обращения к данному инструментарию. Постоянно растущий интерес к возможностям поисковых систем типа Google и недостаточная изученность применения инструментария, предоставляемого современными технологиями, в лингвистических и, в частности, в когнитивных исследованиях определяют актуальность данной работы.

При изучении когнитивных процессов чрезвычайно важными оказываются статистические параметры, учёт которых позволяет интерпретировать эти процессы на другом уровне. К. Фишер одним из первых поставил перед лингвистами-когнитологами вопрос о том, как можно при помощи количественных методов исследовать значение в когнитивной семантике [Fischer, 2010, с. 44]. В результате возник целый ряд количественных методов, которые активно используются в когнитивной семантике. В их число входят, например, метод математической модели, кластерный анализ, корреляционный анализ, логистическая регрессия и другие (подробнее об этом см. [Fischer, 2010]). Объединение статистической обработки данных с результатами применения других экспериментальных методик становится важным элементом исследования в когнитивной науке. Анализ эмпирических данных и их интерпретация позволяют чётче формулировать гипотезу и выявлять противоречия между теоретическим и практическим исследованием явления, особенно в том случае, когда «первичная гипотеза, полученная в ходе интроспективного анализа, противоречит результатам статистических данных корпуса; интерпретируя результаты, мы можем прийти либо к формулированию более точной гипотезы и постановке новых вопросов, которые в последующем могут стать основой для экспериментальных проверок данной гипотезы, либо к новым противоречиям с результатами корпусных данных и т. д.» [Tummers, 2005, с. 233]. Другими словами, анализ статистики корпусных данных выступает таким инструментом исследования, в основе которого заложен значительный потенциал

для интерпретации определённого лингвистического явления и коррекции наших воззрений на данное явление с учётом объективных факторов. Таким образом, новизна настоящего исследования заключается в том, что в нём предпринята попытка выявления возможностей современного инструментария для изучения языковой действительности с позиций когнитивной науки.

Цель настоящей работы – установить, насколько репрезентативен и релевантен языковой материал, полученный на основе поисковой системы Google, для когнитивных исследований языка. Для достижения указанной цели планировалось решить следующие задачи:

1) выявить роль корпусных данных при проведении когнитивных исследований языка;

2) выяснить и описать особенности обработки языкового материала поисковой системой;

3) сравнить результаты поисковых запросов, полученных на основе Национального корпуса русского языка (далее – НКРЯ) и Google, и установить обоснованность применения Google данных как аналога языкового корпуса.

В данной статье мы рассматриваем потенциал поисковой системы Google как базы корпусных данных при проведении когнитивных исследований языка. Языковой корпус текстов выступает как «большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [Захаров, Богданов, 2011, с. 7]. Подробное описание лингвистического корпуса и его видов можно найти, например, в работе Н. В. Козловой [Козлова, 2013]. Однако наряду с существующими языковыми корпусами активно используется и сеть Интернет, поскольку она обладает многими характеристиками данного инструментария. Дискуссия о «вебе как корпусе» становится всё более популярной у исследователей и, несмотря на то, что веб отличается спонтанностью и специфической репрезентативностью, а также отсутствием лингвистического замысла, А. Ю. Мордовин полагает, что лёгкая доступность и высокая объективность данных ввиду отсутствия авторского коллектива в значительной степени компенсируют эти недостатки и функционально приравнивают веб к корпусу, что требует переосмысления онтологического определения последнего. При этом он отмечает, что веб «успешно применяется для решения исследовательских задач корпусными методами, и воспроизводит многие существенные признаки корпуса текстов: аутентичность текстов, их машиночитаемый формат и средства навигации по материалу, репрезентативность материала (в соответствии с идеологией мониторингового типа корпуса), а также более чем достаточный размер» [Мордовин, 2015, с. 171]. Это позволяет говорить о том, что необходимо корректировать современные воззрения о классическом корпусе текстов. Настоящая статья представляет собой попытку рассмотрения этого вопроса в определённом ключе, а именно, в контексте использования таких данных в когнитивных исследованиях языка.

2. Информационная поисковая система как база корпусных данных в лингвистических исследованиях

Многие исследователи уже используют веб как источник аутентичных, естественных, контекстуализированных языковых моделей (конкордансов, устойчивых словосочетаний, лексических комбинаций, фраз, идиом и др.) [Сулейманова, Демченко, 2018 ; Квашина, Ажель, 2017 ; Мордовин, 2015 ; Gatto, 2014 ; Geluso, 2013 и др.]. Благодаря неисчерпаемому объёму Всемирная сеть является уникальным ресурсом для анализа частоты использования тех или иных языковых моделей, а значит, естественного, «живого» языка. Наиболее востребованной поисковой системой является система Google благодаря своей простоте и скорости, функциональности поиска информации, а также репрезентативности. Результат работы Google можно сравнить с интернет-каталогом, отобранным с помощью рейтинговой системы на основе алгоритмов. Более конкретно алгоритм поиска можно описать как «нахождение элемента с заданными свойствами среди списка элементов». В современном обучении иностранному языку эта система используется в качестве инструмента, позволяющего проанализировать частоту употребления и значение той или иной языковой модели [Panah, 2013 ; Квашина, Ажель, 2017].

Информационная поисковая система (далее – ИПС) строится на особом алгоритме обработки языковой информации. При вводе документа в базу данных ИПС его индексируют (поэтому саму базу поисковой системы часто называют индексом). Процесс индексирования связан с определением и выборкой ключевых слов обрабатываемых документов и выражением их формально в виде поискового образа. Так, база данных ИПС состоит из множества индексных поисковых образов. Непосредственно при поиске производится сопоставление сделанного запроса, то есть того, что указано в запросе, с поисковым образом, то есть с тем, что хранится в индексе.

Ключевыми характеристиками информационного поиска, релевантными для обоснованности надежности Google данных, служат такие показатели, как: полнота выдачи информации и точность её выдачи; но при этом имеют место потери информации и информационный шум.

Отношение количества выданных релевантных документов к общему числу релевантных документов, содержащихся в базе информационно-поисковой системы, определяет полноту поиска. Точность поиска определяется количеством выданных системой релевантных документов к общему числу документов в выдаче. Понятие «релевантность» выступает как фундаментальное понятие теории информационного поиска. Документ, центральный предмет или тема которого в целом соответствует смысловому содержанию информационного запроса, называется релевантным, а свойство смысловой близости между документом и информационным запросом – релевантностью.

Важное место в системе представления информации в ИПС играют поисковые тезаурусы (одноязычные или многоязычные). Они представляют собой специальные словари для информационного поиска по массивам естественно-языковых документов, организованные по принципу

сопоставления слов с их понятиями. Их структура и разработка часто стандартизируются (см., напр., ГОСТ 7.25-2001). Такой дескрипторный словарь используется как средство лексического контроля (напр., снятия омонимии, синонимии единиц) при индексировании документов и запросов. Например, при индексировании все синонимы запроса и поискового образа представляются одной и той же лексической единицей – дескриптором (*ср. лингвистика – языкознание языковедение, наука о языке*).

Как правило, тезаурусы относятся к классу семантических словарей. Пример организации семантической информации в поисковом тезаурусе можно продемонстрировать на примере «Русского семантического словаря» под редакцией Н. Ю. Шведовой [Шведова, 1998]. Так, в разделе «Формы и сущностные характеристики реалий и явлений действительности» представлена следующая семантическая структура организации лексической информации: в вершине древа лежат четыре исходных множества: 1) «Время; его ход, периоды и моменты его течения»; 2) «Пространство: его основные характеристики; мера пространства, его величины»; 3) «Движение, изменение местоположения в пространстве» и 4) «Количество. Счёт. Масса (вес), её мера. Степень, предел, граница». Каждое из этих четырёх множеств включает свои собственные подмножества, членящиеся каждое по принципу сужения и конкретизации общего значения и нисходящие, разветвляясь до конечных лексико-семантических рядов, далее разбиваемых только на отдельные лексические единицы или их мельчайшие объединения. Например, лексическое множество «Время: его ход, периоды и моменты его течения» существует в двух подмножествах: «Время, его основные характеристики» и «Отрезки, периоды времени, соотносимые с каким-нибудь состоянием, событием, деятельностью, действием». В первом из них исходный смысл 'время' далее конкретизируется по признакам, организующим предконечные и конечные группировки словозначений; а) ход, течение времени (напр., *давность, длительность, продолжительность*), б) способы, системы исчисления времени (напр., *летосчисление, календарь, стиль*), в) виды сегментации и единицы исчисления времени (напр., *интервал, момент, период; век, год, месяц, день, час*). Второе из названных выше подмножеств – «Отрезки, периоды времени, соотносимые с каким-нибудь состоянием, событием, деятельностью, действием» – включает такие единицы, которые, наряду со смыслом 'время' заключают в своём значении либо обобщённое указание на социальные или физические явления, соотносимые с определённым временным периодом (напр., *будущее, настоящее, прошлое /человечества/, времена, годы /юности/, полоса, сезон /дождей/*), либо конкретную характеристику такого явления, состояния (например, *сеанс, антракт, отпуск, стаж; рассвет, закат; зима, весна, лето; молодость, старость; неолит, палеолит*). Соотношение этих элементов значения в слове определяет собою последующее членение данного подмножества.

Таким образом, у поисковых систем Интернета имеются свои алгоритмы поиска и выдачи информации, которые разработаны на основе компьютерной семантики. Для управления и использования гибкого поиска

в языке поисковых запросов используются, так называемые, операторы – это символы и команды, позволяющие настраивать условия поиска запроса. Сам язык запросов максимально упрощён, чтобы любой пользователь мог с ним справиться и выбрать необходимый для его случая оператор. Подробнее язык запросов и значения операторов можно изучить, задав поиск в браузере поисковой системы. Важной особенностью операторов является возможность их комбинирования. За счёт комбинаций операторов можно выстроить сложные запросы. Например, оператор «» фиксирует порядок слов и слова идут именно в этом порядке: «пламень и лёд». В результате запрос выдает 8,660 результатов, в которых именно такой порядок следования компонентов. Комбинирование операторов позволяет максимально гибко настроить поиск.

Вхождение ключевого слова – это органично вписанный в текст поисковый запрос. Вхождения ключевых слов в текст бывают прямые и разбавленные. К первому типу относятся фразы, вписанные в текст как есть, без изменений, а ко второму – все остальные. Прямые и точные вхождения ключевых слов встречаются не так часто, потому что оставить все словоформы без изменений почти никогда не получается. Чаще всего они выступают текстом анкера для ссылок, используются в контекстной рекламе или в метатегах. Разбавочные вхождения ключевых слов – это вхождения, в которых фразы изменяются разнообразным способом, например, введение дополнительных слов, изменение словоформы, порядка слов. Ключевой запрос: «семантический словарь» выдаёт следующие разбавочные вхождения: «Русский семантический словарь», «семантические словари», «семантический словарь русского языка», «семантические и понятийные словари русского языка» и прочее. Выделяют семь видов вхождения ключей: прямое, разбавленное, морфологическое, прямое вхождение с разбавлением знаками препинания, ключи с ошибками, вхождение на англоязычной раскладке, обратные вхождения¹. По логике роботов-поисковиков формальное прямое вхождение на 100% соответствуют пользовательскому запросу, а значит и вся страница релевантнее. Стоит отметить, что для поисковых роботов разница между прямым и разбавочным вхождением ключевого слова не очевидна. Слова в ключевых фразах могут быть заменены синонимами, роботы-поисковики также учитывают эти вхождения.

Таким образом, частота вхождений в поисковой системе отражает релевантность сайта запросу пользователя. Чем точнее сформулирован запрос, тем точнее получается результат поиска. Другими словами, представление информации по результатам запроса будет отражать некий аспект общей системы знаний через призму компьютерной семантики. Компьютерная семантика является элементом интеллектуальной системы и косвенным образом коррелирует с интеллектуальной системой индивида. Соответственно, результаты исследования семантических явлений через призму интеллектуальной системы дают возможность опосредованно исследовать семантику интеллектуальной системы человека.

---

¹ Подробнее об этом см. <http://kopiraitery.ru/seo-kopirajting/7-vidov-vxozhdeniya-klyuchey.htm>

2.1. Сравнение результатов запросов в НКРЯ и Google при изменении комбинаторики фразы

Для обоснованности использования поисковой системы в когнитивных исследованиях мы обращаемся к механизмам репрезентации действительности в языке посредством биноминальных конструкций, хотя для этих целей могут быть использованы и другие языковые структуры, например, атрибутивные конструкции (см., напр., работу О. А. Сулеймановой и И. М. Петровой [Сулейманова, Петрова, 2018]).

Биноминальные конструкции представляют собой парные образования существительных, объединённых либо антонимическими отношениями, например, *добро и зло*, либо синонимическими *тепло и свет* и тому подобные. Исследование комбинаторики такой фразы или порядка организации подобной синтаксической структуры можно рассматривать как способ указания на распределение значимой для говорящего информации во фразе, то есть актуализации того элемента высказывания, который выступает ведущим в данной паре. Соответственно, изменение этого порядка отражает изменение в ментальной модели репрезентации этих компонентов информации, что сопряжено с индивидуальной интерпретацией данных элементов фразы. Так, если мы говорим *свет и тьма*, *добро и зло*, то очевидно, что в такой репрезентации понятий первый компонент играет ведущую роль в данной дуальной паре. Изменение комбинаторики фразы на *тьма и свет* демонстрирует смещение акцента и актуализацию понятия *тьма* в прагматическом аспекте фразы [Петрова 2018, 2019].

Рассмотрим потенциал Google для когнитивного исследования на примере предоставления эмпирических данных при изучении данного вопроса более подробно. Для этой цели сравним результаты поисковых запросов в НКРЯ и Google на предмет потенциальных возможностей для изучения комбинаторики биноминальных фраз. Возьмём некоторые базовые понятия, выраженные биноминальной фразой, и произведём изменение комбинаторики этой фразы, например, *мужчина и женщина* (1) и *женщина и мужчина* (2); *свет и тьма* (3) и *тьма и свет* (4). В рамках данного эксперимента формирование поискового запроса в Google строилось с помощью двух операторов: (1) фиксирующего порядок слов [] и (2) отмечающего порядок слов «». Для проведения этого эксперимента мы не использовали фильтры расширенного поиска намеренно, поскольку на данном этапе основная задача, стоящая перед нами, состояла в установлении репрезентативности и релевантности отобранного материала в контексте корпусного подхода.

Как видно из представленной таблицы (табл. 1), количество вхождений, определённых оператором «», устанавливающим точный порядок слов, ниже, чем оператором, фиксирующим порядок слов. Это говорит в пользу того, что запрос, оформленный оператором «» представляет релевантную по структуре фразы информацию. Поскольку вхождения фразы могут быть прямыми и разбавочными, приведём примеры такого представления информации.

**Т а б л и ц а 1. Результаты поискового запроса в Google,
оформленного операторами [] и «»**

Русский вариант фразы			
Оператор []	Число вхождений	Оператор «»	Число вхождений
(1) [мужчина и женщина]	121,000,000	(1) «мужчина и женщина»	6,950,000
(2) [женщина и мужчина]	114,000,000	(2) «женщина и мужчина»	694,000
(3) [свет и тьма]	4,710,000	(3) «свет и тьма»	404,000
(4) [тьма и свет]	4,920,000	(4) «тьма и свет»	128,000
Английский вариант фразы			
Оператор []	Число вхождений	Оператор «»	Число вхождений
(1) [man and woman]	21,020,000,000	(1) “man and woman”	41,700,000
(2) [woman and man]	3,250,000,000	(2) “woman and man”	9,340,000
(3) [light and darkness]	891,000,000	(3) “light and darkness”	8,140,000
(4) [darkness and light]	876,000,000	(4) “darkness and light”	6,840,000

Примечание к таблице 1 : Дата обращения 18.11.2018.

1. Примеры прямого вхождения фразы:

– *Мужчина и женщина: в чем отличия* (<http://russian7.ru/post/7-razlichij-mezhdu-muzhchinami-i-zhenshhinami/>);

– *Женщина и мужчина – отношения сквозь века* (<https://foma.ru/zhenshhina-i-muzhchina-otnosheniya-skvoz-veka.html>);

– *The creation and separation of light and dark comes up in the myths of creation or cosmogony—from kosmos, meaning “order,” and genesis, meaning “birth.”* (<https://www.psychologytoday.com/us/blog/myth-the-mind/201805/myths-light-and-dark>)

Данные примеры содержат заданные формы биномов. В основном они были получены в результате использования оператора «».

2. Примеры разбавочного вхождения фразы:

– *A Man And A Woman* by U2 song meaning, lyric interpretation, video and chart position. (<https://www.songfacts.com/facts/u2/a-man-and-a-woman>). – В данной фразе введён артикль «a».

– *Jun 5, 2018 – Showrunner Joe Pokaski and director Gina Prince-Bythewood weigh in on the darkness and light parallels of Freeform’s new show* (<https://www.tvguide.com/news/cloak-dagger-preview-behind-darkness-light-metaphors/>). В этом примере имеет место введение новых элементов в результате запроса, в частности артикль *the* и слово *parallels*, которые дополняют структуру.

Приведённые примеры по большей части представляют собой тексты, выполняющие номинативную функцию. Очевидно, что выданные поисковым роботом данные гораздо разнообразнее. Мы представили указанные примеры исходя из их относительной краткости и возможности продемонстрировать сохранение порядка следования компонентов биномов. В результате анализа экспериментального материала мы пришли к выводу, что оператор «» представляет более точную выборку материала, поскольку процент разбавочных вхождений фраз в этом случае ниже.

Проанализируем полученные данные на предмет процентного соотношения количества фраз (1) *мужчина и женщина* и (3) *тьма и свет*, которые показывают максимальное число вхождений в поисковую систему и представляют собой традиционный порядок следования данных биномов, с фразами (2) *женщина и мужчина* и (4) *тьма и свет*. Для анализа используем данные, полученные с помощью оператора «», который продемонстрировал более релевантную выборку материала (см. табл. 2).

Т а б л и ц а 2. Число вхождений фраз в Google, отобранных с помощью оператора «»

Фраза	Русский язык	Английский язык
(1) мужчина и женщина / man and woman	6,950,000	41,700,000
(2) женщина и мужчина / woman and man	694,000	9,340,000
(3) свет и тьма / light and darkness	404,000	8,140,000
(4) тьма и свет / darkness and light	128,000	6,840,000

Примечание к таблице 2 : Дата обращения 18.11.2018.

Расчёт процентного соотношения производился нами на основании пропорции, в которой количество вхождений фраз (1) и (3) приравнялось к 100%. Цель проведения данных расчётов – установить сколько процентов составляют инвертированные фразы от объёма традиционных фраз. Данный процент наглядно демонстрирует частотность употребления инвертированной фразы пользователями интернета в конкретный временной период. В итоге были получены результаты, представленные в таблице 3.

Т а б л и ц а 3. Процентное соотношение числа вхождений фраз в Google

Соотносимые фразы	Русский язык	Английский язык
(1) мужчина и женщина / man and woman	9,98%	22,39%
(2) женщина и мужчина / woman and man		
(3) свет и тьма / light and darkness	31,68%	84,02%
(4) тьма и свет / darkness and light		

Как видно из полученных данных, для англоязычных пользователей процент инвертированной фразы (2) *женщина и мужчина* / *woman and man* на порядок выше, чем для русскоязычных пользователей. Это может свидетельствовать о том, что значение компонента бинома *женщина* в репрезентации действительности в англоязычной среде Интернета выше по сравнению с русским языком.

Соотношение фраз (3) *свет и тьма* / *light and darkness* и (4) *тьма и свет* / *darkness and light* показывает достаточно высокую частотность инвертированных фраз в обоих языках, но в английском языке мы наблюдаем очень высокий показатель – 84%. Это означает, что компоненты, составляющие этот бином, не представляют существенной разницы для англоязычной среды Интернета, в то время, как для русскоязычной среды эта разница существенна.

Таким образом, проведённый эксперимент демонстрирует возможности поисковой системы Google при исследовании вопросов отражения реальной действительности в языке посредством статистических данных. Разумеется, подобный подход не заменяет других методов лингвистического исследования, но, безусловно, обогащает наше представление о языковой реальности.

Теперь рассмотрим, как представлен результат исследования комбинаторики фразы с помощью НКРЯ². Поиск по НКРЯ строится по принципу поиска точных форм.

Т а б л и ц а 4. Результаты поискового запроса (НКРЯ)

Фраза	Число вхождений
(1) мужчина и женщина	447 (363 документа)
(2) женщина и мужчина	2 (1 документ)
(3) свет и тьма	54 (36 документов)
(4) тьма и свет	19 (14 документов)

Примечание к таблице 4 : Дата обращения 28.11.2018.

Согласно этим результатам, процентное соотношение фраз (1) *мужчина и женщина* и (2) *женщина и мужчина* составляет 0,4%, а фраз (3) *свет и тьма* и (4) *тьма и свет* соответственно 35,1%. Можно отметить, что трудно обнаружить корреляцию в соотношениях фраз (1) и (2) между данными НКРЯ и Google, поскольку мы имеем 0,4% в НКРЯ и практически 10% в Google. Корреляция соотношения фраз (3) и (4) выглядит следующим образом: 35,1% в НКРЯ и 31,68% в Google, что достаточно близко. Из этого

² Для сравнения, в «Библиотеке лексикографа» (которая, к сожалению, носит принципиально оффлайн-характер) А. А. Бурыкина [Бурыкин, 2015] это соотношение выглядит следующим образом: (1) «мужчина и женщина» – 19530 документов; (2) «женщина и мужчина» – 4177 документов; (3) «свет и тьма» – 1298 документов; (4) «тьма и свет» – 428 документов. Эти цифры также демонстрируют значительно более высокую частотность первой биномиальной фразы по сравнению со второй и третьей по сравнению с четвёртой.

следует, что, хотя объём анализируемого материала в НКРЯ несопоставимо меньше, чем в Google, определённая корреляция между данными прослеживается, что свидетельствует об их релевантности.

Приведём несколько примеров из НКРЯ по результатам поискового запроса, чтобы получить представление о том, какого рода языковой материал подвергался обработке.

Пример к фразе (1) *мужчина и женщина*:

*В этот самый момент к нашему столу подошли двое, **мужчина и женщина**, и уселись на свободные места* [Булат Окуджава. Искусство кройки и житья. М., 1985].

Пример к фразе (2) *женщина и мужчина*:

*Вдруг трое – **мужчина, женщина**, старик, только что спокойно с ним разговаривавшие, – кто что-то объясняет, кто советы дает, как пройти, – увидели хлеб и как-то дернулись, словно бы бросились мгновенно и тут же сдержались* [Татьяна Толстая. Сомнамбула в тумане. М., 1989].

Пример к фразе (3) *свет и тьма*:

***Свет и тьма** в едином, в Огне – это и есть Макиавелли* [Юрий Азаров. Подозреваемый. М., 2002].

Пример к фразе (4) *тьма и свет*:

*Поэтому мне кажется, в каждом из нас есть **тьма и свет**, просто не всегда у нас в жизни есть возможность воспитать в себе хорошее* [Юлия Идлис. Красота – это мы, люди // «Русский репортер». № 22 (200), 9 июня 2011].

Анализ приведённых примеров показывает, что языковой материал НКРЯ нуждается в тщательном анализе, поскольку, хотя запрос и выдал 1 документ с фразой (2) *женщина и мужчина*, фактически, в представленном примере порядок следования компонентов бинама был другим (см. выше пример к фразе 2). Тем не менее, и в Google, и в НКРЯ имеет место статистическое отражение изменённой комбинаторики фразы, в частности, количество вхождений фраз (2) *женщина и мужчина* и (4) *тьма и свет* значительно ниже по показателям обеих баз данных. Количество примеров, представленных Google, гораздо больше, чем в НКРЯ.

В результате проведённого исследования можно отметить ряд особенностей рассматриваемых корпусных систем обработки языковых данных, которые мы представляем в таблице 5.

Таким образом, поставленная перед нами задача относительно установления правомерности использования Google данных в качестве аналога корпусной базы данных представляется решённой. Очевидно, что в проведённом эксперименте релевантность примеров относительно комбинаторики фразы требует внимательного рассмотрения, поскольку в поисковой системе может иметь место разбавочное вхождение, а в НКРЯ, например, изменение синтаксической организации фразы. Однако применение опции расширенного поиска в Google открывает перед нами возможность получения данных, ограниченных рядом параметров, например, точной структурой фразы, типом текста, страной употребления и так далее.

Т а б л и ц а 5. Сравнение НКРЯ и Google в контексте лингвистических исследований

Особенности системы	НКРЯ	Google
Объём текстового материала	Ограничен в объёме и временных параметрах	Неограничен
Широта диапазона рассматриваемого текстового материала	Ограничена определёнными текстовыми жанрами и требованиями репрезентативности	Неограниченная, позволяющая обнаружить материалы различных типов и жанров
Лингвистическая обработка текстового материала	Имеет место непосредственная обработка с выделением различных лингвистических аспектов	Имеет место семантическая обработка поисковой системой в процессе подготовки информации к регистрации в сети
Возможность отбора данных, отвечающих определённым критериям	Имеется возможность поиска с учетом лексико-грамматических, морфологических особенностей фразы	Функция расширенного поиска позволяет более точно подобрать материал, отвечающий определённым запросам

Полученные результаты могут служить материалом для дальнейшего более подробного изучения вопросов комбинаторики и вариативности фразы в когнитивном ключе.

Необходимо отметить, что данная поисковая система предоставляет возможность решать задачи не только относительно биномиальных конструкций, но и других синтаксических структур, например, атрибутивных групп, типа *большой красивый дом / красивый большой дом*. Анализ статистических данных вариативности подобных фраз в контексте теории классов позволяет делать важные выводы об изменении когнитивного фокуса высказывания. Кроме того, инструментарий Google представляется полезным и при проведении лингвистических экспериментов, например, на вариативность предикаций, выраженных фразами с фазовыми глагольными биномами такими, как *решил начать учиться / учиться решил начать*. Анализ вариативности таких структур даёт возможность внимательно рассмотреть механизм распределения информации в когнитивном поле через призму категорий Интенция – Фаза – Действие. Таким образом, спектр задач с использованием рассматриваемого экспериментального инструментария может быть достаточно широким.

Вместе с тем необходимо подчеркнуть, что язык запросов Google не поднимается на более высокий уровень абстракции. Это не позволяет работать с частями речи, синтаксическими единицами, модальностями и тому подобными вещами, а только с конкретными лексемами (и ограниченно – с формами этих лексем: числами, падежами, родами и т. п.; а также ограниченно – с опечатками и синонимами (благодаря специаль-

ной системе синонимов, работа над которой велась разработчиками Google 5 лет)). Это означает, что исследователи, использующие Google и другие поисковые системы в сети Интернет, не должны питать иллюзий относительно их аналитических возможностей на уровне языковых абстракций. Эффективное планирование поисковых запросов, систематизация полученных данных и их анализ по-прежнему остаются на плечах лингвистов. Однако грамотное использование сети с ясным пониманием её ограниченности, несомненно, поможет нарастить объёмы данных, что, в сочетании с традиционными методами лингвистического анализа, сделает исследование более объективным и достоверным.

3. Заключение

Проведённое исследование показывает, что статистические данные, полученные на основе корпусов текстов и поисковых систем, могут выступать в качестве материала исследования при изучении когнитивных аспектов языковой действительности. В ходе когнитивных исследований языка с помощью корпусных баз данных важным представляется правильность формулировки исследовательской задачи, позволяющая интерпретировать полученные данные в контексте заданного подхода. Объёмный статистический материал может служить основой для верификации выдвигаемых гипотез. Поисковая система Google может выступать в качестве аналога корпуса текстов при исследовании тех языковых явлений, которые находят отражение в статистических результатах. Данный инструментарий особенно эффективен при изучении комбинаторики и вариативности фразы, поскольку позволяет собрать достаточный по объёму и разнообразию эмпирический материал, который не всегда доступен традиционным корпусам.

Список литературы

1. Бурыкин, А. А. Электронный ресурс для исследований в области русской лексикологии и лексикографии «Библиотека лексикографа»: опыт работы, перспективы пополнения, возможности использования [Текст] / А. А. Бурыкин // Теоретическая и прикладная лингвистика. – 2015. – Вып. 1. – № 4. – С. 5–28.
2. Голубкова, Е. Е. Возможности использования корпусов в целях лингвистического исследования [Текст] / Е. Е. Голубкова. // Магия ИННО: новые измерения в лингвистике и дидактике / [отв. ред. Д. Н. Новиков]. – Т. 1. – М. : МГИМО–Университет, 2017. – С. 394–400.
3. Захаров, В. П. Корпусная лингвистика: учебник [Текст] / В. П. Захаров, С. Ю. Богданов. – Иркутск : ИГЛУ, 2011. – 161 с.
4. Квашина, О. С. Использование Google как системы поиска языковых данных в веб-корпусе: обучение англоязычной письменной речи [Текст] / О. С. Квашина, Ю. П. Ажель // Alma mater (Вестник высшей школы). – 2017. – № 6. – С. 106–109.
5. Козлова, Н. В. Лингвистические корпуса: определение основных понятий и типология [Текст] / Н. В. Козлова // Вестник НГУ. Серия : Лингвистика и межкультурная коммуникация. – 2013. – Том 11. – Вып. 1. – С. 79–88.

6. Национальный корпус русского языка [Электронный ресурс]. – URL : <http://ruscorpora.ru/> (дата обращения: 28.11.2018).
7. Манерко, Л. А. Методологические основы использования корпусных данных в исследованиях по когнитивной лингвистике [Текст] / Л. А. Манерко // Когнитивные исследования языка. – 2018. – № 33. – С. 70–78.
8. Мордовин, А. Ю. «Веб как корпус» или «корпус как веб»: новая реальность корпусной лингвистики [Текст] / А. Ю. Мордовин // Вестник МГЛУ. – 2015. – № 3 (714) : Языкознание и литературоведение. Когнитивные аспекты языка и речи. – С. 163–172.
9. Петрова, И. М. Когнитивный аспект комбинаторики сочинительных парных словосочетаний в английском и русском языках [Текст] / И. М. Петрова // Современная наука: актуальные проблемы теории и практики. – 2018. – № 9. – С. 151–156.
10. Петрова, И. М. Комбинаторика биномиальных конструкций как отражение процесса иконической репрезентации объектов действительности [Текст] / И. М. Петрова // Когнитивные исследования языка. – 2019. – № 37. – С. 621–625.
11. Рахилина, Е. В. Экспериментальная и корпусная лингвистика: рецензия на сборник статей к юбилею О. А. Сулеймановой «Контенсивные аспекты языка: константность и вариативность» [Текст] / Е. В. Рахилина. / отв. ред. Т. Д. Шабанова. – М. : Флинта, 2016 // Вестник Моск. город. пед. ун-та. Сер. : Филология. Теория языка. Языковое образование. – 2017. – № 3 (27). – С. 119–128.
12. Сулейманова, О. А. Использование BIGDATA в экспериментальных лингвокогнитивных исследованиях: анализ семантической структуры глагола shudder [Текст] / О. А. Сулейманова, В. В. Демченко // Когнитивные исследования языка. – 2018. – № 33. – С. 466–472.
13. Сулейманова, О. А. Экспланаторный потенциал теории классов для лингвистического исследования: порядок следования определений [Текст] / О. А. Сулейманова, И. М. Петрова // Филология: научные исследования. – 2018. – № 3. – С. 52–64.
14. Чумарина, Г. Р. Ресурсы и арсенал электронных корпусов в современной лексикографии [Текст] / Г. Р. Чумарина // Филологические науки. Вопросы теории и практики. – 2017. – № 3–1 (69). – С. 173–175.
15. Fischer, K. Quantitative Methods in Cognitive Semantics. Introduction to the volume [Text] / K. Fisher // Quantitative methods in cognitive semantics: corpus-driven approaches / D. Glynn, K. Fischer (eds). – Berlin / N.Y. : Walter de Gruyter, 2010. – P. 43–59.
16. Gatto, M. The Web as Corpus: Theory and practice [Text] / M. Gatto. – London, New York : Bloomsbury Academic. – 2014. – 256 p.
17. Geluso, J. Phraseology and frequency of occurrence on the web: native speakers' perceptions of Google-informed second language writing [Text] / J. Geluso // Computer Assisted Language Learning. – 2013. – № 26:2. – P. 144–157.
18. Panah, E. Google-informed pattern-hunting and pattern-defining: Implication for language pedagogy [Text] / E. Panah, M. Yunus, M. A. Embi // Asian Social Science. – 2013. – № 9 (3). – P. 229–238.

19. Tummers, J. Usage-based approaches in Cognitive Linguistics: A technical state of the art [Text] / J. Tummers, K. Heylen, D. Geeraerts // *Corpus Linguistics and Linguistic Theory*. – 2005. – № 1(2). – P. 225–261.

References

1. Burykin, A. A. (2015). Elektronnyy resurs dlya issledovaniy v oblasti russkoy leksikologii i leksikografii «biblioteka leksikografa»: opyt raboty, perspektivy popolneniya, vozmozhnosti ispol [Electronic resource for studies in the field of Russian lexicology and lexicography «Lexicographer's library»: Experience, outlook for enlarging, possibilities of using]. *Teoreticheskaya i prikladnaya lingvistika* [Theoretical and Applied Linguistics], 1 (4), 5–28.
2. Golubkova, E. E. (2017). Vozmozhnosti ispol'zovaniya korpusov v celyah lingvisticheskogo issledovaniya [Corpora in linguistic research]. In D. N. Novikov, *Magiya INNO: novye izmereniya v lingvistike i didaktike* [The magic of INNO: New dimensions in linguistics and didactics] (Vol. 1, pp. 394–400). Moscow : MGIMO University Press.
3. Zaharov, V. P., Bogdanova, S. Yu. (2011). *Korpusnaya lingvistika: uchebnik* [Corpus linguistics: A coursebook]. Irkutsk : Irkutsk State Linguistic University.
4. Kvashina, O. S., Azhel', Yu. P. (2017). Ispol'zovanie Google kak sistemy poiska yazykovykh dannykh v veb-korpuse: obuchenie angloyazychnoy pis'mennoy rechi [Using Google as the system of search of language data in Web corpus: training in English writing pedagogy]. *Alma mater (Vestnik Vysshey Shkoly)* [Alma Mater (Higher School Herald)], 6, 106–109.
5. Kozlova, N. V. (2013). Lingvisticheskie korpusa: opredelenie osnovnykh ponyatiy i tipologiya [Linguistic corpus: Typology and terms]. *NSU Vestnik. Series : Linguistics and intercultural communication*, 11(1), 79–88.
6. Russian National Corpus [Digital resource]. Retrieved November 28, 2018 from <<http://ruscorpora.ru/>>.
7. Manerko, L. A. (2018). Metodologicheskie osnovy ispol'zovaniya korpusnykh dannykh v issledovaniyakh po kognitivnoy lingvistike [Methodological foundations of corpus data usage in cognitive linguistics research]. *Kognitivnye issledovaniya yazyka* [Cognitive studies of language], 33, 70–78.
8. Mordovin, A. Yu. (2015). «Veb kak korpus» ili «korpus kak veb»: novaya real'nost' korpusnoj lingvistiki [«Web as corpus» or «corpus as web»: The new reality of corpus linguistics]. *Vestnik MGLU. Series : Yazykoznanie i literaturovedenie. Kognitivnye aspekty yazyka i rechi* [Vestnik of Moscow State Linguistic University. Linguistics and Literary Studies. Cognitive Aspects of Language and Speech], 3 (714), 163–172.
9. Petrova, I. M. (2018). Kognitivnyy aspekt kombinatoriki sochinitel'nykh parnykh slovosochetaniy v angliyskom i russkom yazykakh [Cognitive aspect of combinatorics of compositional paired phrases in English and Russian]. *Sovremennaya nauka: aktual'nye problemy teorii i praktiki* [Modern science: actual problems of theory and practice], 9, 151–156.
10. Petrova, I. M. (2019). Kombinatorika binominal'nykh konstruktsiy kak otrazhenie protsessa ikonicheskoy reprezentatsii ob"ektov deystvitel'nosti [Combinatorics of

- binomial phrases as a reflection of the process of iconic representation of objects of reality]. *Kognitivnye issledovaniya yazyka* [Cognitive studies of language], 37, 621–625.
11. Rahilina, E. V. (2017). Eksperimental'naya i korpusnaya lingvistika: retsenziya na sbornik statey k yubileyu O. A. Suleymanovoy «Kontensivnye aspekty yazyka: konstantnost' i variativnost'» [Experimental and corpus linguistics: Review of the collection of articles for the anniversary of O. A. Suleymanova «Content aspects of language: constancy and variability»]. *Vestnik Mosk. gorod. ped. un-ta. Ser: Filologiya. Teoriya yazyka. Yazykovoe obrazovanie* [Vestnik of Moscow City University. Series : Philology. Theory of language. Language education], 3 (27), 119–128.
  12. Souleimanova, O. A., Demchenko, V. V. (2018). Ispol'zovanie BIGDATA v eksperimental'nyh lingvokognitivnyh issledovaniyah: analiz semanticheskoy struktury glagola shudder [Using big data in experimental linguo-cognitive studies: Analysis of the semantic structure of the verb shudder]. *Kognitivnye issledovaniya yazyka* [Cognitive studies of language], 33, 466–472.
  13. Souleimanova, O. A., Petrova, I. M. (2018). Eksplanatornyj potencial teorii klassov dlya lingvisticheskogo issledovaniya: poryadok sledovaniya opredelenij [Explanatory potential of the theory of classes for linguistic research: Word order in attributive group]. *Filologiya: nauchnye issledovaniya* [Philology: Scientific researches], 3, 52–64.
  14. Chumarina, G. R. (2017). Resursy i arsenal elektronnyh korpusov v sovremennoj leksikografii [Resources and arsenal of electronic corpora in contemporary lexicography]. *Filologicheskie nauki. Voprosy teorii i praktiki* [Philological Sciences. Issues of Theory and Practice], 3–1 (69), 173–175.
  15. Fischer, K. (2010). Quantitative Methods in Cognitive Semantics. Introduction to the volume. In D. Glynn, K. Fischer (Eds.), *Quantitative methods in cognitive semantics: corpus-driven approaches* (pp. 43–59). Berlin / N. Y.: Walter de Gruyter.
  16. Gatto, M. (2014). *The Web as Corpus: Theory and practice*. London, New York : Bloomsbury Academic.
  17. Geluso, J. (2013). Phraseology and frequency of occurrence on the web: native speakers' perceptions of Google-informed second language writing. *Computer Assisted Language Learning*, 26 (2), 144–157.
  18. Panah, E., Yunus, M., Embi, M. A. (2013). Google-informed pattern-hunting and pattern-defining: Implication for language pedagogy. *Asian Social Science*, 9 (3), 229–238.
  19. Tummers, J., Heylen, K., Geeraerts, D. (2005). Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory*, 1 (2), 225–261.